Univerzita Palackého v Olomouci

Přírodovědecká fakulta

Katedra geoinformatiky

Studijní program: **P1301 Geografie**

Obor: Geoinformatika a kartografie

PROSTOROVÉ A VÍCEROZMĚRNÉ STATISTICKÉ ANALÝZY EPIDEMIOLOGICKÝCH DAT

Doktorská disertační práce

Mgr. Lukáš MAREK

Školitel: Doc. Mgr. Jiří Dvorský, Ph.D.

Olomouc 2015

Čestné prohlášení

Prohlašuji, že jsem disertační práci doktorského studia oboru Geoinformatika a kartografie vypracoval samostatně pod vedením Doc. Mgr. Jiřího Dvorského, Ph.D.

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

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

V Olomouci 24. dubna 2015 _____

Poděkování

Na tomto místě bych chtěl poděkovat vedoucímu práce doc. Mgr. Jiřímu Dvorskému, Ph.D. za vedení práce a ochotu, se kterou ke mně během studia přistupoval. Dále bych chtěl poděkovat kolegům z katedry geoinformatiky, kteří společně vytváří plodné pracovní i přátelské prostředí, a z nich především Mgr. Vítu Pászto a Mgr. Pavlu Tučkovi, Ph.D. za konzultace a rady během psaní práce a za podporu během celého studia.

Za poskytnutá data děkuji pracovníkům Státního zdravotního ústavu, Krajské hygienické stanici v Olomouci a Českého statistickému úřadu.

Poděkovat chci i své rodině, která mě podporuje a důvěřuje mi ve všech aktivitách a vytvořila mi skvělé prostředí, díky kterému jsem měl možnost studovat. Zvláštní poděkování a věnování patří především mé partnerce Ivi, bez jejíž podpory a schovívavosti by disertační práce nemohla vzniknout.

OBSAH

ÚVOD

Je součástí přirozené lidské povahy hledat povědomé vzory i ve zdánlivě nahodilých situacích. Příkladem mohou být souhvězdí na noční obloze, mraky připomínající beránky nebo spojování teček na papíře, ze kterých vznikne smysluplný obrázek, jsou-li spojeny ve správném pořadí. Je-li tato zvídavost přenesena do vědeckého prostředí, pak formalizované metody připomínající právě zmíněné spojování teček a hledání vzorů mohou poskytnout nástroje vhodné k identifikaci a kvantifikaci prostorových vzorů reálných jevů vyskytujících se v prostředí a případně pomoci k odhalení jejich podmiňujících faktorů (Waller a Gotway, 2004).

V případě zdravotnických či konkrétně epidemiologických dat, je nejčastěji analyzována trojice „čas-osoba-místo" (Elliott a Best, 1998). V epidemiologických studiích nedávné minulosti ovšem převažuje důraz na první dva prvky z trojice, tedy čas a osobu, zatímco geografický aspekt byl dlouho v pozadí (Ostfeld et al., 2005). Situace se však změnila s rostoucí dostupností (prostorových) datových sad, programových prostředků schopných analyzovat prostorová data a samozřejmě i s výkonem výpočetní techniky, která umožňuje urychlit zpracování i výpočetně náročných úkolů (Marek et al., 2012). Lze hovořit o nově vzniklém interdisciplinárním vědním oboru, jehož základem je aplikovaná statistika, epidemiologie a geovědní obory, který může nést různá označení jako geografická epidemiologie, prostorová epidemiologie (Elliott et al., 2000), lékařská geografie nebo dokonce geomedicína (Davenhall, 2012).

Současně žijeme v době, kdy jsou podpora zdraví a s ní související studie, stejně jako pronikání informačních technologií do běžného života i vědy, v popředí zájmu laické i odborné veřejnosti. Díky tomu se i původně úzce specializované vědecké obory, jakým epidemiologie bezpochyby byla, stávají více interdisciplinárními. Geografické informační systémy (GIS) se v důsledku schopnosti efektivně spravovat, analyzovat a zobrazovat prostorová data staly důležitým nástrojem v geovědních oborech a také všude tam, kde je potřeba nebo možnost zpracovávat geodata. Geografické informační systémy proto nejsou vnímány pouze jako nástroje pro tvorbu jednoduchých tematických map, ale jako plnohodnotný analytický nástroj, jehož silná stránka tkví právě ve schopnosti prostorových analýz a odhalování prostorových souvislostí (Rezaeian et al., 2007). Takto se geografickým informačním systémům mimo jiné povedlo proměnit i analýzy zdravotnických dat, které jsou v současnosti jedním z nejaktuálnějších témat v geovědách, což dokazuje i množství nově vznikajících publikací (Davenhall, 2012; Pfeiffer et al., 2008), specializovaných sekcí nebo odborných konferencí (např. Esri Health GIS Conference[1], GEOMED[2], Spatial Statistics Conference[3], …) nebo specializovaného software (Epi Info[4], SpaceStat[5], ClusterSeer[5], SaTScan[6], atd.).

[1] *http://www.esri.com/events/health/index.html*
[2] *http://www.shef.ac.uk/scharr/sections/ph/conferences/geomed2013/2013conf*
[3] *http://www.spatialstatisticsconference.com/*
[4] *http://wwwn.cdc.gov/epiinfo/*
[5] *http://www.biomedware.com/?module=Page&sID=software*
[6] *http://www.satscan.org/*

Předkládaná disertační práce představuje v ucelené formě možnosti aplikací prostorových a statistických analýz v souvislosti s epidemiologickými nálezovými daty. Zvolené teoretické postupy jsou prakticky využity v komplexní případové studii týkající se kampylobakteriózy, která je v současnosti nejrozšířenější bakteriální střevní infekcí v České republice (ÚZIS, 2013).

Hlavní motivací k vypracování této disertační práce byly zejména dvě skutečnosti. První skutečností je fakt, že ačkoliv je zdraví jedním z nejaktuálnějších témat současné společnosti, tak mu v geovědních disciplínách na našem území stále není věnována dostatečná pozornost. I přes existenci specializovaných pracovišť se k laické i odborné veřejnosti dostávají informace často v ne zcela vhodné či úplné podobě a prostorové analýzy a vazby mohou být často považovány za méně podstatné. Mapy či geovizualizace tak mohou být prezentovány a považovány jen za doplňující složku schopnou vyjádřit pouze jednoduché prvky reality, a nikoliv za komplexní výstup, který může poskytnout komplexní popis situace a pomoci pochopit celý proces. Druhou skutečností a hlavním motivem případové studie je kampylobakterióza. Infekční onemocnění, které je vůbec jedno z nejčastějších v rozvinutých zemích, Českou republiku nevyjímaje, a stále je mezi veřejností spíše neznámé. Velkou výzvou bylo také samotné zpracování rozsáhlých (prostorových) dat a operace s nimi.

Vzhledem k faktu, že je disertační práce zejména aplikačního rázu a dodržuje postupy vymezené prostorovou epidemiologií v kombinaci se základními i pokročilými statistickými metodami, a také že je práce tvořena na půdě katedry geoinformatiky, tak bylo nutné konzultovat témata a metody disertační práce s odborníky na epidemiologii a hygienu. Těmi byli zaměstnanci Státního zdravotního ústavu v Praze, Krajské hygienické stanice Olomouckého kraje a Národní referenční laboratoře pro využití GIS v ochraně a podpoře veřejného zdraví, které spadá pod Zdravotní ústav se sídlem v Ostravě.

1 CÍLE PRÁCE

Hlavním cílem disertační práce je **provedení komplexní prostorové analýzy epidemiologických dat s využitím v současnosti dostupných technologií z oblasti geoinformatiky a prostorové statistiky**. Veškeré analýzy budou provedeny v souladu se standardními metodami prostorové epidemiologie a principy geografických informačních systémů. Analyzovaná data týkající se infekčního onemocnění kampylobakteriózy pocházejí z databáze EPIDAT, která je výstupem stejnojmenného programu sloužícího k zajištění povinného hlášení, evidence a analýzy výskytu infekčních nemocí v České republice. Prostorově se data vztahují k území České republiky, časově jsou omezena lety 2008 a 2012.

Disertační práce ve svých dílčích částech hledá odpovědi na otázky týkající se časoprostorové distribuce výskytu kampylobakteriózy, jejich vztahu k případným vnějším environmentálním rizikovým faktorům a identifikace možných zdrojů nákazy. Právě odpovědi na dotazy typu *„Je možné nalézt prostorové trendy a prostorové vzory okolo místa výskytu?“*, *„Je dále možné tyto vzory popsat a kvantifikovat pomocí exaktních metod?“* nebo *„Je riziko onemocnění stejné v celém regionu nebo se prostorově liší?“* či *„Je možné nalézt vztah mezi průměrným počtem nemocných danou chorobou a vlivem okolí?“*, poskytují v doktorské práci představené nástroje, metody a postupy.

Postup řešení, který vede k naplnění hlavního cíle práce, je s ohledem na metody a postupy prostorové epidemiologie rozdělen na následující dílčí cíle (DC):

- Prvním dílčím cílem (DC1) je **mapování a celkový popis charakteristik výskytu kampylobakteriózy** v České republice v letech 2008—2012.
- Druhým dílčím cílem (DC2) je **průzkum, kvantifikace a vizualizace prostorových a časoprostorových vzorů** ve výskytu kampylobakteriózy v České republice v letech 2008—2012 a jejích vlastnostech.
- Třetím dílčím cílem (DC3) je **identifikace a analýza možných vztahů mezi výskytem onemocnění a vnějšími environmentálními, demografickými či socioekonomickými faktory** pomocí vícerozměrné statistiky a statistických modelů a následné hodnocení jejich přesnosti a využitelnosti pro predikci ohrožení vybraných územních jednotek kampylobakteriózou. Současně je cílem také klasifikace územních jednotek do skupin na základě jejich podobných vlastností a atributových vzorů souvisejících s výskytem onemocnění.
- Čtvrtým dílčím cílem (DC4) je **zhodnocení přítomnosti automatů na čerstvé mléko jako potenciálních bodových zdrojů nákazy** kampylobakteriózou v jejich okolí.
- Pátým dílčím cílem (DC5) je **převedení vybraných výsledků jednotlivých DC do podoby vhodné k další interaktivní exploraci v prostoru i čase**.

Výjimečnost disertační práce leží především ve dvou rovinách. První je mezioborový přesah studie a zapojení pokročilých metod prostorové statistiky (hodnocení prostorové autokorelace, geograficky vážené modelování, geoprofiling) a GIS v epidemiologii. Ačkoliv jsou metody a postupy prostorové epidemiologie známy a ve světě používány, tak na našem území se jejich aplikace vyskytují pouze v omezeném množství. Druhou rovinou je potom

měřítko analýz, které probíhají zejména na lokální úrovni (obcí či jejich částí). Dosavadní studie provedené v ČR totiž nejčastěji zahrnují pouze úrovně okresů.

Zvolené postupy, analýzy, metody a jejich výsledky byly průběžně konzultovány s odborníky na hygienu a epidemiologii (MUDr. Michael Vít, Ph.D., Státní zdravotnický ústav; MUDr. Růžena Halířová, Krajská hygienická stanice Olomouckého kraje) a GIS i prostorovou statistiku v této oblasti (Mgr. Hana Šlachtová, Ph.D., Národní referenční laboratoř pro využití GIS v ochraně a podpoře veřejného zdraví Zdravotního ústavu se sídlem v Ostravě).

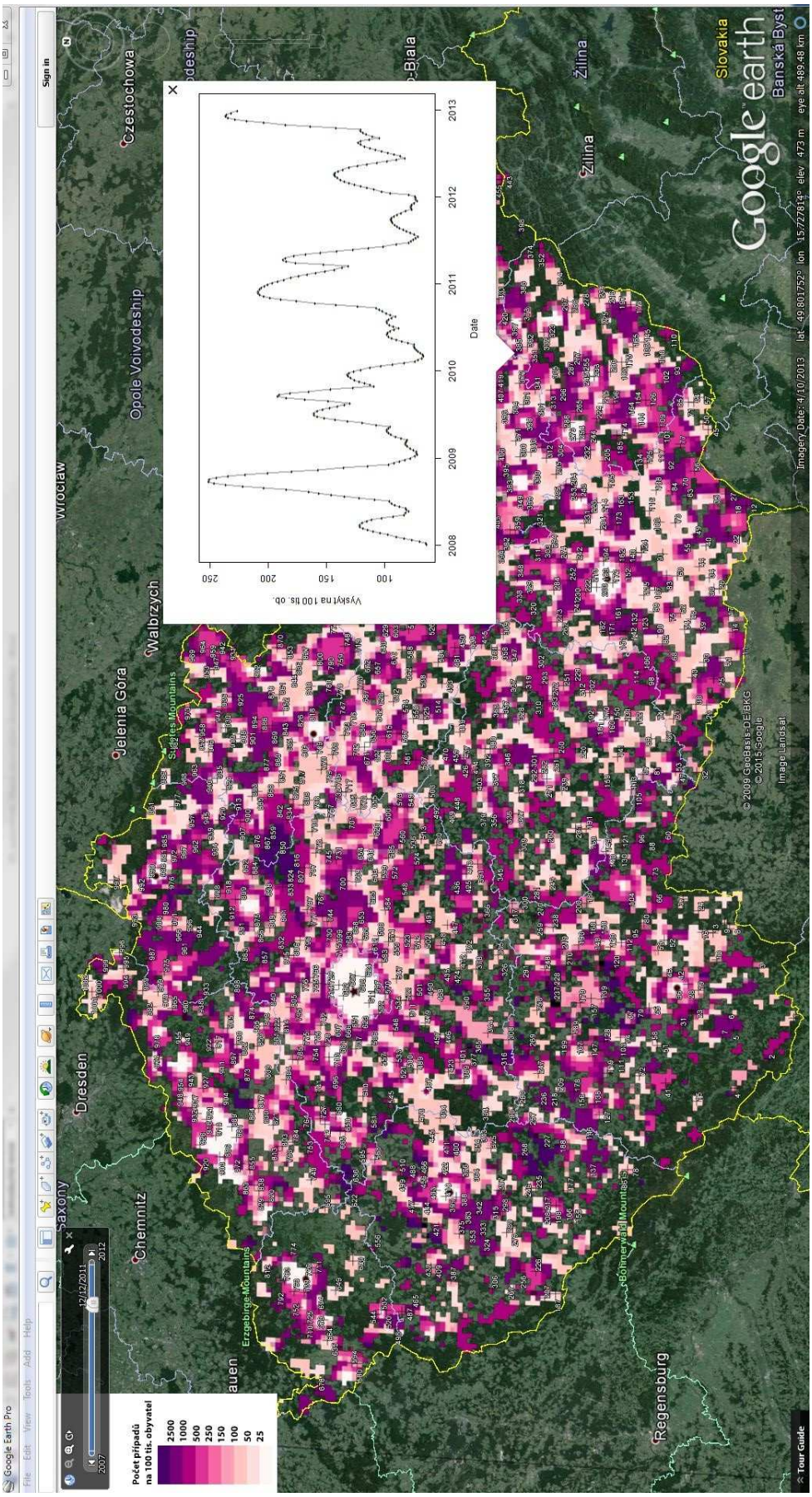
2 METODY A POSTUP ZPRACOVÁNÍ

Hlavním cílem předkládané disertační práce je komplexní prostorová analýza výskytu kampylobakteriózy v České republice v letech 2008—2012 a její podrobný popis. Jednotlivé metody a postup zpracování vychází z logického uspořádání dílčích cílů a je doplňujících aktivit. Je tak možné je shrnout do několika fází, které jsou v souladu s teorií a metodami prostorové epidemiologie, principy GIS a prostorové statistiky (Lawson et al., 1999; Bailey, 2001). Základem pro vhodně provedené analýzy je literární rešerše tématu, díky které je možné se blíže seznámit s danou problematikou, specifickou terminologií (např. morbidita, incidence, prevalence, SMR, riziko), zavedenými postupy, možnostmi a výzvami tématu.

Postup tvorby disertační práce začíná sérií základních statistických analýz, společně s analýzou časové řady výskytu nemocí a základních indexů nemocnosti. Tento krok si klade za cíl přinést všeobecný přehled o průběhu výskytu onemocnění v jednotlivých letech a jeho charakteristikách. Další postup se částečně drží klasifikace prostorových analýz v epidemiologii, kterou zmiňují např. Bailey (2001) nebo Elliott a Wartenberg (2004):

- Mapování nemocí – tvorba map výskytu nemocí, základních epidemiologických charakteristik a standardizovaných měr, rizika onemocnění, bayesovké vyhlazování, vizuální analýza;
- průzkum prostorových vzorů nemocí na území důrazem na odhalení procesů shlukování – výběr sousedství a rozsahu analýz, identifikace typu procesů v území, testy shlukování na globální (myšleno ve smyslu celého území) i lokální úrovni (identifikace shluků), ESDA, časoprostorová explorace;
- prostorová korelační analýza vztahu mezi výskytem vybrané nemoci a rizikovými faktory prostředí – modelování vztahů mezi nemocností a vybranými environmentálními a antropogenními charakteristikami za pomoci (prostorových) regresních modelů, geostatistických metod, příp. simulací a vícerozměrné statistiky.

Tento stručný seznam a jeho jednotlivé kroky jsou ovšem doplněny o další specifické postupy, které jsou také přidanou hodnotou předkládané práce. V rámci mapování nemocí je představeno využití časoprostorového krigingu, průzkum prostorových vzorů je obohacen o složku času a prostorová analýza možného vlivu environmentálních faktorů je doplněna o klasifikaci území na základě podobných vlastností morbidity obyvatelstva a vybraných vlastností území za pomoci vícerozměrné statistiky. Dále je využito geografické profilování (tzv. geoprofiling) jako metoda identifikace potenciálních zdrojů nákazy a interaktivní vizualizace výsledků pomocí jazyka Keyhole Markup Language (KML) a následná geovizuální analýza v prostředí Google Earth. Schéma postupu řešení jednotlivých dílčích cílů a jejich vzájemných interakcí je zobrazeno na Obrázku 1.



Obr. 1 Schéma postupu řešení dílčích cílů disertační práce

2.1 Použitá data, metody a postup řešení cílů práce

2.1.1 Data a jejich příprava

Základní podmínkou pro úspěšné provedení prostorových analýz, vizualizaci jejich výsledků a vyhodnocení závěrů jsou data, jejichž kvalita je stěžejním faktorem pro rozsah, podrobnost a spolehlivost studie. Základní datovou sadou v disertační práci jsou data pocházející z databáze EPIDAT (Epidemiologická databáze), která byla poskytnuta Státním zdravotním ústavem v Praze. Databáze obsahuje kompletní záznamy o výskytu kampylobakteriózy na celém území České republiky v letech 2008—2012. Tato datová sada je anonymizovaná, což znamená, že není známo jméno pacienta s hlášenou nemocí a ani jeho přesná adresa, takže lokalizace je možná do úrovně uliční sítě. Kromě věku a pohlaví pacienta jsou v databázi obsaženy i další atributy jako datum hlášení, zaměstnání, diagnóza, národnost apod. Datová sada obsahuje téměř 100 tisíc záznamů se 78 atributy s různou úplností a kvalitou vyplnění, které je nutné zkontrolovat, vyčistit a opravit. Zkontrolovaným datům byla na základě místa nakažení a/nebo místa bydliště přiřazena poloha pomocí geokódování s využitím API poskytovaným mapovým portálem Mapy.cz.

Dalšími datovými sadami, které byly využity v analýzách, jsou zejména ty týkající se demografické struktury obyvatelstva a jeho ekonomické aktivity. Příkladem jsou např. data ze Sčítání lidí, domů a bytů 2001 a 2011, Registr emisí a zdrojů znečištění ovzduší atd. Dále byla použita data fyzicko-geografická jako digitální modely terénu, klimatické jevy (průměrná teplota vzduch, atmosférické srážky), využití půdy, radonové riziko, znečištění ovzduší apod. nebo data odvozená (např. vzdálenost od vodních toků nebo silnic, apod.). Pro analýzy a vizualizace byla použita data administrativních jednotek, případně pravidelné sítě pro agregaci dat.

2.1.2 Mapování a charakteristika výskytu kampylobakteriózy v České republice v letech 2008—2012 [DC1]

Hlavním úkolem DC1 je s pomocí základních statistických metod a analýzy časových řad poskytnout celkový přehled o výskytu případů kampylobakteriózy a jejich základních charakteristikách na území České republiky ve zkoumaných letech. V kapitole jsou přehledně zobrazeny výčty a statistiky nemocných s ohledem na jejich věk, pohlaví či zaměstnání a to jak v numerické, tak i v grafické podobě.

Současně s tím je provedeno mapování výskytu nemoci v podobě map měr morbidity (nemocnosti). Konkrétně je mapována incidence jako relativní míra onemocnění ve zvolené územní jednotce a standardizovaná incidence, kterou je možno považovat za míru relativního rizika postižení zkoumaného území onemocněním (Bivand et al., 2008). Při výpočtu měr morbidity bylo využito především nepřímé věkové standardizace. Vzhledem k malému rozsahu dat u některých mapovaných jednotek jsou kromě map standardizovaných měr morbidity vytvořeny také jejich shlazené ekvivalenty. Pro výsledné shlazení jsou využity metody globálního empirického Bayesova odhadu, kde je pro vyhlazování měr v jednotlivých areálech aplikován konstantní průměr a rozptyl vycházející z negativně binomického rozdělení. Dále je využito také lokálního (adaptivního) empirického Bayesova vyhlazování,

kde jsou průměrná hodnota a rozptyl využité pro vyhlazování odhadu měr morbidity definovány na základě sousedství 1. řádu typu královna.

Poslední částí DC1 je vytvoření a vizualizace souvislého povrchu incidence v týdenních intervalech, který je oproštěn od zatížení administrativních hranic obcí. Základem pro tvorbu spojitého povrchu je grid hustoty zalidnění a agregovaná data výskytu onemocnění. K vytvoření spojitého povrchu je využito časoprostorového krigingu, který na rozdíl od tradiční varianty krigingu umožňuje zohlednit pro tvorbu spojitého povrchu nejen prostorovou strukturu (varianci) jevu, ale také strukturu změny jevu v čase i v čase a prostoru současně (Pebesma a Gräler, 2014). Zmíněné vztahy jsou vyjádřeny časoprostorovým variogramem, který tvoří základ pro interpolaci spojitého povrchu incidence síti o dvoukilometrové hraně buňky. Výsledný povrch je pro účely vizualizace zobrazen pouze pro obydlená území.

Výsledky statistické analýzy a mapování slouží jako základ pro stanovení hypotéz týkajících prostorových a časoprostorových vzorů v území, které jsou konkrétně kvantifikovány a vyjádřeny numericky i vizuálně v DC2.

2.1.3 Podobnosti onemocnění v prostoru a čase [DC2]

Vizuálním vyhodnocením DC1 je zjištěno, že v rámci České republiky existovaly v letech 2008—2012 spojité oblasti se zvýšeným výskytem kampylobakteriózy a současně i oblasti, které byly touto nemocí téměř nebo zcela nepostihnuty. Druhý dílčí cíl (DC2) přímo navazuje na tato zjištění a snaží se tento jev kvantifikovat, kvantifikaci vizualizovat a následně i popsat. K řešení DC2 je využito několika hlavních metod a postupů prostorové a časoprostorové statistiky.

V prvním kroku je za pomoci metod průzkumu prostorové autokorelace zkoumán pouze prostorový vzor. K tomuto je využito lokálních indikátorů prostorové asociace (LISA) a především lokálního Moranova I (Anselin, 1995), založeném na základě sousedství 1. řádu typu královna. Pro srovnání je použita jak základní varianta tohoto indexu, tak i varianta využívající k lepšímu odhadu skutečného prostorového vzoru jevu lokálního empirického Bayesova vyhlazování v kombinaci s randomizací a určením významnosti výsledků pomocí permutačních testů. Vizualizovány jsou oblasti na hladině významnosti do 5 %. Výsledkem jsou mapy identifikující oblasti shluků vysokých nebo nízkých hodnot, tedy s vyšším/nižším výskytem onemocnění, a pak tzv. *„outliers"* neboli území, která vybočují z trendu v jejich okolí. Tato vybočující území představují buď oblasti nízkých hodnot v blízkosti více nemocí postižených oblastí, nebo naopak území s výskytem vyšším poblíž oblastí, které nejsou výrazně zasáhnuty.

V dalším navazujícím kroku DC2 je k průzkumu prostorového vzoru onemocnění přiřazen i rozměr času. Jednotlivé případy nákazy jsou agregovány do územní jednotky (městských částí) v týdenních intervalech a současně rozlišením věku a pohlaví. Agregovaná data jsou následně srovnávána s podobně strukturovanými demografickými údaji území (počet obyvatel, věková struktura a struktura pohlaví) s pomocí metody *spatio-temporal scan statistics* - SaTScan (Kulldorff et al., 2005). Metoda využívá k odhalení shluků s podobnou strukturou simultánního skenování území v čase i prostoru. Výsledky jsou podbné jako v případě LISA, ovšem s tím rozdílem že je současně indikována primární shluk i shluky sekundární.

V případě této analýzy byla zvolena maximální velikost shluku odpovídající 3 % celkové populace a maximálním dobou trvání 50 % časového rozsahu dat, ovšem s výjimkou shluků vyskytujících se po celé období. Analýza identifikovala tři desítky shluků (14 shluků vysokých hodnot a 16 nízkých).

2.1.4 Analýza vztahů mezi onemocněním a vnějšími faktory prostředí [DC3]

Předchozí dílčí cíle DC1 a DC2 odhalily nenáhodný proces shlukování obcí s vyšší incidencí kampylobakteriózy a lze tak stanovit předpoklad, že v daných územích bude existovat i environmentální, demografický či socioekonomický faktor, který může přispívat ke zvýšené incidenci a relativnímu riziku. V první fázi třetího dílčího cíle jsou z velkého množství zdrojů nashromážděna a sjednocena data, ze kterých je vytvořena datová sada vlastností území obcí obsahující více než stovku charakteristik ke každé obci České republiky. Pro účely modelování však nejsou takto rozsáhlá data vhodná ani potřebná a postupně byla datová sada pomocí korelační analýzy, logických úvah, dostupných literárních zdrojů a konzultací redukována až na 11 spolu téměř nekorelujících charakteristik. Výsledná sada byla ještě pomocí analýzy hlavních komponent redukována na pět nově vytvořených proměnných, které slouží jako vstup do modelů. Jako závislá proměnná, případně jako klasifikační schéma je využito relativní riziko (SIR - standardized incidence ratio), podle kterého jsou obce rozděleny do čtyř skupin (nulové, minimální, průměrné a vysoké riziko).

Analýza vztahů mezi onemocněním a vnějšími faktory prostředí je postupně realizována pomocí metod vícerozměrného regresního modelování, prostorových regresních modelů a metod strojového učení a data miningu. Jednotlivé metody a modely jsou hodnoceny pomocí ROC křivek a AUC[7]. Vzhledem k nízké úspěšnosti většiny modelů je pro srovnání aplikován negativní binomický regresní model pro případy s velkým množstvím nul, kterým je modelován průměrný počet případů v obcích.

Předchozí metody prostorového shlukování představené v DC2 jsou založeny pouze na incidenci v jednotlivých částech České republiky a její geografické podobnosti s okolními územními jednotkami (prostorové autokorelaci). V poslední části DC3 je navíc zkoumána podobnost v atributové rovině. Na základě analýzy důležitosti a vhodnosti doplňkových dat pro modelování relativního rizika onemocnění v obcích jsou vybrány potenciálně nejdůležitější doplňková data a také míry morbidity. Tato data slouží jako vstupní údaje pro shlukovou analýzu. Jejím cílem je identifikovat skupiny obcí s podobnými vlastnostmi v rámci skupiny tak, aby se od sebe jednotlivé skupiny současně co nejvíce odlišovaly (Hebák et al., 2005b). Pro účely nalezení podobných datových struktur je v práci využita kombinace samoučící se neuronové sítě, konkrétně Kohonenovy samoorganizační mapy (SOM) a hierarchického shlukování. Cílem SOM je vytvořit mapu skupin co nejpodobnějších objektů, které je možné vizualizovat ve 2D prostoru a současně co nejvíce zachovat jejich topologii (Kohonen, 1982, 2001). V procesu shlukování může komplexní výstup ze SOM doplnit výpočetně jednodušší míry podobnosti, které tvoří základ hierarchického shlukování. Buňky SOM tak mohou být kategorizovány a následně ohodnoceny i územní jednotky do buněk

[7] Receiver operator characteristic / Area under curve

spadající. Tímto postupem je vytvořeno, popsáno a vizualizováno 7 hlavních skupin územních jednotek podle environmentálních a socioekonomických vlastností a vlastností spojených s onemocněním kampylobakteriózou.

2.1.5 Identifikace možných zdrojů infekce [DC4]

Čtvrtý dílčí cíl případové studie měl jako hlavní motiv identifikaci možných zdrojů nákazy. Po konzultaci s MUDr. Michaelem Vítem, Ph.D. (bývalým hlavním hygienikem ČR a v současnosti vedoucím Centra hygieny práce a pracovního lékařství Státního zdravotního ústavu) byly za možné zdroje nákazy zvoleny automaty na prodej čerstvého mléka (tzv. mlékomaty), které se začaly objevovat právě mezi roky 2008—2012. Informace o umístění mlékomatů pochází z několika zdrojů – z portálu registrovaných subjektů Státní veterinární správy[8], kde ovšem nejsou veškerá historická data, ale pouze v současnosti funkční mlékomaty a dále z Venkovského fóra[9] nebo přímo od provozovatelů (např. společnost TOKO[10]). Celkem bylo testováno 267 automatů na prodej čerstvého mléka, které byly umístěny v letech 2008—2012.

Pro identifikaci mlékomatů jako potenciálního zdroje nákazy byla zvolena metoda geografického profilování, která byla původně využívána zejména v kriminalistice, ale v posledních letech je úspěšně aplikována také v biologii a především v prostorové epidemiologii jako nástroj sloužící k vyhledávání a hodnocení mnohonásobných zdrojů nákazy (Le Comber et al., 2011; Verity et al., 2014). Vyhodnocování je založeno na Dirichletově modelu pro smíšené procesy (DPM - Dirichlet process mixture model) a jde tedy o bayesovské modelování shluků s pomocí Markovových řetězců (MCMC). Vyhodnocování jednotlivých bodových zdrojů následně probíhá na základě tzv. hitscore, které udává velikost oblasti, kterou je potřeba prohledat k nalezení zdroje. Vzhledem k výpočetní náročnosti operace bylo nutné nejdříve redukovat datovou sadu míst nákazy na okolí do 15 km od mlékomatu a dále území rozdělit do 18 oblastí. Pomocí geografického profilování bylo zjištěno, že množství mlékomatů mohlo být v průběhu svého fungování zdrojem nákazy kampylobakteriózou. Současně byl ovšem zjištěn i nedostatek modelu, kdy jsou oblasti s nízkým výskytem případů onemocnění nadhodnocovány jako rizikové. Reálný odhad je tedy nižší a pohybuje se kolem 10 % zkoumaných mlékomatů.

2.1.6 Prezentace a průzkum dat v čase i prostoru [DC5]

Výstupem prostorových analýz bývají nejčastěji statické mapy a geovizualizace, které sice umožnují vhodným způsobem sdělit jejich výsledky, ale problémem může být jejich další využitelnost, případně kombinace s dalšími zjištěními. Náplní posledního dílčího cíle (DC5) je zpřístupnění nejdůležitějších výsledků dříve definovaných v DC1—DC4 v podobě, která umožní jejich snadný interaktivní průzkum jak v prostoru, tak ideálně i v čase s pomocí časových značek. Výsledky analýz jsou upraveny do podoby vhodné k interaktivnímu zobrazení a z původních formátů exportovány do formátu KML[11] (Google,

[8] *http://eagri.cz/public/app/svs_pub/subjekty/mleko.php*
[9] *http://www.venkovskeforum.cz/mlekomaty*
[10] *http://www.tmleko.cz/cerstve-mleko/seznam-automatu/*
[11] *Keyhole Markup Language*

2009), který je standardem pro výměnu a zobrazení geografických dat (Open Geospatial Consortium, 2008). Jako nástroj vhodný k další geovizuální analýze je zvolen Google Earth, který je v současnosti vůbec nejrozšířenější prohlížečkou prostorových dat (Hengl, 2007) a jeho ovládání je natolik známé a intuitivní, že je hojně využíván jak laiky, tak i odborníky. Jeho nevýhodou je ovšem nutná příprava dat v prostředí GIS (extenze *Export to KML* pro ArcMap 10.1 nebo pluginu *GEarthView* pro QGIS), případně s využitím vhodných balíků statistického software **R** (např. *plotKML* a *raster*). Takto připravená data umožňují zmíněné komplexní interaktivní časoprostorové hodnocení pomocí geovizuální analýzy. Pro interaktivní zobrazení jsou zvoleny vizualizace agregovaných nálezových dat, spojitého povrchu týdenní incidence, statistik časoprostorového shlukování, roční incidence a geografického profilování.

2.2 Použité programové vybavení

V průběhu tvorby disertační práce byla využita řada programových prostředků, které byly vybírány dle vhodnosti, použitelnosti pro konkrétní úkol a také dostupnosti. Mezi základní analytické nástroje ovšem budou patřit programy:

- **R**[12] – open source software a GUI pro statistické výpočty a modelování, který díky velkému množství rozšiřujících balíčků a možností progamování funkcí umožňuje analyzovat jevy pomocí statistiky, prostorových metod, epidemiologických postupů, časových řad apod., jako IDE k **R** bylo použito RStudio[13];
- GeoDa[14] – nástroj sloužící pro základní úlohy prostorového modelování a průzkumové analýzy prostorových dat (regresní modely, vícerozměrná statistika, prostorová autokorelace);
- SaTScan[15] – analýza shlukování v čase i prostoru;
- ArcGIS for Desktop 10.1[16] a QGIS 2.6[17] – nástroje pro tvorbu mapových výstupu, operace s vektorovými i rastrovými daty a prostorové analýzy (prostorové shlukování, LISA, …);
- Podpůrný software – nástroje pro psaní a editaci textu a tabelárních dat (kancelářská sada Microsoft Office), úprava a tvorba grafiky a obrázků (balík nástrojů CorelDraw), apod.

[12] *http://www.r-project.org/*
[13] *http://www.rstudio.com/*
[14] *https://geodacenter.asu.edu/projects/opengeoda*
[15] *http://www.satscan.org/*
[16] *http://www.arcgis.com/*
[17] *http://www.qgis.org/*

3 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY

Ačkoliv je prostorová epidemiologie relativně novým oborem, její koncept nový není. Souvislostí mezi výskytem nemocí a environmentálními podmínkami si již ve starověku všimnul např. Hippokrates nebo později John Snow, který je díky jeho geografické analýze cholery v Londýně z roku 1854 považován za jednoho ze zakladatelů moderní epidemiologie (Meade a Emch, 2010). Právě tento okamžik byl jedním z důležitých momentů historie epidemiologických studií, který ukázal potenciál využití geografických analýz pro popis rozšíření nemocí a umožnil, aby se jejich prostorové modelování dostalo mezi všeobecně uznávané operace. Samotný název oboru prostorové epidemiologie tak vyjadřuje dynamickou povahu teorie a analytických metod, které hledají a zkoumají prostorové (příp. časoprostorové) vzory objevujících se nemocí a příčin mortality (Waller a Gotway, 2004).

Společně s představením aplikací standardních metod prostorové analýzy a prostorové statistiky ve zdravotnických studiích je prostorová epidemiologie oblíbeným a v literatuře hojně diskutovaným tématem poslední doby (Ostfeld et al., 2005). Rozšiřující se nasazení metod prostorové epidemiologie je způsobeno zejména možnostmi současných geografických informačních systémů společně s vlastnostmi databází, kde jsou zdravotnická a epidemiologická data ukládána. V důsledku to vše umožňuje relativně snadné a rychlé prostorové hodnocení dat.

Epidemiologické a medicínské studie zahrnující prostorová data využívají velkou šíři možných prostorových metod, od základních technik mapování výskytu nemoci v administrativních jednotkách (EU, 2009) a její lokalizaci s využitím adres pacientů nebo souřadnic (Rushton, 2003) až po pokročilejší metody predikce a rozšíření choroby (Goovaerts, 2005), analýzu prostorových vzorů (Gatrell et al., 1996), rizik, vizualizaci bayesovských a jiných modelů (Aamodt et al., 2006) nebo kombinaci s daty dálkového průzkumu Země (Tatem et al., 2004). Je vhodné si také uvědomit, že se tento druh studií může zaměřovat na různé aspekty zdravotnických dat, jako jsou dostupnost zdravotní péče (Delmelle et al., 2010), demografická struktura pacientů a jejich umístění (Oakes, 2004) nebo komplexní analýza konkrétní nemoci a jejího výskytu, vlivu faktorů životního prostředí nebo socioekonomických faktorů a následné pokusy o kvantifikaci těchto vztahů ve formě modelu vhodných pro predikci (Cromley, 2003).

Z českých autorů se problematikou výzkumu zdravotnických dat, zdraví a jejich propojením s geovědami (mimo specializovaná pracoviště) dlouhodobě zabývá např. doc. Dzúrová a doc. Spilková z Katedry sociální geografie a regionálního rozvoje na Univerzitě Karlově v Praze, které se zabývají především výzkumem demografických a sociálně geografických aspektů zdraví (Dzúrová et al., 2006 a 2010; Spilková et al., 2011 a 2014). Podobnými aspekty se zabývají také s nimi často spolupracující sociální epidemiologové dr. Pikhart a prof. Bobák (Bobak et al., 1998), kteří působí na Katedře epidemiologie a veřejného zdraví na University College London. Využitím geoinformačních technologií a kartografických metod v tomto tématu se v rámci své odborné činnosti a vědecko-

výzkumných projektů zabývali také Konečný, Kubíček, Geryk a kol. v projektu MediCarto[18] (2007-2009) případně Konečný, Geryk, Čada a kol. v rámci projektu Visual Health[19] (2008-2009). Opomenut by neměl být ani společný projekt Jihočeské univerzity v Českých Budějovicích a Univerzity Ludvíka Maxmiliána v Mnichově s názvem Klíšťata a jimi přenášená onemocnění v podmínkách jižních Čech a Bavorska[20] (Hönig et al., 2011), který je pravděpodobně nejznámějším počinem týkajícím se prostorové epidemiologie u nás. Dále lze zmínit, již bez uvedení konkrétních prací, aktivity Národní referenční laboratoře pro využití GIS v ochraně a podpoře veřejného zdraví při Zdravotním ústavu v Ostravě nebo aktivity MUDr. Radima Šráma či dr. Jiřího Preisse.

3.1 Data, jejich typologie a konfidence

Pro úspěšnou studii jsou kromě vhodně zvolených metod klíčovým prvkem také data. Manipulace se zdravotnickými daty s sebou navíc nese specifika spojená s faktem, že se jedná o data osobního charakteru, takže jsou často poskytována v agregované či anonymizované podobě nebo jako neúplná (myšleno ve smyslu chybějících přesných adresních záznamů). Takto upravená data si také žádají specifické metody analytických postupů a je současně potřeba brát v úvahu určitou nepřesnost spojenou s charakterem dat.

Jako základní klasifikaci lze využít rozdělení na data bodová s přímou identifikací místa nákazy nebo bydliště a data četností výskytu, která jsou agregovaná např. na základě zvolené administrativní jednotky nebo pravidelné sítě. Podrobnější dělení poskytuje dále např. Bailey (2001), který představuje i další skupiny dat v závislosti na jejich původu a účelu.

- Data v nepravidelné mřížce – jde o data, která jsou agregovaná na úroveň administrativní či jiné nepravidelné jednotky nebo vyjadřují průměrnou hodnotu jevu v jednotce. Tematicky mohou vyjadřovat četnosti nemocí v jednotce, pravděpodobnost ohrožení obyvatelstva, socioekonomické míry nebo environmentální pozorování apod.;
- data událostí – vyjádřena je poloha (většinou bydliště) jednotlivých případů výskytu nemoci nebo členů vhodné kontrolní skupiny. Současně mohou obsahovat informace o dalších proměnných vztahujících se k tomuto místu;
- geostatistická data – bodová měření (nejčastěji environmentálního charakteru);
- data v pravidelné mřížce - data, která jsou agregovaná do pravidelné mřížky nebo vyjadřují průměrnou hodnotu jevu v buňce.

Konkrétní příklady možných doplňujících environmentálních prostorových dat (Tabulka 1) využitelných pro prostorově epidemiologickou analýzu přehledně uvádí Pfeiffer et al. (2008), který čerpá z několika dalších zdrojů. Podstatnou roli dále hrají zejména primární data sbíraná přímo zdravotními institucemi, která nejsou v přehledu přímo zmíněna. Tabulka 1 však neobsahuje kompletní výčet možných dat (ten je ostatně těžko možný), která vstupují do souvislosti se vznikem nemocí nebo zdravotních problémů obyvatelstva (nebo fauny), ale poskytuje alespoň částečný přehled, který umožňuje vytvořit si představu o množství aktuálně dostupných dat pro vybranou studii.

[18] Kartografická vizualizace a modelování současných trendů vývoje zdravotního stavu a zdravotní péče v ČR

[19] Vizualizace zdravotních dat pro podporu interdisciplinárního vzdělávání a vztahů s veřejností

[20] Mapový portál projektu je dostupný na adrese *gis.vsb.cz/klistata*

Tab. 1 Příklady prostorových dat použitelných pro epidemiologické studie (upraveno podle Pfeiffer et al. (2008))

Obecný druh dat	Proměnné
Poloha	Zeměpisná délka a šířka
Antropogenní	Vzdálenost od cesty Vzdálenost od městského osvětlení
Demografická	Celková populace Hustota zalidnění
Topografická	Nadmořská výška
Pokryv půdy	Využití půdy Půdní pokryv NDVI
Teplota	Teplota zemského povrchu Teplota vzduchu Odrazivost ve středním IR pásmu
Voda a vlhkost	Nasycení vzduchu vodními parami Vzdálenost k vodním tokům Doba trvání studených oblak Potenciální evapotranspirace
Klimatické podmínky	Modelování vegetačních podmínek

3.1.1 Vybraní poskytovatelé zdravotnických dat v ČR

Stejně jako je aktuálním tématem zdraví člověka, tak je otázkou existence a dostupnost zdravotnických dat, a to jak dat primárních pro možné provádění dalších analýz a šetření, tak i dat již zpracovaných ve formách statistických přehledů, zpráv či ročenek. Hlavním problémem tak často není nedostatek či neexistence dat, ale spíše jejich dostupnost ve vhodné a dostatečně podrobné formě (Marek et al., 2013a). Vynikající přehled mezinárodních i domácích zdrojů zdravotnických dat, včetně dat prostorových, a zhodnocení jejich potenciálů uvádí ve svých pracích např. Štampach (2010 a 2013) nebo Štampach a Geryk (2012). Zmíněni by měli být alespoň nejvýznamnější mezinárodní poskytovatelé a správci zdravotních dat – Světová zdravotnická organizace (WHO), Eurostat (statistický úřad EU) nebo Organizace pro hospodářskou spolupráci a rozvoj (OECD), kteří spravují či sbírají velké množství komplexních statistických dat včetně těch s tématikou zdraví a poskytují je v digitálních přehledech (Eurostat, 2009) či interaktivních mapách (Eurostat, 2014; OECD, 2014). Opomenuta by neměla být ani směrnice INSPIRE[21], jejíž třetí příloha (Annex III) nazvaná Lidské zdraví a bezpečnost se zabývá tématy od alergií, přes výskyt onemocnění až po kvalitu životního prostředí (INSPIRE: Thematic Working Group Human Health and Safety, 2013).

Jedním z národních poskytovatelů informací týkajících se zdravotnictví a zdraví obyvatelstva je **Český statistický úřad**[22] (ČSÚ), který je jedním z ústředních orgánů státní správy České republiky. Byl zřízen dne 8. ledna 1969 zákonem č. 2/1969 Sb., o zřízení ministerstev a jiných ústředních orgánů státní správy a zajišťuje výkon státní statistické služby

[21] Infrastructure for Spatial Information in Europe

[22] *http://www.czso.cz/*

(ČSÚ, 2015). Dle zákona č. 89/1995 Sb., o státní statistické službě, je státní statistická služba činnost, která zahrnuje získávání údajů, vytváření statistických informací o sociálním, ekonomickém, demografickém a ekologickém vývoji České republiky a jejích jednotlivých částí, poskytování statistických informací a jejich zveřejňování. ČSÚ sbírá a spravuje mimo jiné informace o zdraví a zdravotnické problematice, které jsou ve veřejné databázi ČSÚ rozmístěny v rámci několika témat (demografie, zdravotnické služby, ekonomika a ICT), zejména pak v oblasti *Práce a sociálních statistik*.

Dalším významným poskytovatelem a správcem zdravotnických dat je **Státní zdravotní ústav**[23] (SZÚ), který je příspěvkovou organizací Ministerstva zdravotnictví ČR. SZÚ je zdravotnickým zařízením pro základní preventivní obory, jakými jsou hygiena, epidemiologie, mikrobiologie a pracovní lékařství. SZÚ je oprávněn zpracovávat za účelem přípravy podkladů pro tvorbu státní zdravotní politiky a sledování dlouhodobých trendů výskytu infekčních a jiných hromadně se vyskytujících onemocnění údaje o zdraví fyzických osob v souvislosti s předcházením vzniku a šíření infekčních onemocnění, ohrožení nemocí z povolání a jiných poškození zdraví z práce, o expozici fyzických osob škodlivinám v pracovním a životním prostředí a o epidemiologii drogových závislostí a předávat je orgánům ochrany veřejného zdraví (SZÚ ČR, 2014). SZÚ poskytuje aktuální data z dříve zmíněných oblastí formou specializovaných časopisů, statistických publikací a ročenek. Veřejně dostupné informace jsou nejčastěji publikovány jako národní statistiky či pro úroveň krajů. Mimo jiné je SZÚ správcem Národního registru nemocí z povolání.

Zřejmě nejvýznamnějším poskytovatelem a správcem zdravotnických dat v České republice je **Ústav zdravotnických informací a statistiky České republiky**[24] (ÚZIS). ÚZIS je organizační složkou státu, jejímž zřizovatelem je Ministerstvo zdravotnictví ČR, které ústav pověřilo správou Národního zdravotnického informačního systému (NZIS). ÚZIS je součástí státní statistické služby (na základě kompetenčního zákona) a tuto činnost vykonává podle zákona č. 89/1995 Sb., o státní statistické službě, ve znění pozdějších předpisů. Spolupracuje s orgány státní statistické služby, především s Českým statistickým úřadem, zajišťuje vazby mezi NZIS a jednotlivými poskytovateli zdravotních služeb a spolupracuje s provozovateli informačních systémů jiných organizací v resortu i mimo něj. ÚZIS spolupracuje s asociacemi nemocnic, sdruženími lékařů, odbornými lékařskými společnostmi, zdravotními pojišťovnami a dalšími organizacemi zejména na zpřesňování obsahu NZIS a využití sbíraných dat. V oblasti zdravotnické statistiky na mezinárodní úrovni spolupracuje Ústav zejména s organizacemi WHO, OECD, OSN, Eurostat a dalšími. Ústav je předkladatelem oficiálních informací z NZIS za Českou republiku (ÚZIS ČR, 2014). ÚZIS spravuje a zpracovává značné množství dat a registrů (např. Národní registry rodiček, novorozenců, potratů, vrozených vad a několika dalších) a to i za menší územní jednotky – obce, správní obvody ORP a okresy. Veřejně je však dostupný jen zlomek těchto dat, zejména v systému prezentace dat, katalogu publikací nebo v několika zdravotnických registrech (Štampach, 2010).

[23] *http://www.szu.cz/*
[24] *http://www.uzis.cz/*

Koordinační středisko pro rezortní zdravotnické informační systémy[25] (KSRZIS) je organizační složkou státu v přímé řídící působnosti Ministerstva zdravotnictví, které vzniklo na základě Opatření o zřízení ke dni 1. ledna 2004 (KSRZIS, 2010). KSRZIS zajišťuje zavádění, rozvoj a provoz informačních systémů. KSRZIS má za úkol zajišťovat metodiku a provoz systémů (datové, komunikační a číselníkové standardy) navazující na informační systém veřejné správy. Zahrnuje v sobě též národní zdravotní registry pro sledování pacientů s vybranými, společensky závažnými chorobami. Shromažďuje a zpracovává data o zdravotním stavu obyvatelstva, o zdravotnických zařízeních, jejich činnosti a ekonomice za účelem řízení, pro výzkum a statistiku a k usměrňování poskytování zdravotní péče. KSRZIS z těchto úkolů zajišťuje oblast národních zdravotních registrů a zaměřuje se též na specializované informační systémy v oblasti zdravotnictví. Mimo zpracování registrů nad nimi středisko provozuje i z velké části neveřejný geoportál (ARCDATA Praha, 2013). Je vhodné podotknout, že KSRZIS je pouze zpracovatelem registrů, jejichž správou a garantováním jsou pověřeny příslušné organizace (např. ÚZIS). Hlavní skupiny registrů a jednotlivé registry jsou (KSRZIS, 2010):

Národní zdravotní registry

- Národní registr kloubních náhrad (NRKN)
- Národní onkologický registr (NOR)
- Národní registr cévní chirurgie (NRCCH)
- Národní registr kardiovaskulárních intervencí (NRKI)
- Národní registr osob nesouhlasících s posmrtným odběrem tkání a orgánů (NROD)
- Národní kardiochirurgický registr (NKCHR)

Registry hygienické služby

- Registr kosmetických prostředků (KOPR)
- Registr chemických látek a prostředků CHLAP)
- Registr akutních respiračních infekcí (ARI)
- Úložiště dat pro EPIDAT (povinné hlášení, evidence a analýza výskytu infekčních nemocí v České republice).
- Registr očkovacích látek (OČKO)
- Informační systém Pandemie
- Pitná voda (IS PiVo)
- Informační systém Rozhodnutí hlavního hygienika (IS RoHy)
- Registr pohlavních nemocí (RPN)
- Registr hygieny dětí a mladistvých (HDM)
- Registr hygieny výživy (HVY)
- Registr kategorizace prací (KaPr)
- Registr předmětů běžného užívání (IS PBU)
- Registr tuberkulózy (RTBC)

Specializované zdravotnické informační systémy

- Oftalmologický registr (OFR)
- Registr intenzivní péče (RIP)
- Registr nozokomiálních infekcí (RNI)

[25] *http://www.ksrzis.cz/*

Ucelený přehled o dalších registrech a jejich obsahu podávají např. Kasal et al. (2011), Hrejsemnou (2009) nebo (Štampach, 2010). Zde je podrobněji popsán pouze registr **EPIDAT**, ze kterého pochází data použitá v disertační práci. Program EPIDAT slouží k zajištění povinného hlášení, evidence a analýzy výskytu infekčních nemocí v České republice. Program je celostátně používán Hygienickou službou ČR od 1. 1. 1993. Hlášení infekčních nemocí je základem pro místní, regionální, národní a nadnárodní kontrolu šíření infekčních nemocí i pro hlášení infekcí z České republiky do Společenství EU a Světové zdravotnické organizaci (SZÚ ČR, 2014). Úložiště dat pro EPIDAT slouží k bezpečné výměně aktuálních datových souborů o výskytu infekcí mezi jednotlivými pracovišti Hygienické služby ČR, Ministerstva zdravotnictví ČR a Státním zdravotním ústavem v Praze (KSRZIS, 2010).

3.1.2 Důvěrnost a měřítko datových sad

Účelem klinických studií je zkoumat jevy a vztahy na individuální úrovni. Každý účastník studie je tedy považován za samostatnou entitu, zatímco skupina osob tvoří datovou sadu. Zachování důvěrnosti dat a ochrany soukromí musí být v případě klinických studií zahrnuto mezi nutné podmínky výzkumu. V případě geograficky orientovaných studií je možnost analyzovat individuální záznamy, včetně veškerých důvěrných informací, často omezená. Adresa bydliště, místo nákazy či dokonce přímo souřadnice jsou sice podstatnou podmínkou pro prostorovou analýzu v lokálním měřítku. Na druhou stranu však často nemusí být podrobná data nutná, ani žádoucí a používá se spíše agregovaných zýázanmů.

Běžně využívané údaje o pacientech a jejich zdravotním stavu, jako jsou jméno či adresa, mohou být snadno použity k jednoznačné identifikaci osob. Stejným způsobem však může být využita i přesná poloha získaná na základě měření pomocí polohových navigačních systémů (Waller a Gotway, 2004). I přes stále větší sdílení osobních informací samotnými uživateli na internetu (např. sociální sítě) poskytuje současná legislativa jak v Evropské unii, tak i ve Spojených státech amerických, stále větší uznání práva jednotlivce na ochranu osobních údajů a to včetně údajů o zdravotním stavu (Elliott a Wartenberg, 2004). Systémy a registry pro vykazování zdravotních údajů však podléhaly důsledné ochraně soukromí jednotlivce již od svých počátků nezávisle na této legislativě (Bell et al., 2006). Právě ochrana soukromí je jedním z klíčových faktorů, které mohou přímo ovlivnit jak kvalitu dat při sběru, tak i jejich poskytování a následnou analýzu. Důvěrnost dat a snaha zabránit jejich možnému zneužití (dat i výsledků) jsou také důvodem nutného kompromisu mezi analýzou v lokálním měřítku a ochranou soukromí zúčastněných subjektů.

Hlavní cíle metod prostorové epidemiologie pro malá území, stejně jako prostorové analýzy zdravotnických dat všeobecně, mohou být rozděleny do třech hlavních oblastí (Elliott a Wartenberg, 2004), kterými jsou mapování nemocí, studium geografických korelací a průzkum shlukování společně s monitoringem vývoje jevu. Volba měřítka dat a výsledných zjištění je důležitým faktorem v každé z vyjmenovaných oblastí a samozřejmě i v komplexních studiích, které využívají kombinací těchto postupů. Zdravotnická data jsou běžně mapována i analyzována (a poskytována) nikoliv jako jednotlivé záznamy, ale spíše jako agregované údaje odpovídající administrativním jednotkám. Změna typu administrativních jednotek nebo měřítka může vést i ke změně vnímaní mapovaného jevu. Nestabilita územního vymezení administrativních jednotek vede často i k potřebě modifikovat tvar areálů těchto jednotek, resp. distribuovat hodnoty z jednoho areálu do druhého (Horák, 2009). Tento

problém je často označován jako MAUP (*Modifiable Area Unit Problem*) či problém areálové interpolace (Openshaw, 1984a, 1984b). Seskupování dat pořízených v různém prostorovém rozlišení nebo jejich agregování do výrazně odlišných jednotek bez použití vhodných postupů může vést k velkému kolísání výsledků, které může vést k chybné interpretaci zjištění (Beale et al., 2008).

3.1.3 Anonymizace dat

Přístup k relevantním datům je nezbytný pro růst každého vědního oboru, prostorovou epidemiologii nevyjímaje. Zavedení vhodných kontrolních mechanismů k uchování důvěrné povahy dat je základní podmínkou k zachování podpory a důvěry veřejnosti (Elliott a Wartenberg, 2004). Vzhledem k tomu, že důvěrná a soukromá povaha dat je v přímém rozporu se snahou (geo)analytiků poskytnout co nejpřesnější analýzu v co nejpodrobnějším měřítku, tak poskytovatelé těchto dat běžně zavádějí opatření, která zajišťují anonymitu jedinců a zabraňují jejich zpětné identifikaci. Metody využívané k zajištění ochrany soukromí lze shrnout následovně (Bell et al., 2006):

- Agregace dat v prostoru a čase;
- nahrazením skutečné polohy jednotlivce environmentálními či etiologickými vlastnostmi prostředí;
- náhodným pozměněním skutečné polohy jednotlivce;
- omezením přístupu k vlastnostem umožňujícím zpětnou identifikaci pomocí omezení uživatelských účtů či funkcionality software

Obsáhlé shrnutí metod využívaných k agregaci dat a randomizaci polohy či atributů zmiňuje např. Armstrong et al. (1999).

Anonymizace jednotlivých záznamů

Jakýkoliv pokus o zachování důvěrnosti poskytnutých dat je jejich významnou modifikací, která je ovšem nutná zejména v případě podrobných analýz, kdy současně existuje snaha mezi uchováním maximálního množství informace a znemožněním zpětné identifikace jednotlivce.

První typ anonymizace záznamů představje *atributové anonymizace*, kterou lze považovat za slabou/mírnou anonymizaci. Představuje situaci, kdy poskytovatel dat umožní zachování přesné polohové informace záznamu (místo bydliště či nákazy), ale kvůli možné zpětné identifikaci neposkytne detailní informace o charakteru záznamu. V praxi to může vypadat tak, že není známo jméno, datum narození či další důvěrné informace, zatímco některé podstatné charakteristiky zůstávají zachovány či jsou nahrazeny příslušností k nějaké skupině (věk, pohlaví, zaměstnání apod.). Tento typ anonymizace je poskytovateli zřídkakdy využíván pro veřejně dostupná data, ale je běžnější v případě užší spolupráce zainteresovaných institucí. Výstupy z takto podrobných analýz slouží spíše interním účelům organizace než veřejnosti, pro kterou bývají výsledky agregovány.

V případě druhého typu anonymizace jde jak o odstranění osobních identifikátorů, tak především o *randomizaci nebo záměrné znepřesnění polohy* záznamu. V tomto případě jsou sice stále zachovány všechny záznamy z primárních dat, ale jejich pozice je buď náhodně pozměněna, případně není distribuován celý atribut nutný k přesnému polohovému určení. V praxi může jít o posunutí záznamu v náhodném směru či vzdálenosti definovanými předem. Stejně tak mohou být data distribuována buď se sníženou přesností (podél uliční sítě,

v katastrálním území apod.), či zcela náhodně. Tento postup sice zabrání reidentifikaci jedince, ale umožní zachovat původní prostorový vzor (Bell et al., 2006). Výstupy z analýz využívajících takto upravená data musí být brány s určitou rezervou, aby bylo zamezeno možným nedorozuměním nebo chybnému porozumění vycházejícím z polohové nepřesnosti dat.

Agregace – anonymizace založená na měřítku

Nálezová data jsou nejčastěji prezentována ve formě tabulek nebo místních agregovaných statistik generovaných pro definovanou územní jednotku jako je okres nebo kraj. Agregovaná data mohou být snadno přetvářena do podoby jednoduchých i komplexních map a geovizualizací, ale vždy je nezbytné zvolit jednotku agregace tak, aby bylo možné vykreslit statistické/epidemiologické inference v datech (Pfeiffer et al., 2008). Stejně tak je důležité rozpoznat, jak měnící se úrovně agregace mohou ovlivnit celkové vnímání prostorového vzoru. Jev známý jako MAUP nebo také jako ekologická chyba (Openshaw, 1984a; Jelinski a Wu, 1996) je široce uznávaným nedostatkem v intepretaci statistických dat, kdy jsou poznatky o jedinci získány na základě agregovaných dat či skupiny, do které tento jedinec spadá (Cressie, 1993). Výsledky získané čistě na základě agregovaných dat by neměly být použity pro formulování předpokladů o chování jedince (Beale et al., 2008).

Analytik by se měl vždy předem rozhodnout, jaké měřítko, a tedy i platnost, bude mít výsledná analýza, a na tomto předpokladu zvolit typ a podrobnost agregace. To znamená rozhodnout, zda je vhodné seskupovat nálezová data do administrativních jednotek či pravidelné sítě. Nebo zda je vhodné agregovat do bodového pole tvořeného místy v úzkém vztahu k původní lokaci, či jen do středu územní jednotky nebo pravidelného rastru (Armstrong et al., 1999). Kromě prostorového aspektu je možné data agregovat současně i do časových intervalů pokud to jejich povaha dovolí. Dalším běžným omezením je nastavení prahové hodnoty agregované jednotky, na základě které jsou data zobrazena (např. nelze pracovat s územím, kde je méně než 5 záznamů či žije méně než 50 osob apod.).

3.2 Oblasti zájmu prostorové epidemiologie a základní terminologie

Analýza prostorového rozložení výskytu nemoci a také vztah k potenciálním rizikovým faktorům hraje důležitou roli v nejrůznějších zdravotnických studiích. Podle Bailey (2001) lze identifikovat čtyři hlavní oblasti zájmu geografické epidemiologie s důrazem na statistické a prostorové analýzy:

- mapování nemocí,
- ekologické studie,
- studie shlukování nemocí,
- environmentální hodnocení a monitoring.

Elliott a Wartenberg (2004) s předchozí klasifikací souhlasí, ale definují pouze tři hlavní směry, kterými se ubírají základní analýzy prostorové epidemiologie v lokálním či regionálním měřítku:

- mapování nemocí,
- studie geografických korelací,
- shlukování, shluky nemocí a jejich sledování.

Obě klasifikace oblastí zájmu jsou si na první pohled velmi podobné, pouze v případě Elliotta a Wartenberga (2004) došlo ke sloučení některých tříd a přejmenování ekologických studií na studie geografické korelace, které názvem lépe vystihují skutečnou podstatu cílů a analytických procesů skupiny. Obě kategorizace doplňuje i poznámka autorů o běžném prolínání postupů mezi skupinami tak, aby byla co nejlépe postihnuta variabilita jevu, často nejen v prostoru, ale také v čase a vlastnostech.

3.2.1 Základní pojmy v (prostorové) epidemiologii

Morbidita *(nemocnost, chorobnost)* je obecný pojem, kterým je označován číselný údaj vztažený pro danou nemoc k určitému časovému úseku a počtu obyvatel. Je nejčastěji udáván jako počet nemocných za rok na 100 tisíc obyvatel (Velký lékařský slovník, 2008).

Prevalence - vyjadřuje počet nemocných osob (buď celkem, nebo s určitou nemocí) vztažený k celkovému počtu osob v populaci (střednímu stavu populace), nejčastěji vyjádřena na 100 tisíc obyvatel (Demografie, 2014). V závislosti na délce období, v němž se výskyt onemocnění zjišťuje, jsou rozeznávány dva základní typy prevalence: okamžiková a intervalová (Bencko et al., 2003).

Incidence - je ukazatel intenzity a dynamiky onemocnění, který kvantifikuje výskyt nově vzniklých onemocnění (s danou diagnózou) ve zvoleném časovém intervalu. Často se počítá roční incidence, specifická incidence, hrubá incidence. Pro srovnání různých populací se doporučuje využít standardizovaných měr (Bencko et al., 2003).

Hrubá incidence - je definována jako podíl počtu nově zjištěných případů onemocnění v dané populaci v daném období a počtu osob v dané populaci v daném období (často jako počet onemocnění na 100 tisíc obyvatel). Hrubá incidence zohledňuje rozsah populace, nezohledňuje však její věkovou strukturu. Pokud se příliš neliší věkové složení srovnávaných populací, poskytuje možnost základního srovnání. Výhodou je dosažitelnost potřebných demografických údajů z veřejně dostupných zdrojů, může selhat při srovnávání odlišných populací (Široký, 1999).

Přímo (věkově) standardizovaná incidence – míra incidence je průměrována s využitím vah založených na věkově specifických mírách incidence specifikované standardní populace. Přímo standardizovaná incidence vyjadřuje hrubou míru incidence, která by nastala v případě, že by studovaná populace měla stejnou věkovou strukturu jako populace standardní (Last a Abramson, 2001).

Nepřímo (věkově) standardizovaná incidence - se používá při srovnání populací s neznámými resp. statisticky nestabilními věkově specifickými mírami incidence. Míry standardní populace jsou průměrovány s využitím distribuce studované populace. Ukazuje, jaký by byl očekávaný výskyt onemocnění ve studované populaci, kdyby frekvence výskytu v ní byla stejná jako ve standardní populaci. Ukazatel nepřímo věkově standardizované incidence (SIR - standardized incidence ratio) udává poměr skutečného počtu nových případů k počtu očekávanému (Široký, 1999). SIR může být považován za jeden z ukazatelů relativního rizika onemocnění (Bivand et al., 2008)

Etiologie - nauka o příčinách determinantách a podmínkách nemocí či původu onemocnění (Last a Abramson, 2001; Velký lékařský slovník, 2008).

3.3 Prostorové analýzy a metody prostorové statistiky v epidemiologických studiích

V kapitole 3.2 byly definovány hlavní oblasti využití prostorových technologií (zejm. GIS) pro analýzu a vizualizaci prostorových dat. Na základě představených klasifikací lze pojmenovat hlavní oblasti výzkumu, kterými jsou mapování nemocí, analýza vztahu mezi výskytem nemoci a vnějšími faktory (ekologické studie či geografické korelace) a průzkum prostorových (příp. časoprostorových) vzorů a shlukování. V následujícím popisu jsou postupně představeny vybrané metody a postupy spadající do jednotlivých okruhů, které jsou současně dány do souvislosti s již provedenými studiemi.

3.3.1 Mapování nemocí

Mapy incidence nemocí a jejich výskytu jsou nejstaršími nástroji prostorové epidemiologie. Mezi nejznámější historické mapy patří již dříve zmíněná mapa cholery v Londýně od Johna Snowa (1854), dále potom mapa žluté zimnice od Valentine Semana (1798) nebo mapa moru v Neapoli (okolo 1690). Podrobně se historií mapování nemocí zabývají např. Koch (2005, 2009) nebo Dorling (1998). I v současnosti je nejčastější aplikací GIS v oblasti epidemiologických analýz právě jeho využití pro mapování nemocí (Rytkönen, 2004).

Popularita map nemocí spočívá zejména v jejich schopnosti podat rychlý vizuální přehled jinak velmi komplexních geografických informací, jejichž správná interpretace může vést k objevení prostorových vzorů a souvislostí, které nejsou z tabelárních dat patrné (Elliott a Wartenberg, 2004). Podle Bivanda (2008) je hlavním účelem epidemiologických map vizualizace rozložení nemoci v prostoru a prostorová variace nákazy, která pomáhá detekovat oblasti s vyšším výskytem nemoci, což může vést k odhalení dříve neznámých rizikových faktorů. Data pro vizualizaci jsou většinou rozdělena podle vybraných atributů (věk, pohlaví, apod.).

Lawson et al. (1999) uvádí jako hlavní cíle mapování nemocí následující důvody:

- popis prostorové variability incidence nemoci a formulování etiologických hypotéz,
- identifikace oblastí s neobvykle vysokým rizikem onemocnění za účelem zásahu,
- poskytnutí spolehlivých map rizika onemocnění kvůli vhodné alokaci zdrojů a vyhodnocení rizika.

Mezi nejčastěji mapované epidemiologické jevy patří zejména absolutní výskyt nemoci (bodově lokalizovaný nebo agregovaný do plochy) (Marek et al., 2012; Bailey, 2001), prevalence a incidence (Naish et al., 2011), standardizovaná úmrtnost a index standardizované úmrtnosti (SIR) (Feng, 2011), který podle Wallera a Gotwaye (2004) odpovídá riziku onemocnění.

Pro zobrazení dat v mapě je většinou využito standardních kartografických metod jako např. (nepravý) kartogram či pseudokartogram, metoda teček (Alexander et al., 2003), dasymetrická metoda, zobrazení výsledků interpolací aj. Více se kartografickými metodami v mapování nemocí zabývá např. Koch (2009) nebo Koch a Denike (2004). Rytkönen (2004) se pak kriticky zamýšlí nad úvahou, proč si nejsou epidemiologické mapy rovny z pohledu různě zvolených standardů a územních jednotek.

Ačkoliv jsou mapy nemocí vizuálně přesvědčivé i intuitivní, stále je potřeba zvýšená pozornost v případě interpretace, aby nedošlo k přecenění či naopak podcenění některých podmiňujících faktorů nebo shluků oblastí s vysokými hodnotami morbidity. Současně je důležité také vhodně zvolit doplňkový tematický obsah a podkladová data, aby nedocházelo k těmto dezinterpretacím.

Mezi největší problémy spojenými s tvorbou tematických epidemiologických map jednoznačně patří přesnost a úplnost získaných dat (Elliott a Best, 1998), měřítko mapy či měřítko provedených analýz (Tatem et al., 2012) a s ním spojený problém s vhodně definovaným sousedstvím (Oakes, 2004). Dále potom jde o problém areálové interpolace (tzv. MAUP) (Openshaw, 1984b) a také o povahu samotných dat, která většinou podléhají Poissonovu rozdělení (Goovaerts, 2006). Právě díky jinému než normálnímu rozdělení dat a také častému využití standardizovaných veličin může nastat problém s nadhodnocením rizika v případě území s malou hustotou zalidnění (Rytkönen, 2004). Pro analýzu standardizovaných epidemiologických dat se za účelem odstranění možnosti nadhodnocení využívají nejrůznější metody shlazení dat pomocí lokálních vážených průměrů, neparametrické regrese nebo bayesovských metod (Bivand et al., 2008; Waller a Gotway, 2004).

3.3.2 Ekologické studie

Ekologické studie, nazývané také jako studie geografické korelace či regresní studie, se soustředí na analýzu asociací mezi pozorovanou incidencí, relativním rizikem či přímo pozorovanými četnostmi a potenciálně rizikovými faktory v prostoru. Jejich cílem je vyhodnotit prostorovou variabilitu ohrožených skupin populace, které jsou vystaveny vnějším environmentálním faktorům (např. vzduch, voda, půda), socioekonomickým a demografickým faktorům (např. příjem, národnost) a/nebo životnímu stylu (strava, kouření, apod.) ve vztahu ke zdraví v určitém geografickém prostoru a měřítku (Elliott a Best, 1998). Přehled možných dat poskytuje výše uvedená Tabulka 1 v kapitole 3.1.

Při výběru vhodných metod pro ekologické studie je jednou z nejdůležitějších podmínek typ zpracovávaných dat a také požadovaný výsledek. V případě agregovaných areálových dat četností výskytu nemocí bude pravděpodobně použito jiných regresních modelů než v případě bodových dat nebo dokonce souvislých povrchů. Stejně tak je důležité i měřítko celé studie. Nejrelevantnější výsledky podávají studie na lokální úrovni nebo v relativně malých oblastech, kde je snížena problematika heterogenity okolních podmínek díky faktu, že analýza je prováděna blíže k úrovni jednotlivce než velké populace (Staessen et al., 1999).

Pro prvotní zjištění neprostorových vztahů mezi výskytem nemoci a jednotlivými faktory prostředí lze použít metody základní statistiky a statistického testování jako jsou např. korelace, kontingenční tabulky nebo analýza rozptylu (Gruebner et al., 2011). Na základě výsledků těchto postupů lze dále vytvářet hypotézy o rizikových vlastnostech prostředí, které jsou následně analyzovány pomocí (prostorových) regresních modelů (Peng a Dominici, 2008).

Elementárními technikami studia asociace rizikových faktorů a prostorové variability onemocnění jsou regresní modely a jejich prostorové alternativy. Mezi základní typy používaných modelů se řadí lineární, Poissonova či logistická regrese, a dále metody

pokročilého modelování jako jsou generalizované aditivní modely, generalizované lineární a lineární smíšené modely nebo víceúrovňové modelování. V případě víceúrovňových modelů je nejdříve zkoumána prostorová autokorelace standardizovaných residuí jednoduchých modelů a následně, pokud je indikována prostorová asociace, je aplikován samotný víceúrovňový model (např. model variačních komponent, modely s náhodnou konstantou, atd.) (Pfeiffer et al., 2008). Příklady použití regresních i víceúrovňových modelů uvádí například Pfeiffer et al. (2008), McLafferty (2003) nebo Bailey (2001). I v případě regresní analýzy lze aplikovat bayesovský přístup k tématu (Aamodt et al., 2006). Logickou nadstavbou regresních modelů jsou potom techniky představující možnosti predikce a rekonstrukce vývoje onemocnění v dané lokalitě a s nimi spojené simulace za využití geostatistických metod nebo algoritmů Monte Carlo, které zmiňují Beneš a Bodlák (2011) nebo (Bergquist, 2011).

Kromě regresních modelů lze úspěšně využít pro průzkum vztahů mezi výskytem či četností výskytu nemoci a rizikovými faktory také metody vícerozměrné statistiky. Ty kromě redukce dimenze slouží také k odhalení latentních proměnných (komponent, faktorů), které mohou právě v regresních modelech nahradit velké množství možných nezávislých vysvětlujících proměnných. Mezi tyto metody můžeme řadit např. analýzu hlavních komponent, faktorovou analýzu, diskriminační analýzu nebo mnohorozměrné škálování. Aplikací vícerozměrných statistických metod v epidemiologii a jejich kombinací s regresními modely seznamuje Armitage et al. (2008).

3.3.3 Prostorové vzory, shlukování nemocí a jejich sledování

Třetí významnou skupinou prostorových analýz ve zdravotnických a epidemiologických studiích je průzkum prostorových (příp. časoprostorových) vzorů v datech, tzv. *spatial patternu*. Zatímco mapování nemocí umožňuje vizuální identifikaci prostorového rozložení výskytu nemoci, analýza prostorového vzoru umožňuje kvantifikovat zóny, kde je riziko výskytu nemoci vyšší než by se dalo očekávat (Bivand et al., 2008). Cílem analýzy prostorového vzoru je zjistit odchylky od náhodných procesů (např. homogenní/heterogenní Poissonův proces (CSR/HEPP), Coxův proces apod.) a zjistit tak, zda je zkoumaný jev rozložen na daném území zcela náhodně, pravidelně nebo ve shlucích (Baddeley, 2010). Pro účely epidemiologie je z těchto procesů nejdůležitější a také nejčastěji zmiňované právě shlukování (Moore a Carpenter, 1999), které je interpretováno jako počet případů onemocnění, které si jsou blízké v prostoru a/nebo čase. Kromě postihnutí typu procesů probíhajících v daném území, je možné dále zkoumat lokální shluky. Důležitou vlastností metod prostorového shlukování je možnost odhadu statistické významnosti procesů nebo shluku nejčastěji pomocí tzv. *p-value*, která může vznikat jak přímo z výpočtu, tak na základě simulací (Rushton, 2003).

Z pohledu celkového měřítka analyzovaných dat a postupů lze rozdělit metody prostorového shlukování do dvou hlavních kategorií (Pfeiffer et al., 2008; Waller, 2009), kterými jsou:

- globální odhady prostorového shlukování;
- lokální odhady prostorového shlukování.

Globální odhady prostorového shlukování (příp. metody obecného shlukování) jsou označovány jako vlastnosti prvního řádu, jejichž cílem je popis způsobu, kterým se očekávané hodnoty procesu (např. průměr) mění v prostoru. Jinými slovy vyjadřují intenzitu jevu v prostoru (Gatrell et al., 1996). Mezi globální metody odhadu pro bodová nebo agregovaná data patří Moranovo *I*, Gearyho *c*, Tangův test, Ripleyho *K*-funkce a mnoho dalších. V případě analýzy časoprostorových shluků lze využít např. Knoxův test, Bartonův test nebo Mantelův test (Waller, 2009). Dalšími nástroji, které zde lze zahrnout jsou jádrový odhad nebo kvadrátové testy.

Lokální odhady prostorového shlukování oproti tomu představují vlastnosti druhého řádu, které popisují lokální variabilitu jevu jako výsledek interakce mezi sousedícími prvky (Gatrell et al., 1996). Lokální metody jsou často známé také jako LISA (lokální indikátory prostorové asociace) a jsou nedílnou součástí exploratorní analýzy prostorových dat (ESDA), jak ji představil Anselin (1994, 1995, 1996). Mezi lokální metody patří lokální Moranův test, Getis-Ordovo *Gi**, Kulldorffova prostorová statistika (Kulldorff, 1999), Openshawův GAM nebo Digglův test. Časoprostorem se zabývá např. Kulldorffova časoprostorová statistika.

Souhrnný přehled metod globálního i lokálního odhadu poskytují např. Moore a Carpenter (1999), Waller a Gotway (2004) nebo Shekhar et al. (2011).

Pokud jsou hlavním tématem analýzy shlukování, je kromě výše zmíněných technik prostorové statistiky možné využít i vícerozměrných statistických metod, zejména pak shlukování, které v poslední době začíná být vhodné i k prostorovým analýzám díky vývoji vhodných aplikací (Carvalho et al., 2009; Esri, 2012; Horák et al., 2012).

Pod pojmem prostorové analýzy epidemiologických dat si velká část odborníků představí zejména dva typy aplikací. Prvním typem jsou mapy rozšíření nemocí, druhým právě shlukování. Jejich kombinací s vhodnými daty může uživatel odhalit rizikové faktory a jejich vliv na zdraví člověka. Stejně jako u většiny jiných analytických metod, i v případě shlukování může dojít k nevhodné interpretaci výsledků, které nemusí vhodným způsobem popisovat skutečnost. Rizika u prostorového shlukování vyplývají především z použití prostorové složky dat, u podobných analýz je nezbytné vhodné nastavení definice sousedství, například počtu k-nejbližších sousedů a jejich schéma nebo velikosti použitého jádrového odhadu (kernelu) apod. Dalším častým problémem je potom již dříve zmíněný MAUP či přítomnost velkého množství řídce osídlených administrativních jednotek s poměrně vysokou prevalencí nemoci, které mohou vyústit v nadhodnocení výsledného rizika. Pro vhodně zvolenou velikost sousedství lze využít např. Ripleyho *K*-funkce (Bivand et al., 2008), grafu Morishitova indexu (Baddeley, 2010) nebo grafu střední kvadratické chyby (Bivand, 2007).

3.3.4 Geografické profilování (Geoprofiling)

Metoda geografického profilování či geoprofilingu je známá především z kriminologie, pro jejíž účely byla vyvinuta Kimem Rossmo. Jde o techniku prostorové statistiky, která slouží k identifikování možných kandidátů ze seznamu podezřelých ze spáchání závažného trestného činu na základě jejich geografického chování (Rossmo, 1999). V rámci hlavních okruhů se může částečně řadit k analýze prostorových vzorů i geografické korelační analýze. Geografické profilování využívá lokalizace událostí (trestných činů, případů nákazy apod.) k identifikaci možných zdrojů na základě dvou klíčových konceptů – obalových zón kolem

bodových událostí a vlivu vzdálenosti (Rossmo, 1995b). Předpokladem metody je, že události se budou vyskytovat spíše poblíž záchytných bodů než dále od nich, a tyto body je potřeba identifikovat a ohodnotit jejich význam (Le Comber a Stevenson, 2012). Kromě svého původního účelu v kriminologii a odhalování nebezpečných zločinců postupně nachází tato metoda uplatnění i v epidemiologii – např. při identifikaci míst lihnutí komárů nakažených malárií v Káhiře (Le Comber et al., 2011) a pro tento účel je i dále zdokonalována (Verity et al., 2014), nebo při studiu rozšíření invazivních druhů (Stevenson et al., 2012) nebo bílých žraloků (Martin et al., 2009).

3.4 Kampylobakterióza

Praktická část disertační práce představuje komplexní výzkumná studie zabývající se prostorovou a časoprostorovou analýzou distribuce kampylobakteriózy v České republice v letech 2008—2012. Přestože není kampylobakterióza mezi laickou veřejností známá tak, jako například salmonelóza, jedná se o nejčastější gastroenteritidu (bakteriální střevní onemocnění) ve vyspělých zemích (Weisent et al., 2011), Českou republiku nevyjímaje (ÚZIS, 2013). Původce onemocnění, nejčastěji bakterie *Campylobacter jejuni* nebo *Campylobacter coli*, se běžně vyskytuje v trávicím traktu různých zvířat, kterým však nemusí způsobit onemocnění (The Center for Food Security & Public Health, 2013). K nákaze člověka dochází nejčastěji kontaminovaným drůbežím masem, ale k přenosu může dojít i kontaktem s nemocnými zvířaty, drůbeží či domácími mazlíčky (koťata, štěňata) nebo kontaminovanou vodou či mlékem a mléčnými výrobky. Možný (byť vzácně) je i interpersonální přenos (Ambrožová, 2011). Dle veterinárních údajů jsou kuřata v tržní síti kontaminována kampylobaktery až v 75 %, větší riziko představují kuřata chlazená než mražená (Ambrožová, 2011). Ačkoliv jsou způsoby přenosu onemocnění známy, tak dostatečně popisují vznik jen přibližně poloviny všech případů onemocnění (Ekdahl et al., 2005). Onemocnění se projevuje průjmem, únavou, horečkou, bolestí břicha a nauzeou nebo zvracením. Častá je příměs krve ve stolici (Lexová et al., 2013). Inkubační doba je obvykle 2—5 dní (rozsah 1—10 dní) a příznaky přetrvávají v průměru jeden týden. Nejvíce ohroženými skupinami obyvatel jsou děti a mladí dospělí do třiceti let nezávisle na pohlaví (Lexová et al., 2013). Incidence kampylobakteriózy v České republice v roce 2013 dosáhla 230 případů na 100 tisíc obyvatel, pro srovnání - incidence salmonelóz se od roku 2008 stabilizovala a nepřesahuje přibližně 100 případů na 100 tisíc obyvatel (ÚZIS, 2013). Zvýšený počet onemocnění kampylobakteriózou je v našich podmínkách zejména v teplejších měsících roku s vrcholem v srpnu (Domasová, 2014; Lexová et al., 2013). Incidence kampylobakteriózy ve světě je však v současnosti stále považována za podhodnocovanou, k čemuž přispívají jak nehlášené případy onemocnění se slabším průběhem, tak často i nedostatečné diagnostické testy. Je odhadováno, že ve skutečnosti se může v České republice ročně objevit až 11,3× více případů než je skutečně hlášeno, v rámci EU to pak může být až 46,7× (Havelaar et al., 2013).

3.4.1 Prostorové analýzy kampylobakteriózy

Geografickými aspekty rozšíření onemocnění, možnými environmentálními i socioekonomickými faktory, které mohou vznik onemocnění podporovat, a etologií onemocnění se v současnosti zabývá, či v nedávné minulosti zabývalo, hned několik mezinárodních studií. Tyto studie a publikace jsou často zaměřeny na jedno téma, které odpovídá jedné z hlavních oblastí zájmu prostorové epidemiologie. Weisent et al. (2011) se zabývala prostorovými vzory kampylobakteriózy v americkém státě Tennessee a kromě

identifikace shluků se autoři zaměřili také na výběr vhodné administrativní jednotky a metody shlazování incidence. V navazující publikaci (Weisent et al., 2012) se stejný autorský kolektiv zabývá průzkumem asociace socioekonomických determinantů, které mohou podpořit či tlumit riziko výskytu onemocnění. Využitím regrese a geograficky vážené regrese byl zjištěno, že zvýšené riziko onemocnění hrozí především v oblastech s vyšší sociální deprivací. Podobnými problémy na příkladu kanadské provincie Quebéc, se zabývala ve svých publikacích také Julie Arsenault (Arsenault, 2010; Arsenault et al., 2012, 2013), která pro modelování rizika doporučuje spíše menší územní jednotky (obec či sčítací obvod), ale současně upozorňuje na problém ekologické chyby a identifikuje teplejší oblasti a zvýšenou hustotu domácích zvířat jako významné faktory zvyšující incidenci onemocnění.

Mullner et al. (2010) se pokusili s pomocí individuálního modelu stanovit prostorové riziko nákazy stanovit na základě věku a pohlaví obyvatelstva, potenciálního zdroje nákazy a genotypu jednotlivých pacientů. S modelováním rizika či incidence kampylobakteriózy se s různou úspěšností identifikace environmentálních, demografických či socioekonomických asociací potýkaly i další studie z Kanady (Green et al., 2006), Švédska (Nygård et al., 2004) či Anglie (Gillespie et al., 2008). Přímý vliv klimatu na geografickou distribuci kampylobakteriózy studovali Sari Kovats et al. (2005) nebo Jagai et al. (2007). S pomocí Poissonovy logistické regrese a Monte Carlo odhalil Gabriel et al. (2010) segregaci postižených bakterií *Campylobacter jejuni* nakažených kuřecím a hovězím masem a také identifikovali zvýšenou incidenci v urbánním prostředí. Nízkým počtem uspokojivě vysvětlených zdrojů onemocnění u pacientů se motivoval Ekdahl et al. (2005), který za jednoho z možných přenašečů označil mouchy. Podhodnocením skutečné incidence kampylobakteriózy se věnoval Havelaar et al. (2013), který s pomocí dat o pohybu švédských turistů odhadl také faktor v té době aktuálního podhodnocení.

České studie se spíše než prostorovými aspekty zabývají onemocněním a jeho původci především z pohledu způsobu přenosu infekce na člověka (Bardon et al., 2009). Využitím prostorové informace (ačkoliv silně agregované) se ve své studii týkající se využití epidemiologického informačního systému na Slovensku věnuje Zeleňáková et al. (2012). Mezi prvními aktivitami, které se týkají prostorové či časoprostorové distribuce tohoto onemocnění na našem území jsou publikace autora disertační práce (Marek et al., 2014, 2015). Na možné spojení mezi kampylobakteriózou a automaty na čerstvé mléko v ČR upozorňuje s odkazem na prohlášení hlavního hygienika např. Andrlová (2011).

4 MAPOVÁNÍ ONEMOCNĚNÍ: CO NÁM MOHOU PROZRADIT MAPY? [DC1]

Základní geografickou úlohou je zkoumání variací jevů v prostoru a toho, jak se tyto jevy chovají v různých měřítcích, díky čemuž je již dlouhou dobu přirozeně využívána i pro popis prostorového rozložení zdraví a nemocí (Meade a Emch, 2010). Dokladem toho jsou jak italské mapy ze 17. století, slavná mapa cholery Johna Snowa z Londýna 19. století, ale i současné statické a dynamické mapy a geovizualizace (Koch, 2005). Jedním z prvních a nejdůležitějších kroků v prostorové epidemiologické analýze je vizualizace dat a jejich charakteristik. Díky tomu je možné vnímat případné prostorové struktury a vzory, identifikovat možné chyby a formulovat hypotézy týkající se faktorů, které mohou pozorované vzory vysvětlit (Pfeiffer et al., 2008). Kromě toho je (geo)vizualizace důležitá také pro komunikaci výsledků a jejich předání širokému spektru možných příjemců. Těmi mohou být jak odborníci, tak i laici, což platí obzvlášť v interdisciplinárních oborech, kde spolupracují různě zaměření experti, takže je vždy potřeba zvolit kompromis mezi způsobem a podrobností výstupu.

Prostorová data jsou nejčastěji poskytována ve dvou formách – jako body zaznamenávající jednotlivé případy, nebo v agregované podobě, která shrnuje celou skupinu případů a přiřazuje jí jednu hodnotu, která ji vhodně reprezentuje (četnost výskytů, průměr skupiny). Podrobnější dělení dat zohledňující jejich původ a účel představuje Bailey (2001) a je zmíněn v kapitole 3.1.

Základní dělení dat předurčuje hlavní typy jejich geovizualizace a využití kartografických metod. Nejjednodušším a často i nejefektivnějším způsobem zobrazení jednotlivých případů onemocnění se známou polohou je jejich vyobrazení s pomocí ***metody teček***. Její výhodou je, že umožňuje přímo zobrazit primární nálezová data, která nejsou zkreslená statistickou analýzou a umožňuje tak odhalit přímou prostorou distribuci jevu (Pfeiffer et al., 2008). Na druhou stranu může přesné umístění bodů umožnit zpětnou identifikaci nakažené osoby a být tak v přímém rozporu s podmínkou důvěrnosti dat. Tato nevýhoda může být odstraněna v případě vážení teček, kdy jedna tečka vyjadřuje několik případů onemocnění. Možnou alternativou a doplňkem metody teček mohou být *mapy hustoty* („heat maps“), které nahrazují konkrétní body intenzitou jevu vypočítanou na základě vhodně zvoleného jádrového odhadu.

Převládající formou poskytování zdravotnických informací je forma (prostorově a časově) agregovaných záznamů, které shrnují četnost onemocnění v časovém intervalu ve zvolené administrativní jednotce (obec, městská část, PSČ, okres, atd.) či v pravidelném gridu (Marek et al., 2013b). Vhodným způsobem zobrazení takových dat jsou ***kartogramy či pseudokartogramy*** měr nemocnosti (incidence, prevalence či relativního rizika), které nejčastěji vyjadřují vztah mezi četností výskytu onemocnění ve zvolené jednotce a počtem (a strukturou) obyvatelstva, na jehož základě jsou jednotlivé jednotky obarveny. Kromě výše zmíněných nejčastějších metod existuje velké množství dalších přístupů, metod a technik, které jsou či byly použity pro mapování výskytu onemocnění v prostoru. Jejich přehled uvádí Koch a Denike (2004), Waller a Gotway (2004), Koch (2005) či Meade a Emch (2010). Za všechny je možné dále zmínit dasymetrickou metodu, isolinie a isopletovou mapu či kartografickou anamorfózu. Přehled základních mapovaných charakteristik je obsahem terminologie v kapitole 3.2.1. Otázkou by také vždy měla být možnost srovnávání sousedních jednotek v případě hrubých měr, které do výpočtu nezahrnují podobnou strukturu obyvatelstva. Ideálně by tedy měly být mapovány standardizované míry, ačkoliv některé

studie týkající se SIR ukázaly, že srovnání hodnot mezi geografickými jednotkami bude zavádějící pouze v případě extrémně odlišných populací, což se např. v rámci jednoho státu či oblasti děje v praxi velmi zřídka (Goldman a Brender, 2000; Jarup, 2004).

4.1 Příprava základní vstupní datové sady

Základní datovou sadou v disertační práci jsou data pocházející z databáze EPIDAT (Epidemiologická databáze), která byla poskytnuta Státním zdravotním ústavem v Praze. Databáze obsahuje záznamy o hlášených případech kampylobakteriózy na celém území České republiky (v případě nakažení i mimo ČR) v letech 2008—2012. Data jsou anonymizována, takže není známo jméno pacienta ani přesná adresa jeho bydliště. Lokalizace záznamů je ve většině případů možná až do úrovně uliční sítě. Kromě věku a pohlaví pacienta jsou v databázi obsaženy i další atributy jako datum přijetí pacienta u lékaře, zaměstnání, národnost, agens (původce onemocnění), pravděpodobný způsob nakažení, apod. Datová sada obsahuje téměř 98 933 záznamů se 78 atributy s různou úplností a kvalitou vyplnění. Jednotlivé záznamy jsou do systému EPIDAT vyplňovány lékaři manuálně, díky čemuž se v nich objevuje množství chyb, překlepů a nekonzistentností (např. různé typy použitých zkratek). K tomu, aby mohla být data použita pro neprostorové statistické i prostorové analýzy, bylo nutné celou datovou sadu zkontrolovat, vyčistit a opravit. Tento proces bohužel nebylo možné automatizovat a jednotlivé záznamy tak musely být kontrolovány především manuálně. Pro kontrolu nebyly vybrány všechny atributy, ale pouze ty, které mohly být dále využity a nebylo v nich příliš mnoho nejasností či nevyplněných záznamů. Ze všech charakteristik byly upravovány pouze atributy – věk / věková skupina, pohlaví, místo nákazy, místo onemocnění, zaměstnání, předpokládaný zdroj nákazy a agens (původce onemocnění). Většina z vybraných atributů měla jasně definovanou strukturu či rozsah hodnot, a tak byla jejich kontrola a případná oprava možná.

4.1.1 Určení polohy záznamů a geokódování

Nejpodstatnější vlastností dat nutnou pro průběh prostorových analýz je prostorová složka jednotlivých záznamů či jejich seskupení. V připravených datech však není poloha záznamů v podobě geografických souřadnic udána. To ostatně ani není možné z důvodu ochrany citlivých údajů o pacientech. Geografické souřadnice však nahrazuje umístění objektu v rámci ulice a/nebo čtvrti, města a okresu. Díky těmto informacím lze geografickou polohu objektů odhadnout. V případě, že by byla dostatečná informace o poloze záznamů v rámci okresu či obce, tak by nebylo třeba dalších kroků a data by mohla být snadno prostorově lokalizována a agregována na základě údajů v nich obsažených s pomocí základních databázových dotazů a operací a propojením s prostorovými daty územní jednotky. Pro podrobnější umístění je nutné data geokódovat, tzn. přiřazení geografických souřadnic záznamům na základě dostupných informací, které umožní tyto záznamy zobrazit v mapě a následně je dále zpracovávat, analyzovat nebo seskupovat na základě jejich prostorových vlastností.

Proces geokódování lze obecně popsat jako porovnávání konkrétní adresy či textem popsané polohy zadané uživatelem s existující databází adresních míst se známými souřadnicemi. Způsobů a prostředků pro geokódování je velké množství a vždy záleží na požadavcích uživatele, kvalitě a množství lokalizovaných dat nebo kvalitě srovnávací

databáze. Důležité jsou i podmínky poskytovatele geokódovacích služeb v případě využití online řešení či API. Geokódovaní může probíhat lokálně s využitím lokálně umístěné databáze adres (např. srovnáváním s lokální databází adresním míst z RÚIAN[26]), lokálně s využitím geokódovacích služeb (např. v QGIS prostřednictvím srovnávání s OSM Nominatim nebo ArcGIS Online geocode service) nebo zcela online pomocí nahrání textového souboru do webové služby či aplikace (nejčastěji cloudové aplikace – např. Google Fusion Tables, MangoMap, CartoDB a další).

Geokódování datové sady hlášených případů kampylobakteriózy v ČR v letech 2008—2012 s 98,5 tisíci záznamů (odmazány byly záznamy, kdy se pacient onemocněl mimo území ČR), která navíc obsahovala neúplný adresní údaj dostupný maximálně do úrovně ulice, je náročným úkolem. Vyžaduje totiž způsob určení adresy, který umožní částečnou aproximaci záznamu, bude vhodná pro území České republiky a zároveň umožní geokódovat velké množství záznamů, aniž by byly porušeny podmínky služby. Pravděpodobně nejvhodnější by bylo využít službu Google Geocode API[27], která splňuje většinu těchto podmínek, ale bohužel licenční podmínky umožňují pouze 2 500 dotazů během 24 hodin. Geokódovací služba OpenStreetMap s názvem Nominatim[28] sice takové omezení nemá, ale nedoporučuje se využívat ji pro velké množství dat a její přesnost má stále svá omezení.

Prog. kód 1 Funkce v programovacím jazyku **R** *pro geokódování pomocí Mapy API*

```
#### funkce geokodovani pres mapy.cz ####
geocode <- function(adr){
  require("XML")
  # generovani parametru do URL
  adr <- gsub(' ','%20',adr)
  # otevreni spojeni a dotaz
  connectStr <- paste("http://api4.mapy.cz/geocode?query=",adr,sep="")
  con <- url(connectStr)
  data.xml <- xmlTreeParse(paste(readLines(con), collapse=""))
  # uzavreni spojeni a nasleduje reseni vyhozenych dat
  close(con)
  # zpracovani ziskanych dat
  koren <-xmlRoot(data.xml)
    if (!is.null(koren[["point"]][["item"]])) {
      x<-xmlGetAttr(koren[["point"]][["item"]], "x")
      y<-xmlGetAttr(koren[["point"]][["item"]], "y")
      return(c(x,y))
    # koren obsahuje hodnoty souradnic
    }
}
```

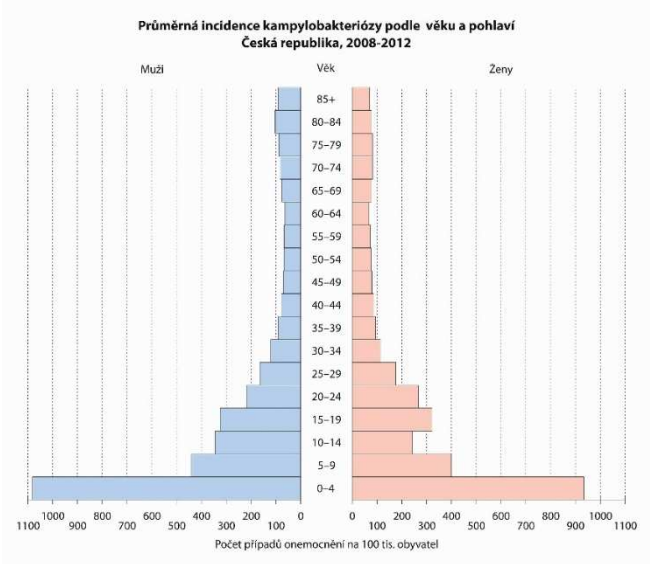
Za nejvhodnější řešení bylo zvoleno geokódování pomocí Mapy API[29] poskytovaného společností Seznam.cz a jejím mapovým portálem Mapy.cz. Toto API má sice omezenou možnost aproximace výsledků, na druhou stranu však umožňuje geokódování velkých souborů (v případě vložení krátké časové prodlevy mezi dotazy, aby nebyla operace

[26] Registr územní identifikace, adres a nemovitostí
[27] *https://developers.google.com/maps/documentation/geocoding/*
[28] *http://wiki.openstreetmap.org/wiki/Nominatim*
[29] *https://api.mapy.cz/*

vyhodnocena jako DoS útok) a poskytované výsledky geokódování mají vyhovující úroveň pro použití v prostorových analýzách. Více informací a srovnání zmíněných API poskytuje Cícha (2013) v diplomové práci vedené autorem disertační práce, v rámci které byl i sestaven skript v programovacím jazyku **R** umožňující geokódování s pomocí Mapy API (Programový kód 1). První geokódovaní s pomocí tohoto skriptu proběhlo ve srovnání s opravou dat velmi rychle a trvalo 85 minut čistého času bez započítání prodlevy mezi dotazy. Chybně lokalizována nebo nelokalizována byla 3 % záznamů, která byla manuálně opravena a lokalizována ve druhém kole geokódování trvajícím zhruba 3 minuty. I přes uspokojivé výsledky geokódování je potřeba mít na paměti, že jeho výsledky s sebou stále nesou jistou nepřesnost v určení, která je způsobena primárními daty a jejich anonymizací. Lokalizovaná data je však možné s dostatečnou přesností využít pro statistické analýzy buď přímo v podobě bodové vrstvy, nebo po agregaci na úroveň administrativní jednotky, případně pravidelné sítě.

4.2 Základní charakteristiky a analýza časových řad

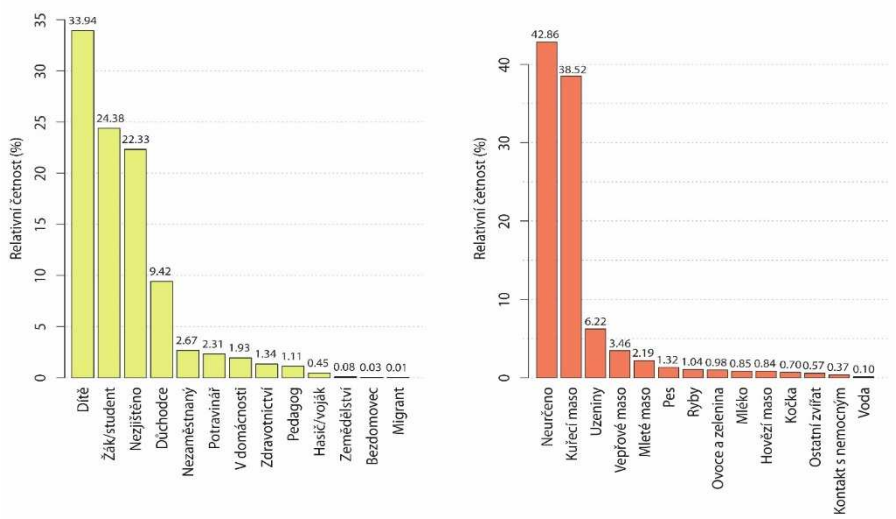
Před samotným mapováním, prostorovými analýzami či modelováním výskytu a rizika kampylobakteriózy v obcích České republiky, bylo pohlíženo na datovou sadu jako na celek a byly prozkoumávány její základní statistické charakteristiky. Účelem bylo zjistit, jak jsou ohroženy jednotlivé věkové skupiny obyvatelstva, jaké jsou nejčastější zaměstnání nakažených a zdroje nákazy a samozřejmě, jaký průběh má onemocnění během roku.



Obr. 2 Průměrná incidence kampylobakteriózy rozdělená podle věkových skupin a pohlaví

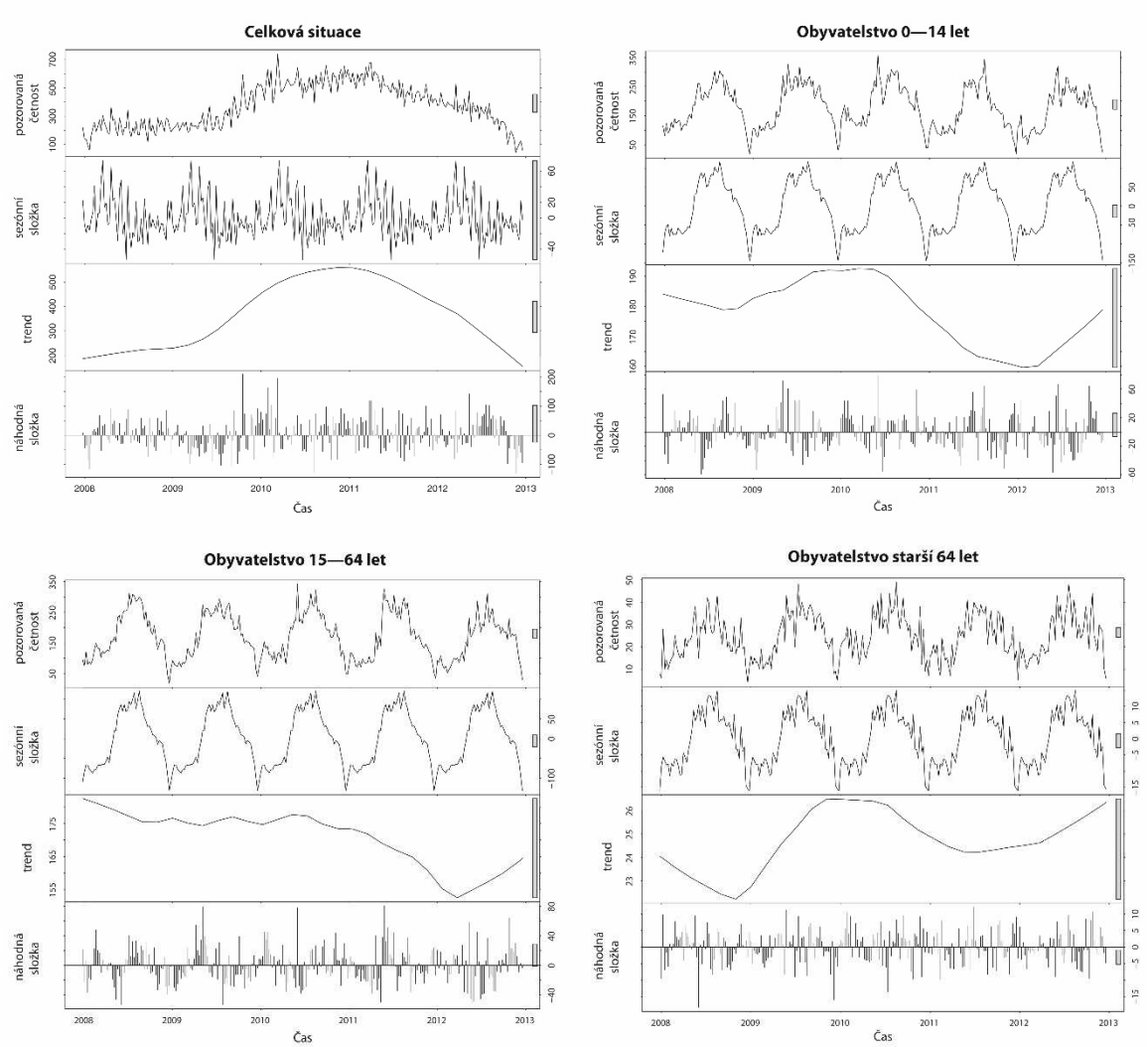
Obrázek 2 zobrazuje průměrnou roční incidenci kampylobakteriózy v české populaci rozdělenou podle věku a pohlaví. Z grafického vyjádření jde vyčíst, že nejohroženější skupinou obyvatelstva jsou děti ve věku do 4 let, přičemž malí chlapci mají mírně vyšší incidenci než děvčata. Dá se také konstatovat, že osoby do 30 let věku jsou více ohroženy než osoby starší 30 let, u kterých incidence postupně klesá až na minimum v rozmezí 60—64 let a ustálí se pod hodnotou 100 případů onemocnění na 100 tisíc obyvatel. Rozdíl mezi pohlavími není patrný. Se zjištěním, že jsou děti a mladí lidé častěji nakaženi korespondují i zjištění plynoucí ze sloupcového diagramu v levé části Obrázku 3, který znázorňuje nejčastějších zaměstnání nakažených pacientů. Ve výrazné převaze jsou i zde děti následované žáky

a studenty. Tyto dvě kategorie dohromady tvoří 58,32% podíl všech případů zaznamenaných ve studovaném období. Data zaznamenávaná do systému EPIDAT jsou v některých ohledech plněna výrazně neúplnými informacemi, což dokládá fakt, že u více než 22 % případů se nepovedlo zjistit jejich zaměstnání. Třetí nejčastější skupinu pacientů tvoří důchodci s téměř 10 % všech případů. Nejčastěji nakaženými mezi konkrétními profesemi jsou pracovníci v potravinářském průmyslu s 2,31 % případů. Pravá část Obrázku 3 obsahuje informace o pravděpodobných nejčastějších příčinách onemocnění. V některých zahraničních studiích lze najít tvrzení o tom, že identifikovat zdroj infekce je možné pouze u zhruba poloviny případů (Nylen et al., 2002). V případě České republiky je uspokojivě nalezen (či oznámen) pravděpodobný primární zdroj nákazy v 57 % případů. Jako jednoznačně nejčastější zdroj nákazy je uváděno kuřecí maso (38,52 % případů), mezi zdroje s více než 2% zastoupením dále patří uzeniny, vepřové maso a mleté maso. Méně častou, ale stále relevantní příčinou je kontakt se psy (1,32 %), kočkami (0,70 %) a dalšími zvířaty (0,57 %) či přímý kontakt s nemocným (0,37 %). Dalšími rizikovými potravinami jsou pak ryby (1,04 %), čerstvé ovoce a zelenina (0,98 %), mléko (0,85 %) a hovězí maso (0,84 %).



Obr. 3 Nejčastější zaměstnání pacientů postižených onemocněním (vlevo) a jejich pravděpodobné příčiny onemocnění (vpravo) kampylobakteriózou

Je-li zaměřena pozornost na průběh a četnost týdně hlášených případů onemocnění během sledovaného období (Obrázek 4), pak v případě pozorování celkové situace (Obrázek 4 – vlevo nahoře) je jasně patrný trend nejdříve postupného zvyšování počtu hlášených případů od roku 2008 s maximem v roce 2010 a poté následným snižováním. Sezónní složka již nemá v případě celé datové sady tak jednoznačný průběh, ale náhodná složka vykazuje nejvyšší rozptyl hodnot zejména během roku 2010. Zjištěná tvrzení ohledně celkového trendu potvrzují i základní statistické charakteristiky standardizované roční incidence a četnosti hlášení onemocnění v jednotlivých letech v Tabulce 2. Jsou-li data dále rozčleněna podle ekonomické aktivity obyvatelstva do kategorií 0—14 let (Obrázek 4 – vpravo nahoře), 15—64 let (Obrázek 4 – vlevo dole) a starší 64 let (Obrázek 4 – vpravo dole), pak je sezónnost výskytu onemocnění mnohem více patrná a vrchol četnosti se pravidelně opakuje během letních měsíců. Trend u nejmladších osob a osob v produktivním věku lze označit jako klesající se změnou průběhu během roku 2012, zatímco u osob starších 64 let lze spatřit mírně stoupající tendence.



Obr. 4 Rozklad časové řady výskytu kampylobakteriózy v České republice v letech 2008—2012. Postupně je zobrazena celková situace (vlevo nahoře) a rozdělení četností dle ekonomické aktivity obyvatelstva – předproduktivní věk (vpravo nahoře), ekonomicky aktivní (vlevo dole) a postproduktivní věk (vpravo dole)

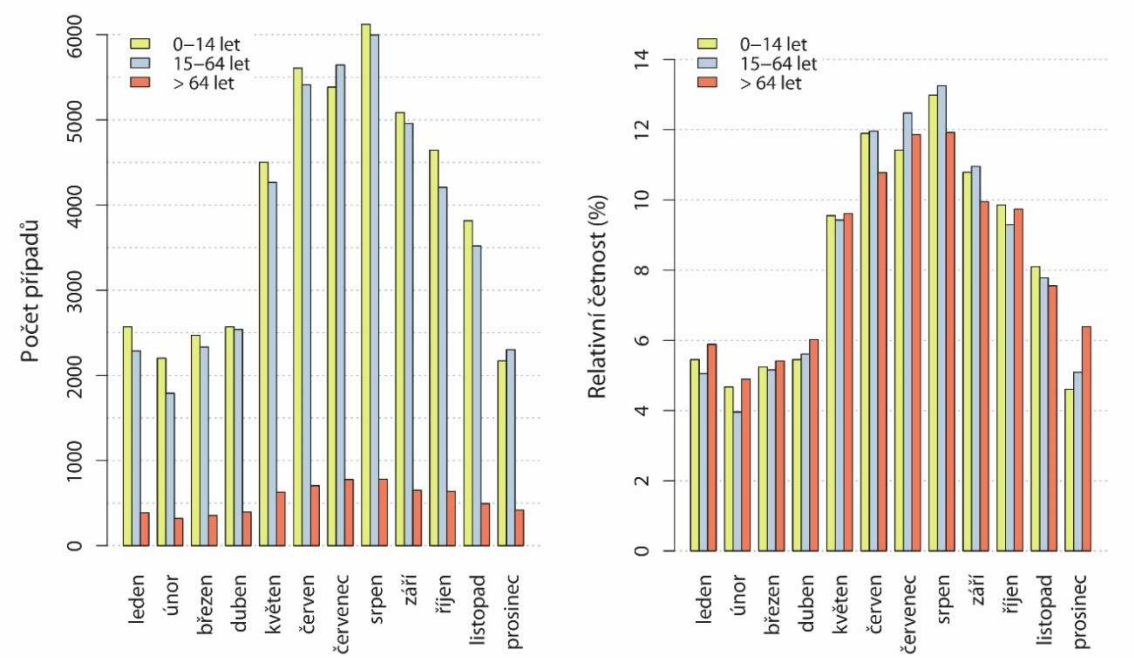
Tab. 2 Základní statistické charakteristiky průměrné roční incidence (počet případů na 100 tisíc obyvatel) a absolutní četnosti výskytu kampylobakteriózy v obcích České republice v letech 2008—2012.

	2008		2009		2010		2011		2012		2008–2012	
	Frek	Inc	Frek	Inc	Frek	Inc	Frek	Inc	Frek	Inc	Frek	Inc
Min	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Max	480,0	7750,5	412,0	4049,6	403,0	3532,5	339,0	7472,9	363,0	7892,2	396,0	6605,4
Medián	0,0	142,1	0,0	153,5	0,0	157,0	0,0	141,9	0,0	145,4	0,4	144,4
Průměr	3,1	164,6	3,2	171,7	3,3	179,4	2,9	156,7	2,9	161,5	3,1	161,7
SD	16,9	165,6	15,8	145,7	16,5	145,0	14,1	149,2	13,2	150,6	15,1	138,9
Suma	20076		20348		21150		18797		18393		19752	

Frek – absolutní četnost, Inc – průměrná roční incidence; Statistické charakteristiky: Min – minimum, Max – maximum, medián, průměr, SD – směrodatná odchylka, Suma – celkový součet. Průměrná roční incidence je počítána pro obce ČR s výjimkou obcí, které jsou tvořeny městskými částmi, pro které je výpočet proveden; jde o standardizovanou míru.

Sloupcový graf na Obrázku 5 popisuje průběh onemocnění v jednotlivých měsících roku a skupinách obyvatelstva dle ekonomické aktivity. V levé části Obrázku 5 jsou znázorněny absolutní počty případů v jednotlivých měsících, kde je patrný podobný průběh všech u všech kategoriích s největší četností během června—září s vrcholem v srpnu. U nejstarších osob je

průběh během roku více vyrovnaný a jednotlivé měsíce vykazují menší rozptyl četnosti případů na rozdíl od dalších dvou skupin, kde četnost během roku vzrůstá i trojnásobně. Pravá část Obrázku 5 pak zobrazuje stejnou situaci, ale tentokrát jako relativní četnosti jednotlivých tříd, čímž došlo ke srovnání velikosti sloupců a možnosti lepšího srovnání mezi třídami. Opět je možné si všimnout vrcholu výskytu onemocnění v letních měsících a menších změn v případě průběhu u obyvatel starších 64 let. Zajímavé u této věkové kategorie je také chování onemocnění během prosince a částečně i ledna—dubna, kdy je relativní podíl onemocnění v kategorii vyšší než u zbylých skupin. Letní vrchol onemocnění je běžným jevem u průběhu onemocnění popsaným i v dalších zemích (Nylen et al., 2002; Sari Kovats et al., 2005), který je často spojován s letním grilováním.



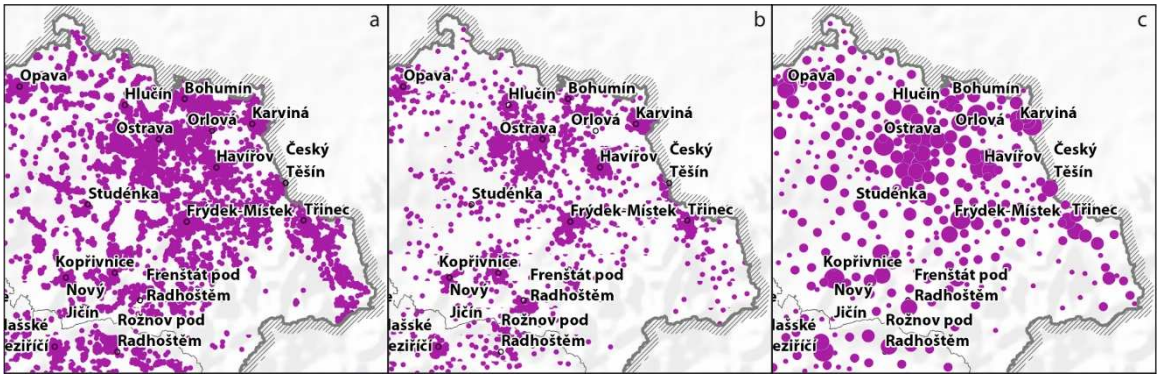
Obr. 5 Měsíční absolutní počet hlášených případů (vlevo) a relativní počet hlášených případů (vpravo) kampylobakteriózy s rozdělením dle ekonomické aktivity

4.3 Mapování výskytu kampylobakteriózy v České republice

Pravděpodobně nejběžnější kartografickou metodou používanou pro zobrazení areálových dat je (pseudo)kartogram[30]. Kartogram je mapa s dílčími územními celky, do kterých jsou plošným způsobem znázorněna statistická data (relativní hodnoty) většinou geografického charakteru (Kaňok, 1999). Přestože nejsou kartogramem cíleně vyjádřeny spojité hodnoty pro celé území, ale pouze statistická veličina shrnující některou z vlastností území, tak často umožní vnímat prostorovou distribuci a proměnlivost jevu v sousedních územních celcích (Rushton, 2003). Kromě běžného zobrazení mapovaného jevu vztažného k územní jednotce (část obce, obec, katastrální území, apod.) je možné i zobrazení skutečnosti formou pravidelné mřížky/gridu (nejčastěji čtvercového či šestiúhelníkového tvaru buňky), kterým je studované území překryto. Výhodou postupu je odstranění vlivu administrativních

[30] Pro zjednodušení terminologie je dále využíván pojem kartogram i pro mapové vyjádření jevů formou pseudokartogramu (nepravého kartogramu) či metody plošných znaků

hranic na vnímání mapovaného jevu a je vhodný zejména u jevů spojitých v prostoru. Naopak nevýhodou může být v případě, že je potřeba znát konkrétní údaje takto definovaných buněk gridu (např. věkovou strukturu apod.), která musí být odvozena od jednotek nadřazených či jinak definovaných a může tak snadno dojít k ekologické chybě. Pokud chce tvůrce mapy zdůraznit některé skutečnosti, tak je možné podpořit vyjádření kartogramem využitím metody kartografické anamorfózy. Kartografická anamorfóza představuje metodu založenou na geometrické přeměně vybraného parametru jevu (např. plochy území) podle určitých pravidel při zachování neměnného tematického prvku mapy (např. tvar území) pro jeho zvýraznění (Voženílek et al., 2011). Díky netradičnímu pojetí je kartografická anamorfóza atraktivní metodou geovizualizace, která díky své podstatě vážení prostorových jevů pomocí plochy/populace současně umožnuje správnější nahlížení na mapovaný jev.



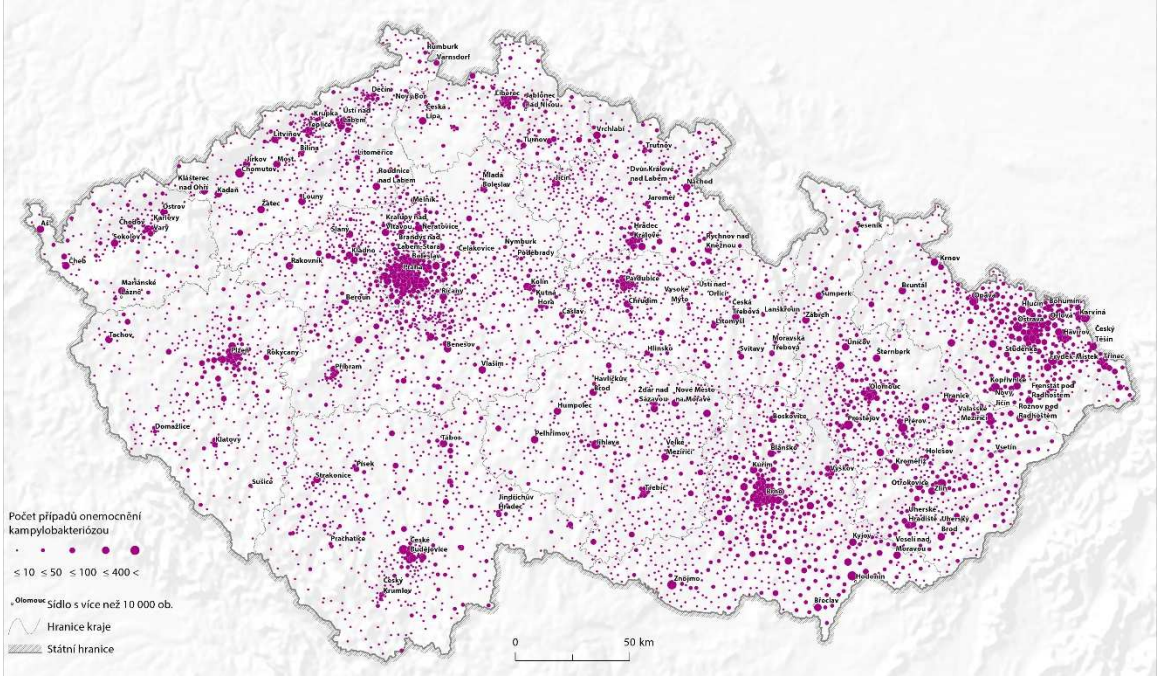
Obr. 6 Ukázka tečkové metody nálezových dat: a) topografický způsob, b) topografický způsob - vážené tečky – 1 tečka odpovídá 10 případům, c) kartogramový způsob - vážené tečky – velikost tečky odpovídá kategorii množství případů v území

Kromě kartogramů vztahujících se nejčastěji k celé ploše administrativní jednotky nebo pravidelného gridu, je často pro mapování distribuce onemocnění případů v prostoru využíváno i metody teček. Metoda teček je vhodná pro znázornění absolutních hodnot a četností (např. absolutní četnost onemocnění v daném místě), protože umožňuje okamžitě rozpoznat prostorovou distribuci, tedy popsat intenzitu a variabilitu jevu v prostoru. Je-li zvolen topografický způsob umístění teček, pak jsou případy v mapě zobrazeny nejčastěji jednotlivě nebo jako agregovaný počet, pokud se na jednom místě objevila nemoc vícekrát ve studovaném období. V případě kartogramového způsobu je umístění teček podmíněno výskytem v celém hodnoceném území. Tečky mohou vyjadřovat jeden nebo i více případů pokud jde o vážené tečky (např. jedna tečka v mapě může představovat 10 skutečných případů onemocnění) a přibližují se tak metodě kartodiagramu a kartogramu (Obrázek 6). Tečky mohou být v území rozmístěny pravidelně či na základě nějakého dalšího pravidla (např. ve středu území).

4.3.1 Mapování výskytu a incidence onemocnění

Mezi první a současně nejjednodušší postupy při mapování prostorové distribuce nemocí patří tečková metoda. Ve své nejjednodušší podobě jde vlastně jen o zobrazení polohy jednotlivých případů do mapy, kdy jedna (topograficky umístěná) tečka odpovídá jednomu případu onemocnění. Vytvoření map pomocí tečkové metody většinou není náročné a vizuální hodnocení takto vytvořených map je velmi intuitivní, protože tečky přirozeně

znázorňují hustotu i intenzitu mapovaného jevu. Nevhodným způsobem ovšem může tento přístup být v případech, kdy buď nelze zobrazovat individuální záznamy, nebo je množství případů tak velké, že kombinace hustoty teček a jejich velikost neumožní podrobné zkoumání jevu, a tečky buď splynou, nebo jsou velmi malé. V těchto případech je vhodnější data seskupit na základě polohy a zobrazovat tak ne jednotlivé záznamy, ale skupinu záznamů pomocí vážení teček. Metoda vážených teček vycházející z agregace záznamů do částí obcí je použita i pro tvorbu mapy souhrnné distribuce kampylobakteriózy v České republice v letech 2008—2012, která je znázorněna na Obrázku 7.

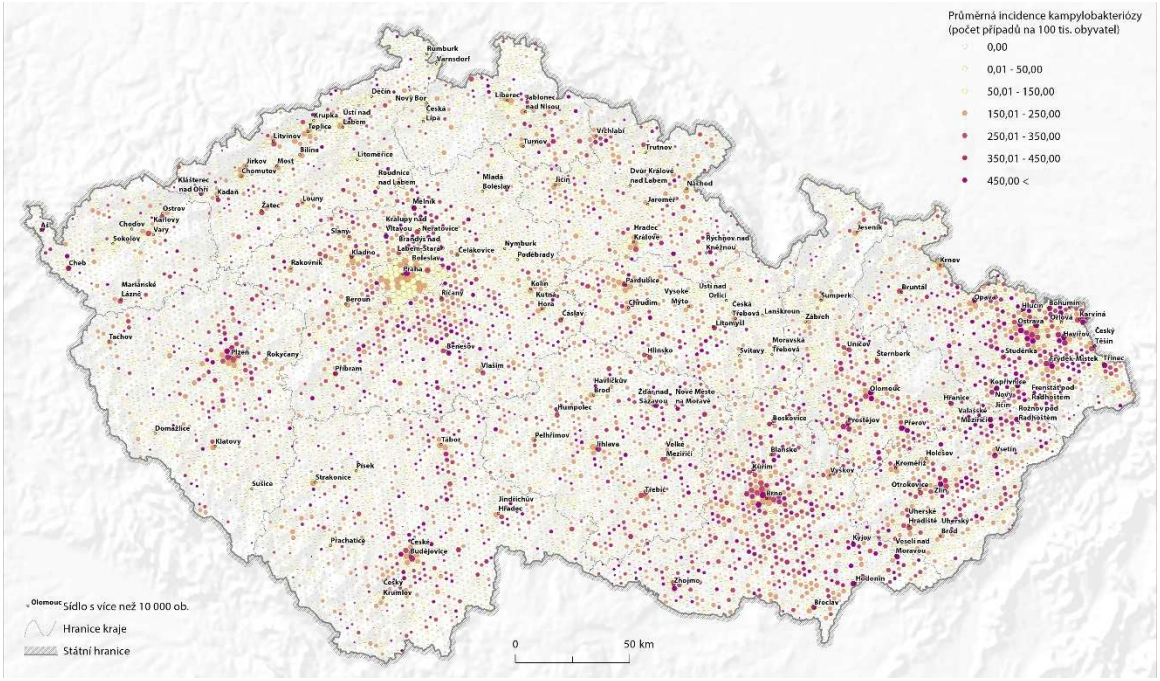


Obr. 7 Případy kampylobakteriózy zobrazené pomocí metody teček s využitím vážených teček

Je patrné, že největší množství případů je koncentrováno v hustě zalidněných oblastech v okolí velkých (zejm. krajských) měst. Patrné také je, že na Moravě a především ve Slezsku je hustota teček vyšší než ve zbytku republiky (s výjimkou Prahy, Plzeňska a Pardubicka). Ačkoliv využití tečkové metody odhaluje určitý prostorový vzor a oblasti, kde je možná zvýšená incidence kampylobakteriózy, tak je stále potřeba si uvědomit, že jsou vizualizovány absolutní četnosti onemocnění, a ty jsou logicky vyšší v oblastech s vysokou koncentrací obyvatelstva (tedy ve větších městech) než v oblastech s nižší hustotou zalidnění. Pro představu, alespoň jeden případ onemocnění kampylobakteriózou byl zaznamenán v 72 % obcí ČR, pokud jsou ale hodnoceny části obcí, pak je to pouze v 41 % částí obcí.

Vhodný postup pro srovnávání výskytu nemoci v jednotlivých místech tedy není pouze na základě absolutních četností případů, ale je nutné zahrnout do výpočtu i údaje o populaci a její struktuře. Tyto údaje jsou nejčastěji poskytovány národními statistickými úřady a stejně tomu je i v případě České republiky, kdy je poskytovatelem Český statistický úřad. Nejsnadnějším postupem pro zahrnutí populace je využít dat, která odpovídají některé z administrativních jednotek – ideálně obci, pro které jsou známy údaje o obyvatelstvu i jeho struktuře. Pokud je požadované rozlišení ještě podrobnější a nelze využít existující data, pak je častým postupem pro zjištění přepočítaných měr nemocnosti odhad struktury obyvatelstva na základě nadřazených nebo podobných administrativních jednotek. Často využívané jsou

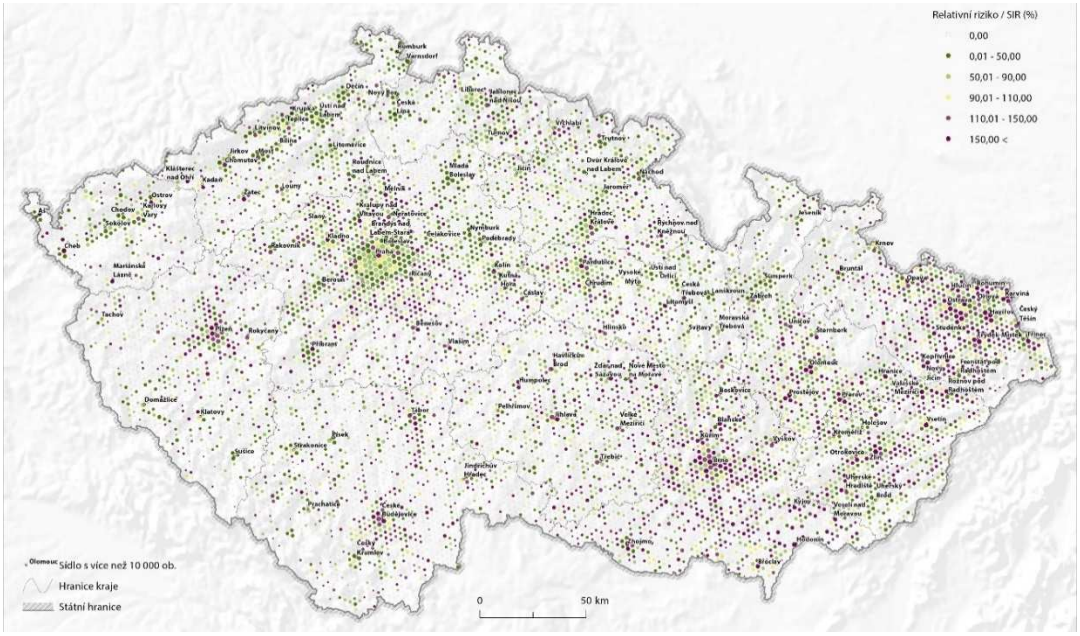
metody mapování v rámci pravidelné sítě (gridu) tvořeného čtverci či šestiúhelníky, díky čemuž je eliminován efekt hranic administrativních jednotek a jejich vnitřní nehomogenity, nicméně hrozí výskyt tzv. ekologické chyby – omezené platnosti charakteristik jednoho typu rozdělení území v typu jiném.



Obr. 8 Průměrná hrubá incidence kampylobakteriózy v České republice vizualizovaná prostřednictvím hexagonů

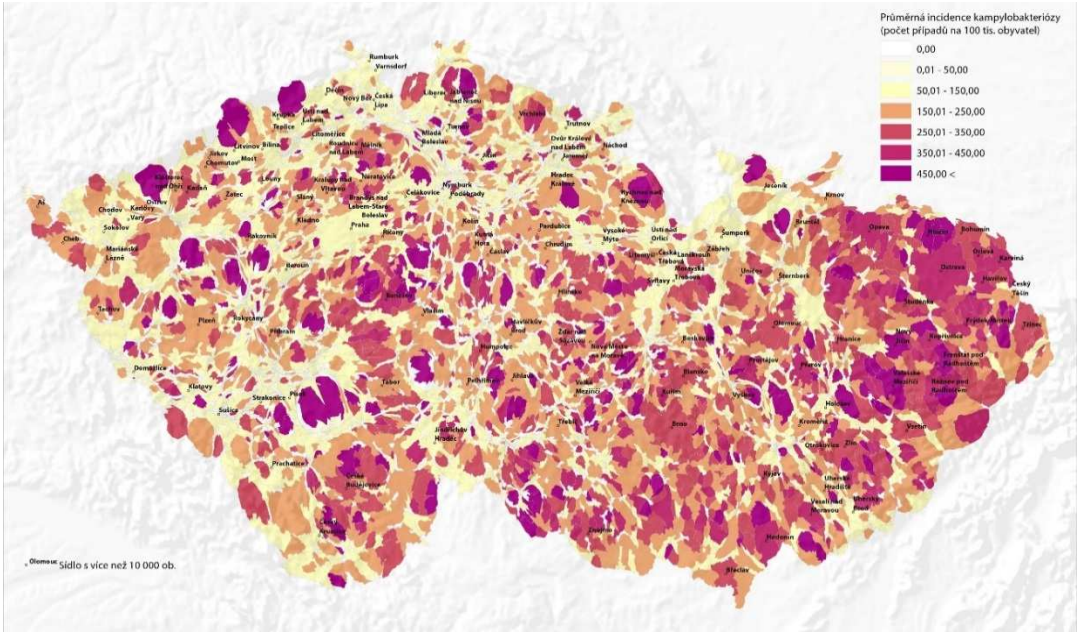
Mapy na Obrázku 8 a Obrázku 9 zobrazují průměrnou hrubou roční míru incidence kampylobakteriózy v ČR, respektive relativní riziko (SIR) vztažené nikoliv k území administrativní jednotky, ale k pravidelné síti 13 488 šestiúhelníků (včetně jejich částí v příhraničních oblastech) pokrývající celé území České republiky. Tyto vypočtené míry v sobě už zahrnují alespoň informaci o počtu obyvatel v šestiúhelnících. Velikost šestiúhelníku byla zvolena tak, aby jeho plocha odpovídala ploše průměrného katastrálního území obce (6,05 km^2) a data o počtu obyvatel pochází z datové sady *GEOSTAT 1 km^2 population grid 2011*[31]. Obě mapy popisují celkovou situaci v ČR vhodněji než absolutní četnosti u mapy vytvořené tečkovou metodou. Při bližším pohledu také obě mapy podávají podobné informace – nemocí pravděpodobně nejvíce zasaženou oblastí je oblast severovýchodní Moravy a Slezska, kde je jak vysoká incidence, tak i relativní riziko nakažení kampylobakteriózou (tmavě purpurové oblasti na obou mapách). Zvýšený výskyt a částečně i riziko lze identifikovat také na Brněnsku a v části jižní Moravy. Na rozdíl od mapy na Obrázku 7 již není Praha identifikována jako oblast s vyšším výskytem nemoci či jejím rizikem, což ovšem neplatí o oblasti jihovýchodně od Prahy, Plzeňsku nebo Českobudějovicku. Stejné charakteristiky (průměrná roční incidence a relativní riziko), ale vztažené k částem obcí je možné prohlédnout si a vizuálně porovnat na Obrázku 11 a Obrázku 14.

[31] *http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography*



Obr. 9 Relativní riziko (SIR) vizualizované prostřednictvím hexagonů

Na Obrázku 10 je představena ukázka kartografické anamorfózy průměrné incidence kampylobakteriózy v obcích České republiky. Díky kartografické anamorfóze se na úkor Čech zvětšila plocha velké části Moravy a Slezska, kde je zaznamenána vysoká incidenci onemocnění. Zvětšila se také plocha Českobudějovicka a oblasti jihovýchodně od Prahy. Vnímání anamorfózy tak více odpovídá skutečnému postižení jednotlivých částí České republiky. Výhodou tohoto typu anamorfní mapy je, že zůstává zachováno přímé sousedství obcí, ačkoliv se nerovnoměrně mění plocha a tvar obcí i měřítko mapy. Pro tvorbu anamorfózy byl použit program Scape Toad[32]. Ten k deformaci původních dat na základě zvolených charakteristik využívá Gaster-Newmanův algoritmus, který vytváří souvislou neradiální anamorfózu (Gastner a Newman, 2004).



Obr. 10 Zobrazení průměrné incidence kampylobakteriózy formou kartografické anamorfózy

[32] *http://scapetoad.choros.ch/index.php*

4.3.2 Bayesovské vyhlazování incidence

Prezentace měr nemocnosti v ploše pomocí (pseudo)kartogramů může vést k neúmyslnému poskytnutí zavádějících informací, a to především v malých oblastech a územních jednotkách, které mohou být díky své velikosti a hustotě osídlení zdrojem zvýšené variability onemocnění v důsledku jeho náhodného výskytu. Typicky mohou být mylně generovány extrémní hodnoty SIR v řídce osídlených oblastech s malou (či nulovou) četností onemocnění. To je způsobeno nepřímou úměrou SIR a očekávaného počtu případů, v důsledku čehož je u malých populací zjištěna velká variabilita odhadnutých měr (Elliott a Wartenberg, 2004) a odhad měr samotných je spíše nestabilní.

Řešení k upravení takto nestabilních odhadů měr nemocnosti mohou poskytnout Bayesovské metody odhadu a vyhlazování. Ty využívají modelů pravděpodobnosti k získání vyhlazených odhadů skládajících se z kompromisu mezi skutečně pozorovanou (vypočítanou) hodnotou v území a odhadu vycházejícího z větší datové základny, tzn. z měr vypočítaných na základě celého území či několika sousedních oblastí (Waller a Gotway, 2004). Základním principem bayesovských metod je posílení dat s vyšší nejistotou pomocí apriorní pravděpodobnosti – předem známé nebo předpokládané informace. V případě empirických bayesovských odhadů míry nemocnosti proměnlivé v prostoru může být aposteriorní (výsledné) riziko v oblasti odhadnuto na základě lokální míry (lokální věrohodnosti) a rizika získaného z okolních sousedních oblastí, které reprezentují apriorní informaci (Clayton a Bernardinelli, 1996).

Výpočet empirického bayesovského odhadu rizika onemocnění vychází z následujících vztahů (Bailey a Gatrell, 1995),

$$\widehat{\theta_i} = w_i r_i + (1 - w_i)\gamma_i \qquad [4.1]$$

kde θ_i představuje empirický bayesovský odhad míry v oblasti i, w_i jsou váhy pro výpočet lokálního a okolního (sousedského) odhadu, r_i je lokální riziko v oblasti i a γ_i je průměr apriorní distribuce. Míru rizika lze vyjádřit jako

$$r_i = \frac{y_i}{n_i} \qquad [4.2]$$

kde y_i představuje četnost případů n_i počet obyvatel v oblasti i. Váhy w_i ve vzorci 4.1 jsou určeny jako

$$w_i = \frac{\phi_i}{\left(\phi_i + \frac{\gamma_i}{n_i}\right)} \qquad [4.3]$$

kde ϕ_i je rozptyl prioru, y_i je průměr prioru a n_i je populace v oblasti i. Samotný odhad je proveden pomocí pravděpodobnosti a integrálové aproximace (Lawson et al., 2003). Odhad vypočítaný na základě [4.1] tíhne ke globálnímu průměru. Metoda umožňuje odhadnout také lokálně vyhlazené odhady tak, že do výpočtu zahrne pouze sousední oblasti a globální průměr bude nahrazen průměry lokálními. Lokální tzv. adaptivní vyhlazování snižuje nestabilní riziko vůči lokálnímu průměru, z čehož vyplívá, že odhady v oblastech s větší populací (větším množstvím informace) jsou vyhlazeny méně než odhady v oblastech méně zalidněných, kde je současně vyšší variabilita jevu (Beale et al., 2008). Další informace o mapování nemocí s využitím bayesovských metod a včetně jejich využití pro průzkum prostorového vzoru poskytují Lawson et al. (2003), Waller (2005) nebo Gelfand et al. (2010).

Vyhlazení průměrné míry incidence a SIR u kampylobakteriózy v České republice v letech 2008—2012

Data využitá v tomto kroku jsou vztažena k částem obcí České republiky. Jako střední populace byl zvolen počet obyvatel vycházející ze SLDB 2011, věková struktura využitá ve standardizaci dat pak vychází ze z věkové struktury nadřazených obcí uvedených rovněž ve výsledcích SLDB 2011. Každá část obce tak obsahovala kumulovaný počet onemocnění ve studovaných letech a dopočítanou průměrnou četnost onemocnění. Z důvodu zobrazení ve formě kartogramu a vzájemné srovnatelnosti měr morbidity byla využita standardizace vycházející z informace o četnosti výskytu onemocnění v celé ČR. Na základě takto definované standardní populace bylo možné odhadnout očekávaný průměrný počet případů onemocnění v částech obcí a zjistit tak relativní riziko (SIR) onemocnění kampylobakteriózou.

Výpočet standardizovaných měr, očekávaného počtu případů i relativního rizika v územních jednotkách proběhl v prostředí **R** s využitím balíků *epiR* (Stevenson, 2015), *epitools* (Aragon, 2012) *a SpatialEpi* (Chen et al., 2014). Vytvořená data byla následně připojena k polygonové vektorové vrstvě části obcí a takto bylo možné dále data shlazovat s pomocí programu GeoDa. Standardizace samotná je procesem, který již umožní shlazení zjištěných měr morbidity. Dále bylo na vyhlazení vypočítané průměrné incidence a relativního rizika použito globálního Bayesova vyhlazení vycházejícího z negativního binomického rozdělení a také jeho lokální obdoba založená na sousedství 1. řádu typu královna, tzn. sousedy jsou všechny územní jednotky, které sdílejí alespoň část hranice územní jednotky.

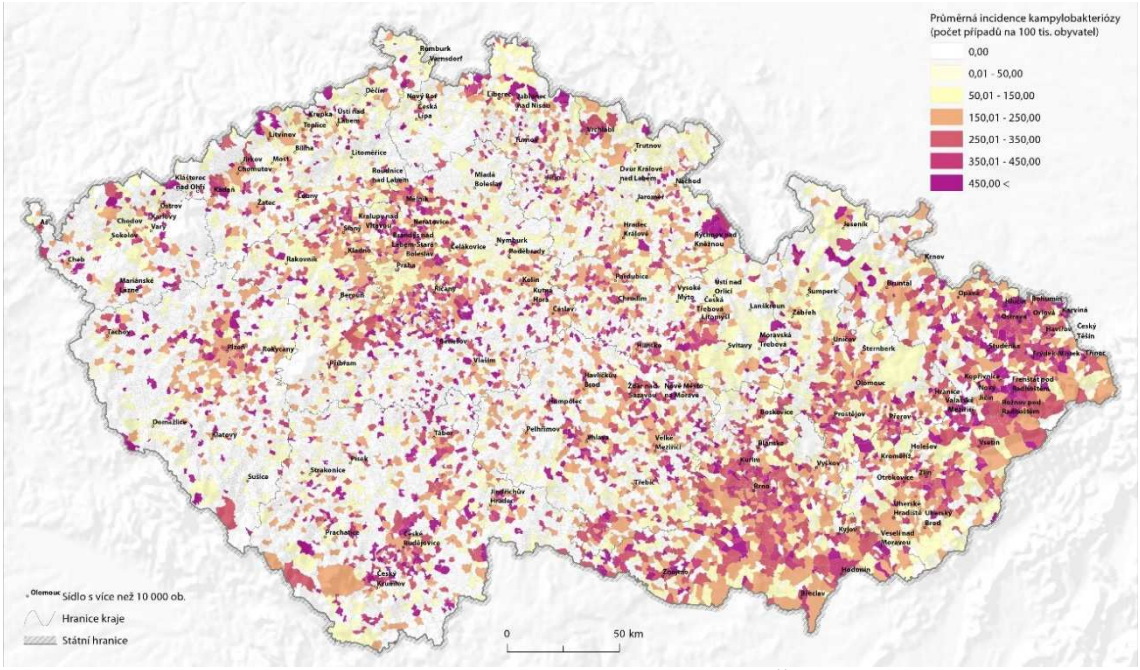
Základní statistické charakteristiky, které shrnují absolutní pětiletou četnost onemocnění v jednotlivých městských částech i její vyhlazenou podobu, jsou uvedeny v Tabulce 3. Průměrná hodnota a směrodatná odchylka nejsou díky některým odlehlým hodnotám právě zcela charakteristikami popisujícími vhodně absolutní četnost onemocnění v částech obcí, která je silně asociována především s populací a její distribuce neodpovídá normálnímu rozdělení. Ve velkém množství části obcí navíc nebyl ve sledovaném období zaznamenán žádný výskyt nemoci. Pro zhodnocení měr morbidity jsou však vhodnými charakteristikami polohy a variability medián a mezikvartilové rozpětí. Kromě statistických charakteristik úplné datové sady jsou v Tabulce 3 zobrazeny i statistické charakteristiky odpovídající pouze obcím se zaznamenaným výskytem onemocnění, které se výrazně odlišují. Z tabulky je patrné, že změnou strategie vyhlazování se příliš nemění statistické charakteristiky v rámci ČR.

Tab. 3 Statistické charakteristiky pozorované, standardizované a vyhlazené pětileté kumulativní četnosti výskytu kampylobakteriózy v částech obcí v ČR v letech 2008—2012

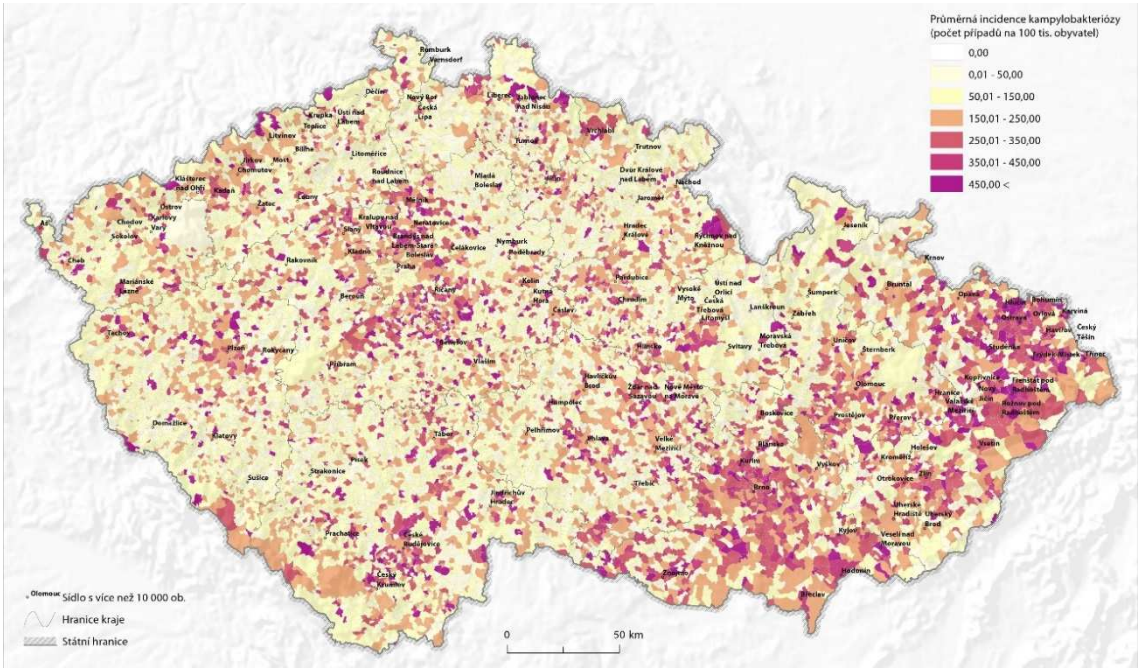
Četnost	Průměr		Medián		Směrodatná odchylka		Mezikvartilové rozpětí	
	Úplná	Redukce	Úplná	Redukce	Úplná	Redukce	Úplná	Redukce
Pozorovaná	6,54	15,99	0,00	4,00	34,81	53,05	3,00	8,00
Standard.	6,54	14,39	1,23	4,11	26,16	39,40	3,14	7,63
EBB	6,60	15,84	0,33	3,69	34,72	52,97	2,12	8,60
LEBQ1	6,64	15,65	0,63	3,56	34,66	52,92	2,52	8,42

Označení Úplná označuje charakteristiky pro úplnou datovou sadu; označení Redukce se vztahuje pouze pro části obcí s alespoň jedním případem. Pozorovaná – zaznamenané případy; Standard. – populací standardizovaný počet případů; EBB – četnosti vycházející z globálního Bayesova vyhlazení založeném na negativním binomickém rozdělení; LEBQ1 – četnosti vycházející z lokálního Bayesova vyhlazení založeném na negativním binomickém rozdělení a sousedství typu královna 1. řádu

Na Obrázku 11 je zobrazena průměrná hrubá incidence kampylobakteriózy v částech obcí v ČR v letech 2008—2012. Incidence kampylobakteriózy je zde prezentována jako počet případů onemocnění v části obce na 100 tisíc obyvatel. Tím, že nebylo využito žádné standardizace ani vyhlazení, tak mapa zdůrazňuje řídce osídlené oblasti, kde jakýkoliv případ nemoci výrazně zvýší hrubou incidenci - nejtmavší oblasti. Příkladem mohou být menší obce jihovýchodně od Prahy.



Obr. 11 Hrubá průměrná incidence kampylobakteriózy v částech obcí v ČR v letech 2008—2012 v počtech případů na 100 tisíc obyvatel

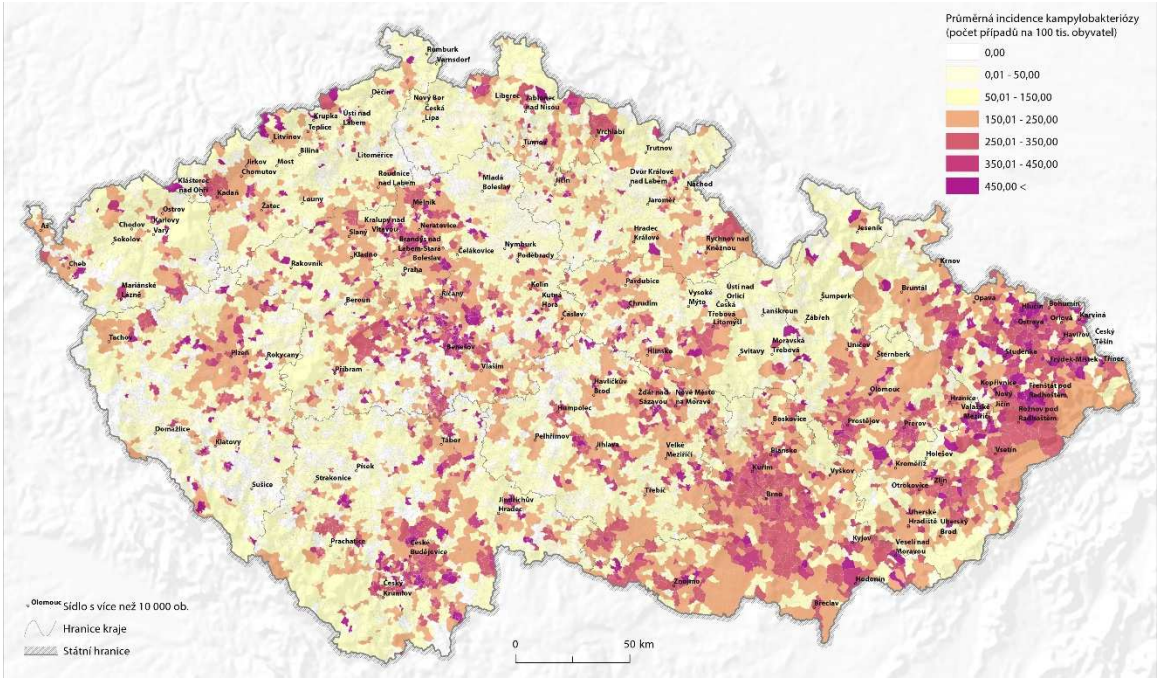


Obr. 12 Vyhlazená standardizovaná průměrná incidence kampylobakteriózy v částech obcí v ČR v letech 2008—2012 v počtech případů na 100 tisíc obyvatel, která vznikla pomocí globálního Bayesova vyhlazení založeného na negativním binomickém rozdělení

Následující kartogram (Obrázek 12) prezentuje standardizovaná data upravená pomocí globálního Bayesova vyhlazení vycházejícího z negativního binomického rozdělení, které k vyhlazování využívá globálního průměru. Na mapě je jasně patrné, že i oblastem bez zjištěného výskytu onemocnění je přiřazena nízká incidence. Na druhou stanu více než dvě stovky obcí zařazených původně mezi obce s nejvyšší incidencí nyní spadají do kategorií s její nižší hodnotou.

Třetí mapa (Obrázek 13) vykresluje průměrnou standardizovanou incidenci kampylobakteriózy vyhlazenou pomocí lokálního Bayesova vyhlazení vycházejícího z negativního binomického rozdělení a sousedství 1. řádu typu královna. To znamená, že tentokrát není využit celorepublikový průměr, ale pouze adaptivně se měnící lokální průměr vypočítaný z nejbližšího okolí dané části obce. V porovnání s předchozími případy představuje poslední mapa největší míru shlazení, kdy jsou vytvořeny prostorově spojité oblasti. Tento typ vyhlazení přiřadil části obcí bez výskytu onemocnění malou incidenci a naopak v částech obcí s vysokou incidencí byla incidence snížena. I přes velkou míru vyhlazení (nebo právě proto) poskytuje poslední metoda nejméně fragmentovaný pohled na prostorovou distribuci incidence onemocnění a umožňuje nejjasněji vnímat jeho prostorový vzor s převahou míst s vysokou incidencí vyskytujících se především na Moravě a ve Slezsku, případně ve Středních Čechách a v okolí Českých Budějovic a Plzně. Četnosti částí obcí spadajících do jednotlivých kategorií v mapě jsou uvedeny v Tabulce 4.

Ačkoliv z pohledu základních statistických charakteristik celého území se různé typy vyhlazování dat výrazně nelišily, tak po zobrazení do mapy podávají různé informace. Pro srovnání sousedních obcí je obecně vhodnější vyhlazování a/nebo standardizace využít, což zajišťuje lepší srovnatelnost měr.



Obr. 13 Vyhlazená standardizovaná průměrná incidence kampylobakteriózy v částech obcí v ČR v letech 2008—2012 v počtech případů na 100 tisíc obyvatel, která vznikla pomocí lokálního Bayesova vyhlazení založeného na negativním binomickém rozdělení a sousedství typu královna 1. řádu

Tab. 4 Počty částí obcí spadajících do jednotlivých kategorií průměrné incidence (počet případů na 100 tisíc osob) v závislosti na typu vyhlazování.

	Průměrná incidence						
	0,00	0,01−50,00	50,01−150,00	150,01−250,00	250,01−350,00	350,01−450,00	450,00 <
Hrubá	8955	321	1973	1642	932	463	837
EBB	0	2657	7170	3139	1022	499	636
LEBQ1	1491	2064	5863	3545	1282	413	465

Hrubá – hrubá průměrná incidence; EBB – průměrná standardizovaná incidence vycházející z globálního Bayesova vyhlazení založeného na negativním binomickém rozdělení; LEBQ1 – průměrná standardizovaná incidence vycházející z lokálního Bayesova vyhlazení založeného na negativním binomickém rozdělení a sousedství typu královna 1. řádu

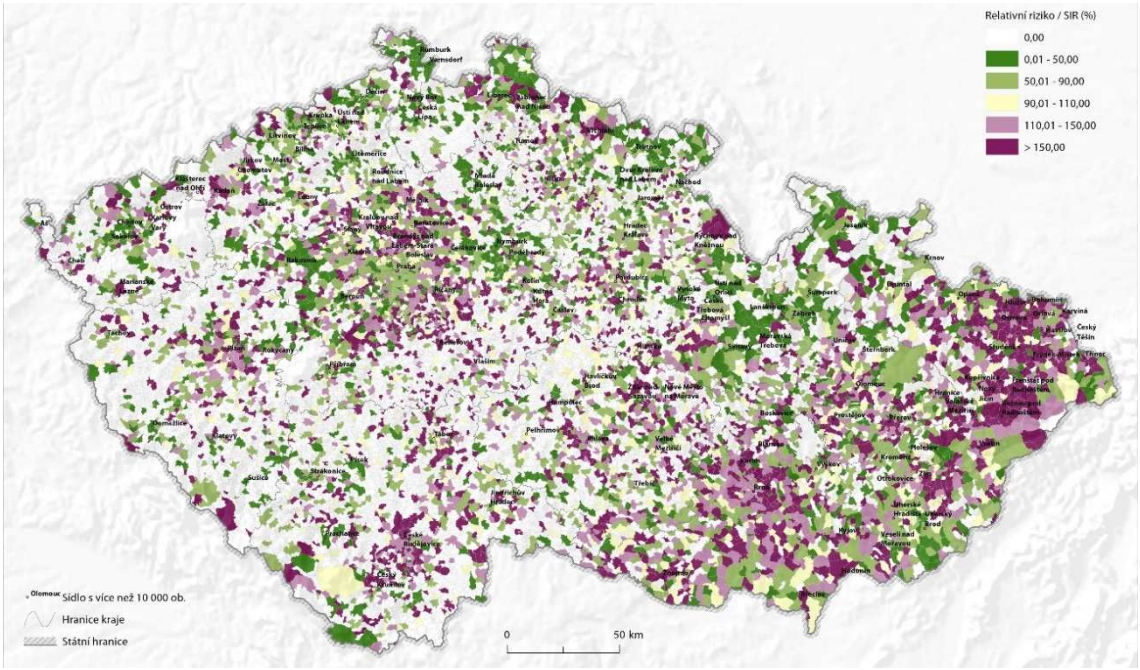
Tabulka 5 numericky shrnuje představené typy vyhlazování prostřednictvím základních statistických charakteristik incidence. Zatímco nelze spatřit výrazný rozdíl mezi statistickými charakteristikami hrubé průměrné incidence a globálního empirického Bayesova vyhlazení (EBB), tak v případě lokálního shlazení (LEBQ1) je rozdíl patrný zejména u průměru a směrodatné odchylky úplné datové sady. Jeho medián a mezikvartilové rozpětí je naopak s předchozími postupy srovnatelný a zůstává stabilní. Při výpočtu hrubé incidence nedošlo k žádnému shlazení, a tak regiony, kde nebyla nemoc přítomna, mají tuto incidenci nulovou a nulový je i medián. Zbylé postupy však velké množství regionů bez výskytu onemocnění vyhladily díky okolním hodnotám, a tak je i medián vyšší. Nejmenší variabilitu datové sady má globálně vyhlazená incidence. S postupnou aplikací složitější metody vyhlazování je patrné i zvyšování průměrné hodnoty incidence.

Tab. 5 Statistické charakteristiky průměrné hrubé incidence a průměrné standardizované incidence vyhlazené pomocí globálního a lokálního Bayesova vyhlazení. Průměrná incidence kampylobakteriózy v částech obcí ČR v letech 2008—2012 je uvedena v počtech případů na 100 tisíc obyvatel

Incidence	Průměr		Medián		Směrodatná odchylka		Mezikvartilové rozpětí	
	Úplná	Redukce	Úplná	Redukce	Úplná	Redukce	Úplná	Redukce
Hrubá	141,68	347,37	0,00	194,20	826,58	1266,44	155,20	202,40
EBB	160,02	262,88	112,60	194,00	275,91	400,63	116,60	177,45
LEBQ1	310,56	274,22	116,00	174,80	8184,65	1016,72	139,40	143,00

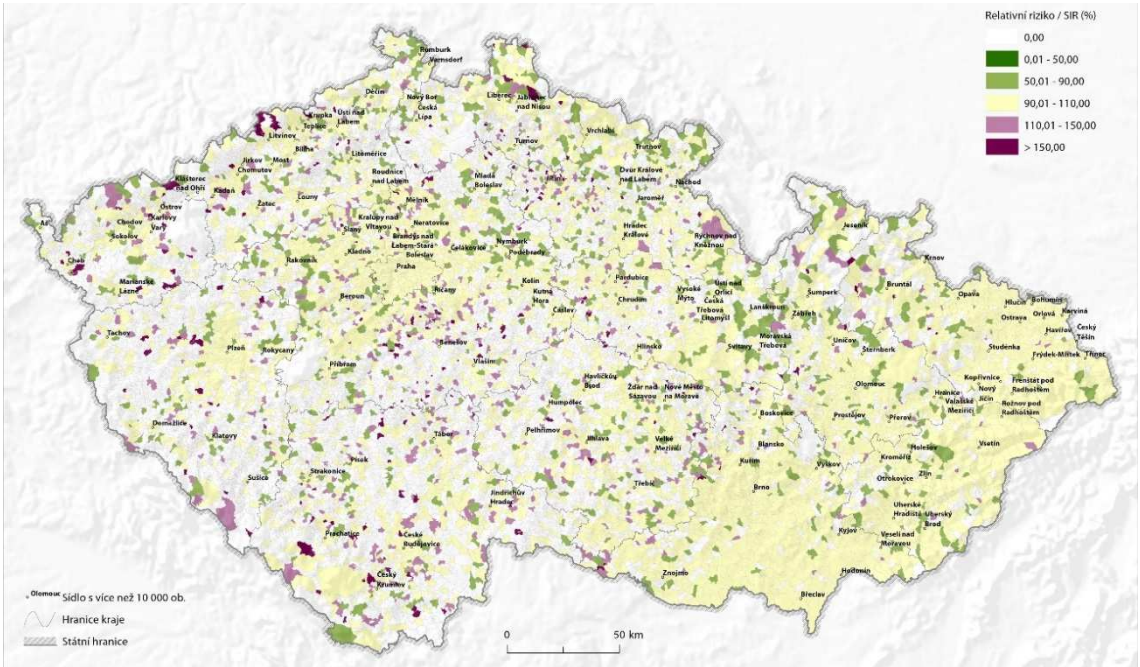
Sloupec Úplná označuje charakteristiky pro úplnou datovou sadu; označení Redukce se vztahuje pouze pro části obcí s alespoň jedním případem. Hrubá – hrubá průměrná incidence; EBB – průměrná standardizovaná incidence vycházející z globálního Bayesova vyhlazení založeného na negativním binomickém rozdělení; LEBQ1 – průměrná standardizovaná incidence vycházející z lokálního Bayesova vyhlazení založeného na negativním binomickém rozdělení a sousedství typu královna 1. řádu

Kromě informace o reálně pozorovaných a průměrných četnostech případů onemocnění v částech obcí, které jsou získány přímo, je možné díky předpokládanému rozdělení pravděpodobnosti, využití standardizace a také bayesovskému vyhlazování získat různé odhady předpokládaných četností onemocnění v území. Díky pozorované a předpokládané hodnotě je možné zjistit relativní riziko kampylobakteriózy ve zvoleném území, které je vyjádřeno jako poměr reálně pozorované a předpokládané hodnoty. Relativní riziko je často udáváno v procentech. Relativní riziko odpovídající 100 % představuje shodu mezi oběma hodnotami a hodnota 50 % představuji území, kde je pouze polovina teoreticky odhadovaných případů a jde tak o území zdravější (méně rizikové) než bylo předpokládáno. Relativní riziko také umožnuje přímé srovnání mezi zkoumanými jednotkami, protože je do jeho výpočtu možné zahrnout informaci o struktuře obyvatelstva.



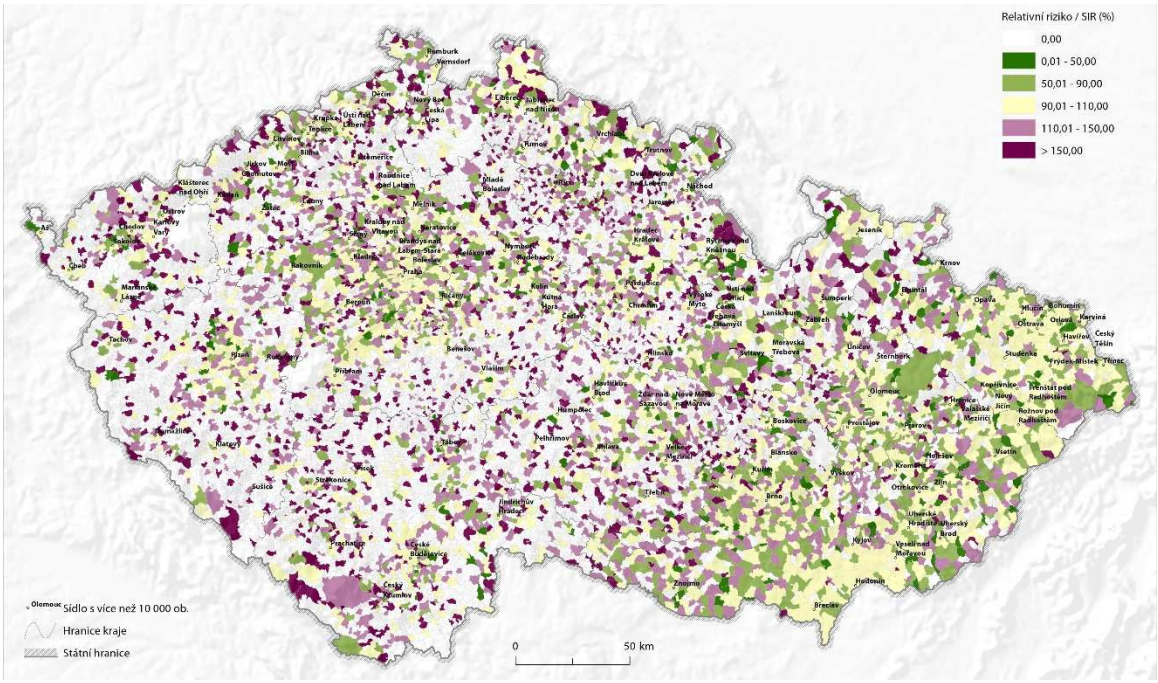
Obr. 14 Relativní riziko kampylobakteriózy (v %) v částech obcí v ČR v letech 2008—2012

První mapa (Obrázek 14) zobrazuje relativní riziko, u kterého očekávané hodnoty odpovídají pouze struktuře obyvatelstva v částech obcí a u kterého není aplikováno žádné vyhlazování. Je možné si všimnout, že ve shodě se zjištěními vyplývajícími z map průměrné incidence jsou Morava a zejména Slezsko výrazně rizikovějšími oblastmi než zbytek České republiky s výjimkou Plzeňska, Českobudějovicka a okolí Prahy. Vychází-li výpočet relativního rizika z globálně bayesovsky vyhlazeného odhadu průměrné incidence (Obrázek 15), pak je výsledný vjem výrazně odlišný a výsledkem je extrémně shlazený povrch, který tíhne ke shodným hodnotám očekávaných a pozorovaných četností onemocnění, a tedy globálně průměrně zdravým částem obcí. Pouze malé množství částí obcí se výrazně odchyluje od průměrného rizika.



Obr. 15 Relativní riziko kampylobakteriózy (v %) v částech obcí v ČR v letech 2008—2012 získané na základě globálního Bayesova vyhlazení založeného na negativním binomickém rozdělení

Nakonec bylo vypočítáno a vizualizováno relativní riziko vycházející z lokálně bayesovsky vyhlazeného odhadu průměrné incidence (Obrázek 16). Výsledná mapa opět podává zcela jiný pohled na relativní riziko onemocnění kampylobakteriózou v porovnání s přechozími dvěma mapami. Nejzajímavější je fakt, že celkový vjem relativního rizika je zcela v rozporu s mapou průměrné incidence (Obrázek 13), která zobrazuje výrazně vyšší relativní riziko v oblasti Čech než na Moravě a to především v místech menších měst a obcí. Díky tomuto faktu vzniká situace, při které lze lokální bayesovské vyhlazování doporučit pro vizuální průzkum incidence, ale nelze jej doporučit pro zkoumání a vizualizaci relativního rizika vzhledem k nestabilním odhadům mimo hustěji osídlená území. Pro neshlazené hodnoty je pak situace přesně opačná – lze jej vhodně využít u relativního rizika (vzhledem k faktu, že jde o nepřímo standardizované hodnoty), ale neposkytuje zcela srovnatelné hodnoty v případě hrubé incidence. Konkrétní četnosti částí obcí spadajících do jednotlivých kategorií relativního rizika v mapách jsou uvedeny v Tabulce 6.



Obr. 16 Relativní riziko kampylobakteriózy (v %) v částech obcí v ČR v letech 2008—2012 získané na základě lokálního Bayesova vyhlazení založeného na negativním binomickém rozdělení a sousedství typu královna 1. řádu

Tab. 6 Počty částí obcí spadajících do jednotlivých kategorií relativního rizika / SIR (%) v závislosti na typu vyhlazování.

	Relativní riziko					
	0,00	0,01–50,00	50,01–90,00	90,01–110,00	110,01–150,00	150,00 <
Hrubá	8955	2187	517	693	981	1790
EBB	8955	38	599	4374	649	208
LEBQ1	8955	902	573	1909	1540	1244

Hrubá – relativní riziko; EBB – relativní riziko vycházející z globálního Bayesova vyhlazení založeného na negativním binomickém rozdělení; LEBQ1 – relativní riziko vycházející z lokálního Bayesova vyhlazení založeného na negativním binomickém rozdělení a sousedství typu královna 1. řádu

Základní statistické shrnutí pro všechny tři metody výpočtu relativního rizika je zobrazeno v Tabulce 7. Globálně vyhlazené odhady relativního rizika (EBB) vykazují nejnižší střední hodnoty i variabilitu (s výjimkou mezikvartilového rozpětí úplné datové sady). To potvrzuje i mapa na Obrázku 15, kde převládají buď obce bez výskytu onemocnění (a tedy i relativního rizika) nebo hodnoty kolem 100 %. Nejvyšší směrodatná odchylka i průměr je zaznamenána v případě nevyhlazeného relativního rizika (RR). Zajímavá je hodnota mediánu, která se v případě úplné datové sady u všech třech metod pohybuje na hodnotě 0 % z důvodu velkého množství území bez záznamu onemocnění. V případě redukované datové sady je hodnota kolem 100 % a prokazuje se tedy vliv standardizace. Výhodou mapování relativního rizika je zachování oblastí bez záznamu onemocnění jako oblastí bez rizika, avšak místní rozdíly mezi metodami při výpočtu relativního rizika poukázaly v rámci této studie na omezenou použitelnost vyhlazování s pomocí bayesovských metod.

Tab. 7 Statistické charakteristiky relativního rizika (SIR) a relativního rizika vyhlazeného pomocí globálního a lokálního Bayesova vyhlazení. Relativní riziko kampylobakteriózy v částech obcí ČR v letech 2008—2012 je uvedeno v procentech (%), hodnota 100 % vyjadřuje shodu pozorované a očekávané četnosti případů onemocnění.

Relativní riziko	Průměr		Medián		Směrodatná odchylka		Mezikvartilové rozpětí	
	Úplná	Redukce	Úplná	Redukce	Úplná	Redukce	Úplná	Redukce
RR	73,32	179,40	0,00	100,27	427,78	654,79	80,32	105,15
EBB	43,09	105,44	0,00	100,01	67,74	68,21	98,95	8,46
LEBQ1	51,52	126,06	0,00	106,25	79,78	78,59	100,06	46,51

Sloupec Úplná označuje charakteristiky pro úplnou datovou sadu; označení Redukce se vztahuje pouze pro části obcí s alespoň jedním případem. RR – relativní riziko – poměr pozorovaného a očekávaného množství případů (%); EBB – relativní riziko vycházející z globálního Bayesova vyhlazení založeného na negativním binomickém rozdělení; LEBQ1 – relativní riziko vycházející z lokálního Bayesova vyhlazení založeného na negativním binomickém rozdělení a sousedství typu královna 1. řádu

4.4 Tvorba spojitého časoprostorového povrchu incidence

Zatímco kriging je velmi známou a dobře popsanou geostatistickou metodou (Bivand et al., 2008; Hengl, 2009), která je v geovědách používaná již několik desetiletí, tak implementace jejího časoprostorového rozšíření je záležitostí mnohem současnější. Samotná myšlenka o zahrnutí časové složky do procesu interpolace prostorových dat se postupně objevovala od konce 20. století, kdy byla vytvářena i její teorie (Rouhani a Myers, 1990; Wikle a Cressie, 1999; Kyriakidis a Journel, 1999). Úspěšné implementaci a rozšíření časoprostorového krigingu však bránila jeho výpočetní náročnost. Svého rozvoje se tak dočkal až v současnosti. Časoprostorový kriging využívá korelace a autokorelace dat, které jsou vyhodnocovány pomocí časoprostorového variogramu popisujícího prostorové, časové i časoprostorové vztahy v datech zároveň (Gräler et al., 2012). Díky své novosti není metoda příliš využívána v kontextu zdravotnických dat, příkladem však mohou být práce Gethinga (2006, 2007). Komplexní popis metody včetně jejích dalších alternativ, modelů a aplikací lze najít v Kyriakidis a Journe (1999), Myers (2004), Heuvelink a Griffith (2010) a Hengl et al. (2012).

4.4.1 Koncept časoprostorového krigingu

V případě modelování časoprostorových interakcí je potřeba si uvědomit některé speciální vlastnosti, které obvykle časoprostorová pole (spojité povrchy) a proměnné mívají. První

z nich je jednosměrnost času – tzn. události v minulosti mohou ovlivnit současnost a budoucnost, ale nikdy naopak. Další důležitou vlastností je anizotropie času, která představuje jednotku času odpovídající jednotce vzdálenosti tak, aby byly vzájemně ekvivalentní a mohly úspěšně popsat závislosti ve všech dimenzích (Gräler et al., 2012).

Varianci časoprostorového pole *Z(s, t)* lze pomocí geostatistické funkce – variogramu - popsat vztahem (Gräler, 2012),

$$\gamma(h,u) = E\big(Z(s,t) - Z(s+h,t+u)\big)^2 \qquad [4.4]$$

kde γ je variance v kroku vzdálenosti h a kroku času u, která je odhadem čtverce rozdílů hodnoty pole v místě s a čase t a hodnoty pole v místě $(s + h)$ a čase $(t + u)$. V jakémkoliv místě a čase (s, t) pak lze vyjádřit empirickou verzi variogramu,

$$\hat{\gamma}(h,u) = \frac{1}{2|N_{h,u}|} \sum_{(i,j)\in N_{h,u}} \Big(Z(s_i,t_i) - Z\big(s_j,t_j\big)\Big)^2 \qquad [4.5]$$

kdy pomocí indexů i, j jsou vyjádřena určení místa v prostoru a čase,

$$N_{h,u}\left\{(i,j)\middle|\begin{matrix} h-\ \epsilon_s \le \|s_i - s_j\| \le h+\epsilon_s \\ u-\epsilon_t \le t_i - t_j \le u+\epsilon_t \end{matrix}\right\} \qquad [4.6]$$

Metrický kriging, který je použit pro interpolaci i v této práci, se snaží přirozeně rozšířit koncept geografického prostoru zahrnutím času jako dalšího rozměru, tzn. vytvořením časoprostoru. Z tohoto důvodu je nutné transformovat časový údaj tak, aby odpovídal prostorovému údaji, tj. formovat korekci časoprostorové anizotropie κ. Jinými slovy jde o to určit, jaká vzdálenost odpovídá časovému intervalu. Díky tomu je možné pracovat s časem, prostorem i jejich kombinací shodným způsobem a získat tak společný metrický časoprostorový kovarianční model C_m (Gräler, 2013):

$$C_m(h,u) = C_j\left(\sqrt{h^2 + (\kappa\cdot u)^2}\right) \qquad [4.7]$$

Metrický časoprostorový variogram je pak definován jako

$$\gamma_m(h,u) = \gamma_j\left(\sqrt{h^2 + (\kappa\cdot u)^2}\right) \qquad [4.8]$$

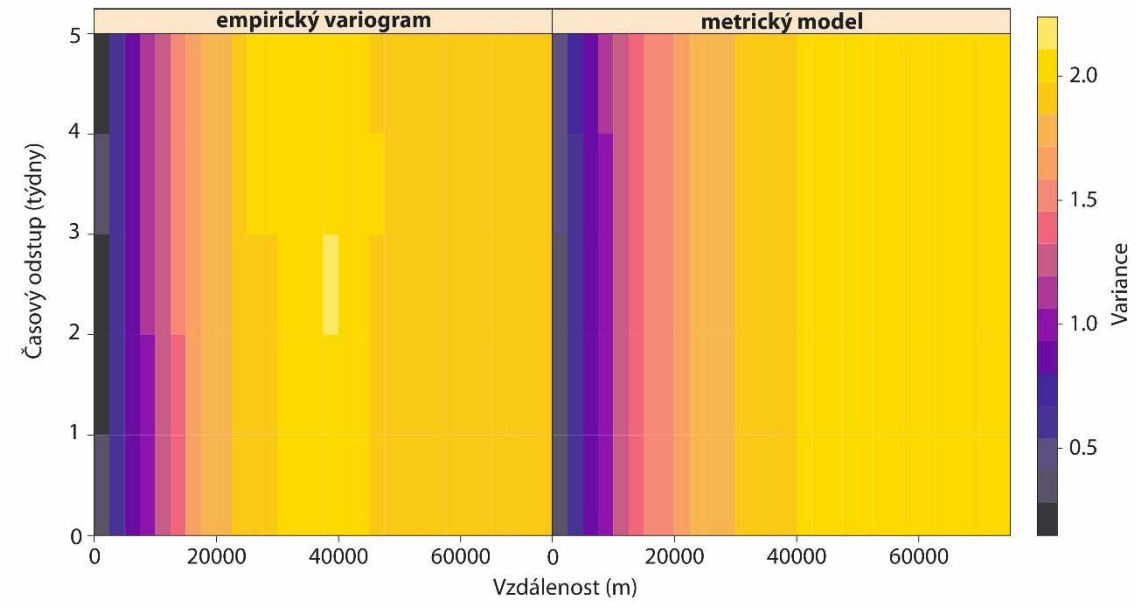
Kde γ_j může být kterýkoliv známý model variogramu včetně modelu s nugget efektem.

4.4.2 Časoprostorový kriging jako nástroj tvorby povrchu spojitého v čase i prostoru

Hlavním cílem využití časoprostorového krigingu v rámci studia kampylobakteriózy v letech 2008—2012 je vytvoření spojitého povrchu průměrné týdenní incidence v osídlených místech České republiky. Výhodou týdenního spojitého povrchu incidence oproti pouhé agregaci dat do společných bodů, pravidelného rastu či administrativních jednotek je především eliminace uměle vytvořených hranic, a tedy postupně se měnící hodnoty povrchu a s tím spojené přirozené porozumění uživatele. Při prohlížení výstupu časoprostorového krigingu v prostředí umožňujícímu i interakci uživatele s časovou složkou je možné pozorovat postupné změny a opakující se časoprostorové vzory, které mohou pomoci upozornit na různé typy chování onemocnění v různých částech zkoumaného území a různých obdobích roku. Ačkoliv je kriging nejběžněji používán jako nástroj pro interpolaci a predikci prostorových jevů, tak v této studii slouží zejména jako způsob areálové interpolace odstraňující vliv

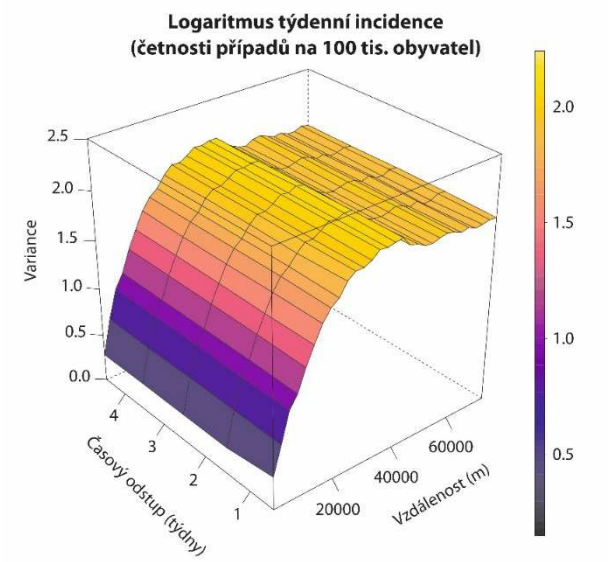
administrativních (či jiných) hranic a vytvářející spojitý povrch vhodný pro následnou vizualizaci. Jde tak o variantu uplatnění dasymetrické metody, která pro rozptýlení vlivu administrativních hranic využívá síly vztahu mezi lokalitami v prostoru i čase.

Data vstupující do interpolace jsou tvořena týdenními odhady hrubé incidence v územích o velikosti 2×2 km. Území České republiky bylo překryto pravidelným čtvercovým gridem o velikosti strany 2 km, do kterého byly postupně agregovány záznamy o jednotlivých případech onemocnění v týdenních intervalech. Tímto postupem vzniklo 261 gridů, pro které byla vypočítána hrubá incidence. Odhad počtu obyvatel v buňkách gridu, který vstupoval do výpočtu týdenní hrubé incidence, vycházel z datové sady *GEOSTAT 1 km² population grid 2011*, která je společným dílem statistických úřadů států EU pod hlavičkou Eurostatu. Následně byl vytvořen grid logaritmované týdenní hrubé incidence, který na rozdíl od týdenní hrubé incidence splňoval podmínku normality dat. Takto připravená data sloužila jako základ pro výpočet experimentálního (empirického) časoprostorového variogramu (Obrázek 17 – vlevo) popisujícího chování incidence v prostoru, čase i jejich kombinaci. Vpočetní čas průběhu zjišťování empirického variogramu pro oblast České republiky čítal 35 hodin a 26 minut (Intel Core i7-3770 CPU 3.90 GHz, 8 GB RAM). Na průběh experimentálního variogramu je pak nasazen vhodný teoretický model (Obrázek 17 – vpravo), který již slouží k interpolaci. Z obrázku 17 je patrné, že k největší interakci (časoprostorové závislosti) dochází v oblastech, které jsou si blízké jak v čase (vertikální osa), tak i v prostoru (horizontální osa) – důkazem jsou sobě podobné vzory na levé straně empirického i metrického modelu.



Obr. 17 Experimentální (empirický) (vlevo) a teoretický (vpravo) model časoprostorového variogramu

Pro interpolaci byl konkrétně zvolen metrický časoprostorový model krigingu, který je určen na základě exponenciálního modelu variogramu s parametry: *nugget = 0,15; práh = 1,94; rozsah = 14150,46 m* a *časoprostorová anizitropie = 544,58 m/den,* které byly zvoleny na základě empirického zkoumání variogramu v kombinaci s optimalizací. Grafická podoba teoretického modelu variogramu je zobrazena na Obrázku 18. Model předpokládá, že maximální prostorová závislost je mezi body vzdálenými od sebe do 14 km, s tím, že ohnisko nemoci se za týden může přesunout až o zhruba 3,8 km. Největší časová interakce je mezi po sobě následujícími týdny, ale teoreticky až jeden měsíc.



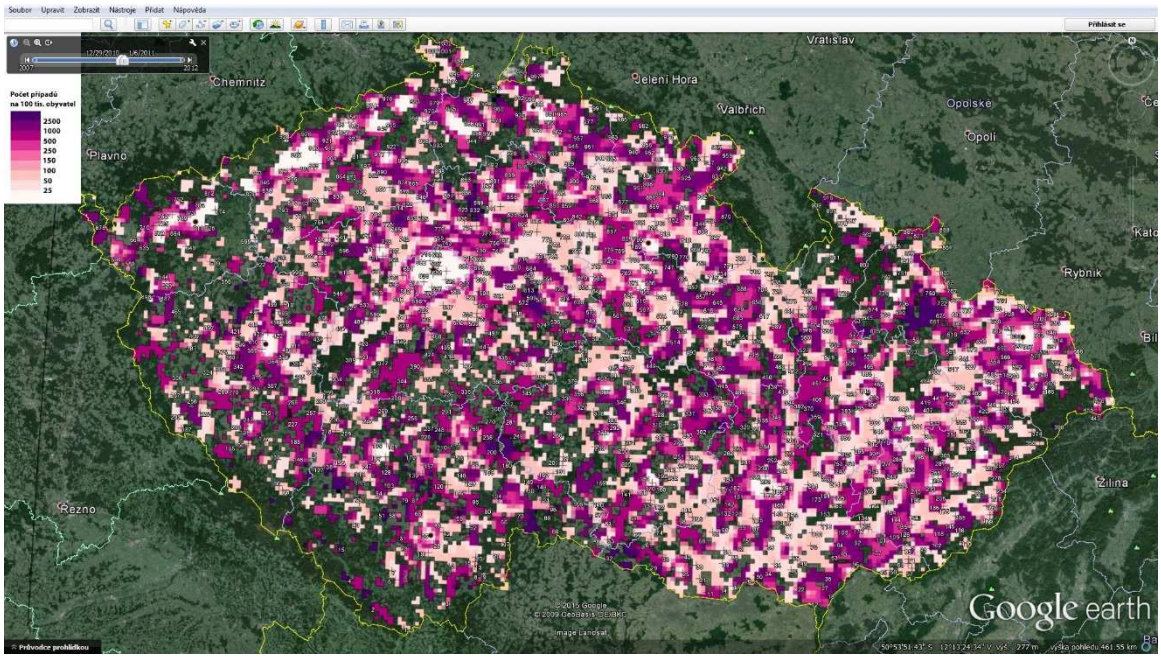
Obr. 18 Časoprostorový teoretický model variogramu použitý pro interpolaci

Stejně jako byl časově a výpočetně náročné získání empirického variogramu, tak byl náročný i výpočet interpolace samotné. Prostředí **R** je známo svou náročností na výpočetní paměť a právě z toho důvodu nebyla možná interpolace týdenní hrubé incidence pro celé území ČR. Celé území tak muselo být rozděleno na 1372 menších ploch, které už mohly být efektivně interpolovány. Celý proces výpočtu trval 14 hodin a 7 minut. Výsledné dílčí plochy, tedy spíše interpolované časové řezy v dílčím území, byly spojeny dohromady a vztaženy pouze na osídlená území ČR, která byla získána z datové sady *CORINE Land Cover 2006*[33]. Navazujícím krokem byla kategorizace výsledků podle incidence (<25; 25—50; 51—100; 101—150; 151—250; 251—500; 501—1000; 1001—2500; >2500 případů na 100 tisíc obyvatel) a tvorba KML souboru, který umožní prohlížení spojitého povrchu týdenní hrubé incidence jak v prostoru, tak i v čase v prostředí Google Earth (Obrázek 19). Interpolovaný povrch je navíc doplněn 1000 body, které nesou informaci o přesných hodnotách v podobě grafu časové řady v daném místě. Tento počet byl zvolen z důvodu čitelnosti výstupu, kterou vyšší množství bodů snižuje.

Vizuálním hodnocením výsledného KML souboru, který díky podpoře časové složky dat umožňuje snadné prohlížení a (geo)vizuální analýzu, je možné identifikovat několik informací. Některé z nich potvrzují obecně známá fakta o kampylobakterióze, jejíž incidence se zvyšuje v teplejších měsících (květen—září) s vrcholem v srpnu. Sezónní cyklus je většinou méně patrný v hustě osídlených oblastech a naopak více patrný ve venkovských oblastech a také v zázemí velkých měst, která často slouží jejich obyvatelstvu k rekreaci. Zvýšenou incidenci je možné pozorovat také v zimních měsících, a to především na úpatí Krkonoš a Jeseníků během zimní turistické sezóny/jarních prázdnin.

Celý proces probíhal s pomocí programovacího jazyka a prostředí **R**, nejdůležitějšími balíky pak byly *rgdal* (Bivand et al., 2014), *raster* (Hijmans, 2014) , *spacetime* (Pebesma, 2012), *gstat* (Pebesma a Gräler, 2014) a *plotKML* (Hengl et al., 2014). Celý proces je podrobněji popsán v kapitole 8.

[33] *http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster*



Obr. 19 Zobrazení časoprostorového krigingu v prostředí Google Earth

4.5 Shrnutí a formulace hypotéz o shlukování

Na základě shrnutí datové sady jako celku pomocí základních statistických charakteristik a grafů a také na základě mapování výskytu, incidence a relativního rizika kampylobakteriózy lze vyřknout první zjištění a případně i hypotézy s ohledem na navazující kroky, které se týkají především shlukování v čase a prostoru, ale také možných environmentálních faktorů, které mohou podmiňovat výskyt onemocnění.

První pracovní hypotézy dalších dílčích cílů lze stanovit s přihlédnutím k základním charakteristikám nakažených osob. Skupinou obyvatelstva s nejvyšší průměrnou incidencí onemocnění v letech 2008—2012 jsou nejmenší děti ve věku do 4 let. Zvýšená incidence je patrná také u starších dětí, mládeže a dospělých do 30 let. Vliv pohlaví není výrazně viditelný. Na základě těchto zjištění je tedy možné očekávat, že obce s nižším průměrným věkem obyvatelstva (případně vyšším podílem dětí na celkové populaci) jsou více ohroženy než obce s vyšším průměrným věkem. Vzhledem k tomu, že nejvíce ohroženy jsou děti a mladí lidé, tak i nejčastější profesí nakažených je kategorie dítě či student, následovaná nezjištěnými údaji, důchodci a nezaměstnanými. Z toho pohledu tak sice nelze usuzovat na přímý vliv pracovního prostředí na incidenci kampylobakteriózy, ale je možné zkoumat například procentuální podíl pracujících v zemědělství či potravinářském (zejména masozpracovatelském) průmyslu, jehož možný vliv byl indikován i v zahraničních studiích (Arsenault, 2010). Nejčastějším zdrojem nákazy v našich podmínkách je kuřecí maso, následované dalšími druhy mas a masných výrobků. Společně s informací, že počet případů v průběhu roku je vyšší v teplejší části roku s vrcholem výskytu onemocnění v srpnu je možné předpokládat, že četnost onemocnění a incidence souvisí s rostoucí popularitou grilování a venkovní úpravy masa (Ambrožová, 2011). Je také možné, že vyšší riziko onemocnění mohou mít i území s vyššími průměrnými teplotami (např. delší grilovací sezóna, rekreační oblasti apod.).

Pro prezentaci ve formě map incidence a relativního rizika i (geo)vizuální analýzu lze doporučit využití kartogramů, a to jak pro administrativní jednotky, tak i pro údaje v pravidelné síti. V případě map standardizované incidence za územní jednotky je výhodné použít i lokální bayesovské vyhlazování, díky kterému je možné snadno identifikovat prostorové vzory a globální chování v celém území. V případě SIR (relativního rizika) je vhodné data standardizovat, ale vyhlazování samotné už tak výhodné není, protože dochází k velkým změnám vnímání celkového obrazu situace na území.

Při sledování prostorové distribuce na základě kartogramů je vizuálně patrná asociace hustoty zalidnění a četnosti případů onemocnění. Výrazně více postižená je oblast Moravy a Slezska (především severovýchod Moravy, Slezsko a Brněnsko), v Čechách jde potom o jižní Čechy, Plzeňsko a oblast jihovýchodně od Prahy (Benešovsko). Při současném sledování času a prostoru s využitím interpolovaného spojitého povrchu je potvrzena sezónnost, která je u tohoto onemocnění běžná. Sezónní změny jsou méně patrné ve městech a více patrné v zázemí měst a na venkově, tedy oblastech, které mohu sloužit k rekreaci městského obyvatelstva během letní sezóny. V podhorských oblastech (především Jeseníků a Krkonoš) je možné pozorovat zvýšení incidence i během zimní sezóny příp. jarních prázdnin.

V rámci DC1 byl časoprostorový kriging využit jako alternativa k dasymetrickému mapování. Jeho využití je sice náročnou operací, ale je velkým zlepšením pro zkoumání časoprostorových vztahů a vizualizaci průběhu onemocnění. Časoprostorový variogram umožňuje exploraci časoprostorového vzoru a pomohl identifikovat vzdálenosti v obou těchto dimenzích, kde dochází k největším interakcím mezi jednotlivými místy (14—15 km vzdálenost a 1—4 týdny v čase). V interaktivní podobě je možné zkoumat výsledky DC4 díky KML souboru vytvořenému v rámci DC5 (kapitola 8.2), konkrétně jde o kartogram incidencí, agregované počty četností v pravidelné síti a týdenním intervalu a spojitý povrch týdenní incidence.

5 PODOBNOSTI VÝSKYTU ONEMOCNĚNÍ V ČASE I PROSTORU [DC2]

5.1 Identifikace shluků v prostoru

Během analýzy prostorového rozložení nemocí, a zejména v případě agregovaných dat, je často podstatné zaměřit se na lokální proměnlivost výskytu nebo měr nemocnosti raději než ji studovat pouze globálně na celém území. Postupy, které tímto způsobem zkoumají prostorovou distribuci případů onemocnění, jsou označovány jako detekce prostorových shluků onemocnění. Vzhledem k tomu, že identifikace a analýza prostorových vzorů nejrůznějších onemocnění je jedním z hlavních zájmů prostorové epidemiologie, existuje velké množství přístupů a metod sloužících k tomuto účelu. Jejich přehled a konkrétní využití, včetně tradičních i bayesovských přístupů, uvádějí mimo jiné Haining (1998, 2004), Lawson (2002, 2009) nebo Waller (2009).

V geovědách je často pod pojmem prostorové shlukování označena některá z metod analýzy prostorové autokorelace. Teorie prostorové autokorelace je modifikací konceptu autokorelace při studiu časových řad a jako jedni z prvních se jím zabývali Cliff a Ord (1973). Prostorová autokorelace je korelací mezi hodnotami jedné proměnné, která přímo odpovídá jejich vzájemné relativní poloze v rovině a představuje prostorovou obdobu tradičního statistického předpokladu odchylek od nezávislého pozorování (Griffith a Arbia, 2010). Ve své podstatě jde o obdobu a kvantifikaci tzv. *prvního zákona geografie*, který říká, že „*vše souvisí se vším, ale věcí blízké spolu souvisí více než věci vzdálené*“ (Tobler, 1970). Kladná autokorelace odkazuje na prostorový vzor, kde jsou si blízké či sousední hodnoty podobné, zatímco záporná autokorelace popisuje prostorový vzor, kdy jsou sousední hodnoty velmi rozdílné. Z hlediska rozsahu lze popsat dva základní typy autokorelace, a to autokorelaci globální, která popisuje celkový převládající prostorový vzor ve zkoumaném území, a autokorelaci lokální, která odhaluje konkrétní polohu a rozsah shluků (Pfeiffer et al., 2008). Metody identifikace prostorové autokorelace jsou modifikací tradičních statistických metod, proto jsou z nich vycházející zjištění doprovázena odhadem statistické významnosti výsledku (např. v podobě *p-value*). Nulovou hypotézou, se kterou jsou postupy srovnávány, je tvrzení, že hodnoty v hodnoceném území jsou rozmístěny náhodně a neexistuje tam tedy žádné prostorové shlukování či jiný pozorovatelný prostorový vzor. Náhodné rozmístění hodnot pro nulovou hypotézu často vychází z homogenního Poissonova procesu pro danou oblast, případně je možné využít i jiné procesy (heterogenní Poisonův proces, Coxům proces apod.) či využít simulací (Horák, 2011). Techniky průzkumu prostorové asociace bývají označovány také jako metody průzkumové analýzy prostorových dat (ESDA – Exploratory Spatial Data Analysis), jejich lokálně zaměřené varianty pak jako lokální indikátory prostorové asociace (LISA – Local Indicators of Spatial Association).

Pravděpodobně nejčastěji využívanými metodami pro globální i lokální analýzu prostorové autokorelace jsou Moranovo I kritérium, Gearyho C kritérium, případně Getis-Ordovo G kritérium. K jejich používání přispívá kromě dobré možnosti interpretace výsledků také fakt, že jsou implementovány v GIS programech jako je ArcGIS for Desktop nebo nástrojích k provádění explorační analýzy prostorových dat jako je GeoDA. V této kapitole je využito Moranovo I kritérium. Moranovo I kritérium je obdobou Pearsonova korelačního koeficientu a kvantifikuje podobnost zvolených proměnných v oblastech, které jsou

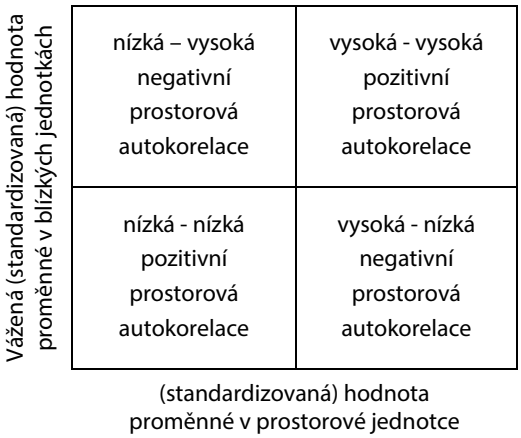
definovány jako prostorově příbuzné (Moran, 1950). Moranovo I kritérium je definováno vztahem:

$$I = \frac{n \sum_i \sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i \sum_j w_{ij} \sum_k (z_k - \bar{z})^2} \quad [5.1]$$

kde z_i je residuum (rozdíl mezi pozorovanou a očekávanou hodnotou) nebo standardizovaná míra nemocnosti v konkrétním místě, $\bar{z}$ je průměrná hodnota, w_{ij} je míra blízkosti (prostorové váhy) mezi místy i a j, k je celkový počet studovaných míst. Matice vah je v analýze využita k tomu, aby definovala prostorové vztahy mezi všemi místy tak, aby místa blíže v prostoru měla větší váhu na výpočet kritéria než místa vzdálenější. Moranovo I kritérium je globální míra autokorelace a jejím výsledkem je jedna hodnota indexu popisující převládající vzor v území. Proměnná vykazuje pozitivní, resp. negativní prostorovou autokorelaci, pokud je hodnota Moranova I kritéria kladná, resp. záporná, přesněji větší, resp. menší než očekávaná hodnota $I = -\frac{1}{n-1}$, kde n je počet analyzovaných jednotek (Fotheringham et al., 2000; Netrdová a Nosek, 2009). Pro případ hodnocení lokálního prostorového vzoru je používáno lokální Moranovo I kritérium, které je jedním z indexů LISA a umožňuje odlišit shluky podobných hodnot (clusters) od skupin hodnot nepodobných (outliers). Na rozdíl od globálního indexu umožnuje určit umístění těchto shluků na základě definované matice vah. Lokální Moranovo I kritérium je definováno jako:

$$I_i = r_i \sum_j (w_{ij} r_j) \quad [5.2]$$

kde r_i a r_j jsou standardizované hodnoty veličiny z a w_{ij} je matice vah (Horák, 2011). Další vysvětlení k hypotézám a teorii podávají Anselin (1995) nebo Scott a Janikas (2010). K hodnocení výsledků lokálního Moranova I je kromě vyjádření vlastních kategorií shluků vhodné vyjádření pomocí Moranova diagramu (Obrázek 20), kde levý dolní a pravý horní kvadrant vyjadřují prostorové shluky stejných hodnot (*hot spots/cold spots*) a levý horní a pravý dolní kvadrant odlehlé hodnoty (*outliers*).



Obr. 20 Moranův diagram (upraveno podle Spurná (2008))

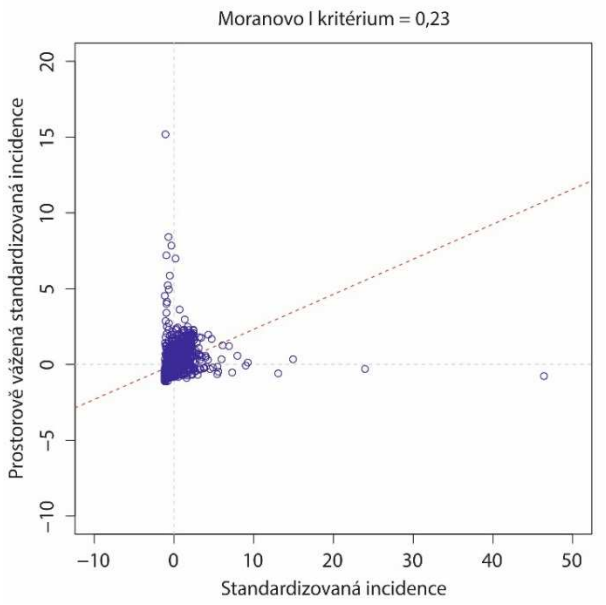
Problémy s nestabilitou odhadovaných hodnot a rozptylu měr nemocnosti, které byly jedním z hlavních důvodů k vyhlazování kartogramů v kapitole 4.3.2, mohou ovlivnit i inference Moranova I testu prostorové autokorelace (Anselin, 2003). Jedním z řešení, jak se s problémem nestability odhadu a rozptylu hodnot vyrovnat, je implementace vyrovnávacího

postupu s využitím empirického bayesovského principu, kterou navrhují Assuncao a Reis (1999). Tento postup vytvoří novou proměnnou, která je obohacena o informaci srovnávající zkreslení rozptylu, které vzniklo v důsledku rozdílných velikostí ohrožených populací (Anselin, 2003).

5.1.1 Prostorový vzor kampylobakteriózy v České republice

Ačkoliv je často možné identifikovat shluky více ohrožených nebo naopak zdravějších obcí již na základě vizuálního hodnocení mapových výstupů (map vytvořených tečkovou metodou, kartogramů či anamorfózy), tak bez použití prostorových statistických metod k průzkumu prostorové autokorelace je jen obtížně možné tyto shluky popsat a ohodnotit jejich významnost v rámci okolí. Jednou z nejpoužívanějších metod k hodnocení prostorových vzorů, a z nich především identifikaci prostorových shluků na globální i lokální úrovni, je Moranovo I kritérium.

Při průzkumu prostorových vzorů je vhodné postupovat v navazujících krocích. Nejdříve byl popsán globální prostorový vzor, na základě kterého bylo zjištěno, jak silná je prostorová autokorelace ve studovaném území a jakým způsobem jsou v prostoru distribuovány prostorové jevy, zda se jeví jako náhodné, shlukované nebo pravidelně rozmístěné. Pro výpočet Moranova I kritéria (stejně jako pro většinu ostatních metod hodnotících autokorelaci) je klíčová volba nastavení prostorových vah, které většinou vychází z definování typu sousedství, počtu sousedů, maximální vzdálenosti působení autokorelace nebo jejich kombinace. V případě hodnocení globálního prostorového vzoru průměrné incidence kampylobakteriózy v České republice v letech 2008—2012 bylo zvoleno sousedství typu královna 1. řádu, tzn. sousedy jsou všechny územní jednotky, které sdílejí alespoň část hranice územní jednotky. Průměrně každá obec/městská část sousedí s šesti okolními územními jednotkami (*min = 1; max = 35; medián = 6; průměr = 5,92, směrodatná odchylka = 2,26*). Po definování sousedství byl proveden výpočet Moranova I kritéria podle vzorce 5.1. Výpočet i definování typu sousedství v územních jednotkách byl uskutečněn v prostředí programů GeoDA a **R**. Výsledná hodnota Moranova I kritéria je *I = 0,23* a jde tedy o pozitivní prostorovou autokorelaci, která poukazuje na existenci shlukování ve studované oblasti. Významnost hodnoty kritéria je testována proti očekávané hodnotě (*I = 0,00*) v případě náhodného procesu (CSR) pomocí randomizovaných permutací (*p-value = 0,0001* při 10 tisících permutacích), která prokázala statistickou významnost pozorovaného vzoru. Grafické znázornění výsledků Moranova I kritéria je na Obrázku 21. Jednotlivé sektory, tak jak jsou popsány na Obrázku 20, vyjadřují typ autokorelace dat, kdy červená linie znázorňuje globální prostorovou autokorelaci. Na základě Obrázku 21 bylo v diagramu odhaleno několik odlehlých hodnot. Jde zejména o menší obce s vysokou hodnotou incidence kampylobakteriózy. Pokud by tyto obce byly z výpočtu odstraněny, pak je možné, výsledná hodnota Moranova I kritéria mírně vzroste, jejich množství však výrazně ovlivnilo výslednou hodnotu.

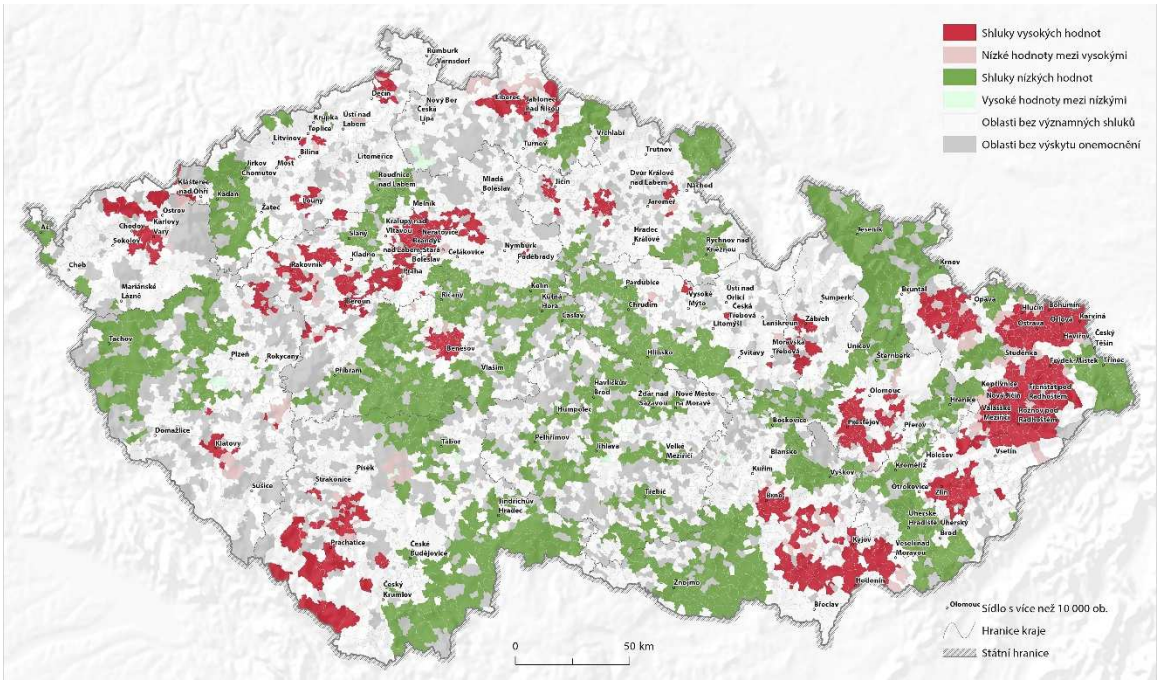


Obr. 21 Moranův diagram pro průměrnou standardizovanou incidenci v obcích a městských částech v letech 2008—2012

Globální Moranovo I kritérium prokázalo, že ve studovaném území existují tendence ke shlukování vysokých a nízkých hodnot. K odhalení jejich umístění slouží lokální Moranovo I kritérium (vztah 5.2), které vyjadřuje úroveň a sílu vazeb mezi sousedními případně blízkými jednotkami. Stejně jako u globálního hodnocení prostorové autokorelace, tak i zde je nutné definovat sousední regiony a z nich vycházející prostorové váhy. Pro hodnocení lokální prostorové autokorelace průměrné incidence kampylobakteriózy v České republice na úrovni obci/městských částí je opět využito prostorové sousedství typu královna a analýza tak přímo navazuje na globálního Moranovo I kritérium. Výsledkem lokálního hodnocení v prostředí programu GeoDA je jednak mapa shluků a odlehlých hodnot se stejnými kategoriemi jako v případě globálního Moranova I kritéria (viz Obrázek 20) a pak také doplňující mapa významnosti identifikovaných prostorových shluků. Obrázek 22 představuje výsledek lokální analýzy prostorové autokorelace pomocí LISA. Červenou barvou jsou znázorněny lokální shluky vysokých hodnot průměrné incidence, tedy oblasti, kde je hodnota incidence vyšší než v jejich okolí. Zelené shluky představují naopak obce a městské části, kde je průměrná hodnota incidence nižší než v jejich okolí. Světle červeně a světle zeleně jsou dále vyobrazeny odlehlé hodnoty (*outliers*), které jsou výjimkami v jinak homogenních shlucích – jde tak o nízké hodnoty obklopené shluky nízkých hodnot a naopak. Šedou barvou jsou navíc vyznačena území bez výskytu onemocnění. Zobrazeny jsou pouze shluky, které i po 10 tisících permutacích vykazují statistickou významnost (*p-value* < *0,05*), a proto není nutné současně prezentovat i mapu významnosti shluků.

Výsledky hodnocení lokálního prostorového vzoru částečně potvrzují zjištění získaná vizuální analýzou kartogramů a map z kapitoly 4. Největší i nejvíce homogenní shluky obcí s vysokou hodnotou průměrné incidence se nacházely na severovýchodě Moravy (Valašsko a Lašsko), v části Slezska (Ostravsko, Karvinsko a Opavsko) a částečně také na jižní Moravě, v okolí Brna a na Hané. Na Moravě byly ovšem identifikovány i homogenní shluky nízkých hodnot. Oproti tomu působí situace v Čechách rozdílně, shluky vysokých hodnot jsou méně homogenní a vyskytují se především na Prachaticku a Klatovsku, v oblastech na sever a západ

od Prahy (Mělnicko, Kladensko) a kolem Benešova. Stejně jako v případě Moravy i v oblasti Čech se nachází množství homogenních shluků nízkých hodnot, tedy zdravých oblastí. Ve srovnání s mapami incidence není jako významný shluk identifikováno Plzeňsko a také oblast Českých Budějovic. Celkem bylo do shluků vysokých hodnot incidence zařazeno 505 obcí/městských částí (8 % územních jednotek), do shluků nízkých incidencí 1222 obcí/městských částí (19 %), a jako hodnoty odlehlé bylo identifikováno 108 obcí/městských částí (2 %). U 2697 oblastí (42 %) nebyl identifikován statisticky významný shluk a v 1853 případech (29 %) nebyla kampylobakterióza ve sledovaném období zaznamenána. I přesto, že je absolutně téměř 2,5× více obcí a městských částí v oblastech shluků nízkých hodnot než v oblastech shluků vysokých hodnot, tak populačně je situace srovnatelná. V oblastech shluků nízkých hodnot žije odhadem 2,27 mil. osob (21,8 % populace) a v oblastech shluků s vysokou incidencí je to dokonce 2,44 mil. osob (23,4 % populace).



Obr. 22 Prostorové shluky vysokých a nízkých hodnot relativního rizika získané pomocí lokálního Moranova I kritéria s využitím empirického bayesovského principu a randomizace

5.2 Časoprostorové skenování: Identifikace shluků v prostoru i čase

Při hodnocení prostorových dat je v současnosti časové určení často bráno pouze jako jedna z mnoha vlastností, ačkoliv by mělo být považováno za neoddělitelnou složku dat, která umožní komplexní pohled na daný problém. Důvodem je omezená implementace časoprostorových analýz v nejčastěji používaných GIS programech a současně také vyšší nároky na uživatele při přípravě dat, analýze i vizualizaci výsledků. Z těchto důvodů je tak časoprostorová analýza geografických jevů omezena buď na oddělené časové a polohové složky nebo v lepším případě na hodnocení geografických jevů v časových řezech. Pomocí nich je sice možné popsat vývoj jevu, ale jednotlivé časové kroky jsou v podstatě zpracovávány nezávisle na sobě a jde tak spíše o sérii samostatných analýz.

Kromě oddělených přístupů však existují i metody, které umožní provádět časoprostorovou analýzu téměř v pravém slova smyslu. Velmi rozšířeným způsobem, který

současně hodnotí prostorové vzory v prostoru i čase, je metoda časoprostorového skenování – spatio-temporal scan statistics, jejímž autorem je Martin Kulldorff, který ji implementoval do software SaTScan (Kulldorff a Information Management Services Inc, 2009). Metoda časoprostorového skenování využívá koncept okna válcového tvaru s kruhovou či eliptickou základnou a výškou válce odpovídající času. Válcové okno se pohybuje v prostoru i čase a současně mění i své parametry (velikost základny i výšky). Takto navštíví všechny možné kombinace míst a časů a vznikne teoreticky nekonečné množství překrývajících se válců o různých velikostech, kdy každý válec představuje možný shluk (Kulldorff et al., 2005). Pro každý rozměr válce je vypočítán test poměru věrohodností, který vychází z reálně pozorovaného počtu případů onemocnění a počtu očekávaného v oblasti uvnitř a vně válce. Testové kritérium je následně srovnáno s věrohodností vycházející z nulové hypotézy. Funkce věrohodností vycházející z alternativní hypotézy předpokládá Poissonovo rozdělení pravděpodobností případů, které odpovídá (Kulldorff a Nagarwalla, 1995; Kulldorff, 1999):

$$\left(\frac{c}{e}\right)^{c}\left(\frac{C-c}{C-e}\right)^{C-c} I() \qquad [5.3]$$

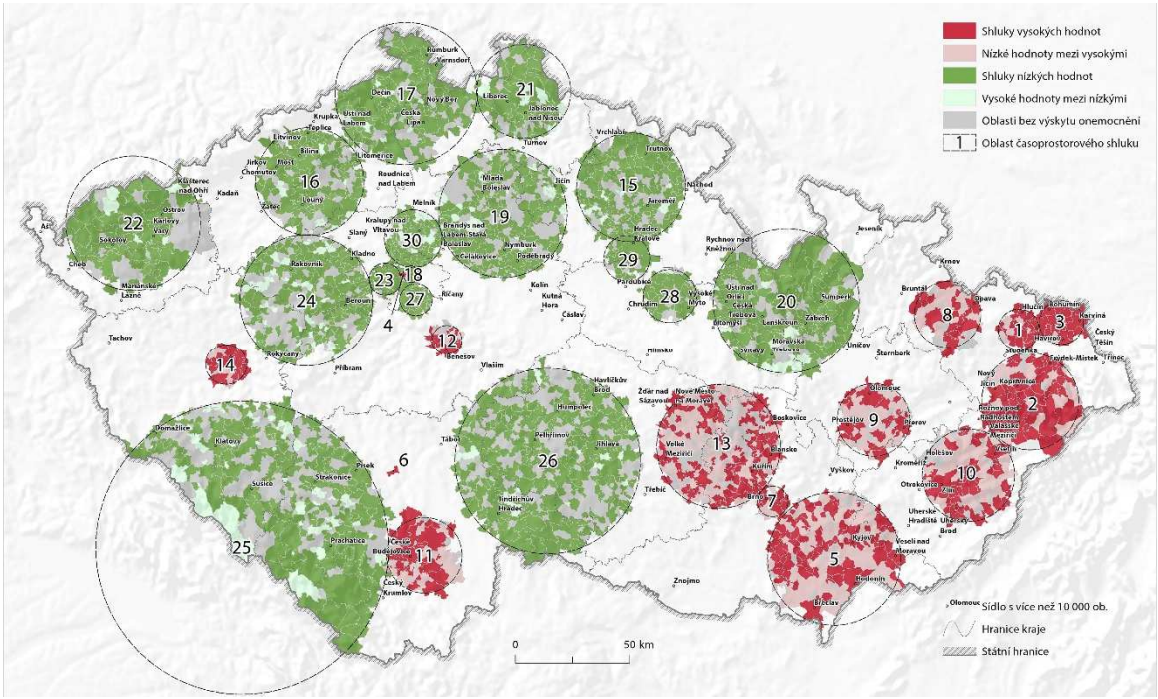
kde *C* je celkový počet případů, *c* je pozorovaný počet případů uvnitř válce, *e* je očekávaný počet případů uvnitř válce. *I()* je rovno 1 v případě, že $c > e$, v opačném případě je rovno 0. Válce s nejvyšší hodnotou poměru věrohodností jsou identifikovány jako potenciální shluky a jejich významnost (*p-value*) je vypočítána na základě Monte Carlo simulací. Dle významnosti je následně vyhodnocováno, zda jsou případy v prostoru a čase uspořádány náhodně (H_0) či nenáhodně a shlukují se (H_A) (Aamodt et al., 2006). Výsledkem analýzy jsou vyhodnoceny buď shluky vysokých hodnot relativního rizika (rizikové oblasti), nebo shluky nízkých hodnot relativního rizika (zdravé/nerizikové oblasti). Současně je vyhodnocována významnost shluků. *Primární shluk* je ten shluk, který byl vyhodnocen jako nejvíce věrohodný. *Sekundární shluky* vznikají v důsledku pohybujícího se válce, a ačkoliv jsou dostatečně významné (*p-value* < *0,05*), tak nemohou být shlukem primárním. Právě díky této významnosti by neměly být zanedbávány pro další zkoumání (Chen et al., 2008). Výhodou časoprostorového skenováním, jak ho definoval Kulldorff je, že přístup je současně deterministický (umožňuje lokalizovat shluky) i inferenční (testuje a vyhodnocuje významnost shluků) (Osei, 2014).

Vzhledem k množství parametrů byla data agregována do časových období a územních jednotek, případy byly stratifikovány dle věku a pohlaví a byly zvoleny parametry nastavení samotné analýzy, která je citlivá na změny. Mezi základní nastavení patřilo maximální procentuální počet populace v jednom shluku, doba trvání shluků, manipulace s časovým trendem a nastavení počtu simulací. Nastavení bylo provedeno na základě povahy vstupních dat s ohledem na zahraniční studie zabývající testováním vlivu nastavení na výsledky (Kulldorff et al., 2005; Aamodt et al., 2006; Green et al., 2006; Chen et al., 2008; Weisent et al., 2011).

5.2.1 Časoprostorové vzory kampylobakteriózy v České republice

Cílem časoprostorového skenování je identifikovat a vyhodnotit shluky vysokých a nízkých hodnot relativního rizika, tzn. shluky *ohrožených* a *zdravých* obcí jak v prostoru, tak i v čase. K výpočtu statistiky bylo použito prostředí software SaTScan 9.3. Celý postup umožnil potvrdit či vyvrátit, zda jsou pozorované prostorové vzory statisticky významné reálné situace nebo pouze realizace náhodných procesů ve studovaném území. Vstupní data se sestávají z případů rozdělených dle pohlaví a věku, které jsou agregovány prostorově dle jejich příslušnosti k obci (nebo městské části v případě, že je do nich obec dělena) a časově v týdenním intervalu, který je typickou délkou průběhu onemocnění. Doplňující datovou sadou, která slouží jako vstup do analýzy, je demografická struktura (počet obyvatel dle věku a pohlaví) a geografické středy územních jednotek. Časoprostorová retrospektivní analýza shluků s vysokou, či nízkou mírou rizika je založena na takto stratifikovaných datech a jejich kombinaci s Poissonovým rozdělením pravděpodobnosti. Časoprostorové skenování bylo nastaveno tak, aby do jednoho shluku zahrnul maximálně 3 % populace a délka trvání shluku byla definována dvěma způsoby: (1) maximálně 50 % celkové zkoumané délky nebo (2) shluk se v oblasti vyskytuje po celé studované období (100 %), hodnoty byly zvoleny na základě empirického pozorování a obdobných studií (např. Weisent et al., 2011). Do analýzy byla z důvodu zajištění srovnatelnosti mezi jednotlivými obdobími zahrnuta také neparametrická úprava časového trendu založená na časově stratifikované randomizaci záznamů (Kulldorff, 1999). Shluky byly identifikovány v případě, že jejich významnost určená pomocí p-value byla nižší než 0,05 i po ověření pomocí 999 simulací typu Monte Carlo. Následně byly identifikovány shluky a pro v nich se nacházející obce a městské části bylo vypočítáno relativní riziko onemocnění v obci jako poměr pozorovaných a očekávaných případů (Green et al., 2006).

Výsledkem časoprostorového skenování pomocí programu SaTScan jsou indexované kruhy, které představují jednotlivé typy shluků (shluky vysokých či nízkých hodnot) a dále identifikátory administrativních jednotek, které do shluků spadají. V rámci zobrazení výsledných shluků byly tyto výsledky opět zkombinovány s původními daty a vznikla tak mapa časoprostorových shluků kampylobakteriózy v České republice v letech 2008—2012 (Obrázek 23). Tvorba její interaktivní podoby je pak obsahem kapitoly 8. Při vizuálním vyhodnocování shluků je kromě polohy identifikovaných shluků důležitá také jejich vnitřní (ne)homogenita. Na Obrázku 23 jsou původci heterogenity shluků zobrazeni světlejšími barvami, jde-li o odlehlé hodnoty nebo šedou, jde-li o území, kde během studovaného období nebyl zaznamenán výskyt kampylobakteriózy. Světle červeně jsou v mapě znázorněny hodnoty odchýlené od shluků vysokých hodnot, tedy oblasti s průměrným či nižším rizikem než je v okolí běžné ($RR \leq 1{,}50$). Světle zeleně jsou naopak vykresleny oblasti odchýlené od shluků nízkých hodnot, u kterých se jedná o oblasti s průměrným či zvýšeným rizikem ($RR > 0{,}80$). Nevybarvené oblasti znázorňují obce nebo jejich části, které nespadají do žádného shluku.



Obr. 23 Časoprostorové shluky onemocnění kampylobakteriózou v letech 2008—2012

V průběhu sledovaného období bylo v České republice identifikováno 30 statisticky významných shluků (*p-value* < *0,001*), jejichž charakteristiky jsou shrnuty v Tabulce 8. Čtrnáct z nich představuje shluky vysokých hodnot, které označují území se zvýšeným relativním rizikem výskytu onemocnění (shluky vysokých hodnot, RR > 1,50), a tedy oblasti kampylobakteriózou více ohrožené. Primární a nejvíce věrohodný shluk (v mapě i tabulce označen číslem 1, RR = 2,16) zahrnuje město Ostrava a přilehlé obce. Shluk je složen z 31 obcí a/nebo městských částí Ostravy, ve kterých žije téměř 293 tisíc obyvatel. Ostatní identifikované shluky jsou označovány jako sekundární. Shluky 2, 3 a 10 přímo navazují na primární shluk 1 a v případě analýzy zahrnující vyšší maximální počet obyvatel by tvořily jednu oblast. Shluky 1 a 3 jsou společně se shlukem 4 těmi vůbec nejhomogennějšími mezi shluky vysokých hodnot, nejsou-li započítány prostorové struktury zahrnující pouze jednu obec / městskou část. Naopak velká heterogenita je viditelná u prostorově rozsáhlých shluků 5, 9, 10 a 13, u kterých je také průměrné relativní riziko menší než 2,00. Průměrné relativní riziko shluku vyšší nebo rovno 2,00 je pouze v případě šesti shluků, které jsou zařazeny mezi ty více pravděpodobné. Obecně lze vypozorovat, že shluky vysokých hodnot jsou více heterogenní než shluky s nízkými hodnotami relativního rizika.

Devět ze čtrnácti shluků vysokých hodnot je lokalizováno na Moravě či ve Slezsku a pouze pět jich je v Čechách. Tento fakt potvrzuje jeden z předpokladů odhadovaných na základě vizuální analýzy kartogramů v kapitole 4.3. Většina ze shluků vysokých hodnot se vyskytuje po celé studované období od ledna 2008 do prosince 2012 a pouze pět jich je časově více specifických (5, 10, 11, 13, 14). Tři z těchto pěti shluků (5, 13, 14) ohraničují teplejší polovinu jednotlivých let, zatímco čtvrtý shluk – Valašsko (10) - je rok a půl trvající období od jara 2009 do podzimu 2011 a pátý shluk – Českobudějovicko (11), který má vůbec nejvyšší relativní riziko (RR = 5,41), představuje šestitýdenní období během zimy roku 2010. Dva z identifikovaných shluků vysokých hodnot pokrývají pouze jednu administrativní jednotku. Prvním je centrum Prahy (RR = 4,13) a druhým je jihočeská obce Dražíč (RR = 4,19).

Tab. 8 Časoprostorové shluky vysokých a nízkých hodnot relativního rizika ohrožení kampylobakteriózou v České republice v letech 2008—2012

Shluk	T[1]	Období trvání shluku[2]	Region[3]	Č[4]	P[5]	O[6]	RR[7]	Populace[8]
1*	V	2008/01/01 – 2012/12/31	Ostrava	31	5975	2861	2,16	292978
2	V	2008/01/01 – 2012/12/31	Lašsko	70	5414	2788	2,00	277236
3	V	2008/01/01 – 2012/12/31	Havířov a Karviná	16	4773	2534	1,93	256657
4	V	2008/01/01 – 2012/12/31	Praha - střed	1	1006	245	4,13	29948
5	V	2008/05/13 – 2010/11/01	jižní Morava	167	2274	1432	1,60	292885
6	V	2008/01/01 – 2012/12/31	Dražíč	1	7	2	4,19	214
7	V	2008/01/01 – 2012/12/31	Brno - město	19	3951	2590	1,55	271742
8	V	2008/01/01 – 2012/12/31	Opava	37	1714	877	1,97	87203
9	V	2008/01/01 – 2012/12/31	Haná	66	3828	2526	1,54	256721
10	V	2009/04/14 – 2011/09/05	Valašsko	90	1596	932	1,72	196522
11	V	2010/01/12 – 2010/02/22	Českobudějovicko	60	194	36	5,41	157425
12	V	2008/01/01 – 2012/12/31	Benešovsko	15	640	313	2,05	31115
13	V	2010/04/06 – 2010/10/04	Blanensko - Novoměstsko	224	568	286	1,99	284346
14	V	2011/05/03 – 2011/11/14	Plzeň	22	394	201	1,96	197263
15	N	2008/01/01 – 2012/12/31	Podkrkonoší	128	997	1841	0,54	182641
16	N	2008/01/01 – 2012/12/31	Chomutovsko - Mostecko	108	1853	2930	0,63	290222
17	N	2008/01/01 – 2012/12/31	Ústecko - Děčínsko	93	1266	2958	0,42	288203
18	N	2008/01/01 – 2012/12/31	Praha - východ	4	1591	2530	0,62	280780
19	N	2008/01/01 – 2012/12/31	Mladoboleslavsko	173	1124	2590	0,43	256738
20	N	2008/01/01 – 2012/12/31	Třebovsko - Šumpersko	138	1482	2571	0,57	253941
21	N	2008/01/01 – 2012/12/31	Liberecko	59	1302	2320	0,56	230360
22	N	2008/01/01 – 2012/12/31	Karlovarsko	82	1284	1992	0,64	202256
23	N	2008/01/01 – 2012/12/31	Praha - západ	16	1805	2961	0,60	305103
24	N	2008/01/01 – 2012/12/31	Kladno – Beroun - Rakovník	172	1950	2684	0,72	268391
25	N	2008/01/01 – 2012/12/31	Pošumaví	211	1578	2667	0,58	268701
26	N	2010/11/23 – 2011/04/25	Vysočina	252	84	247	0,34	294203
27	N	2008/01/01 – 2012/12/31	Praha – jihovýchod	21	1821	2961	0,61	318958
28	N	2008/01/01 – 2012/12/31	Vysoké Mýto	31	271	509	0,53	49304
29	N	2010/11/09 – 2012/06/11	Královéhradecko	25	173	348	0,50	113501
30	N	2008/01/01 – 2012/12/31	Neratovicko	71	2000	2490	0,80	249994

** označuje primární shluk; p-value všech shluků < 0,001; [1] typ shluku – V označuje shluk vysokých hodnot (vysoké relativní riziko), N označuje shluk nízkých hodnot (nízké relativní riziko); [2] období trvání shluku; [3] jméno oblasti, kde se shluk vyskytuje; [4] počet obcí / městských částí zahnutých ve shluku; [5] počet reálně pozorovaných případů ve shluku; [6] počet očekávaných případů ve shluku; [7] relativní riziko; [8] odhadovaný počet obyvatel ve shluku*

Z třiceti identifikovaných shluků se v šestnácti případech jedná o shluky nízkých hodnot, které tvoří oblasti s nižším relativním rizikem (RR ≤ 0,80) nakažení kampylobakteriózou než je očekáváno. Všechny shluky nízkých hodnot jsou lokalizovány v Čechách (shluky 20 a 26 jsou na pomezí Čech a Moravy). Byly definovány dva převládající typy shluků nízkých hodnot. První typ se nalézá v především v horských a podhorských oblastech se spíše nižší hustotou zalidnění (15, 17, 20, 21, 22, 25, 26). Druhý typ je opakem toho prvního a nachází se zejména v hustěji zalidněných oblastech v okolí středně velkých měst. I v případě shluků nízkých hodnot převládají, až na dvě výjimky (Vysočina (26) a Královéhradecko (29)), shluky

trvající během celého studovaného období. Mezi oblasti s nejnižším relativním rizikem patří Vysočina (RR = 0,34), Ústecko - Děčínsko (RR = 0,42) a Mladoboleslavsko (RR = 0,43). Celkem žije v identifikovaných shlucích vysokých hodnot až 2,6 mil. osob. Přibližně 25 % obyvatelstva České republiky tak žije v oblastech se zvýšeným rizikem nakažení kampylobakteriózou, zatímco 3,9 mil. obyvatel (37 % populace ČR) žije v oblastech, kde je relativní riziko nakažení nižší.

5.3 Shrnutí výsledných zjištění

Analýza globálního prostorového vzoru i lokálních prostorových vzorů pomocí Moranova I kritéria potvrdila předpoklady, které byly definovány na základě (geo)vizuálního hodnocení map v kapitole 4. Na území České republiky byly ve sledovaném období zjištěny statisticky významné prostorové shluky, z nichž ty nejhomogennější se vyskytují na severovýchodní Moravě a ve Slezsku. Morava je, v souladu se zjištěními vycházejícími z vizuální analýzy kartogramů, obecně ohrožena více než oblast Čech, kde jsou shluky vysokých hodnot méně homogenní a více rozptýlené. Na rozdíl od vizuální analýzy nebyly tendence ke shlukování případů identifikovány na Plzeňsku a Českobudějovicku, místo těchto oblastí však byly identifikovány shluky na Prachaticku, Karlovarsku a také v okolí Prahy. Shluky nízkých hodnot průměrné incidence, kterých je výrazně více, a jsou také velmi homogenní, jsou umístěny většinou v oblastech s nižší hustotou zalidnění. Proto je také srovnání počtu obyvatel spadajících do shluků nízkých a vysokých hodnot velmi vyrovnané – 21,8 % populace žije ve shlucích nízkých hodnot, 23,4 % ve shlucích vysokých hodnot.

Analýza časoprostorových vzorů byla provedena pomocí časoprostorového skenování. Tato metoda bere v úvahu vzájemné umístění případů v prostoru i v čase a současně umožnuje integrovat do analýzy i vliv demografické struktury obyvatelstva (věk a pohlaví) případně i dalších souvisejících faktorů. Pomocí časoprostorového skenování bylo identifikováno celkem 30 shluků (14 shluků zvýšeného rizika onemocnění kampylobakteriózou a 16 rizika nižšího). Primární shluk byl opět umístěn do oblasti Ostravska a bezprostředně na něj navazovaly i další shluky zvýšeného rizika. Diverzifikace Čech a Moravy je patrná ještě více než v případě prostorového shlukování. Většinu ze shluků je možné pozorovat na místě po celé studované období, ačkoliv několik jich trvalo pouze po omezenou dobu, což je případ Plzeňska, Českých Budějovic, Blanenska, jižní Moravy a Valašska, resp. Královéhradecka a Vysočiny. V interaktivní podobě je možné zkoumat výsledky prostorového skenování díky KML souboru vytvořenému v rámci DC5 (kapitola 8.2).

6 ANALÝZA VZTAHŮ MEZI ONEMOCNĚNÍM A VNĚJŠÍMI FAKTORY PROSTŘEDÍ [DC3]

6.1 Zdroje dat a jejich příprava

Vliv environmentálních a socioekonomických faktorů na výskyt, incidenci či relativní riziko onemocnění kampylobakteriózou je s různými úspěchy hodnocen v množství zahraničních studií. Mezi nejčastěji zmiňované významné faktory patří hustota zalidnění (Arsenault et al., 2013); vliv klimatu (teplota a srážky) (Sari Kovats et al., 2005; Manitz a Höhle, 2013); přítomnost jatek a chovů hospodářských zvířat (zejm. kuřat) (Arsenault et al., 2012); venkovskost území (Green et al., 2006) či výskyt povodní (McBride a Mittinty, 2007; Hashizume et al., 2008). Dále jsou zmiňovány některé faktory týkající se demografických a socioekonomických charakteristik obyvatelstva, jako je rasa (Gillespie et al., 2008), věk, pohlaví či sociální deprivace (Spencer et al., 2012; Weisent et al., 2012) apod. Stále je však nutné si uvědomit, že až polovina všech hlášených případů zůstává bez uspokojivého vysvětlení (Ekdahl et al., 2005), a proto je pravděpodobné, že mohou existovat i další faktory prostředí, které mohou podporovat rozšíření onemocnění a jeho zvýšenou incidenci u obyvatelstva.

U některých ze zmiňovaných faktorů je pravděpodobné, že mohou určitou měrou přispívat k nárůstu či utlumení kampylobakteriózy i v prostředí České republiky. Pro účely modelování bylo proto nutné získat vhodná data pro vybrané charakteristiky, které budou potřebám lokální analýzy odpovídat kvalitou i prostorovým rozlišením. Datová sada sestavená za účelem identifikace významných faktorů se skládala ze 103 primárních i odvozených charakteristik území. Za prostorovou jednotku určenou pro modelování byly zvoleny obce, a to ze dvou důvodů: (1) velké množství statistických charakteristik, záznamů a dalších vlastností je dostupných právě pro obce jako pro nejmenší územní jednotku a (2) obce jsou ve většině případů vhodným kompromisem mezi podrobností analýzy a zachováním anonymity jednotlivých případů onemocnění. Vybrané charakteristiky obcí a jejich poskytovatelé jsou uvedeni v Tabulce 9.

Nejdůležitějším poskytovatelem dat pro tuto část disertační práce je Český statistický úřad, především data pocházející ze Sčítání lidu, domů a bytů v roce 2011 (SLDB 2011), které svou kvalitou i prostorovou podrobností vyhovují bezezbytku požadavkům pro jejich využití v prostorových a geostatistických analýzách. Tato data byla hlavním zdrojem informací o demografické struktuře obyvatelstva i o některých socioekonomických charakteristikách. Současně je SLDB 2011 hlavním zdrojem pro výpočet odvozených charakteristik (např. index socioekonomické deprivace). Kromě SLDB 2011 je ČSÚ také poskytovatelem Registru ekonomických subjektů, z něhož byly odvozeny charakteristiky týkající se zemědělství – počty podnikatelských subjektů v obcích, které se věnují zemědělské činnosti, zpracování masa nebo chovu a zpracování drůbeže. Informace o mlékomatech pochází od Státní veterinární správy, která ovšem eviduje pouze mlékomaty fungující. Data proto byla doplněna záznamy od provozovatelů mlékomatů a vybranými online zdroji[34]. Informace o území, kde jsou překročeny imisní limity, jsou založeny na datech Českého hydrometeorologického ústavu,

[34] např. *http://www.venkovskeforum.cz/mlekomaty*

zatímco oblasti záplavových zón dvacetileté vody pochází z databáze DIBAVOD (Digitální báze vodohospodářských dat). Data o počtech hospodářských zvířat pochází z oficiálních statistik Ministerstva zemědělství ČR a také z ČSÚ. Informace o charakteristikách podnebí (průměrné teplotě a srážkách) pochází z globální datové sady WorldClim[35].

Tab. 9 Vybrané environmentální, demografické a socioekonomické charakteristiky obcí. Tučně označené charakteristiky byly využity pro hodnocení možné souvislosti s výskytem kampylobakteriózy.

	Charakteristika	Zdroj dat		Charakteristika	Zdroj dat
Demografická struktura	Počet obyvatel	ČSÚ	Socioekonomické charakteristiky	**Obyvatelstvo se základním vzděláním a bez vzdělání**	ČSÚ
	Hustota zalidnění	Odvozeno z dat ČSÚ		Obyvatelstvo s vysokoškolským vzděláním	ČSÚ
	Změna počtu obyvatel 2001—2011	ČSÚ		Míra nezaměstnanosti	ČSÚ
	Obyvatelstvo do 15 let	ČSÚ		Míra ekonomické aktivity	Odvozeno z dat ČSÚ
	Obyvatelstvo 15—64 let	ČSÚ		Vyjížďka za prací	ČSÚ
	Obyvatelstvo nad 64 let	ČSÚ		**Socioekonomická deprivace**	Odvozeno z dat ČSÚ
	Průměrný věk	Odvozeno z dat ČSÚ	Zemědělství	**Společnosti zabývající se živočišnou výrobou**	MF ČR / ČSÚ
				Ekonomické subjekty v zemědělství	MF ČR / ČSÚ
	Index stáří	Odvozeno z dat ČSÚ		Zemědělská půda	EEA (CORINE)
	Počet mužů	ČSÚ		**Odhad počtu drůbeže**	Odvozeno z dat MZe ČR / ČSÚ
	Počet žen	ČSÚ		Odhad počtu přežvýkavců	Odvozeno z dat MZe ČR / ČSÚ
Klima a ovzduší	**Průměrná teplota**	WORLDCLIM		**Masozpracující společnosti**	MF ČR / ČSÚ
	Průměrné srážky	WORLDCLIM		**Mlékomaty**	SVS ČR*
	Poměr záplavových území (Q_{20})	VÚV TGM	Onemocnění	Průměrná vzdálenost případu onemocnění od mlékomatu	Odvozeno z dat SZÚ
	Území s překročením imisních limitů	ČHMÚ		Průměrný věk případu onemocnění	Odvozeno z dat SZÚ

ČSÚ – Český statistický úřad, WorldClim – Global Climate Data, VÚV TGM – Výzkumný ústav vodohospodářský TGM, ČHMÚ – Český hydrometeorologický ústav, MF ČR – Ministerstvo financí ČR, EEA – Evropská agentura pro životní prostředí, MZe ČR – Ministerstvo zemědělství ČR, SVS ČR – data pocházejí především od Státní veterinární správy*

6.1.1 Zpracování vybraných charakteristik

K dalšímu využití, analýzám a srovnání byla převedena na odvozené/relativní charakteristiky, které zohledňují vlastnosti obce, především její plochu a/nebo populaci. Z tohoto pohledu byla nejsnadněji upravitelná data z ČSÚ, která byla přímo vztažena k jednotce obce a absolutní hodnoty (četnosti) byly převedeny na hodnoty na relativní, což platilo pro změnu počtu obyvatel mezi lety 2001—2011, kde byla základem populace roku

[35] *www.worldclim.org*

2001, i pro jednotlivé kategorie obyvatelstva (0—14, 15— 64, 64+, pohlaví). Index stáří vyjadřuje poměr obyvatel starších 64 let a dětí do 14 let věku. Na počet obyvatel v obci byly relativizovány také charakteristiky vzdělání a vyjížďky. Míra ekonomické aktivity udává podíl ekonomicky aktivních obyvatel (zaměstnaných i nezaměstnaných) na počtu osob starších 15 let. Komplexnější mírou, než jsou podíly charakteristiky z celkového počtu obyvatel, je index socioekonomické deprivace, který zahrnuje několik socioekonomických charakteristik. Pro účely zkoumání byl využit index sestavený pro použití s daty českých cenzů. Ten je součtem normalizovaných hodnot charakteristik podílu osob se základním nebo neukončeným vzděláním či bez vzdělání, podílu osob vyjíždějících za prací, míry nezaměstnanosti, míry ekonomické aktivity, vývoje počtu obyvatel a indexu stáří (Klufová, 2009).

Komplexnější zpracování vyžadovala data z Registru ekonomických subjektů (RES). Z registru byly nejdříve vybrány pouze záznamy, které obsahovaly požadovanou třídu definovanou kódem CZ-NACE[36]. Byly filtrovány kategorie - Živočišná výroba a Smíšené hospodářství, v dalších krocích pak i Výroba masných výrobků a výrobků z drůbežího masa. Z datové sady ekonomických subjektů, která obsahovala jak fungující, tak i ty zaniklé, byly v dalším kroku vybrány pouze subjekty aktivní v průběhu let 2008—2012. Výsledné vybrané ekonomické subjekty byly agregovány podle identifikátorů obcí. Kromě agregace počtů ekonomických subjektů byl také proveden odhad množství zaměstnanců, který vycházel ze tříd počtu zaměstnanců uvedených v RES. Získané počty ekonomických subjektů byly převedeny na relativní hodnotu jako podíl vybraných ekonomických subjektů a celkového počtu ekonomických subjektů v obci. U subjektů zpracovávajících maso byl přepočítán jejich počet na počet obyvatel v obci. Data o mlékomatech pochází z databáze Státní veterinární správy, která je doplněna o některé online zdroje obsahující i již zaniklé mlékomaty. Mlékomaty byly geokódovány pomocí adresy a přiřazeny k obci. Pro vyhodnocení jejich vlivu na přítomnost kampylobakteriózy byla využita jejich poloha i jejich absolutní a relativní četnosti v obcích.

Největší nejistota spojená s daty panuje u počtu hospodářských zvířat v obcích. Důvodem jsou data, která jsou poskytovaná pouze na úrovni okresů, a po přímém dotazu na ČSÚ bylo autorovi disertační práce sděleno, že takovými daty nedisponují. Vzhledem k faktu, že drůbeží maso je nejčastější příčinou onemocnění a drůbeží chovy jsou často zmiňovány v zahraničních studiích v souvislosti se zvýšenou incidencí onemocnění, tak bylo nutné provést odhad počtů zvířat. Odhady počtu drůbeže i přežvýkavců byly založeny na počtu ekonomických subjektů v obcích, které se zabývají živočišnou výrobou. Takto byl proveden vážený rozklad počtu hospodářských zvířat v okresech do úrovně obcí.

Klimatická data, imisní oblasti, záplavová území a zemědělská půda jsou poskytovány v podobě rastru. Tyto rastry byly postupně proloženy vektorovou vrstvou obcí, do které byly hodnoty z rastru agregovány. Pro teplotu vzduchu a atmosférické srážky byla vygenerována průměrná hodnota na území obce, pro zbylé charakteristiky pak procentuální podíl na rozloze obce. Stejným způsobem bylo připraveno i množství dalších dat podobného charakteru, které ale nakonec nebyly pro hodnocení jejich významu pro výskyt kampylobakteriózy využity. Výběrem reprezentativních charakteristik se zabývá následující podkapitola 6.2.

[36] Klasifikace ekonomických činností

6.2 Korelace, prostorová korelace a autokorelace

Analyzovat vztah mezi onemocněním a vnějšími faktory prostředí je možné několika způsoby. V této části disertační práce jsou analyzovány asociace mezi onemocněním a faktory prostředí dvěma hlavními na sebe navazujícími směry. Prvním je hodnocení korelace, prostorové korelace a také hodnocení prostorové autokorelace dvou proměnných. Druhým směrem je hodnocení asociací na základě modelů, které se rekonstruují a predikují ohrožení obyvatelstva kampylobakteriózou.

Nejdříve jsou na základě statistické (neprostorové) korelace vybrány reprezentativní vlastnosti, které vhodně zastupují co nejširší počet dalších charakteristik, ale současně mezi sebou co nejméně korelují. Vybrané vhodné charakteristiky jsou v dalších krocích hodnoceny prostorovými metodami a dány do souvislosti s relativním rizikem (SIR) onemocnění kampylobakteriózou v obci.

6.2.1 Korelace a výběr charakteristik

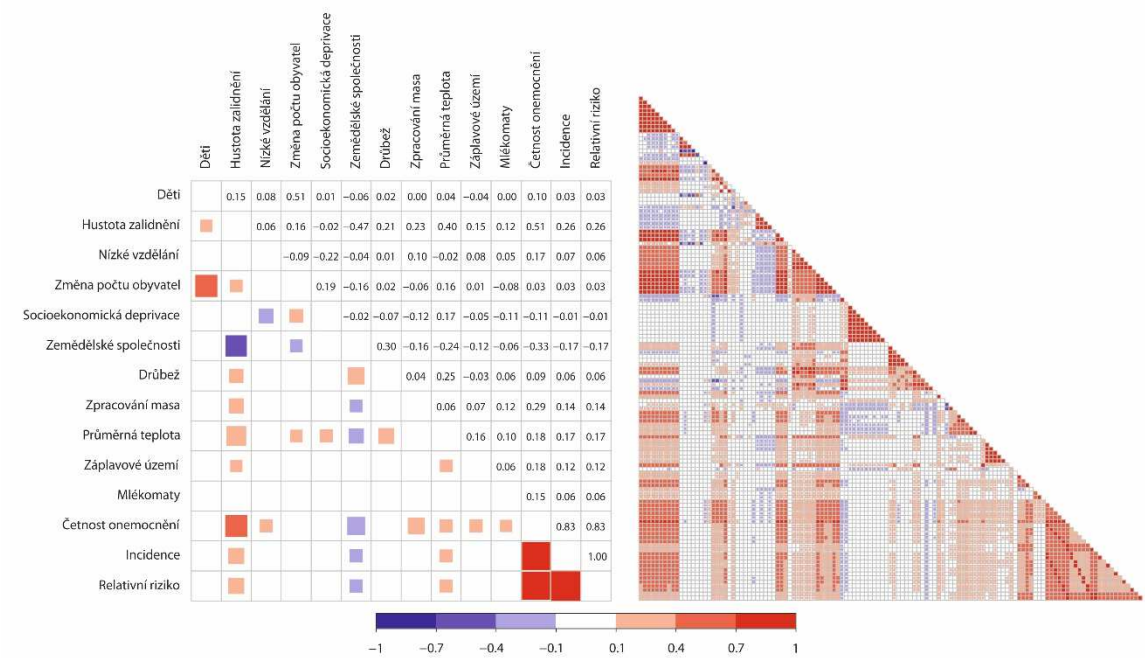
Korelací je nazývána statistická lineární závislost nebo lineární vztah mezi dvěma událostmi, charakteristikami či proměnnými. Její hodnota se pohybuje v $\langle -1; 1 \rangle$, kde -1 vyjadřuje silnou negativní asociaci, 1 silnou pozitivní asociaci a 0 nezávislost. V případě korelace je popisován oboustranný vztah a na jejím základě nelze stanovit vzájemnou kauzalitu jevů (ačkoliv je často pravděpodobná). Nejčastějšími variantami jsou Pearsonův koeficient korelace (r) a Spearmanův koeficient pořadové korelace (r_s), který umožnuje zkoumat i nelineární vztahy či veličiny, které nemají normální rozdělení pravděpodobnosti. Korelaci (Pearsonův koeficient korelace) je možné vyjádřit vztahem:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad [6.1a] \qquad r_s = 1 - \frac{6\sum_{i=1}^{n}(p_i - q_j)}{n(n^2 - 1)} \quad [6.1b]$$

kde x a y jsou hodnocené proměnné, $\bar{x}$ a $\bar{y}$ jsou jejich průměrné hodnoty a n je počet pozorování. Spearmanův koeficient pořadové korelace nahrazuje skutečné hodnoty pořadím v rámci dat (p, q).

Hodnocení korelace proběhlo současně pro zjištění síly lineárních vztahům mezi jednotlivými charakteristikami prostředí a také pro průzkum asociací mezi charakteristikami prostředí a mírou incidence, relativním rizikem a četností případů onemocnění v obcích. Výsledkem je rozsáhlá korelační matice 126 charakteristik. Obrázek 24 (vpravo) je vizuální prezentací této korelačním matice. Ačkoliv v korelační matice neobsahuje popisky, tak upozorňuje na celkový vzor asociací datové sady a výrazné korelace v rámci skupin charakteristik. Právě díky tomu, že spolu ze své podstaty velké množství charakteristik koreluje (silné korelace jsou především v rámci skupin charakteristik, jak jsou představeny v Tabulce 9), tak je možné postupně redukovat jejich počet až na přijatelné minimum, které může sloužit jako sada nezávislých vysvětlujících proměnných při modelování morbidity v obcích. Samotné ukazatele morbidity v obcích nejvíce korelují s průměrným věkem pacientů ($r = 0{,}70$). Mírná těsnost vztahu ($0{,}3 < |r| < 0{,}5$) je přítomna ještě v souvislosti s demografickými charakteristikami (věk a pohlaví) a ekonomicky aktivním obyvatelstvem. Výraznější korelace lze nalézt u absolutní četnosti případů onemocnění kampylobakteriózou, kdy je velká těsnost vztahu ($|r| > 0{,}7$) identifikována v souvislosti s většinou demografických charakteristik.

Postupně byly z původní korelační matice odebírány ty charakteristiky, které mezi sebou nejvíce korelovaly, a současně bylo jejich vzájemné zastupování smysluplné. Z původní korelační matice zůstalo po redukci dat korelační matice o 11 charakteristikách prostředí (tučně označené v Tabulce 9) a třech charakteristikách onemocnění. Vizualizace výsledné korelační matice je na Obrázku 24 (vlevo).



Obr. 24 Vizualizace redukované (vlevo) a původní (vpravo) korelační matice; čím sytější je barva čtverce, tím statisticky významnější je korelace mezi charakteristikami (červená – pozitivní korelace, modrá – negativní korelace)

Nejsilnější vzájemný vztah mezi vybranými charakteristikami zůstal mezi podílem obyvatel do 15 let na celkové populaci a změně počtu obyvatel mezi roky 2001 a 2011 (r_s = 0,51), významná korelace je dále mezi hustotou zalidnění a průměrnou teplotou vzduchu (r_s = 0,43) nebo mezi hustotou zalidnění a podílem firem zabývajících se živočišnou výrobou (r_s = -0,47). Zbylé korelace jsou $|r_s| < 0{,}4$. Výjimkou jsou mezi sebou korelující hodnoty standardizované incidence, relativního rizika a četnosti onemocnění v obci. Stav, kdy mezi vybranými charakteristikami není významná korelace a současně tyto proměnné reprezentují skupinu původních dat, je výhodný zejména s ohledem na analýzu morbidity s využitím regresních a jiných modelů, u kterých je kolinearita překážkou.

6.2.2 Prostorové korelace a autokorelace mezi relativním rizikem a vybranými charakteristikami

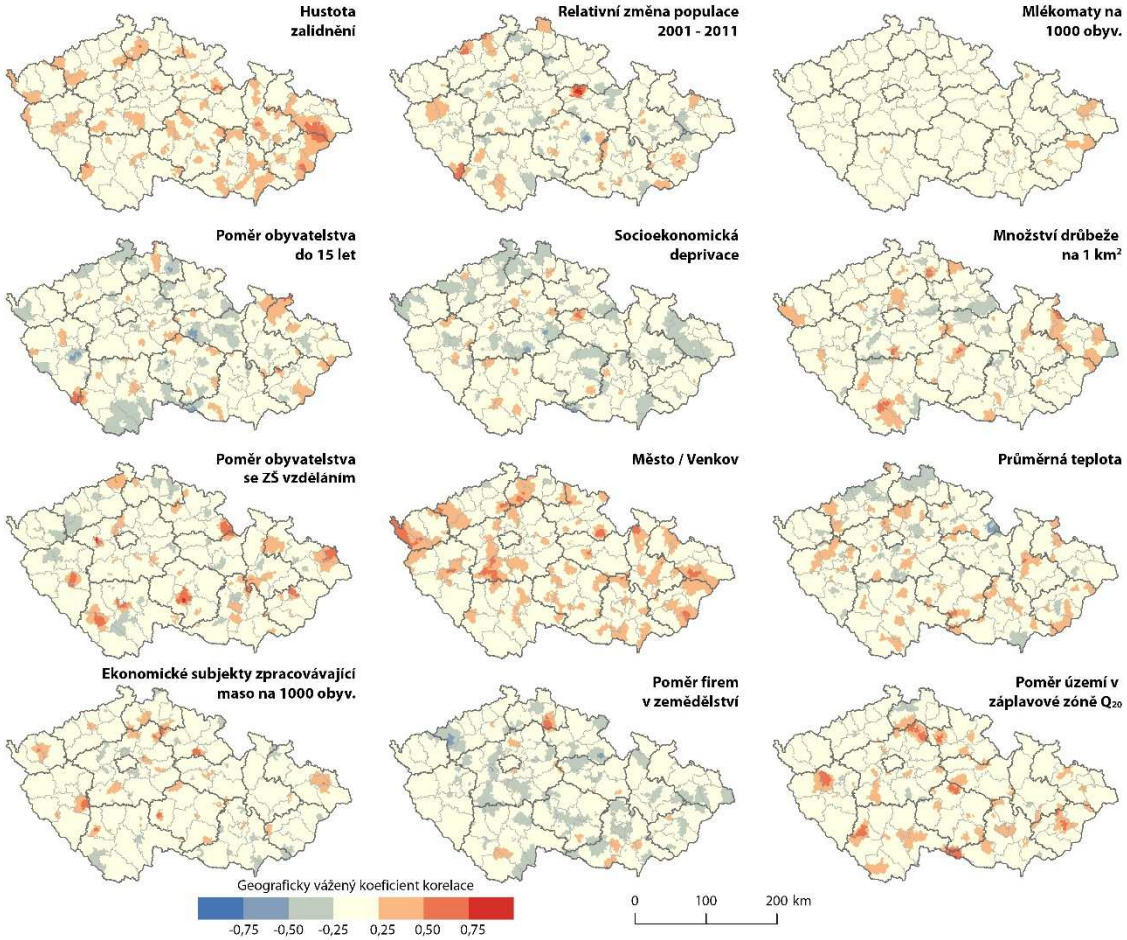
Pomocí korelace lze hodnotit sílu asociace mezi charakteristikami v globálním měřítku, tedy bez přihlédnutí k prostorové variabilitě jevu. Má-li být zohledněna i lokální proměnlivost vybraných charakteristik, pak je nutné využít i lokálních či geograficky vážených měr a s nimi spojených inferencí. Vývoj a využití lokálních ekvivalentů tradičních statistických charakteristik jsou významným příspěvkem prostorové statistiky a geoinformatiky všeobecně k jejich uplatnitelnosti v širokém množství vědních oborů – v geografii či politologii (Maškarinec, 2013), ekonometrii (Bernard et al., 2014; Blažek a Netrdova, 2012; Netrdová

a Nosek, 2009), ekologii (F. Dormann et al., 2007) nebo odkonce zkoumáním struktury v textilní výrobě (Xie, 2008). Korelační koeficient je možné modifikovat do podoby lokálního indexu pomocí stanovení vhodných (prostorových) vah nebo s využitím jádrových odhadů. Podle Kalogirou (2011) je možné vztahu 6.1 využít k výpočtu lokálního Pearsonova korelačního koeficientu (r_l), který je možné zapsat jako:

$$r_l = \frac{\sum_{j=1}^{k}(x_j - \overline{x_l})(y_j - \overline{y_l})}{\sqrt{\sum_{j=1}^{k}(x_j - \overline{x_l})^2}\sqrt{\sum_{j=1}^{k}(y_j - \overline{y_l})^2}} \qquad [6.2]$$

kde *k* je počet nejbližších sousedů okolo bodu *i*; $\overline{x_l}$ a $\overline{y_l}$ jsou průměrné hodnoty proměnných *x* a *y* v *k* nejbližších sousedech *i*, např. $\overline{x_l} = \sum_{j=1}^{k}\frac{x_j}{k}$. Počet sousedů může být zvolen empiricky nebo pomocí optimalizace dostupné například při výpočtu GWR pomocí adaptivního jádrového odhadu (Fotheringham et al., 2002; Charlton a Fotheringham, 2009). Analogicky je možné podobným způsobem modifikovat téměř jakoukoliv statistickou charakteristiku. Výsledkem lokální analýzy je v ideálním případě některá z forem geovizualizace, která na rozdíl od číselných hodnot, tabulek či grafů, dokáže přehledně zobrazit požadovanou prostorovou variabilitu vztahu. Výsledkem výpočtu korelací tak je mapová korelační matice (Obrázek 25).

Pomocí lokálního Pearsonova korelačního koeficientu byla hodnocena těsnost lokálních asociací mezi relativním rizikem (SIR) a reprezentativními charakteristikami prostředí vybranými na základě globálních korelací, konzultací a zahraničních studií. Lokální korelační koeficient byl počítán pro jednotlivé charakteristiky na úrovni obcí. K jeho získání byl využit balík *lctools* (Kalogirou, 2015) dostupný pro **R**. K výpočtu bylo potřeba kromě nahrání charakteristik také stanovit velikost sousedství obcí. Pro volbu sousedství byly uvažovány tři možnosti – již dříve využité sousedství typu královna, dále pak byla zvažována adaptivní sousedství definovaná počtem možných sousedů (1 % a 5 % obcí). Z těchto možností byla vybrána střední varianta – sousedství obce je tvořeno maximálně 1 % všech obcí (n = 63). Sousedství typu královna nebylo vybráno z důvodu nízkého průměrného počtu sousedů (*n=6*), a tudíž ne zcela vypovídajících statistik, varianta s 5 % obcí pak nebyla využita z opačného důvodu, kdy zahrnovala příliš velké množství obcí a výsledkem byly příliš vyhlazené odhady korelačního koeficientu. Výsledky lokálního korelačního analýzy jsou přehledně shrnuty v mapové korelační matici na Obrázku 25, kde jsou vyobrazeny lokální korelace mezi relativním rizikem onemocněním kampylobakteriózou v obci a jedenácti zvolenými charakteristikami. Dvanáctou doplňující charakteristikou je příslušnost obce k venkovskému venkovskému/městskému prostoru jak ji definoval Pászto et al. (2014), kterou jako významnou zmiňují Arsenault et al. (2012). Mezi výslednými charakteristikami později využitými při analýze asociací pomocí modelů není tato charakteristika z důvodu vysoké (globální) korelace s hustotou zalidnění. Je vhodné také zmínit, že v případě hodnocení asociace relativního rizika a relativního počtu mlékomatů není využita lokální Pearsonova korelace, kterou nebylo možné vypočítat z důvodu nízkého počtu mlékomatů vzhledem k počtu obcím. Místo ní byl modifikován původní skript pro účely výpočtu robustnějšího lokálního Spearmanova koeficientu pořadové korelace.



Obr. 25 Geograficky vážené korelace mezi relativním rizikem (SIR) onemocněním kampylobakteriózou v obcích a vybranými charakteristikami prostředí

Z pohledu do mapové korelační matice na Obrázku 25 lze vysledovat, že vazby mezi relativním rizikem (SIR) a vybranými charakteristikami jsou opravdu spíše lokálního než globálního rázu. Převahu mají spíše mírně těsné pozitivní a negativní vazby ($0{,}25 < |r_i| < 0{,}50$). Jako nejvýznamnější je možné hodnotit asociaci s hustotou zalidnění a s ní související příslušností k městskému prostoru. Oproti obecnému očekávání jde výsledek hodnocení lokální korelace mezi SIR a socioekonomickou deprivací i poměrem mladých osob na celkové populaci. Oba dva faktory byly často v zahraničních studiích hodnoceny jako významné vzhledem k nárůstu incidence/relativního rizika (pozitivní korelace), zatímco v případě ČR na významné části území převládá sice mírná, ale negativní lokální korelace – existuje tedy možnost, že se zvyšováním podílu mladých osob a socioekonomické deprivace obyvatelstva se sníží i incidence kampylobakteriózy v dané obci. Nejméně významná lokální korelace v rámci ČR byla identifikována v případě mlékomatů, jejichž význam je rozeznatelný pouze na severovýchodní Moravě. Slabá vazba výrazně lokálního charakteru je patrná také u ekonomických subjektů zabývajících se živočišnou výrobou. V případě zbylých charakteristik nelze určit jejich jednoznačný příspěvek ve vztahu k SIR kampylobakteriózy, protože se mění velmi lokálně a kolísá mezi pozitivní a negativní lokální korelací.

Kromě lokálního koeficientu autokorelace bylo k hodnocení asociací na lokální úrovni možné využto i generalizovaného lokálního Moranova I kritéria. Lokální Moranovo I kritérium pro dvě proměnné rozšiřuje původní koncept průzkumu prostorové autokorelace

a umožnuje využít ho pro analýzu vzájemného dvourozměrného prostorového vzoru. Generalizovaná forma kritéria je definována jako:

$$I_{ko}^{i} = y_k^i \sum_j w_{ij}\, y_o^i \qquad [6.3]$$

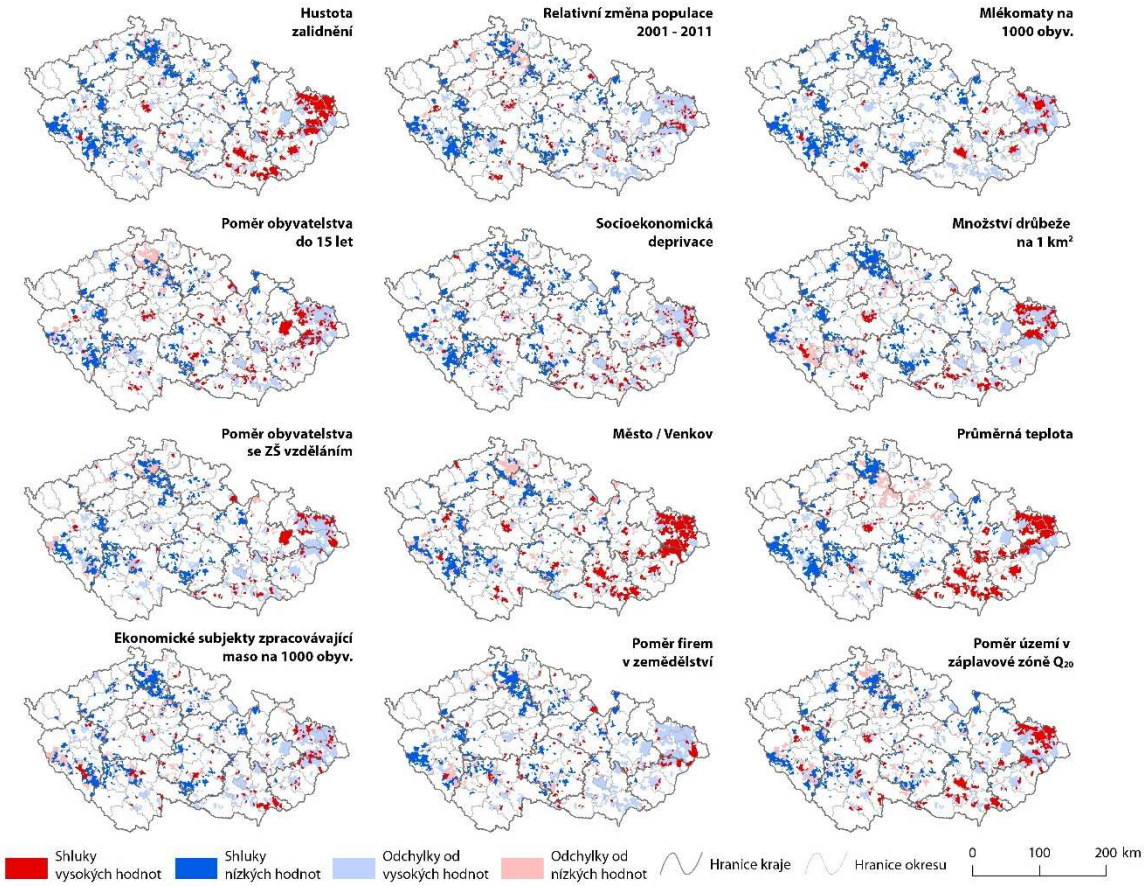
Statistická charakteristika vyjadřuje stupeň lineární asociace mezi hodnotami jedné proměnné y_k v místě i, y_k^i a průměrem hodnot dalších proměnných y_o v sousedních místech j, y_o^j (Gruebner et al., 2011). Vyšší než očekávaná hodnota indexu předpokládá výskyt podobných shluků obou zkoumaných veličin, a naopak vysoká nepodobnost znamená silnou negativní autokorelaci. Významnost síly vazby je vyhodnocena pomocí permutačních testů (Anselin et al., 2002).

Pomocí globálního a lokálního Moranova I kritéria pro dvě proměnné byly, stejně jako v předchozím případě, hodnoceny prostorové asociace mezi relativním rizikem (SIR) a vybranými charakteristikami prostředí v obcích ČR. Analýza proběhla v prostředí programu GeoDA, ve kterém je zvolená metoda implementována. Základním požadavkem pro analýzu je kromě vstupních proměnných také definování matice sousedství, které bylo zvoleno jako sousedství 1. řádu typu královna. Dále byly postupně vypočítány globální a lokální Moranova I kritéria. Výsledky globálního Moranova I kritéria pro dvojici proměnných jsou shrnuty v Tabulce 10. Z ní je patrné, že ve všech případech kromě dvou (nízké vzdělání a mlékomaty) lze mezi hodnocenými statistikami nalézt asociaci, která ač je slabá, tak je statisticky významná. V sedmi případech se jedná o pozitivní asociaci, která předpokládá při nárůstu některé z charakteristik i nárůst relativního rizika onemocnění (a naopak). Ve třech případech jde o asociaci negativní, tedy při nárůstu jedné charakteristiky klesá druhá (a naopak). Negativní prostorovou asociaci lze předpokládat v případě relativní změny obyvatel (zde pravděpodobně souvisí s pozitivní autokorelací socioekonomické deprivace) a oproti předpokladům také v případě poměru zemědělských firem a především podílu obyvatelstva do 15 let na celkové populaci. Tyto výsledky však potvrzují hodnocení pomocí lokálního Pearsonova koeficientu korelace (Obrázek 24).

Tab. 10 Vyhodnocení globálního Moranova I kritéria pro dvě proměnné mezi relativním rizikem (SIR) a vybranými charakteristikami prostředí

Charakteristika	Moranovo I	Očekávaná hodnota Moranova I	p-value	Globální asociace
Hustota zalidnění	0,040	0,000	< 0,001	ANO
Obyvatelstvo do 15 let	-0,020	0,000	< 0,001	ANO
Obyvatelstvo s nízkým vzděláním	0,000	0,000	0,434	NE
Zpracování masa	0,040	0,000	< 0,001	ANO
Změna populace	-0,004	0,000	< 0,001	ANO
Socioekonomická deprivace	0,026	0,000	< 0,001	ANO
Příslušnost k městu/venkovu	0,047	0,000	< 0,001	ANO
Zemědělské firmy	-0,031	0,000	< 0,001	ANO
Mlékomaty	0,005	0,000	0,110	NE
Drůbež	0,023	0,000	0,001	ANO
Průměrná teplota	0,063	0,000	< 0,001	ANO
Záplavové území Q_{20}	0,014	0,000	0,010	ANO

Výsledky lokální analýzy prostorového vzoru dvou proměnných jsou formou mapové matice korelací zobrazeny na Obrázku 26. Podobné lokální prostorové vzory lze pozorovat v případě hustoty zalidnění, příslušnosti k městskému či venkovskému prostoru a průměrné teploty vzduchu, kde je patrná výrazná diferenciace shluků vysokých a nízkých hodnot. Vysoké hodnoty se nalézají především na Moravě (severovýchod, jih) a ve Slezsku, zatímco v Čechách jde spíše o shluky nízkých hodnot. Výjimkou jsou oblasti kolem Benešova, Českých Budějovic a případně i Plzně. Na území Čech se vyskytují shluky nízkých hodnot u všech hodnocených charakteristik, ačkoliv lokálně se objevují i odchylky od těchto shluků, tak jako je tomu v oblasti severních Čech, kde se vyskytují významné odchylky od nízkých hodnot. Prostorový vztah relativního rizika s poměrem osob s nízkým vzděláním, poměrem zemědělských firem, mlékomaty, socioekonomickou deprivací a ekonomickými subjekty zpracovávajícími maso má (zejména na Moravě) podobnou lokální strukturu, kde se kolem menších shluků vysokých hodnot objevují odchylky od těchto hodnot (outliers), které představují místa negativní asociací.



Obr. 26 Lokální Moranovo I kritérium pro dvojici proměnných – hodnocení prostorové autokorelace relativního rizika (SIR) onemocněním kampylobakteriózou ve vztahu k vybraným charakteristikám prostředí

6.3 Analýza asociací mezi kampylobakteriózou a vybranými faktory prostředí pomocí klasifikačních modelů

Pomocí korelace či prostorové korelace lze postupně studovat souvislosti mezi charakteristikou onemocnění a zvolenými faktory prostředí. Nevýhodou v analýze korelací je fakt, že jsou studovány pouze lineární vztahy mezi dvojicí nezávislých charakteristik a není tak definována závislá a nezávislá proměnná (či více proměnných). Korelační koeficient není kvantifikací vztahu mezi proměnnými, ale pouze určuje sílu a směr lineárního vztahu mezi proměnnými. Tyto nedostatky je možné odstranit pomocí statistických modelů a včetně jejich prostorových modifikací. Díky využití modelů je možné nejen analyzovat vztah mezi konkrétní charakteristikou morbidity a potenciálně souvisejícími faktory, ale i predikovat hodnoty do budoucna nebo v místech, kde nebyla charakteristika zjišťována. Často není nutné, aby byla známa konkrétní hodnota charakteristiky onemocnění, která je současně i náročná na přesnost predikce. Místo ní postačí území vhodně ohodnotit vzhledem k jeho zranitelnosti vůči dané nemoci a úloha modelování takto transformovat na úlohu klasifikační. Právě klasifikací území z pohledu jeho náchylnosti k onemocnění kampylobakteriózou, která je vyjádřena relativním rizikem (SIR), se zabývá tato podkapitola.

6.3.1 Redukce dimenze a klasifikace relativního rizika

Z původní více než stovky charakteristik bylo postupným odstraňováním za pomoci globální korelace vybráno jedenáct, které buď vhodně reprezentovaly původní datovou sadu, nebo vyjadřovaly některou z významných vlastností územní jednotky identifikovanou v zahraničních studiích. Výběr těchto jedenácti charakteristik zredukoval původní datovou o téměř 90 %. Konkrétně byly pro další využití vybrány tyto charakteristiky:

- Hustota zalidnění (počet obyvatel na 1 km^2 obce)
- Podíl obyvatelstva se základním vzděláním a bez vzdělání na celkové populaci v obci (%)
- Relativní změna počtu obyvatel 2001—2011 (%)
- Podíl obyvatelstva do 15 let na celkové populaci (%)
- Socioekonomická deprivace (bezrozměrný index)
- Podíl ekonomických subjektů v zemědělství na všech ekonomických subjektech v obci (%)
- Odhad počtu drůbeže na plochu obce (počet kusů drůbeže na 1 km^2 obce)
- Průměrná teplota (°C)
- Podíl plochy obce v záplavovém území (Q_{20}) (%)
- Společnosti zabývající se zpracováním masa na 1000 obyvatel
- Počet mlékomatů na 1000 obyvatel

Jako závislá proměnná (či zdroj klasifikace) většiny modelů a postupů sloužilo relativní riziko ohrožení kampylobakteriózou v obcích ČR (SIR). Často není nutné, aby bylo riziko v obci zobrazováno a modelováno s velkou přesností na jednotky procent či ještě podrobněji, ale vhodnějším postupem může být stanovení tříd rizika pro jednotlivé obce. Stejný postup byl zvolen i v případě popisované případové studie. Obce České republiky byly rozděleny do

čtyř tříd na základě jejich nepřímo standardizované incidence (SIR neboli reaktivního rizika). Do první kategorie spadají obce, kde dosud nebyl zaznamenán výskyt onemocnění. V druhé kategorii jsou obce, kde $0 < RR < 0{,}8$, zatímco ve třetí se nachází obce $0{,}8 \leq RR < 1{,}5$. Čtvrtou a nejvíce ohroženou kategorii, tvoří obce s $RR \geq 1{,}5$.

Analýza hlavních komponent a GW PCA

I přesto, že bylo množství charakteristik území v datové sadě zmenšeno na 10 % její původní velikosti, tak i jedenáct vybraných nezávislých proměnných je stále příliš mnoho pro účely modelování a klasifikace. Přístupem využívaným k redukci dimenze datové základny je analýza hlavních komponent (PCA), která kromě snížení počtu charakteristik (dimenzí) umožňuje nalézt správný rozměr datového souboru, který se značně odlišuje od rozměru původního (Hebák et al., 2005b). Současně s redukcí rozměrů proběhla identifikace nových proměnných nazývaných komponenty. Tyto nové proměnné vznikají vždy, jejich smysluplnost však nebývá vždy jednoznačná a je spíše výjimkou, že jdou vhodně interpretovat. Jednotlivé charakteristiky v datech byly před vstupem do PCA standardizovány, aby byla zajištěna jejich srovnatelnost a nezávislost na jejich měřítku/rozsahu. PCA transformuje původní data K pozorovaných proměnných Z_k na K hlavních komponent F_k, které jsou vzájemně nezávislé (Wang, 2009)

$$Z_k = l_{k1}F_1 + l_{k2}F_2 + \cdots + l_{kK}F_K \qquad [6.4]$$

současně mohou být jednotlivé komponenty F_j vyjádřeny lineární kombinací původních proměnných Z_k.

$$F_j = a_{1j}Z_1 + a_{2j}Z_2 + \cdots + a_{Kj}Z_K \qquad [6.5]$$

Komponenty F_j jsou konstruovány jako vzájemně nekorelované a seřazené podle jejich relativního příspěvku k celkové varianci souboru. Vzájemná nekorelovanost komponent je výhodou při využití komponent jako nezávislých (vysvětlujících) proměnných do regresních modelů, kde bývá jednou z podmínek nekolinearita vstupních dat. Díky původní standardizaci se pak rozdělení pravděpodobnosti jednotlivých komponent přibližuje normálnímu rozdělení.

Rozšířením metody PCA pro účely analýzy prostorových dat je geograficky vážená metoda hlavních komponent (GW PCA) (Lloyd, 2010). V GW PCA dochází k výpočtu série lokalizovaných PCA, při kterých jsou postupně vypočítány a mapovány lokální komponenty, variance a skóre, což umožňuje zachytit lokální změny či prostorově se měnící vlastnosti mnohorozměrných dat (Lu et al., 2014).

V souladu s doporučenými postupy pro průběh PCA byla nejdříve data jedenácti vybraných reprezentativních proměnných normalizována na rozsah $\langle -1; 1\rangle$ využitím vztahu $\frac{x - \bar{x}}{\max(|x - \bar{x}|)}$. V dalším kroku proběhla samotná PCA pomocí funkce *prcomp* v balíku *stats* v **R**. Výsledkem je jednak matice zátěží (či vah) a také rotované hodnoty každé komponenty pro každou obec. Jednotlivé zátěže prvních pěti komponent z celkových jedenácti jsou shrnuty v Tabulce 11. Pět komponent bylo zvoleno jako vhodná redukce a náhrada původní datové sady díky tomu, že dohromady shrnují 89 % jejího celkového rozptylu. Jednoznačná interpretace komponent bohužel není možná. V první a druhé komponentě dominuje vliv obyvatelstva do 15 let a průměrné teploty vzduchu, třetí komponenta je zemědělská (podíl

zemědělských firem a počty drůbeže), ve čtvrté a páté komponentě je významný vliv socioekonomické deprivace a podílu záplavových zón na ploše obce. Vzhledem k tomu, že PCA vysvětluje velkou většinu variability dat, tak je možné ji dále používat i přes fakt nejednoznačné intepretace jejích komponent.

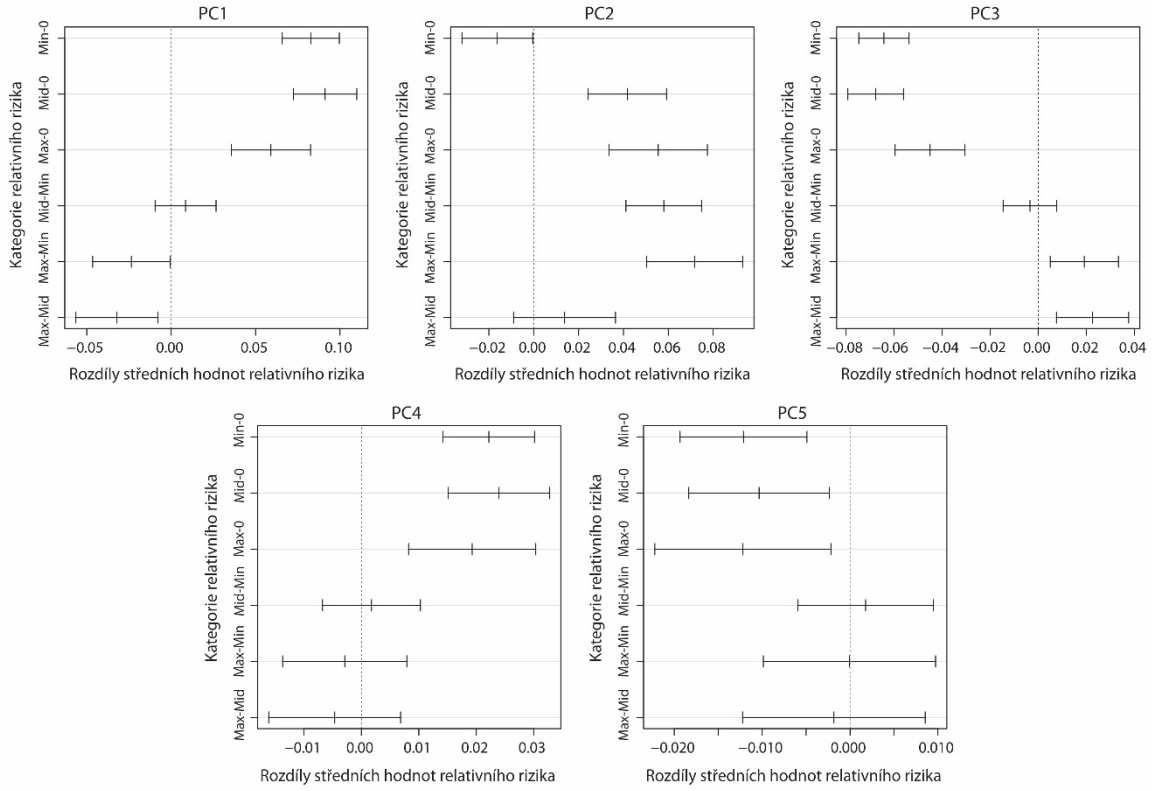
Tab. 11 Shrnutí matice zátěží vzešlé z analýzy hlavních komponent (vyobrazeno je prvních 5 komponent), tučně jsou označeny nejvýznamnější zátěže v komponentě

	PC1	PC2	PC3	PC4	PC5
Obyvatelstvo do 15 let	**0,788**	**-0,609**	0,055	-0,010	0,035
Hustota zalidnění	0,057	0,047	-0,111	0,152	-0,114
Obyvatelstvo s nízkým vzděláním	-0,009	-0,001	-0,006	0,031	-0,011
Změna populace	0,054	-0,020	-0,004	-0,072	0,023
Socioekonomická deprivace	0,032	0,086	0,109	**-0,806**	**0,383**
Zemědělské firmy	-0,167	-0,123	**0,916**	0,119	0,196
Drůbež	0,085	0,090	**0,306**	0,177	**-0,491**
Zpracování masa	-0,010	0,000	-0,008	-0,007	-0,024
Průměrná teplota	**0,576**	**0,765**	0,164	0,001	-0,014
Záplavové území Q_{20}	0,064	0,099	-0,111	**0,525**	**0,747**
Mlékomaty	0,002	-0,001	0,009	0,009	-0,010
Podíl celkového rozptylu (%)	33,650	28,900	13,070	7,314	5,959
Kumulativní součet rozptylu (%)	33,650	62,550	75,620	82,934	88,893

Pro vyhodnocení rozdílnosti jednotlivých komponent s ohledem na příslušnost ke skupině byla použita analýza variance s následným post-hoc porovnáváním pomocí Tukeyho metody mnohonásobného porovnávání. Konkrétně jde o metodou Tukey HSD[37], která je vhodná i pro srovnávání nevyvážených skupin. Vzhledem k tomu, že analýza variance prokázala v případě všech pěti komponent signifikantní rozdílnost, s p-value výrazně pod hladinou 0,05, tak bylo využito Tukey HSD ke zjištění, kde konkrétně je možné nalézt rozdíly mezi skupinami.

Výsledky mnohonásobného porovnávání jsou zobrazeny na Obrázku 27. Horizontální linie představují rozdíly středních hodnot jednotlivých skupin s konfidenčními intervaly, vertikální čárkovaná šedá linie představuje 0, tedy situaci, kdy mezi středními hodnotami není žádný rozdíl. Situace, kdy je horizontální linie rozdílu skupin umístěna mimo vertikální linii, indikuje statisticky významný rozdíl mezi středními hodnotami skupin (p-value < 0,05). Pokud horizontální linie protínají vertikální, pak nemohou být skupiny pokládány za rozdílné. Na základě PC1 (první hlavní komponenty) je možné od sebe teoreticky rozlišit všechny skupiny až na rozdíl mezi minimálním (Min, 0 < RR < 0,80) a středním (Mid, 0,80 ≤ RR < 1,50) relativním rizikem. Na základě PC2 je pak opět možné rozlišit od sebe veškeré skupiny až na skupiny středního a vysokého (RR ≥ 1,50) relativního rizika. Situace v případě PC3 je stejná jako u PC1, PC4 s PC5 umožní určit rozdíl mezi oblastmi s nulovým a nenulovým relativním rizikem.

[37] Tukey Honest Significant Difference

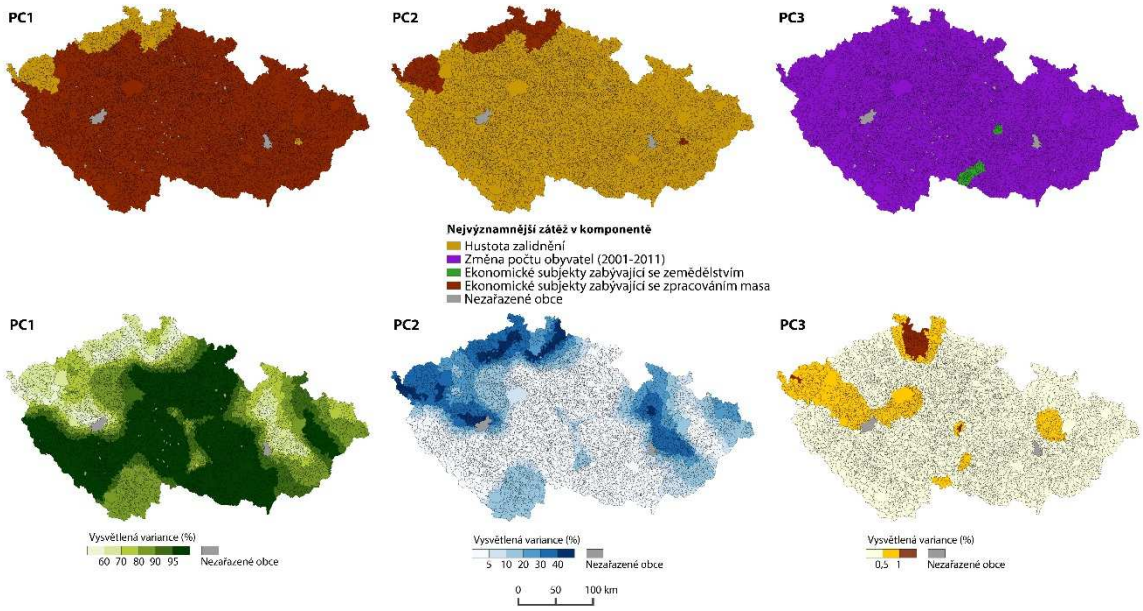


Obr. 27 Vizualizace mnohonásobného porovnávání

Kromě PCA byla data analyzována pomocí GW PCA, která na rozdíl od první jmenované umožní zkoumat a zobrazit lokální proměnlivost komponent v obcích. Analýza byla provedena zejména s pomocí balíku *GWmodel* v **R** (Gollini et al., 2015). Stejně jako v případě dříve zmíněných prostorových analýz je i zde nutné nejdříve stanovit prostorové váhy, na základě kterých budou počítány jednotlivé lokální PCA. Balík GWmodel obsahuje optimalizační funkce založené na krosvalidaci zvolených parametrů, které mohou pomoci s výběrem vhodného typu sousedství. Na výběr je buď adaptivní sousedství definované na základě počtu sousedů, nebo sousedství fixované na vzdálenost. Kromě těchto voleb je nutná i volba tvaru funkce pro výpočet jádrového odhadu hustot, na jehož základě je ohodnocen vliv okolních obcí na výpočet vlastností lokálního PCA. Na výběr jsou jádra typu gausovské, exponenciální, bisquare, tricube a boxcar. Pro GW PCA vybraných charakteristik obcí České republiky bylo na základě krosvalidace pro 5 hlavních komponent vybráno adaptivní sousedství o 320 sousedech. Tento počet na jednu stranu zajistí možnost lokálního průzkumu, ale na druhou stranu poskytne výrazně shlazený výsledek GW PCA. Z tohoto důvodu byla pro srovnání provedena také GW PCA s adaptivním sousedstvím definovaným 54 nejbližšími obcemi, které byly pomocí optimalizace označeny sice jako méně vhodné, ale umožní lepší vnímání lokálních změn. GW PCA tedy byly vypočteny pro adaptivní sousedství o 54 a 320 obcích s využitím jádrové funkce typu bisquare. Data byla před vstupem do analýzy standardizována stejným postupem jako při PCA.

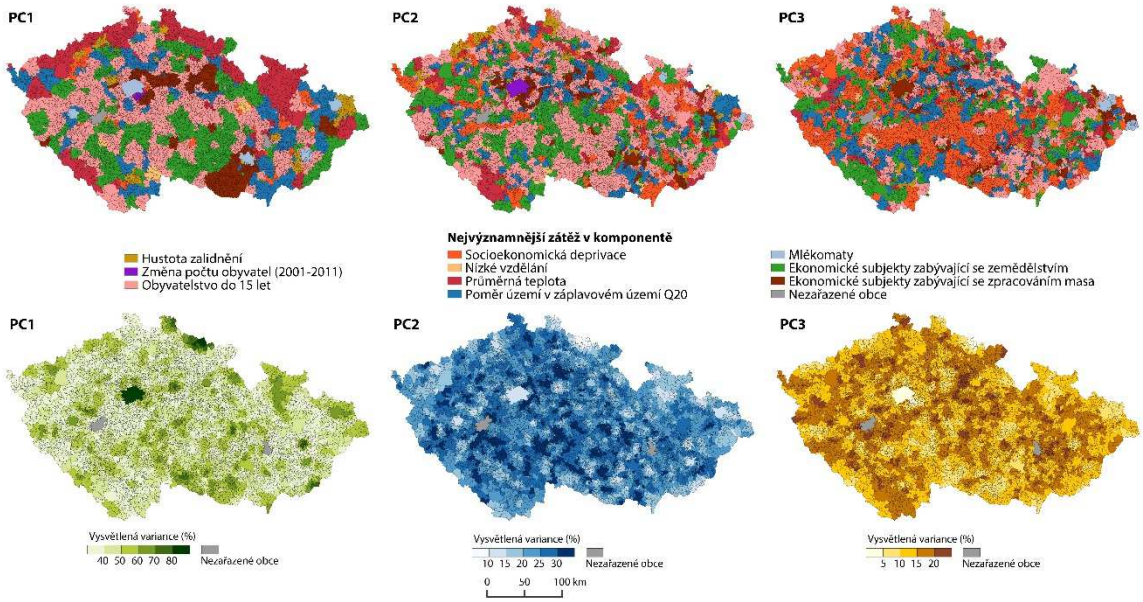
Výsledky obou provedených GW PCA jsou na Obrázcích 28 a 29. GW PCA s prostorovými váhami složenými z 320 nejbližších obcí (Obrázek 28) se svým pojetím výrazně přibližuje PCA. Velké množství variance datové sady je vysvětleno pomocí první komponenty (PC1), téměř celou varianci ve většině obcí pak společně vysvětlují PC1 a PC2. V prostorovém vzoru prvních dvou hlavních komponent je patrný vliv doplňujících se hustoty zalidnění

a relativního počtu ekonomických subjektů zpracovávajících maso. Ve třetí komponentě převládá vliv relativní změny počtu obyvatel obcí mezi lety 2001—2011, který je místy doplněn poměrem zemědělských ekonomických subjektů. Výsledky první GWPCA jsou výrazně shlazeny, ačkoliv adaptivní kernel považoval za sousední pouze 5 % z celkového počtu obcí. Přestože jsou v prvních třech komponentách jasně rozeznatelné nejvýznamnější vlastnosti, tak se může stát, že jsou tyto vlastnosti v jednotlivých komponentách doplněny dalšími, které mohou mít jen o málo nižší hodnotu zátěže v komponentě. Toto je potvrzeno i velkým množstvím variance datové sady vysvětleným pomocí prvních dvou hlavních komponent.



Obr. 28 Geograficky vážená analýza hlavních komponent, sousedství 320 obcí; horní řada zobrazuje nejvýznamnější vlastnost v komponentě, dolní řada podíl komponentou vysvětlené variance

GW PCA s prostorovými váhami složenými z 54 nejbližších obcí (Obrázek 29), které tvoří necelé 1 % celkového počtu obcí, vykazuje výrazně větší prostorovou variabilitu než předchozí případ. Kromě větší prostorové variability je patné také širší zastoupení nejvýznamnějších charakteristik i rozdělení vysvětlené variance v rámci v rámci území i komponent. Největší vliv v rámci první komponenty vykazují podíl obyvatelstva do 15 let, poměr zemědělských ekonomických subjektů, teploty či průměrná teplota. Na Ostravsku a Českobudějovicku je významným činitelem také hustota zalidnění nebo mlékomaty (tam kde jsou přítomny). Druhá komponenta má podobné významné přispěvatele a je doplňkem první komponenty. Ve třetí komponentě je nejvýznamnější charakteristikou socioekonomická deprivace. Celková variance vysvětlená pomocí prvních 3 prostorových komponent se lokálně pro jednotlivé obce pohybuje v rozsahu 65—99 %.



Obr. 29 Geograficky vážená analýza hlavních komponent, sousedství 54 obcí; horní řada zobrazuje nejvýznamnější vlastnost v komponentě, dolní řada podíl komponentou vysvětlené variance

6.3.2 Regresní a prostorové regresní modely

V předchozích kapitolách byly identifikovány charakteristiky, které potenciálně mohou ovlivňovat výskyt kampylobakteriózy v České republice. Jedenáct charakteristik bylo zredukováno pomocí metody PCA na pět hlavních komponent. Jejich výhodou je, že mezi sebou nekorelují a díky standardizaci se také rozdělení pravděpodobnosti jednotlivých komponent výrazně přiblížilo normálnímu rozdělení. Mají-li vybrané charakteristiky, resp. z nich vzniklé komponenty, výrazně přispívat k formování výskytu kampylobakteriózy, pak je na jejich základě možné vhodně klasifikovat známé míry morbidity a případně predikovat pravděpodobný vývoj chování onemocnění při změně charakteristik území. V případě popisu modelování morbidity (relativního rizika či incidence) často není nutné a efektivní, zaměřovat se na konkrétní hodnoty měr morbidity, ale je výhodné zaměřit se na kategorie definované těmito mírami (Miller a Franklin, 2002). Proto není modelována konkrétní hodnota, ale spíše příslušnost k určité předem definované skupině, například zda je relativní riziko onemocnění v obci nulové, minimální, průměrné či nadprůměrné. K popisu onemocnění pomocí vybraných charakteristik je možné využít (regresních) modelů a samotný proces modelování je současně transformován na klasifikační úlohu. Stejný postup je zvolen i v této části disertační práce, ve které jsou využity (zobecněné) regresní modely včetně prostorových variant a diskriminační analýza.

Modely a klasifikační postupy jsou hodnoceny pomocí ROC (Receiver Operating Characteristic) křivek, pomocí celkové přesnosti zařazení a přesnosti zařazení do tříd. ROC křivky jsou užitečným vizuálním nástrojem k hodnocení výkonu modelu, techniky či klasifikátoru. Hodnotí vztah mezi senzitivitou – poměrem skutečně pozitivních pozorování k součtu skutečně pozitivních a falešně negativních případů, a specificitou – poměrem skutečně negativních pozorování a součtu skutečně negativních a falešně pozitivních klasifikací (Metz, 1978). Na ose x je znázorněna relativní četnost falešně pozitivních případů (pravděpodobnost nesprávného zařazení), na ose y pak relativní četnost skutečně pozitivních

případů (pravděpodobnost správného zařazení). Křivka srovnává zisky (správně zařazené případy) a ztráty (nesprávně zařazené) způsobené klasifikátorem (Fawcett, 2006). Křivka je vhodná především pro binární klasifikátory, ale postupným testováním je možné ji rozšířit i na více než dva. Na křivce je hodnoceno tzv. AUC (Area Under Curve), tedy plocha pod křivkou. Přesněji řečeno jde o poměr plochy pod křivkou z celkové plochy grafu s hodnotami $\langle 0; 1\rangle$. Obecně se dá říci, že s rostoucí AUC roste i schopnost klasifikátoru odlišit od sebe jednotlivé třídy. Pokud je ovšem AUC < 0,50, pak je klasifikátor zcela nevhodný.

Multinomická a ordinální logistická regrese

První využitou skupinou modelů jsou generalizované lineární modely, konkrétně multinomická regrese (logistická regrese s vícekategoriální vysvětlovanou proměnou) a ordinální logistická regrese. Oba dva modely jsou vhodné k modelování a klasifikaci dat, kde modelovaná (závislá nebo také vysvětlovaná) proměnná je kategorická a může nabývat více než dvou hodnot. Hlavním rozdílem mezi nimi je fakt, že ordinální logistická regrese umožňuje pracovat se daty, jejichž kategorie je možno seřadit. **Multinomická logistická regrese** je definována jako (Middel, 2007):

$$z_{ij} = x'\beta_j \qquad [6.6]$$

kde x je vektor prediktorů (nezávislých proměnných) - $x' = [1, x_1, x_2, \dots, x_k]$ a β označuje regresní parametry těchto proměnných obsahující odhady koeficientů, kde $\beta = [\beta_0, \beta_1, \dots, \beta_k]$. S vypočítanou sadou koeficientů β_j je zjšťována pravděpodobnost jednotlivých tříd P_j pro počet skupin $s \geq 2$ jako:

$$P_j = \frac{\exp(x'\beta_j)}{\sum_{j=0}^{s-1}\exp(x'\beta_j)} \qquad [6.7]$$

kdy pro referenční kategorii ($j = 0$) $\beta_0 = 0$ a pro bazické logity platí (Pecáková, 2007):

$$ln\frac{P_j}{P_0} = x'\beta_j, j = 0, 1, \dots, s-1 \qquad [6.8]$$

Základní vyjádření pro **ordinální logistický model** je definováno stejně jako v případě multinomiálního dle vztahu 6.6. Rozdílné je ovšem pojetí vyjádření pravděpodobností. Pro ordinální logistický model založený na kumulativních logitech, lze kumulativní logit zapsat jako:

$$ln\frac{F_j}{1-F_j} = ln\frac{P(Y \leq y_j)}{P(Y > y_j)} = ln\frac{P_0 + P_1 + \cdots + P_j}{P_{j+1} + P_{j+2} + \cdots + P_{s-1}}, j = 0, 1, \dots, s-2 \qquad [6.9]$$

a regresní funkci s užitím kumulativního logitu jako:

$$ln\frac{F_j}{1-F_j} = x'\beta_j = \beta_{0j} - \sum_{i=1}^{k}\beta_{ij}x_i\,, j = 0, 1, \dots, s-2 \qquad [6.10]$$

Parametry β_{0j} jsou prahové parametry pro jednotlivé kategorie veličiny Y, které představují logaritmus šance, že Y nabývá nejvýše j-té kategorie a nikoliv vyšší (Pecáková, 2007). Je důležité si uvědomit, že výsledkem logistických modelů nejsou většinou přímo pravděpodobnosti, ale hodnota logitu – šance (či poměru šancí), ze kterých lze odvodit pravděpodobnosti transformací.

Využitím prostorových vah (w_{ij}) je možné ordinální model transformovat na prostorový (lokální) ordinální regresní model, který na jednu stranu umožní zohlednit lokální variabilitu hodnoceného jevu, na druhou stranu je však obtížnější jeho interpretace a hodnocení. Koncept tohoto typu modelu, který částečně vychází z Nur Aidi a Purwaningsih (2013) lze definovat jako

$$ln\frac{F_j}{1-F_j} = x'\beta_j w_{ij} = \beta_{0j} w_{ij} - \sum_{i=1}^{k} \beta_{ij} x_i w_{ij} \qquad [6.11]$$

Model takto provede sérii lokalizovaných výpočtů na základě definovaného sousedství a jeho výsledkem jsou regresní zjištěné poměry šancí pro každou územní jednotku, které se na rozdíl od tradičního modelu mohou výrazně odlišovat v souvislosti s prostorovými změnami charakteristik.

Poissonova a negativní binomická regrese četností výskytu

K modelování absolutních četností výskytu případů onemocnění, a obecně k modelování nálezových dat agregovaných do různých prostorových či neprostorových jednotek, je využíván další typ generalizovaného lineárního modelu – obecný poissonovský model. Předpokladem k jeho využití je, že popisovaná data mají Poissonovo rozdělení parametrizované střední hodnotou a stejně velkým rozptylem. Požadavek na shodnou velikost střední hodnoty a rozptylu však nelze často dodržet. Běžnějším případem je stav, kdy je rozptyl vyšší než střední hodnota a z pohledu poissonovské regrese dochází k tzv. *overdispersion*. Alternativou k poissonovským regresním modelům, pokud je rozptyl větší než střední hodnota, jsou modely, které využívají negativně binomického rozdělení. To lze z poissonova rozdělení odvodit tím, že je náhodně (dle gamma rozdělení) měněna jeho střední hodnota a negativně binomický model tak má oproti původnímu poissonovskému navíc jeden parametr popisující overdispersion, což zvyšuje jeho flexibilitu (Pekár a Brabec, 2009). Negativní binomický model je možné zapsat jako (da Silva a Rodrigues, 2014):

$$y_j \sim NB\left[t_j\, exp\left(\sum_k \beta_k x_{jk}\right), \alpha\right] \qquad [6.12]$$

jde t_j je offset modelu, α je parametr pro overdispersion, β_k je parametr nezávislé proměnné x_k, y_j je závislá proměnná a NB představuje negativní binomické rozdělení. Pro prostorová data je možné modifikovat negativní binomický model pomocí principu GWR (da Silva a Rodrigues, 2014).

Při analýze dat agregovaných do relativně malých územních jednotek (např. obce) je přirozené, že se i v případě častých jevů objevuje v datech velké množství záznamů s četností nula, tzn. bez výskytu sledovaného jevu. Pro tyto případy jsou vhodné tzv. *zero-inflated modely* (modely s nadbytečnými nulami), které rozdělí postup modelování na dvě nezávislé části, jednu pro jednotky s nulovým výskytem a druhou pro jednotky s výskytem větším než nula. Jde tak o zobecněné lineární smíšené modely. Negativní binomický model s nadbytečnými nulami (NBZI) může být parametrizován jako:

$$f(y_j|B_i, G_i, \beta, \gamma, \theta) = \begin{cases} p_i + (1 - p_i)q(0|\mu_i, \theta) & pokud\ y_i = 0 \\ (1 - p_i)q(y_i|\mu_i, \theta) & pokud\ y_i > 0 \end{cases} \quad [6.13]$$

kde $q(0|\mu_i, \theta)$ je dáno jako

$$q(y|\mu, \theta) = \frac{\Gamma(\theta + y)}{\Gamma(a)\Gamma(y + 1)}\left(\frac{\theta}{\theta + \mu}\right)^{\theta}\left(\frac{\mu}{\theta + \mu}\right)^{y}, \qquad y = 0, 1, 2, \ldots \quad [6.13]$$

kde μ a θ jsou střední hodnota a parametr velikosti pro smíšené rozdělení pravděpodobnosti - negativní binomické rozdělení s nadbytečnými nulami. $B_i\ a\ G_i$ jsou nezávislé proměnné sloužící k odhadu nulových, resp. nenulových četností a β, γ, θ jsou získány maximalizací funkce věrohodnosti $L(\beta, \gamma, \theta|y, B, G) = \sum_{i=1}^{n} \log f\ (y_i|B_i, G_i, \beta, \gamma, \theta)$ s ohledem na β, γ a θ (Minami et al., 2007).

Prostorová modifikace tohoto modelu založená na GWR je značně náročná a ještě obtížnější je případná interpretace jeho koeficientů a výsledků. Důvodem je fakt, že jde v podstatě o smíšení dvou modelů a v kombinaci s lokálními průběhy NBZI by k datové sadě o *n* prostorových jednotkách bylo výstupem *2n* NIZB modelů. Modifikace tohoto modelu pro použití s prostorovými daty však může proběhnout nepřímo díky úpravě vstupních nezávislých proměnných. Nezávislé proměnné jsou transformovány pomocí prostorové matice vah v závislosti na zvoleném typu sousedství. Konkrétní hodnoty jsou pak váženými průměry sousedních objektů. Pokud jsou transformovaná data použita k modelování, pak je možné modely označit za prostorové autoregresní modely (Anselin, 1988).

Diskriminační analýza GW DA

K určení, zda zvolené charakteristiky mohou vhodně identifikovat skupinu obce dle relativního rizika onemocnění kampylobakteriózou, jsou kromě zobecněných regresních modelů využity také vícerozměrné statistické metody nazvané diskriminační analýza. Diskriminační analýza je postup, který hledá vhodné pravidlo (diskriminační/rozhodovací funkci) umožňující na základě zadaných charakteristik objektu zařadit objekt do některé z dříve definovaných skupin tak, aby byla minimalizována pravděpodobnost chybných rozhodnutí (Tvrdík, 2003). Na základě diskriminačního pravidla jsou pro každý objekt zjištěny aposteriorní pravděpodobnosti ke každé skupině. Objekt je následně přiřazen do třídy s maximální aposteriorní pravděpodobností. Detailní popis metody je možné nalézt v Hebák et al. (2004) nebo Meloun (2011). V případě volby lineární diskriminační analýzy (LDA) je diskriminační funkce lineární kombinací prediktorů, podobně jako je tomu u lineární regrese.

Stejně jako ostatní geograficky vážené statistické analýzy a modely i geograficky vážená diskriminační analýza (GW DA) je vhodným nástrojem v případě, že je předpokládána prostorová proměnlivost charakteristik či vztahu mezi závislou a nezávislými proměnnými. GW DA je kromě toho také vhodnou alternativou k (prostorové) logistické regresi a multinomiální logistické regresi. Postup pro GW DA je obdobný GW PCA, a jde tedy o sérii lokálních diskriminačních analýz, které zohledňují prostorové vztahy mezi územními jednotkami, které jsou definované na základě prostorových vah a typu jádrového odhadu. Optimální volba sousedství je možná například pomocí krosvalidace. Podrobný popis metody poskytují Brunsdon et al. (2007).

Tvorba a hodnocení regresních metod pro analýzu kampylobakteriózy

Pro prozkoumání vztahů mezi výskytem kampylobakteriózy a možnými vysvětlujícími charakteristikami pomocí regresních modelů byly zvoleny dva různé postupy. První z těchto postupů se nesnaží analyzovat konkrétní počty případů kampylobakteriózy, ale pokouší se vysvětlit vztah mezi třídami obcí určenými relativním rizikem a vybranými prediktory či z nich vycházejícími hlavními komponentami. K tomu jsou použity zejména generalizované regresní modely – logistický, ordinální či jejich prostorová modifikace. Druhý postup je pak zaměřen právě na vysvětlení konkrétních průměrných ročních četností onemocnění v obcích. Kvůli vysokému poměru mezi rozptylem a průměrnou hodnotou je využita negativní binomická regrese, negativní binomická regrese s nadbytečnými nulami a také negativní binomická regrese s nadbytečnými nulami využívající prostorově transformované vysvětlující proměnné. Veškeré výpočty i hodnocení proběhly v **R** s využitím příslušných balíků:

- Multinomická regrese – *glm* (R Core Team, 2014), *glmnet* (Friedman et al., 2010), nnet (Venables a Ripley, 2002);
- Ordinální regrese a negativní binomická regrese – *MASS* (Venables a Ripley, 2002);
- Negativní binomická regrese s nadbytečnými nulami – *pscl* (Zeileis et al., 2008);
- ROC křivky – balík *ROSE* (Lunardon et al., 2014).

Prvním modelem k popisu vztahu mezi skupinami obcí a prediktory je multinomická regrese. Jde o zobecněný lineární model pro vícekategoriální data. V případě tohoto modelu je nutné upustit od podmínky možného seřazení skupin na základě relativního rizika, tzn. skupiny jsou brány jako nominální. Jako predikovaná proměnná sloužily skupiny obcí dle relativního rizika, jako prediktory buď reprezentativní proměnné, nebo pět hlavních komponent. Stejným postupem byl připraven i ordinální model, který zohledňuje seřazení tříd (a díky tomu jsou v něm i prosté členy reprezentující hodnoty pro přechody mezi třídami). V případě lokálního ordinálního modelu byla postupně počítána série ordinálních modelů pro každou obec zvlášť na základě obcí definovaných jako sousedních. Konkrétně bylo zvoleno sousedství o velikosti pásma 50 km, které bylo vyhodnoceno jako optimální pro geograficky vážené diskriminační analýzy. Výsledkem modelů jsou kromě jeho schopnosti predikce i koeficienty náležící vysvětlujícím charakteristikám (prediktorům) a skupinám. Ty ovšem nejsou tak přímočaré jako v případě lineární regrese, ale jedná se o logaritmovaný poměr šancí. Ten lze interpretovat jako hodnotu, o kterou se změní logaritmovaný poměr pravděpodobností mezi třídami v případě, že se hodnota prediktoru zvýší o jednu jednotku při zachování hladiny ostatních prediktorů. Po odlogaritmování poměru šancí je získáno relativní riziko, které již představuje velikost změny poměru pravděpodobností pro jednotlivé třídy při změnách prediktorů. Souhrn koeficientů (logaritmů poměru šancí) i jejich odlogaritmované formy (relativního rizika) pro sestrojené modely jsou shrnuty v Tabulce 12. Modely jsou hodnoceny na základě residuální sumy čtverců (Residual deviance), AIC (Akaikeho informační kritérium) a McFaddenova pseudo-R^2, které kombinuje informaci o vysvětlené variabilitě a informaci o zlepšení modelu vůči modelu s prostým členem.

Tab. 12 Koeficienty a vyhodnocení multinomického a ordinálních modelů pro původní datovou sadu i hlavní komponenty

Prediktor	Lineární kombinace reprezentativních charakteristik									
	Multinomický						Ordinální		*Lokální ordinální	
	Logaritmus poměru šancí			Relativní riziko			Logaritmus poměru šancí	Relativní riziko	Logaritmus poměru šancí	Relativní riziko
	MIN	MID	MAX	MIN	MID	MAX				
Prostý člen	0,717	0,296	-0,389	2,048	1,344	0,678	–	–	–	–
0|MAX	–	–	–	–	–	–	-1,642	0,194	-2,894	0,055
MAX|MID	–	–	–	–	–	–	-1,087	0,337	-1,477	0,228
MID|MIN	–	–	–	–	–	–	-0,068	0,934	1,072	2,920
Obyvatelstvo do 15 let	1,167	0,387	-0,424	3,212	1,473	0,655	0,078	1,081	-0,001	0,999
Hustota zalidnění	39,505	41,534	40,939	$1{,}09*10^{18}$	$6{,}02*10^{17}$	$1{,}43*10^{19}$	0,001	1,001	0,001	1,001
Nízké vzdělání	-0,184	0,139	-3,461	0,832	1,149	0,031	-0,011	0,989	0,156	1,168
Změna populace	-1,965	-4,000	-5,057	0,140	0,018	0,006	0,001	1,001	-0,001	0,999
Socioekonomická deprivace	-1,950	-1,300	-1,011	0,142	0,272	0,364	-0,835	0,434	-0,068	0,934
Zemědělské firmy	-2,170	-2,221	-0,729	0,114	0,108	0,482	-0,051	0,950	-0,100	0,905
Drůbež	0,013	-0,558	-0,077	1,013	0,572	0,926	0,000	1,000	0,000	1,000
Zpracování masa	-0,079	-0,219	-1,271	0,924	0,803	0,280	0,000	1,000	-0,085	0,919
Průměrná teplota	-0,927	0,322	0,475	0,396	1,380	1,609	-0,067	0,935	-0,109	0,897
Záplavové území Q_{20}	1,072	1,167	0,684	2,921	3,213	1,982	0,032	1,033	0,043	1,043
Mlékomaty	1,629	1,807	0,265	5,100	6,093	1,303	0,159	1,173	-0,225	0,799
Residual deviance	15131,230 / 15140,49						15829,710 / 15833,120		1191,170	
AIC	15203,230 / 15188,490						15857,710 / 15853,120		1219,337	
**pseudo-R^2	7,467 / 7,410						3,195 / 3,174			
	Lineární kombinace hlavních komponent									
	Logaritmus poměru šancí			Relativní riziko			Logaritmus poměru šancí	Relativní riziko	Logaritmus poměru šancí	Relativní riziko
	MIN	MID	MAX	MIN	MID	MAX				
Prostý člen	0,226	-0,197	-0,871	1,254	0,821	0,419	–	–	–	–
0|MIN	–	–	–	–	–	–	-0,883	0,414	-1,032	0,356
MIN|MID	–	–	–	–	–	–	0,689	1,992	-0,410	0,664
MID|MAX	–	–	–	–	–	–	2,140	8,498	0,714	2,041
PC1	1,993	2,161	1,274	7,338	8,676	3,574	1,281	3,601	2,188	8,914
PC2	-0,405	1,187	1,577	0,667	3,278	4,841	0,961	2,615	0,658	1,931
PC3	-3,766	-4,193	-2,283	0,023	0,015	0,102	-2,689	0,068	-5,233	0,005
PC4	2,679	2,690	2,232	14,575	14,732	9,320	1,602	4,965	-0,960	0,383
PC5	-1,490	-1,370	-1,295	0,225	0,254	0,274	-1,225	0,294	-0,902	0,406
Residual deviance	15625,890						15906,920		1235,234	
AIC	15661,890						15922,920		1251,234	
**pseudo-R^2	4,441						2,723			

** Hodnoty jsou průměrnými hodnotami koeficientů všech lokálních ordinálních modelů; ** McFaddenovo pseudo-R^2; šedě označené prediktory mohou být vypuštěny na základě významnosti v modelu (p-value) a hodnocení pomocí postupné regrese; šedě označené charakteristiky připadají zjednodušeným modelům*

Pokud jsou modely hodnoceny pro své predikční schopnosti na základě McFaddenova pseudo-R^2, pak ani jeden z modelů není výrazně vhodný. Při srovnání AIC vychází jako nejvhodnější multinomický model s původními daty, který je ale současně modelem nejkomplikovanějším vzhledem k množství koeficientů pro každou třídu. Přímé srovnání s lokálním prostorovým ordinálním modelem není možné kvůli faktu, že jde o sérii ordinálních modelů s různými koeficienty i počty vstupních dat závislých na definované matici sousedství. Multinomický i ordinální model s původními daty byly zjednodušeny pomocí postupné regrese a bylo tak v hodnocení vypuštěno několik prediktorů bez výrazné ztráty popisné schopnosti modelu. Při hodnocení koeficientů multinomické regrese je pracováno s faktem, že i ordinální proměnná je zde brána jako nominální a teoreticky je možné měnit neomezeně stavy obcí ve skupinách. Nejvýraznější charakteristikou obce je v případě multinomického modelu hustota zalidnění, která v případě zvýšení hodnoty o jednotku (1 obyv./km^2) skokově zvyšuje logaritmus poměru pravděpodobností mezi jednotlivými skupinami a skupinou s nulovým výskytem. Významný je i podíl obyvatelstva do 15 let, kdy v případě nárůstu o 1 % se pravděpodobnost výskytu obce vůči skupině s nulovým výskytem zvyšuje trojnásobně pro skupinu MIN, jeden a půl násobně vůči MID. Zajímavé je, že se poměr pro skupinu MAX snižuje o 0,66. Průběh pro změnu populace, socioekonomické deprivace a podíl zemědělských firem odhaluje, že v případě nárůstu těchto charakteristik se snižuje logaritmus poměru pravděpodobností. Jako možný faktor přírůstku onemocnění se na základě hodnocených modelů jeví průměrná teplota vzduchu a záplavové území, které při jednotkovém nárůstu při zachování ostatních charakteristik zvyšují pravděpodobnost výskytu území v jiné skupině než té s nulovým výskytem onemocnění. Koeficienty ordinální regrese prozrazují, že většina z charakteristik nemá výrazný vliv na přesun obcí ze skupiny bez onemocnění do skupiny s některou z vyšších kategorií (zobecněně zvýšení relativního rizika onemocnění). Riziko se mírně zvyšuje pouze s nárůstem podílu obyvatelstva do 15 let a poměru záplavových zón v území. Při srovnání koeficientů s průměrnými hodnotami lokální ordinální regrese lze předpokládat výrazný lokální vliv socioekonomické deprivace – průměrná hodnota odlogaritmovaného poměru pravděpodobností se více než zdvojnásobila. Význam zbylých efektů zůstává bez výrazných změn.

Druhou skupinou modelů, které se odlišovaly predikovanou proměnou a částečně i prediktory, byly negativní binomické modely. Jako predikovaná proměnná sloužily průměrné četnosti výskytu kampylobakteriózy v jednotlivých obcích a jako prediktory byly zvoleny vybrané charakteristiky populace a prostředí v těchto obcích. Skupina negativních binomických modelů byla upřednostněna před poissonovskými modely z důvodu velkého rozdílu mezi rozptylem a střední hodnotou, což narušuje jednu jeho z významných podmínek. Nejdříve byl sestrojen negativní binomický model, který jako prediktory zahrnoval obyvatelstvo do 15 let, hustotu zalidnění, socioekonomickou deprivaci, průměrnou teplotu a přítomnost mlékomatů a masozpracovatelských společností. Do tohoto modelu byla zahrnuta i populace obcí jako tzv. offset, tedy pevně daný efekt v modelu. Vzhledem k velkému množství nul byl testován i negativní binomický model s nadbytečnými nulami, který je vlastně smíšeným modelem pro nuly a nenulové hodnoty. Druhý zmíněný model byl nakonec sestaven i pro tzv. *spatially lagged* prediktory, které vzniknou násobením původní matice prediktorů a prostorovými váhami definujícími prostorovou blízkost mezi obcemi (stanovenou na 50 km). Konkrétní hodnoty jsou pak váženými průměry sousedních objektů

a modely, které takto transformovaná data používají, je možné označit za prostorové autoregresní modely (Anselin, 1988). I v případě těchto modelů jsou výstupem koeficienty, které poukazují na velikost změny logaritmu počtů onemocnění v případě jednotkového nárůstu prediktoru a zachování efektu ostatních prediktorů. Souhrn koeficientů (logaritmu poměru šancí) i jejich odlogaritmované formy (relativního rizika) pro sestrojené modely je jsou v Tabulce 13.

Tab. 13 Koeficienty a vyhodnocení negativních binomických modelů

Prediktor		Negativní binomický		Negativní binomický s nadbytečnými nulami		Negativní binomický s nadbytečnými nulami a transformovanými prediktory	
		Logaritmus poměru šancí	Relativní riziko	Logaritmus poměru šancí	Relativní riziko	Logaritmus poměru šancí	Relativní riziko
Prostý člen		-3,255	0,039	1,780	5,927	-6,540	0,001
Obyvatelstvo do 15 let		0,021	1,021	-0,008	0,992	0,165	1,180
Hustota zalidnění		0,002	1,002	0,004	1,004	0,009	1,010
Socioekonomická deprivace		-1,026	0,358	-1,769	0,171	2,774	16,025
Zpracování masa		0,110	1,116	0,097	1,103	-0,103	0,902
Průměrná teplota		0,191	1,210	0,146	1,158	-0,034	0,967
Mlékomaty		0,120	1,127	0,223	1,255	-0,741	0,476
offset		log(populace)		–		–	
ZI*	Prostý člen	–	–	3,395	29,803	1,62	5,087
	populace	–	–	-0,010	0,990	-0,007	0,993
Residual deviance		5286,900		5004,44		21829,14	
AIC		15876,000		16418,250		21538,87	
McFaddenovo pseudo-R^2		13,061		26,576		3,660	

** Část náležící modelu pro nadbytečné nuly; šedě označené charakteristiky lze vyřadit na základě postupné regrese*

Skupina negativních binomických modelů, zejména pak negativní binomický model s nadbytečnými nulami, poskytuje na základě McFaddenova pseudo-R^2 o poznání lepší predikční schopnost. V případě této skupiny modelů poukazují koeficienty na změnu v logaritmu četností onemocnění v obci a nevztahují se tak k hodnocení skupin jako u multinomické a ordinální regrese. U negativního binomického modelu s nadbytečnými nulami nebyla ve výpočtu využita proměnná obyvatelstvo do 15 let, která byla pomocí stepwise regrese vyhodnocena jako nevýznamná. Jediným prediktorem, který podle modelu mírně snižuje počet případů kampylobakteriózy v obci, je socioekonomická deprivace. Ostatní proměnné přispívají spíše k nárůstu absolutní četnosti. I tento nárůst je však pouze mírný – v průměru o jeden případ na zvýšení hodnoty prediktoru o jednotku – např. přibude-li v obci jeden mlékomat očekává se o 1,255 případů v obci více. Zajímavé je srovnání mezi prvními dvěma negativními binomickými modely a negativním binomickým modelem s nadbytečnými nulami a prostorově transformovanými prediktory. Koeficienty prvních dvou modelů jsou si velmi podobné až na prostý člen, zatímco třetí model se výrazně odlišuje. Hodnocení třetího modelu jako nástroje pro exploraci dat poukazuje na možné lokální diference v prostorové distribuci onemocnění. Nejvýrazněji je tento efekt viditelný u socioekonomické deprivace, která je hodnocena jako velmi výrazný faktor – zvýšení deprivace obyvatelstva o jednotku znamená 16 případů onemocnění navíc. Změna je viditelná

i u relativní přítomnosti mlékomatů a ekonomických subjektů zpracovávajících maso, kdy oba prediktory snižují frekvenci onemocnění v obci, zatímco nárůst hustoty zalidnění a obyvatelstva do 15 let může znamená i nárůst četnosti případů onemocnění.

Hodnocení přesnosti klasifikačních postupů pro analýzu kampylobakteriózy

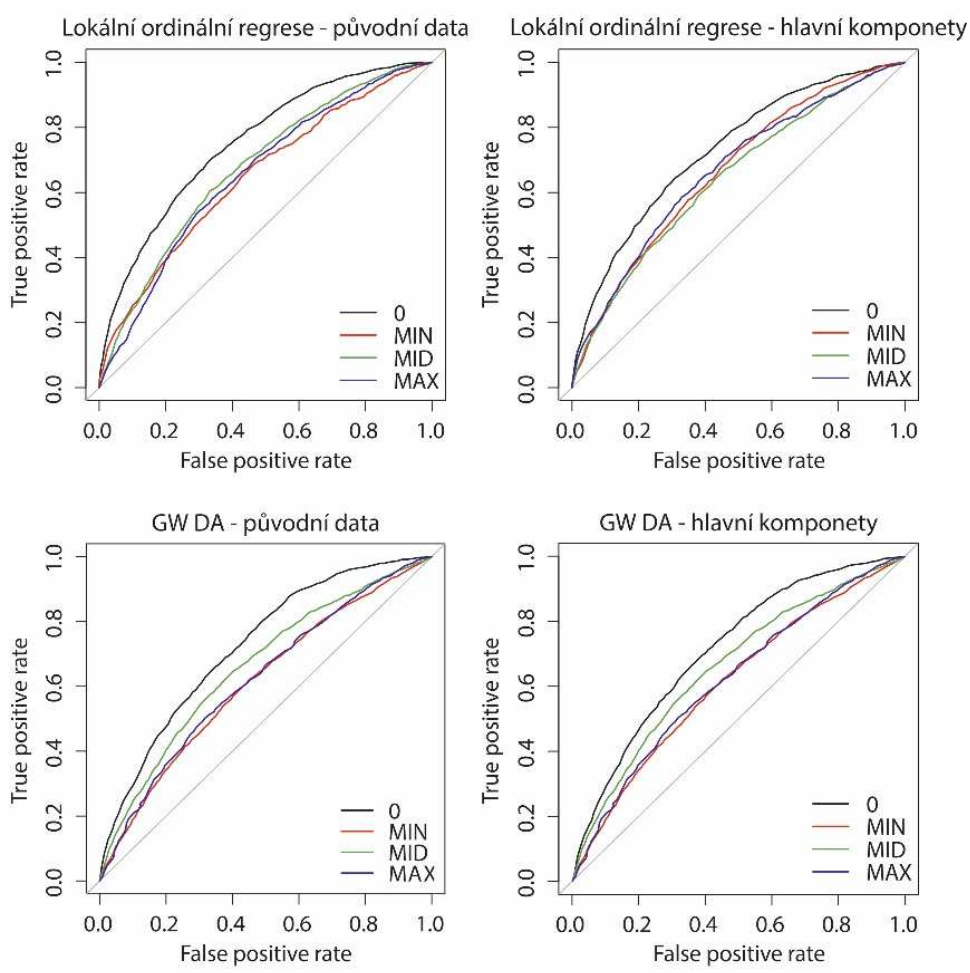
Ke klasifikaci pomocí regresních postupů a diskriminační analýzy bylo přistoupeno dvěma způsoby. Pro klasifikaci bylo jako prediktorů skupin využito jednak jedenácti reprezentativních charakteristik obcí, a pak také pěti hlavních komponent vzniklých redukcí dimenze v PCA. Do klasifikace vstupovaly charakteristiky i hlavní komponenty v lineární formě bez využití dalších interakcí. Pro výpočet jednotlivých klasifikací byla původní datová sada rozdělena na trénovací data (80 % obcí) a data testovací (20 % obcí) sloužící k validaci vytvořeného algoritmu. Výběr testovací a trénovací množiny proběhl s ohledem na klasifikované třídy tak, aby byly obě sady dat co nejvíce vyvážené. Zastoupení jednotlivých tříd v trénovací i validační sadě odpovídalo poměru 29,7/35,2/23,5/11,6.

Hodnocení testovaných metod proběhlo srovnáním výsledků predikce nad testovacími daty a skupinami z původní datové sady odpovídajícím obcím z trénovacích dat. Výkon a kvalita klasifikačních schopností byly hodnoceny pomocí přesnosti klasifikace jednotlivých tříd, průměrné a průměrné vážené přesnosti klasifikace a také pomocí AUC v ROC křivkách. Vzhledem k tomu, že ROC vyhodnocuje binární klasifikátory, tak byla hodnocena vhodnost správnost zařazení do jedné skupiny vůči všem ostatním skupinám.

Postup tvorby pro jednotlivé metody regrese byl popsán v předcházející kapitole, proto bude blíže popsaána diskriminační analýza (DA) a její geograficky vážená modifikace (GW DA). Kroky pro obě metody jsou velmi podobné, hlavním rozdílem je nutnost definování prostorových vah v případě GW DA. Celý postup byl opět proveden v prostředí **R** s využitím funkcí z balíku *MASS* (Venables a Ripley, 2002) a *GWmodel* (Gollini et al., 2015). Definování DA je v **R** velmi podobné zadávání regrese, je nutné zadat pouze vztah mezi popisovanou proměnnou a prediktory v lineární (či kvadratické formě), trénovací množinu, případně další požadované vlastnosti. V případě dat onemocnění kampylobakteriózou byly popisovanou vlastností jednotlivé třídy relativního rizika a prediktory hlavní komponenty nebo charakteristiky obce. Pro diskriminační analýzu byla zvolena lineární diskriminace. V případě GW DA je před definováním vztahu diskriminace potřeba definovat prostorové váhy a z nich šířku pásma a/nebo počet sousedů, kteří budou využiti pro výpočet lokálních diskriminací. Na základě krosvalidace bylo zvoleno pásmo o šířce 50 km, které poskytovalo vhodný kompromis mezi lokálním provedením analýzy a její vypovídající hodnotou. Pro výpočet GW DA byly rovněž využity lokální varianty průměru a kovarianční matice, dále byl upřesněn typ jádrové funkce jako bisquare.

Výsledky klasifikace pomocí regresních metod a diskriminační analýzy jsou přehledně zobrazeny v Tabulce 14. Z Tabulky 14 je zřejmé, že prostorové metody podávají lepší výsledky než metody neprostorové, což je zřejmé zejména u modelů využívajících hlavní komponenty. Celkově ovšem nedokáže ani jeden z klasifikačních modelů správně klasifikovat ani polovinu obcí. Tento stav plně souhlasí s některými tvrzeními ze zahraničních studií, které uvádí, že polovina případů kampylobakteriózy stále zůstává nevysvětlena (Nylen et al., 2002; Ekdahl et al., 2005). Z pohledu celkového množství vysvětlených případů je nejúspěšnější v práci navržený koncept lokální ordinální regrese s úspěšností 47,2 %, resp. 45,6 %. Tato metoda také

vykazuje nejlepší výsledky z pohledu AUC (68,4 % a 67,6 %). Je potřeba zmínit, že využití lokální ordinální regrese je podle výsledků vhodnější s daty redukovanými do hlavních komponent. Druhou nejvhodnější metodou klasifikace vykazující nejlepší vážený průměr správně zařazených obcí je GW DA. Jako nejméně vhodná se ukázala neprostorová ordinální regrese, ačkoliv jde o metodu, která je pro podobné úlohy přímo vytvořena. Obrázek 30 vykresluje ROC křivky pro obě prostorové metody, tedy lokální ordinální regresi a GW DA, v obou případech (a stejně v podstatě u všech hodnocených modelů) jsou výrazně nejlépe klasifikovány obce bez výskytu kampylobakteriózy.



Obr. 30 ROC křivky pro lokální ordinální regresi a GW DA

Tab. 14 Vyhodnocení klasifikačního výkonu regresních modelů a diskriminační analýzy, hodnocená jako celek i pro jednotlivé skupiny; Acc – přesnost hodnocení (%), AUC – plocha pod křivkou (%); tučně jsou označeny nejvyšší hodnoty Acc a AUC pro jednotlivé třídy i celkově

	Multinomická regrese		Ordinální regrese		Lokální ordinální regrese		Diskriminační analýza		Prostorově vážená diskriminační analýza	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
	Lineární kombinace reprezentativních charakteristik									
0	**67,9**	75,0	53,9	68,6	59,7	**75,3**	47,5	70,6	49,2	72,4
MIN	65,7	63,3	**74,5**	60,0	1,8	65,0	73,6	60,8	44,5	**65,9**
MID	10,6	64,9	0,0	55,0	**74,4**	**67,7**	13,4	63,3	42,3	66,1
MAX	0,0	57,7	0,0	51,2	0,3	**65,7**	0,0	57,1	**29,7**	63,4
Celková	45,7	65,2	42,2	58,7	**47,2**	**68,4**	43,2	63,0	45,3	67,0
Vážená	36,1		32,1		34,1		33,7		**41,5**	
	Lineární kombinace hlavních komponent									
0	45,9	68,5	36,3	68,3	**54,1**	**72,9**	45,4	68,5	53,3	71,4
MIN	74,6	60,8	**82,7**	53,3	64,3	**66,4**	75,3	60,6	60,1	61,2
MID	8,2	62,3	4,5	62,1	26,5	64,3	10,8	62,0	**32,4**	**66,0**
MAX	0,0	58,1	0,0	53,7	5,5	**66,6**	0,0	57,6	**6,4**	62,1
Celková	41,8	62,4	40,9	59,4	**45,6**	**67,6**	42,5	62,2	45,3	65,2
Vážená	32,2		30,9		37,6		32,9		**38,0**	

0 – obce bez zaznamenaného výskytu onemocnění; MIN – SIR ≤ 0,80; MID - 0,80 < SIR < 1,50; MAX - SIR ≥ 1,50; Celková – počet správně klasifikovaných obcí / průměrné AUC; Vážená – vážená průměrná přesnost určení zohledňující četnosti obcí ve skupinách

6.3.3 Možnosti strojového učení a data miningu ke klasifikaci územních jednotek podle relativního rizika onemocnění kampylobakteriózou

Kromě tradičních metod statistického modelování, jeho modifikací a vícerozměrné analýzy je možné pro klasifikaci obcí podle relativního rizika onemocnění kampylobakteriózou a vyhodnocení vhodnosti výběru charakteristik využít i komplexních (a mnohdy i komplikovanějších) metod strojového učení a data miningu. Tyto metody vycházejí ze základních statistických metod a klasifikačních úloh, které však zdokonalují pomocí jejich schopnosti adaptace k nastalým podmínkám. Kromě klasifikace jsou v praxi využívány k predikci nebo shlukování (kategorizaci) zadaných objektů datového souboru. Právě díky tomu jsou metody využity i v této části disertační práce. Vzhledem k množství metod, šíři tématu a možnostem jejich využití k průzkumu a analýze prostorových dat, jsou v následující části disertační práce jednotlivé metody pouze krátce představeny. Komplexní popis jednotlivých metod je možné nalézt v Wu et al. (2007), Williams (2011), Zhao (2013) nebo Larose a Larose (2014). Využití metod v analýze prostorových dat představují Ester et al. (1997), Koperski et al. (1996) či Miller a Han (2009).

Jednoduché klasifikátory

Cílem ***naivního bayesovského klasifikátoru*** (naive Bayes) je nejdříve sestrojit pravidlo, na základě kterého je daný objekt přiřazen do třídy na základě sady jeho vlastností. V případě, že je později dodána pouze sada vlastností, pak je podle ní možné objekt klasifikovat objekt

do vhodné třídy. Jde tedy o problém řízené klasifikace. Naivní bayesovský klasifikátor je významný z několika důvodů – (1) je jednoduchý na sestrojení, díky čemuž je možné ho rychle nasadit i na obsáhlé datové sady; (2) jeho výsledky jsou snadno interpretovatelné i pro laika; (3) poskytuje spolehlivé odhady (Wu et al., 2007). Naivní bayesovský klasifikátor využívá nejjednodušší možné implementace Bayesova teorému k výpočtu aposteriorních pravděpodobností příslušnosti objektu k třídě na základě nezávislých proměnných. Pro dvě třídy je definován klasifikátor jako

$$P(y_i|X) = \frac{P(X|x_i)P(y_i)}{P(X)} \qquad [6.14]$$

kde y_i představuje třídy, do kterých jsou objekty X klasifikovány.

Druhým jednoduchým klasifikátorem je ***klasifikátor nejbližšího souseda*** (kNN). Jak vyplývá již z názvu, kNN využívá ke klasifikaci objektu k jemu nejbližších objektů. Klasifikátor tedy vybere trénovací množinu (k nejbližších sousedů) a přiřazuje klasifikovanému objektu hodnotu, která v jeho nejbližším okolí převládá (Zafarani et al., 2014). S ohledem na tento klasifikátor vyvstávají zejména dva problémy – 1) volba metriky, pomocí které bude vyhodnocována vzdálenost k okolním bodům a 2) volba k, tzn. počtu sousedů tak aby byl na jednu stranu vzorek dost široký a na druhou ne příliš variabilní.

Neuronové sítě

Umělé neuronové sítě se pro vyřešení problému či analýzu dat napodobují strukturu a funkci lidského mozku. Toho se snaží docílit pomocí simulace základních částí a procesů mozku – buněk neuronů, synaptických spojení, dendritů, vzruchy, aktivitou neuronů apod. (Casas, 2009). Základní stavební jednotkou neuronové sítě jsou neurony svázané do sítě, pomocí které si předávají signály, přičemž jednotlivé neurony tento signál transformují. Neuronové sítě jsou schopny učení, které probíhá adaptací jednotlivých propojení mezi neurony (Dvorský a Dražilová, 2011).

Jedním z nejčastěji používaných přístupů pro učení pomocí zpětné propagace chyb (back propagation) u neuronových sítí je ***Multi Layer Perceptron*** (MLP). Typické MLP se skládá z trojice vrstev – vstupní, skryté (nebo několika skrytých) a výstupní, ve kterých jsou umístěny úplně propojené neurony. Jedná se o síť s učitelem, tudíž je nutné mít trénovací množinu vstupů a výstupů. Trénování i klasifikace probíhají v neuronech jako blackbox model.

Support vector machine

Mezi současnými metodami strojového učení jsou za jeden z nejefektivnějších způsobů považovány algoritmy ***Support Vector Machine*** (SVM) a to zejména díky své robustní povaze a výsledné přesnosti klasifikace. Dalšími důvody častého využití jsou pevné teoretické základy metody, necitlivost k množství dimenzí (vlastností) a potřeba pouze malého množství trénovacích dat (Wu et al., 2007). SVM jsou sice především binárním klasifikátorem, ale je možné ho snadno transformovat od podoby vhodné i pro data s více než dvěma třídami. SVM umožňuje rozlišení dat a jejich klasifikaci pomocí konstrukce oddělující linie (2D), roviny (3D) či nadroviny (n-D), díky které je možné od sebe odlišit jednotlivé třídy. SVM jsou původně druhem lineárního klasifikátoru, který se však díky generalizaci a modifikacím v současnosti využívá i na problémy, kde není lineární klasifikace možná či vhodná. Takovou modifikací

může být například využití kernelů založených na *funkcích s radiální bází* (RBF), které mohou nahradit lineární klasifikaci a učinit tak SVM flexibilnějším (Smola et al., 1998). Popis aplikací SVM využitý k řešení environmentálních problémů poskytují Kanevski et al. (2009).

Regresní stromy a náhodné lesy

Klasifikační a regresní stromy jsou metody stromového učení a analytické nástroje, které mohou být využity k průzkumu vztahů mezi zdravím či nemocností a determinanty s nimi spojenými (Speybroeck, 2012). Modely pro klasifikaci a predikci vznikají pomocí rekurzivního dělení původních dat a jejich postupného prokládání jednoduchými modely. Stavba klasifikačního stromu začíná rodičovským uzlem (kořen stromu), který obsahuje všechny jedince (záznamy) z datové sady. Rodičovský uzel je dále rozdělen pomocí optimalizovaných podmínek, takže postupně vznikají homogennější potomci, kteří mohou být dále děleni (Loh, 2011). Právě tomuto postupu se říká rekurzivní dělení. Metodu je možné využít pro kvalitativní data (klasifikační stromy) i kvantitativní data (regresní stromy). Výhodou regresních stromů je kromě jejich univerzálního využití i možnost jejich přehledného zobrazení, ve kterém je možné názorně zobrazit vztahy. Mezi nejčastěji používané metody patří stromy typu CART (Breiman, 1984), které jsou implementovány v **R** balíku rpart (Therneau et al., 2014)

Klasifikační lesy jsou modely tvořené více klasifikačními stromy, které o zařazení pozorování do některé ze tříd rozhodují společně většinovým hlasováním. Tak je tomu v případě metody ***Random forests*** (náhodné lesy), jejíž jednotlivé stromy jsou randomizovanými obdobami CART stromů. Výhodou klasifikačních stromů je spíše jednoduchost než extrémní přesnost, ale nevýhodou je jejich neprůhlednost z pozice uživatele (Klaschka, 2011). Kromě klasifikačních lesů je možné zlepšit klasifikační schopnosti stromů také pomoci boostingu (zesilování) tak jak ho popsal Friedman (2001).

Hodnocení přesnosti klasifikačních postupů

Ke klasifikaci pomocí představených metod strojového učení a data miningu bylo přistoupeno stejně jako v případě regrese a diskriminační analýzy. Byla hodnocena schopnost klasifikace vycházející z prediktorů tvořených původními daty i hlavními komponentami. Stejně jako v předchozím případě byla datová sada rozdělena na vyváženou trénovací (80 % obcí) a testovací (20 % obcí) množinu sloužící pro validaci postupů. Samotné hodnocení testovaných metod proběhlo srovnáním výsledků predikce nad testovacími daty a skupinami z původní datové sady. Pomocí přesnosti klasifikace jednotlivých tříd, průměrné a průměrné přesnosti klasifikace a průměrné vážené přesnosti klasifikace (zohledňuje četnosti v jednotlivých třídách) a hodnot AUC v ROC křivkách byl hodnocen výkon a kvalita klasifikace. Veškeré výpočty i hodnocení proběhly v **R** s využitím příslušných balíků:

- nejbližší sousedé – balík *kknn* (Schliep a Hechenbichler, 2014),
- naivní bayesovský klasifikátor – balík *e1071* (Meyer et al., 2014),
- neuronová síť – balík *nnet* (Venables a Ripley, 2002),
- Multi Layer Perceptron – balík *monmlp* (Cannon, 2012),
- Support Vector Machine a Support Vector Machine s radiální bazickou funkcí – balík *kernlab* (Karatzoglou et al., 2004),

- Regresní stromy– balík *rpart* (Therneau et al., 2014),
- Random Forest (Náhodný les) – balík *randomForest* (Liaw a Wiener, 2002),
- ROC křivky – balík *ROSE* (Lunardon et al., 2014).

Výsledky klasifikačního výkonu jednotlivých metod pro původní i redukovaná data jsou shrnuty v Tabulce 15. Na základě tohoto zhodnocení nelze zcela jednoznačně identifikovat metodu, která by byla výrazně lepší než všechny ostatní. V případě klasifikačních metod vycházejících z původních charakteristik je tato diferenciace velmi patrná vzhledem k tomu, že ani jedna z metod neumožňuje ideálně klasifikovat více než jednu skupinu. Výjimkou je SVM s RBF, která při hodnocení na základě AUC nejlépe identifikuje dvě skupiny. Jako nejlepší metodu pro využití s lineární kombinací původních dat je možné na základě celkového výkonu vyhodnotit Random Forest, která vykazuje nejvyšší celkovou i váženou přesnost i nejvyšší AUC (Obrázek 31 - vlevo nahoře). I přesto ovšem nedokáže stejně vvhodně klasifikovat do všech skupin. Největší problém u všech klasifikačních metod (s původními i redukovanými daty) byla identifikace obcí s nejvyšším relativním rizikem. Z tohoto pohledu poskytuje nejlepší výkon neuronová síť (Obrázek 31 - vpravo nahoře), která současně hodnotí celou sadu nejvyrovnaněji.

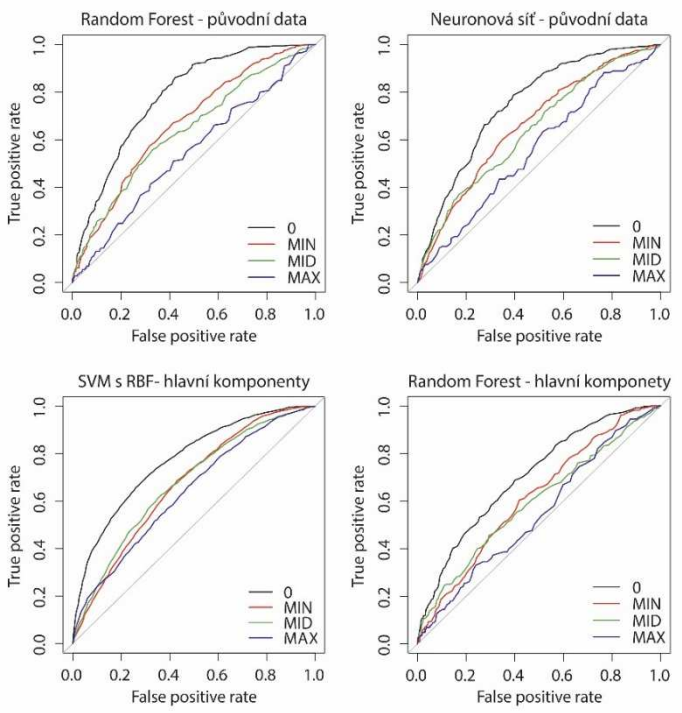
Tab. 15 Vyhodnocení klasifikačního výkonu metod strojového učení a data miningu, hodnocená jako celek i pro jednotlivé skupiny; Acc – přesnost hodnocení (%), AUC – plocha pod křivkou (%); tučně jsou označeny nejvyšší hodnoty Acc a AUC pro jednotlivé třídy i celkově

	Nejbližší sousedé		Naivní bayes		Neuronová síť		MLP		SVM		SVM s RBF		Regresní stromy		Random Forest	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
	Lineární kombinace reprezentativních charakteristik															
0	49,7	72,3	59,8	73,0	48,2	75,1	66,0	78,3	61,7	82,8	63,3	**83,1**	**71,7**	76,0	67,9	78,0
MIN	65,4	62,2	28,8	60,8	46,2	**66,3**	64,5	64,8	**78,8**	54,9	64,1	72,6	50,8	64,2	62,3	64,2
MID	22,9	61,6	**43,8**	62,2	47,7	64,4	14,0	62,0	1,0	**71,0**	18,1	70,6	22,3	65,1	24,0	65,1
MAX	0,0	58,1	4,1	55,8	**27,3**	56,0	1,3	56,9	0,0	51,0	2,8	**68,6**	0,0	56,8	0,7	56,8
Celková	43,2	63,6	38,6	63,0	47,0	**65,5**	44,6	**65,5**	46,2	64,9	45,9	73,7	44,4	**65,5**	**48,0**	**65,5**
Vážená	34,5		34,1		38,0		36,4		35,4		37,1		36,2		**38,9**	
	Lineární kombinace hlavních komponent															
0	46,7	69,9	38,6	69,1	53,0	74,5	59,3	72,0	50,2	74,4	48,1	**77,2**	43,5	63,7	**54,1**	63,7
MIN	62,3	60,4	78,0	60,2	66,4	65,1	65,1	60,6	77,1	63,8	71,9	**67,0**	**83,5**	57,8	54,5	57,8
MID	21,2	60,7	12,3	62,5	17,5	66,5	13,6	62,9	4,1	65,3	16,1	**67,4**	0,0	57,0	**26,4**	57,0
MAX	0,0	56,1	0,7	57,3	0,7	62,1	0,0	61,4	0,0	59,4	**1,4**	**63,8**	0,0	52,1	**1,4**	52,1
Celková	41,0	61,8	41,9	62,3	43,2	67,1	42,8	64,2	43,0	65,7	**43,5**	**68,9**	42,3	57,7	41,5	57,7
Vážená	32,7		32,4		34,4		34,5		32,9		**39,5**		31,8		34,1	

0 – obce bez zaznamenaného výskytu onemocnění; MIN – SIR ≤ 0,80; MID - 0,80 < SIR < 1,50; MAX - SIR ≥ 1,50; Celková – počet správně klasifikovaných obcí / průměrné AUC; Vážená – vážená průměrná přesnost určení zohledňující četnosti obcí ve skupinách; MLP – Multi Layer Perceptron; SVM – Support Vector Machine; SVM s RBF - Support Vector Machine s radiální bazickou funkcí

Při pohledu na stejné klasifikátory využívající místo původních dat pět hlavních komponent (Obrázek 31 – dole), je situace na hodnocení jednodušší. Z pohledu přesnosti hodnocení je nejúspěšnějším klasifikátorem Random Forest, která nejlépe identifikuje hned tři třídy (Obrázek 31 - vpravo dole). Z pohledu hodnocení pomocí AUC je nejúspěšnějším klasifikátorem Support Vector Machine s radiální bazickou funkcí (Obrázek 31 - vlevo dole),

který podává nejlepší výkon ve všech skupinách i celkově. Celková přesnost žádné z metod však nedokázala správně klasifikovat více než polovinu obcí – nejvíce to bylo 48,0 % v případě Random Forest nad původními daty, pro váženou přesnost je tato hodnota nejvyšší u SVM s RBF nad hlavními komponentami (39,5 %). ROC křivky nejúspěšnějších klasifikátorů pro původní data i PCA data jsou umístěny na Obrázku 31.



Obr. 31 ROC křivky nejúspěšnějších klasifikátorů

6.4 Shlukování: Vzory a podobnosti v atributovém prostoru

V kapitole 5 byly obce a jejich části seskupovány (shlukovány) na základě měr morbidity, jejich polohy a/nebo času. V případě, že je cílem najít skupiny obcí na základě podobnosti jejich environmentálních a socioekonomických charakteristik a jejich vztahu k mírám morbidity, tak je vhodným způsobem využití vícerozměrné statistické analýzy zvané shlukování. Jako shlukování je označována skupina metod, jejímž cílem je nalézt v neuspořádaných datech podmnožiny podobných objektů. Jde o metody klasifikace, která vede k vytvoření systému tříd. Každá třída by měla být na základě podobných vlastností definována tak, aby objekty v rámci třídy byly co nejpodobnější a současně, aby se od sebe jednotlivé třídy co nejvíce odlišovaly (Hebák et al., 2005b).

Metody shlukování jsou nejčastěji rozdělovány na hierarchické (aglomerační nebo divizní) a nehierarchické (optimalizační, k-means, PAM) (Horák, 2011).

Při shlukové analýze nastávají dvě základní situace, podle kterých je volena nejvhodnější shluková metoda. První možností je předem známý počet tříd, do kterých má být množina objektů zařazena (např. nehierarchická metoda k-means). Druhou možností je na začátku neznámý počet tříd, který je optimálně stanovován postupně v průběhu procesu nebo podle vhodnosti až na konci analýzy (hierarchické metody). Samotné určování podobnosti a odlišnosti uvnitř shluků bývá uskutečněno na základě pravidel definovaných jako míra podobnosti v atributovém prostoru. Hierarchické shlukování, které bylo použitov této části

disertační práce, nepředpokládá předem definovaný počet shluků, ale vytváří hierarchii možných řešení shlukování na základě matice vzdáleností (podobnosti) mezi objekty (Arentze, 2009). Vytvořenou shlukovou strukturu je možné vizualizovat pomocí dendrogramu, díky kterému je určován optimální počet skupin pro zkoumaná data.

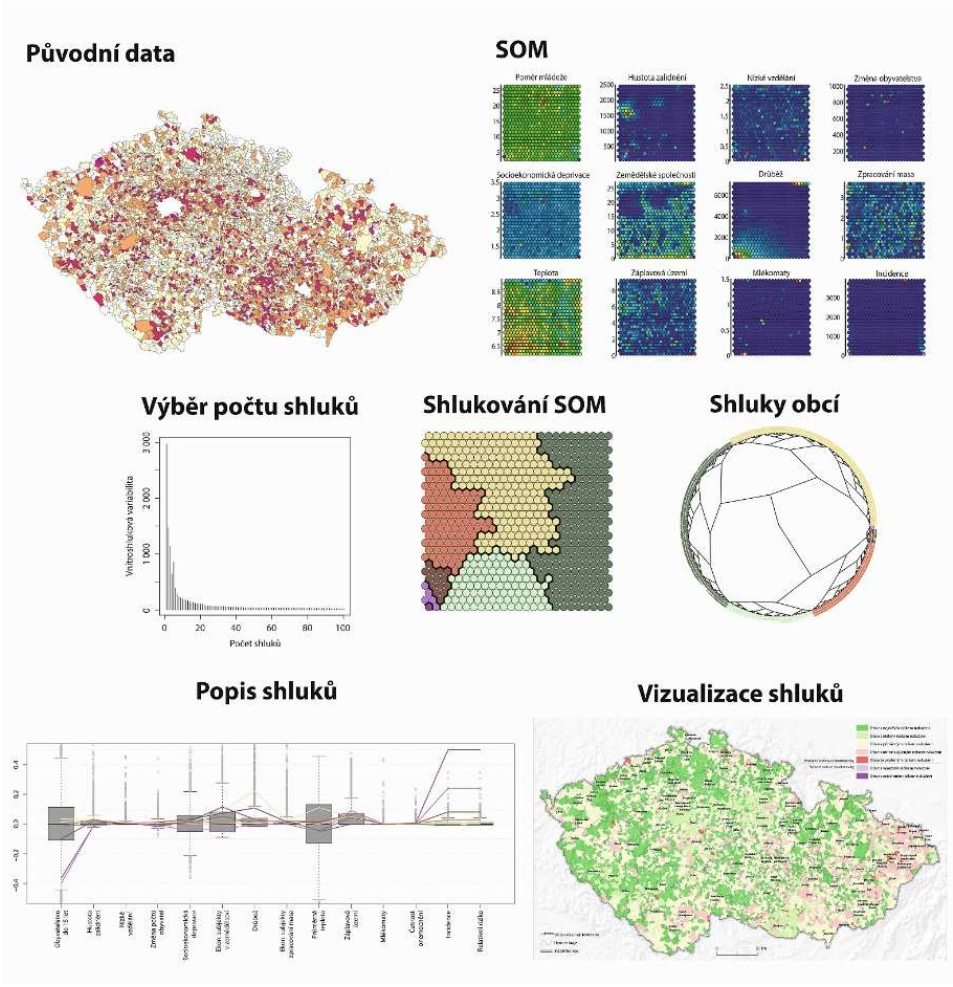
Detailní informace o shlukování prezentují Hebák et al. (2005a), Everitt a Hothorn (2011), Meloun a Militký (2004) nebo Timm (2002). O možnostech využití shlukování ke klasifikaci prostorových dat se zachováním prostorové spojitosti informují Carvalho et al. (2009), Miller a Han (2009) nebo Horák et al. (2012).

Proces shlukování v této kapitole byl rozšířen o data, která byla připravena pomocí samoorganizačních map dříve, než proběhlo samotné hodnocení jejich podobnosti. Samoorganizační mapy (SOM) nebo také Kohonenovy mapy jsou jedním z typů neuronových sítí často používaných pro analýzu dat, klasifikaci dat nebo tvorbu sémantických map (Kohonen, 2001). SOM jsou založeny na projekci vstupních dat z vícerozměrného atributového prostoru do prostoru s nižší dimenzí – typicky do dvourozměrného prostoru, kde může dojít k zobrazení do mřížky či hexagonů. Podstatou vlastností SOM je, že projektovaná data částečně zachovávají (atributovou) topologii, takže sobě podobné body v původním prostoru se budou blízko sebe nacházet i po projekci (Dvorský a Dražilová, 2011). Využitím SOM v geografických analýzách se zabývali Kauko a Goetgeluk (2005) nebo Spielman a Thill (2008).

6.4.1 Shlukování obcí České republiky podle vybraných charakteristik ve vztahu k morbiditě

Cílem této kapitoly bylo klasifikovat obce České republiky podle reprezentativních charakteristik vybraných v minulých kapitolách a doplněných o charakteristiky nemocnosti obce, konkrétně o absolutní četnost, standardizovanou incidenci a relativní riziko (SIR). Celý proces od přípravy datové sady, přes shlukování až po vizualizaci dat (mimo mapu) proběhl v prostředí **R.** Nejdůležitějšími balíky v této části práce byly v *kohonen* (Wehrens a Buydens, 2007), *clusterSim* (Walesiak a Dudek, 2007).

Data pro shlukovou analýzu obsahují 11 dříve představených charakteristik obcí a tři charakteristiky morbidity kampylobakteriózy. Připravené charakteristiky se lišily rozsahem hodnot i jejich jednotkami. Pro srovnání dat a provedení shlukové analýzy, která je citlivá na měřítko dat (Meloun a Militký, 2004) byla data standardizována do intervalu $\langle -1; 1\rangle$ pomocí vztahu $\frac{(x-\bar{x})}{max(|x-\bar{x}|)}$. Takto upravená data vstupovala do vyhodnocení podobnosti obcí. Zde však byl vložen jeden mezikrok, a to klasifikace dat a zjištění souvislostí v nich pomocí SOM. SOM je sama o sobě metodou vhodnou pro roztřízení dat do skupin, jeho nevýhodou ovšem je, že data v klasifikovaných buňkách neuronové sítě jsou agregovaná a klasifikace je provedena kontinuálním způsobem, kdy není znám výsledný počet tříd. Standardizovaná data o 14 rozměrech (atributech) byla projektována do hexagonové sítě o 625 buňkách (průměrný počet obcí v buňce odpovídá 10). S využitím 2 000 iterací a kruhového sousedství byla neuronová síť SOM trénována a následně vyhodnocena. Výsledky SOM je již možné zobrazit v dvourozměrném prostoru. Každý výsledný uzel nese informaci o své pozici, váze spojené se vstupními proměnnými a informaci o objektech, které jsou v uzlu agregovány. Typické zobrazení SOM je pomocí tzv. *heatmaps*, které zobrazují charakteristické váhy uzlů (Obrázek 32 – nahoře vpravo).



Obr. 32 Analytický postup použitý během shlukování obcí

V dalším kroku již došlo k samotnému shlukování nad daty projektovanými pomocí SOM. Cílem shlukování SOM je identifikovat izolované skupiny uzlů s podobnými metrikami, tzn. podobnými kombinacemi vah jednotlivých charakteristik v uzlech. Výsledky ze SOM sloužily jako vstupní hodnoty pro shlukování a byla pro ně zjištěna míra podobnosti mezi všemi objekty (uzly) pomocí vzdálenosti typu *maximum*. Pro shlukování byla zvolena Wardova metoda, jejímž cílem je minimalizace heterogenity jednotlivých shluků. Vyhodnocení vhodného počtu skupin proběhlo na základě empirického odhadu z dendrogramu, simulace shlukování, které umožňuje balík *clusterSim* a sutinového grafu (scree plot – Obrázek 32 – uprostřed vlevo) srovnávajícího počty shluků s jejich vnitroshlukovou variabilitou. Zvoleno bylo rozdělení datové sady do 7 skupin. V tomto kroku jsou však obce stále shlukovány v agregovaných počtech v SOM uzlech. Příslušnost obce ke konkrétnímu shluku byla následně odvozena na základě příslušnosti obce ke konkrétnímu SOM uzlu. Posledním úkonem shlukování je vyhodnocení statistických charakteristik vlastností obcí ve skupinách a jejich popis společně s vizualizací formou mapy. Kroky celého procesu jsou zobrazeny na Obrázku 32.

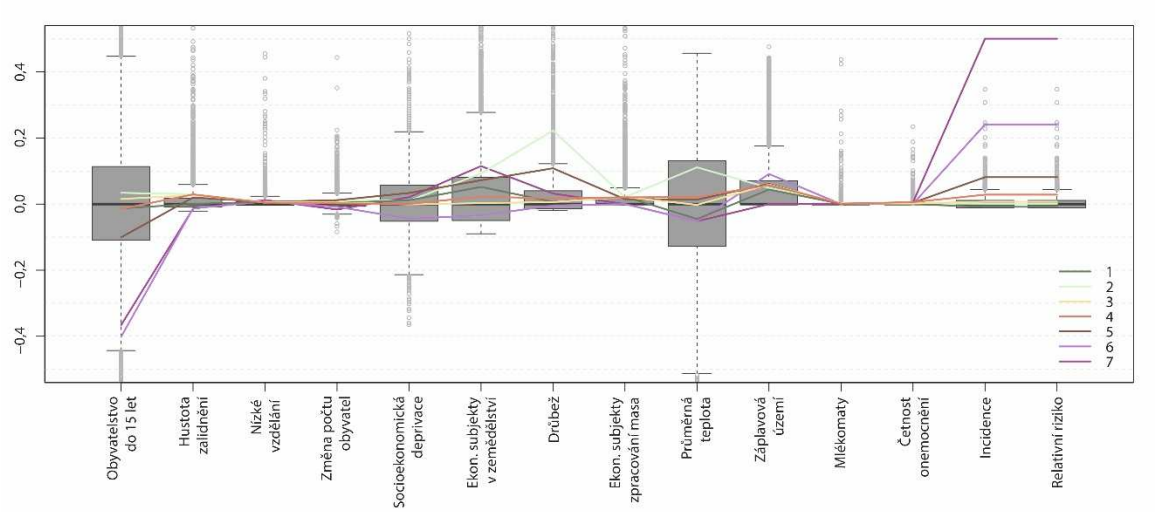
Na základě popsaného postupu analýzy shlukování byly obce České republiky klasifikovány do 7 skupin v závislosti na jejich environmentálních a socioekonomických charakteristikách. Každá ze skupin je popsána pomocí průměrné hodnoty, směrodatné odchylky a mediánu, jejichž konkrétní hodnoty jsou zobrazeny v Tabulce 16.

Tab. 16 Statistické charakteristiky vlastností obcí v identifikovaných obcích ($\bar{x}$ – průměr, s – směrodatná odchylka, $\tilde{x}$ - medián; tučně jsou označeny nejvyšší průměrné hodnoty, šedě jsou označeny nejnižší), barvy skupin odpovídají barvám v boxplotu (Obr. 33) i mapě (Obr. 34)

	Bez rozlišení skupin (n = 6251)			1 (n = 2672)			2 (n = 691)			3 (n = 1942)		
	$\bar{x}$	s	$\tilde{x}$	$\bar{x}$	s	$\tilde{x}$	$\bar{x}$	s	$\tilde{x}$	$\bar{x}$	s	$\tilde{x}$
Obyvatelstvo do 15 let	15,2	3,2	15,2	15,0	3,6	15,1	**15,7**	3,1	15,6	15,4	2,6	15,3
Hustota zalidnění	91,6	142,7	52,0	53,2	48,1	39,1	**124,7**	184,8	70,7	118,5	138,9	69,6
Obyvatelstvo s nízkým vzděláním	0,3	0,8	0,1	0,3	1,1	0,0	0,2	0,5	0,1	0,3	0,7	0,2
Změna populace	109,5	25,5	105,3	108,1	17,6	105,3	115,8	40,3	107,7	110,1	30,0	105,1
Socioekonomická deprivace	**1,6**	0,2	1,6	**1,6**	0,2	1,6	**1,6**	0,2	1,6	**1,6**	0,2	1,6
Zemědělské firmy	7,4	7,4	5,2	8,3	7,6	6,3	10,6	9,8	7,5	5,4	5,6	3,7
Drůbež	373,7	642,5	136,2	184,1	217,2	96,1	**1750,4**	1015,5	1439,1	179,8	211,9	98,3
Zpracování masa	0,7	1,8	0,0	**0,7**	2,1	0,0	0,6	1,5	0,0	0,6	1,4	0,0
Průměrná teplota	7,6	0,8	7,6	7,4	0,8	7,4	**8,1**	0,7	8,1	7,6	0,9	7,6
Záplavové území Q_{20}	1,2	2,0	0,0	1,0	1,9	0,0	1,1	2,0	0,0	1,3	2,1	0,2
Mlékomaty	0,0	0,2	0,0	0,0	0,2	0,0	0,0	0,4	0,0	0,0	0,2	0,0
Počet případů kampylobakteriózy	15,8	161,4	2,0	1,1	3,9	0,0	23,7	331,0	2,0	16,4	41,9	5,0
Průměrná hrubá incidence	672,4	1088,0	491,8	111,0	161,5	0,0	534,1	484,4	470,6	808,1	241,0	792,5
Relativní riziko	0,7	1,1	0,5	0,1	0,2	0,0	0,6	0,5	0,5	0,9	0,3	0,8

	4 (n = 875)			5 (n = 61)			6 (n = 8)			7 (n = 2)		
	$\bar{x}$	s	$\tilde{x}$	$\bar{x}$	s	$\tilde{x}$	$\bar{x}$	s	$\tilde{x}$	$\bar{x}$	s	$\tilde{x}$
Obyvatelstvo do 15 let	15,0	2,6	15,0	13,6	3,7	14,0	9,1	5,4	8,1	9,6	0,9	9,6
Hustota zalidnění	123,2	242,1	61,3	100,9	172,8	42,0	19,9	9,6	17,5	17,8	2,7	17,8
Obyvatelstvo s nízkým vzděláním	0,2	0,5	0,1	0,3	0,5	0,0	0,3	0,5	0,0	**0,5**	0,7	0,5
Změna populace	107,2	15,5	104,4	**116,1**	47,6	105,6	97,7	15,7	102,0	90,1	2,5	90,1
Socioekonomická deprivace	**1,6**	0,2	1,6	**1,6**	0,2	1,6	1,5	0,3	1,5	**1,6**	0,1	1,6
Zemědělské firmy	6,6	6,4	4,8	9,4	10,0	5,9	3,2	5,9	0,0	**12,0**	2,8	12,0
Drůbež	260,2	290,8	147,1	921,5	1093,6	495,9	104,6	224,9	0,0	366,3	82,1	366,3
Zpracování masa	0,6	1,5	0,0	0,4	1,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Průměrná teplota	7,7	0,8	7,8	7,7	0,9	7,7	7,4	0,6	7,4	7,4	0,8	7,4
Záplavové území Q_{20}	1,3	2,1	0,0	1,4	2,1	0,1	**2,0**	3,3	0,6	0,0	0,0	0,0
Mlékomaty	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Počet případů kampylobakteriózy	**51,7**	305,2	9,0	36,8	95,6	8,0	7,1	5,2	7,5	47,0	35,4	47,0
Průměrná hrubá incidence	1765,8	501,7	1627,2	4151,2	1189,0	3972,1	11320,4	2717,7	11015,4	**39713,0**	8204,3	39713,0
Relativní riziko	1,9	0,5	1,7	4,4	1,3	4,2	11,9	2,9	11,6	**41,8**	8,6	41,8

Kromě výpisu v Tabulce 16 jsou výsledky shlukování graficky vyobrazeny na boxplotu (krabicovém diagramu) na Obrázku 33, který poskytuje snadno srozumitelný celkový náhled na jednotlivé skupiny a umožňuje jejich snadné vizuální srovnání. Z důvodu srovnatelnosti charakteristik a jejich současného vykreslení do jednoho grafu byla data standardizována do intervalu $\langle -1; 1\rangle$ pomocí vztahu $\frac{(x-\bar{x})}{max(|x-\bar{x}|)}$. Z důvodu větší srozumitelnosti byla osa y omezena na rozsah $\langle -0{,}5; 0{,}5\rangle$. Mimo tento rozsah se sice vyskytovalo několik málo odlehlých hodnot, ale čitelnost grafu se výrazně zvýšila. Na Obrázku 33 je možné si všimnout rozdílného průběhu linií (nebo alespoň jejich částí), které charakterizují jednotlivé skupiny. Některé z charakteristik (nízké vzdělání, počet mlékomatů na 1000 obyvatel či absolutní četnost onemocnění) se na diferenciaci skupin výrazně nepodílí. Výrazně se na klasifikaci podílí zejména charakteristiky morbidity (incidence a relativní riziko) a dále podíl obyvatelstva do 15 let, podíl zemědělských ekonomických subjektů, počet drůbež na plochu obce). Zbylé charakteristiky je možné využít k upřesnění vybraných skupin.



Obr. 33 Boxploty standardizovaných charakteristik obcí s liniemi znázorňujícími průměrnou hodnotu charakteristiky v každé skupině

První skupinu (tmavě zelená barva v mapě i vizualizacích, 2672 obcí/12,880 % populace ČR) tvoří oblasti s nejnižším rizikem nakažení kampylobakteriózou. Jde o nejpočetnější skupinu obcí, kde je současně výjimečný výskyt onemocnění. Většina hodnocených charakteristik se jeví průměrných s ohledem ke globálním hodnotám pro celé území. Výjimkou je nadprůměrný podíl zemědělských ekonomických subjektů a nižší průměrná teplota vzduchu, což je výjimečná kombinace v rámci ostatních obcí (podobnou má sedmá skupina). V případě první skupiny jde o obce s průměrným nejnižším zalidněním.

Druhou skupinu (světle zelená, 691 obcí/21,516 % populace ČR) tvoří obce s nízkým relativním rizikem onemocnění. Jde o obce s nejvyšší hustotou zalidnění a také s nejvyšším podílem obyvatelstva do 15 let a nejvyšší průměrnou teplotou vzduchu. S teplotou vzduchu je spojen i vysoký počet ekonomických subjektů působících v zemědělství a vůbec nejvyšší odhadovaný průměrný počet drůbeže k ploše obce. V obcích této skupiny také dochází k nárůstu počtu obyvatel (průměrně o 15 % za 5 let).

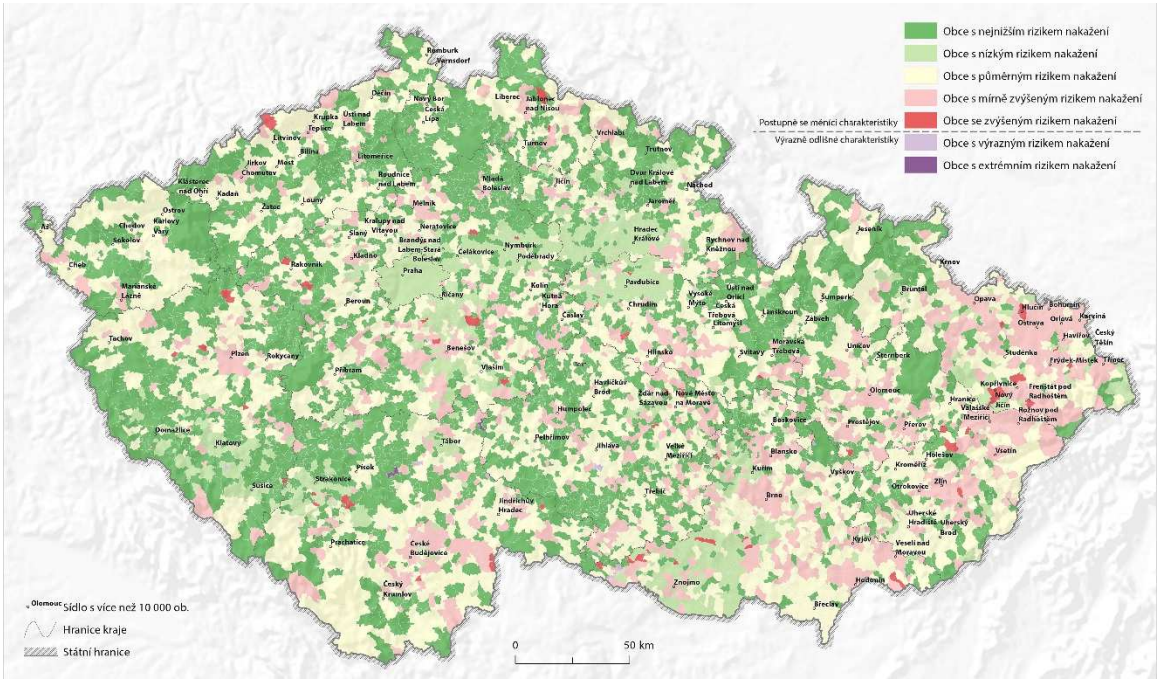
Třetí skupinu (světle žlutá, 1942 obcí/37,183 % populace ČR) jsou obce s průměrným rizikem nakažení nemocí. Průměrnými hodnotami se vyznačuje i většina dalších vlastnosti obcí. Výjimkou je druhá nejvyšší průměrná hodnota podílu obyvatelstva do 15 let.

Čtvrtou skupinu (světle červená, 875 obcí/27,690 % populace ČR) tvoří hustě zalidněné obce s mírně zvýšeným rizikem nakažení kampylobakteriózou, ačkoliv je v těchto obcích nejvyšší absolutní průměrný výskyt případů.

Pátou skupinu (tmavě červená, 61 obcí/0,720 % populace ČR) tvoří obce s vysokým rizikem onemocnění. Jde o hustě zalidněné obce, které ačkoliv mají nižší podíl mladého obyvatelstva do 15 let, tak vykazují nejvyšší průměrnou změnu obyvatelstva a také nejvyšší socioekonomickou deprivaci obyvatelstva. V obcích této skupiny je nadprůměrný podíl zemědělských ekonomických subjektů a s tím související nadprůměrný počet drůbeže na plochu obcí.

Nejvýrazněji odlišně se jeví **šestá skupina** (světle fialová, 8 obcí/0,008 % populace ČR) a **sedmá skupina** (tmavě fialová, 2 obce/0,000 % populace ČR), které se vyznačují výrazně podprůměrným podílem osob do 15 let a vysokou průměrnou incidencí i relativním rizikem. Skupiny se od sebe odlišují zejména socioekonomickou deprivací obyvatelstva a množstvím

zemědělstvím se zabývajících ekonomických subjektů. Na rozdíl od předchozích skupin obsahují tyto dvě velmi málo obcí. Jejich odlišnost od ostatních obcí však byla tak výrazná, že se obě skupiny vyprofilovaly i při využití jiných druhů klasifikace.



Obr. 34 Mapa příslušnosti obcí k identifikovaným skupinám

Geografické rozložení provedené klasifikace je viditelné na Obrázku 34. Na první pohled je patrné, že obce s nejnižším rizikem nakažení nemocí jsou zejména menší obce v oblastech, kde se neočekává vysoká hustota zalidnění (např. horské a podhorské oblasti). Tato skupina také tvoří nejspojitější struktury. U obcí s nízkým rizikem pak jde spíše o populačně větší obce s vyšší hustotou obyvatelstva. V souladu se zjištěními z kapitol 4 a 5 je možné si všimnout obcí s mírně zvýšeným rizikem nakažení a několika obcí se zvýšeným rizikem nakaženým v oblasti (severo)východní Moravy a Slezska, v okolí Brna, na Českobudějovicku a Plzeňsku.

6.5 Shrnutí

Šestá kapitola, která shrnuje třetí dílčí cíl disertačního výzkumu, se postupně zabývala analýzou faktorů prostředí, které mohou mít vliv na prostorové rozšíření kampylobakteriózy v České republice. Nejdříve byla sestavena datová sada 103 charakteristik obcí České republiky, ze kterých bylo pomocí korelací, logického úsudku, konzultací a srovnáváním se zahraničními studiemi vybráno 11 vhodných charakteristik. Tyto charakteristiky buď mohou podmiňovat prostorovou distribuci onemocnění, nebo mohou reprezentovat širší skupinu příčin. Charakteristiky byly podrobeny analýze lokální prostorové korelace s relativním rizikem onemocnění (SIR) a také analýze lokální autokorelace pro dvě proměnné. Bylo zjištěno, že mezi zvolenými charakteristikami a relativním rizikem onemocnění existují lokální prostorové vazby. Nejsilnější vazby byly vyhodnoceny v souvislosti s hustotou obyvatelstva a průměrnou teplotou vzduchu, dále pak s rozsahem záplavových území a ekonomickými subjekty v oblasti zemědělství a zpracování masa. Dimenze datové sady vybraných charakteristik území byla redukována pomocí metody hlavních komponent a také její prostorové modifikace, která hodnotila lokální nejvlivnější charakteristiky v obcích ve

dvou měřítcích definovaných pomocí sousedství (320 a 54 sousedů). Více vyhlazená varianta (320 sousedů) ukázala na nejvýraznější charakteristiky obcí, kterými byly hustota obyvatelstva, masozpracující ekonomické subjekty, zemědělské subjekty a relativní změna počtu obyvatel. Lokálnější varianta (54 sousedů) poukázala na možné lokální zdroje variability charakteristik.

Vybrané charakteristiky byly dále hodnoceny pomocí regresních metod, jejich prostorových alternativ a také pomocí klasifikačních algoritmů. Jako vysvětlovaná proměnná vstupovaly do hodnocení nejčastěji obce rozdělené do čtyř skupin podle míry relativního rizika onemocnění kampylobakteriózou (0—0,01—0,80—1,50—1,51 a více). Jako prediktory (vysvětlující) proměnné sloužily vybrané charakteristiky a/nebo pět hlavních komponent identifikovaných PCA. Klasifikací se sice zjednodušila úloha pro predikci z pohledu přesnosti, ale díky tomu bylo nutné použít metod generalizované lineární regrese a postupů klasifikace. Testován byl i model pro výpočet konkrétní četnosti případů v obcích. Regresními modely byla identifikována souvislost mezi hustotou zalidnění a průměrnou teplotou a distribucí onemocnění. Lokálně byl také předpokládán vliv socioekonomické determinace a nízkého vzdělání obyvatelstva. Nepříliš výrazný se ukázal efekt obyvatelstva do 15 let. Jako nejvhodnější model se na základě McFaddenova pseudo R^2 ukázal být negativně binomický model s nadbytečnými nulami pro průměrnou absolutní četnost výskytu onemocnění v obcích. Multinomický a ordinální model se neukázaly jako vhodná varianta pro predikci, ale ukázaly se jako výhodné pro exploraci síly vztahů mezi charakteristikami a distribucí onemocnění. V této části byl také navrhnut postup pro lokální ordinální logistickou regresi (**R** kód viz vázaná příloha Příloha 15:15). Kromě regresních modelů byly testovány i klasifikační metody strojového učení, data miningu a neuronových sítí pro původní data charakteristik i hlavní komponenty. Nejlépe byly hodnoceny metody lokální ordinální regrese, geograficky vážené diskriminační analýzy z metod jednodušších a dále z komplexnějších postupů Random Forest, neuronová síť a Support Vector Machine s radiální funkcí báze. Klasifikací i regresními postupy se potvrdila skutečnost, že téměř polovina případů kampylobakteriózy u nás i ve světě zůstává nevysvětlena.

Posledním úkonem tohoto dílčího cíle byla klasifikace obcí České republiky do skupin podle charakteristik prostředí a charakteristik nemocnosti. K tomuto kroku byla využita shluková analýza, která v kombinaci se samoorganizačními mapami identifikovala na území ČR sedm skupin obcí. Tyto skupiny byly popsány a vizualizovány v mapě na Obrázku 34. V pěti ze sedmi skupin jde o pozvolnou proměnu charakteristik a nemocnosti v obcích, ale dvě velmi malé skupiny jsou výrazně odlišné. Stejně jako v případě výsledků a hodnocení v DC1 a DC2 je patrná diferenciace mezi východní (více ohroženou) a západní (zdravější) částí České republiky.

7 GEOGRAFICKÉ PROFILOVÁNÍ: IDENTIFIKACE MOŽNÝCH ZDROJŮ INFEKCE [DC4]

Začátkem roku 2010 se spolu s rostoucím počtem automatů na přímý prodej čerstvého mléka rozvinula i diskuze o možných zdravotních problémech, které jeho konzumace může způsobit i přesto, že jsou dodavatelé i automaty pravidelně kontrolováni příslušnými orgány (Agrární komora ČR, 2010; SZÚ ČR, 2010). Původní velmi kladné přijetí možnosti nákupu čerstvého mléka bylo ovlivněno vyjádřením tehdejšího hlavního hygienika upozorňujícího na tyto problémy (Ministerstvo zdravotnictví ČR, 2010). Prohlášení mělo dopad na prodej čerstvého mléka z automatů a v důsledku toho i zpomalení růstu množství automatů a u některých i postupné ukončení jejich činnosti (Andrlová, 2011). Diskuze během jara 2010 natolik zesílila, že Česká agrární komora zvažovala podání trestního oznámení na tehdejšího hlavního hygienika MUDr. Michaela Víta, Ph.D. (Český rozhlas, 2010).

Hlavním motiv DC4 byl přímo inspirován výše zmíněnou situací a jeho cílem je zjištění, zda opravdu mohly být některé z mlékomatů potenciálním zdrojem lokálního zvýšeného počtu případů onemocnění kampylobakteriózou, nebo zda šlo jen o pravidelný sezónní vzestup onemocnění. Zvláštní důraz byl kladen na situaci na Českobudějovicku během ledna a února roku 2010, která byla zmíněna přímo v tiskové zprávě Ministerstva zdravotnictví ČR. Pro hodnocení možné souvislosti mezi mlékomaty a kampylobakteriózou byla zvolena metoda geografického profilování.

7.1 Geografické profilování

Geografické profilování (geoprofiling) je metodou původně vyvinutou pro statistické hodnocení v kriminologii, kde je s velkou úspěšností využíváno pro identifikaci pravděpodobného místa bydliště pachatele na základě umístění sériových trestných činů (Stevenson et al., 2012). Díky tomu byla v nedávné době metoda aplikována pro úlohy s jejím původním účelem zdánlivě nesouvisejícími a to zejména v biologii při studii strategie hledání potravy u žraloků (Martin et al., 2009) či čmeláků (Raine et al., 2009). Později byla metoda geografického profilování použita v prostorové epidemiologii, kde lze často předpokládat jeden nebo více zdrojů nákazy, které je nutné identifikovat (Buscema et al., 2009; Le Comber et al., 2011). Od původního konceptu metody navrhnuté Kimem Rossmo (Rossmo, 1995a) prochází metoda neustálými vylepšeními díky podpoře komerční sféry[38] a v posledních letech i univerzitního výzkumu (Stevenson, 2015).

Prvním předpokladem metody je, že zkoumané jevy mají společnou skrytou geografickou strukturu a vykazují podobné chování. V případě sériových trestných činů je to pohyb ve známém prostředí, v případě živočichů to jsou vzory hledání potravy a v případě prostorové epidemiologie to může být shodný zdroj nákazy infekčního onemocnění. Druhým předpokladem je vhodný počet zkoumaných událostí, aby bylo možné dosáhnout věrohodného odhadu (Snook et al., 2005).

Stále nejčastěji používaným modelem geografického profilování je původní model Criminal Geographic Targeting (CGT), který pro každý umístěný případ konstruuje funkci

[38] Komerčně dosptupný software Rigel Analyst společnosti ECRI, které je Rossmo spoluzakladatelem

oslabení závislou na vzdálenosti. Jde tak o obdobu jádrového odhadu. Nejjistější odhad se nachází ve středu funkce. V kriminologii je běžně využíváno rozdělení se dvěma vrcholy, kdy v blízkosti středu dochází k rychlému nárůstu a dále dochází k pozvolnému poklesu funkce. Kombinací funkcí pro celé území oslabení vzniká povrch priority vyhledávání. V případě, že se objekt vyskytuje v části nejvyšší konfidence více funkcí, tak je mu přidělena vyšší priorita. Výsledkem modelu je shrnutí veškerých funkcí a jejich kombinací do jediného povrchu označovaného jako povrch ohrožení (Rossmo, 1999). Efektivita vyhledávání modelu je vypočítána s využitím hitskóre („*skóre správného určení*"), které je tvořeno podílem plochy, která musí být prohledána, aby byl nalezen zdroj, z celkové plochy studovaného území. Čím nižší je procento hitskóre, tím přesnější je sestrojený geoprofil. Hitskóre dosahující hodnoty 50 % (0,5) představuje buď náhodné, nebo pravidelné rozdělení vyhledávání. V praxi je nejčastěji využívána hraniční hodnota 2 % (0,02) (Rossmo, 2000). Rozdílným přístupem ve vytváření geoprofilů je využití jednoduchých bayesovských principů. K tomu je potřeba odhadnout apriorní rozdělení pravděpodobností, které je kombinováno s polohovou informací a funkcí oslabení, což vyústí ve výsledný povrch aposteriorních odhadů (O'Leary, 2010).

Pro účely DC4 je však využita nejnovější modifikace geografického profilování pomocí Dirichletova modelu pro smíšené procesy (DPM). Tento model je navržen pro vhodnější identifikaci mnohočetných zdrojů, což byla oblast, ve které zůstávají CGT i jednoduchý bayesovský model za očekáváním (Verity et al., 2014). Model DPM ovšem na oba zmíněné modely navazuje. Na rozdíl od modelů shlukování nevyžaduje DPM dopředu znát počet shluků a tím pádem i možných zdrojů. Protože DPM primárně s informací o shlucích nepracuje, tak je vhodný pro situace, kdy není počet shluků znám dopředu nebo není možné ho úspěšně předpokládat. V prvním kroku DPM odhaduje počet shluků (≈ množství zdrojů) ve zkoumané oblasti na základě nejistoty. Ve druhém vypočítá pomocí bayesovských pravidel jejich rozdělení pravděpodobnosti v závislosti na tom, jak jsou k nim přiřazena původní data. V posledním kroku jsou jednotlivé potenciální zdroje vyhodnoceny simulací pomocí Markovových řetězců (MCMC) a je sestrojen geoprofil. Detailní popis DPM je uvádí Verity et al. (2014). Metody geografického profilování předchází omezením analýz agregovaných dat tím, že využívá přímo lokace případů. Umožňuje tím vytvoření spojitého povrchu rizika a odhalení prostorových vzorů, které mohly být agregací skryty (Le Comber et al., 2011).

7.2 Mlékomaty a kampylobakterióza v geografických profilech

Pro hodnocení situace okolo mlékomatů a zjišťování, zda některé z nich byly potenciálními zdroji lokálního zvýšení počtu případů, byla použita zmíněná metoda geografického profilování pomocí DPM, která je vhodná k identifikaci vícenásobných zdrojů. Proto nebylo nutné hodnotit každý mlékomat zvlášť, ale byly současně vytvářeny geografické profily pro větší oblasti s více mlékomaty.

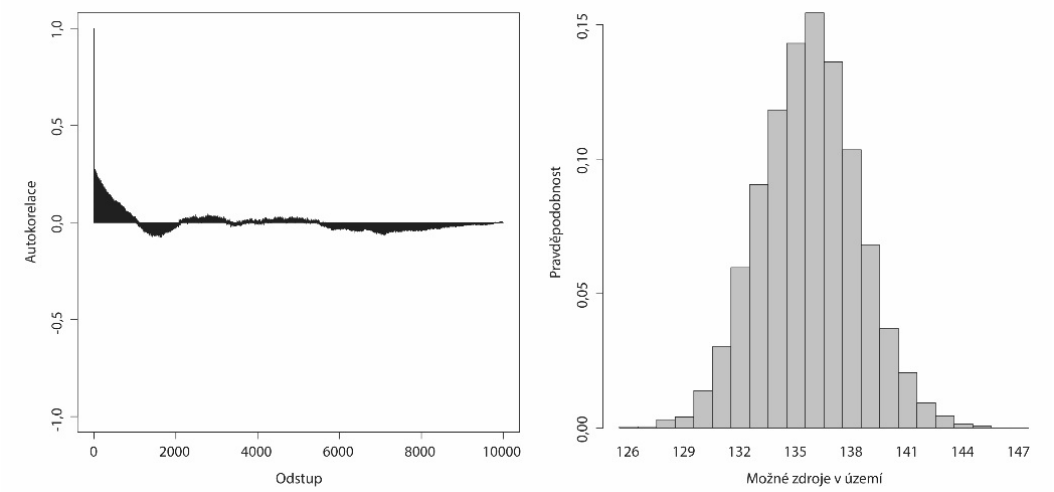
Adresy 158 aktuálně umístěných mlékomatů, který je 158, jsou včetně data schválení přístupná na stránkách Státní veterinární správy ČR[39], kde jsou mimo jiné u údaje o prodejcích čerstvého mléka. S historickými daty nutnými k retrospektivní analýze je situace složitější.

[39] *http://eagri.cz/public/app/svs_pub/subjekty/mleko.php*

Státní veterinární správa na dotaz o data sice zareagovala, ale data samotná poskytnuta nebyla. Datová sada mlékomatů proto byla doplněna z méně oficiálních zdrojů, kterým bylo Venkovské fóru[40] a provozovatelé automatů na čerstvé mléko, např. společnost TOKO[41]. Výsledná datová sada mlékomatů kombinovaná ze všech zdrojů obsahovala 267 automatů pravděpodobně fungujících před rokem 2012.

Data případů kampylobakteriózy byla geokódována již během přípravných prací předchozích dílčích cílů. Vzhledem k faktu, že jsou DPM modely založeny na MCMC, tak jsou v případě objemnějších datových sad velmi náročné na výpočetní výkon i čas. To byl jeden z důvodů, proč byla individuální data případů redukována. Druhým důvodem byla lokální povaha geografického profilování. Do vstupní datové sady byly vybrány pouze případy vzdáleny do 15 km od nejbližšího mlékomatu, což je vzdálenost, která odpovídá průměrné době dojížďky do zaměstnání a škol v Moravskoslezském kraji (Ivan a Tvrdý, 2007). Je vhodné zmínit, že do analýzy vstupovala všechna data výskytu případů bez ohledu na příčinu. Důvodem bylo více než 40 % neurčených příčin v původních datech. Díky tomu je potřeba počítat s tím, že geografické profily zobrazují spíše „skeptickou" variantu odhadu s možnými nadhodnocenými lokalitami. Z důvodu výpočetně a časově náročného procesu byla data rozdělena do 18 nestejnorodých oblastí, které jsou viditelné i ve výsledných vizualizacích.

Dalším krokem už bylo vytvoření geoprofilů. K tomu bylo využito balíku **R** *Rgeoprofile* (Stevenson a Verity, 2014). Výhodou implementace DPM v prostředí **R** je relativní nenáročnost jeho nastavení, nevýhodou je ovšem dlouhý výpočetní čas. Po nahrání vstupních dat probíhal proces geoprofilování v **R** semiautomaticky. Nejdříve byly zvoleny parametry výstupní grafiky, parametry modelu a simulace a následně byla zahájena simulace.



Obr. 35 Graf autokorelace (vlevo) a odhadu počtu zdrojů (vpravo) pro oblast jižních Čech

Nejdůležitějším z volených parametrů byla vzdálenost vstupující do geoprofilování jako parametr funkce oslabení. Tento parametr byl nastaven na 0,01°, tedy přibližně jeden kilometr (pro oblast ČR). Dále byly voleny parametry MCMC, ze kterých je pro sestrojení výsledného geoprofilu nejdůležitější tzv. thinning udávající stupeň redukce pro vzorkování (čím nižší thinning, tím více vzorků zůstane zachováno). Thinning má významný vliv na průběh

[40] *http://www.venkovskeforum.cz/mlekomaty*

[41] *http://www.tmleko.cz/cerstve-mleko/seznam-automatu/*

výsledného geoprofilu, především na to jak bude výsledný povrch hladký. Hodnota thinningu byla odhadována na základě grafu autokorelace (Obrázek 35). Podle Stevensona (2014) odpovídá velikost thinningu odstupu, v jakém je v grafu překračována nulová hodnota autokorelace. Po nastavení thinningu vznikly geoprofily a jim odpovídající hodnoty hitskóre. Posledním krokem je vizualizace v prostředí **R** a tvorba rastru pro geovizualizaci. Pro každou z 18 oblastí byl sestrojen individuální geoprofil. Počty bodů, doba trvání simulace a tvorby geoprofilu, volba thinningu a počet mlékomatů identifikovaných jako potenciální zdroje je uveden v Tabulce 17.

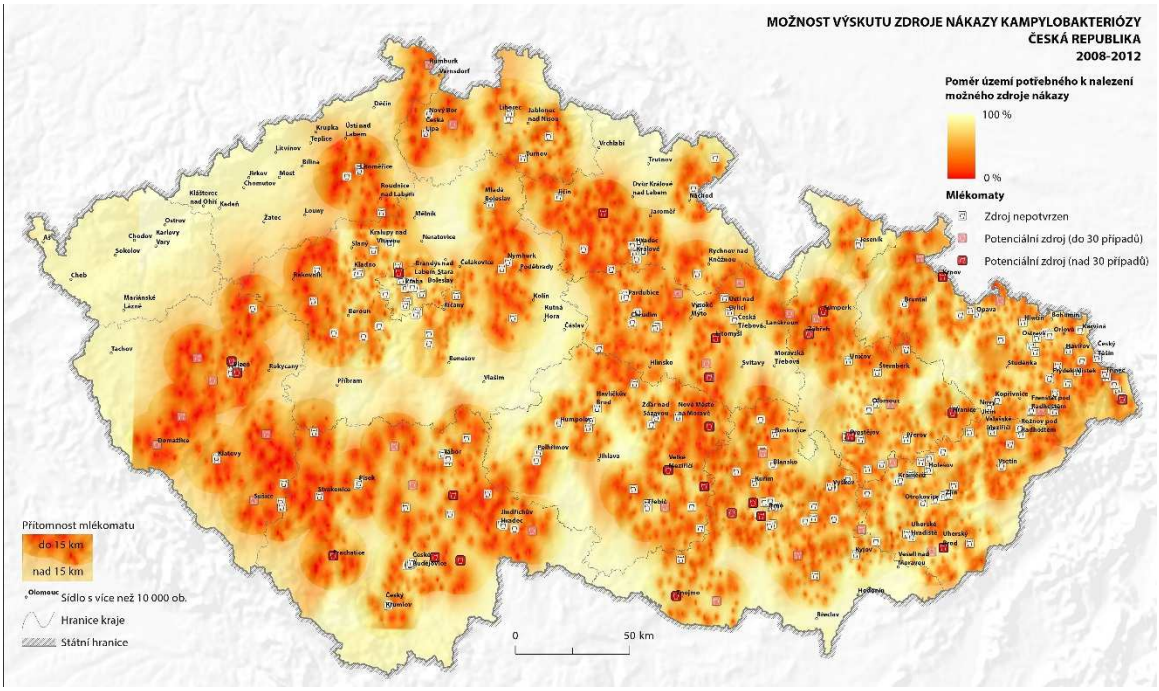
Tab. 17 Tvorba geoprofilů a mlékomaty identifikované jako potenciální zdroje lokální nákazy

Oblast	Případy	Thinning	Mlékomaty celkem	Potenciální zdroj	čas
Brněnsko	6116	100	23	4	36,00
Haná	3180	100	13	3	11,13
jižní Čechy	3046	1000	26	7	9,63
jižní Morava	1494	100	11	5	2,85
Opavsko	1599	100	6	2	3,03
Ostravsko	4657	1500	8	0	13,30
Plzeňsko	3082	500	24	8	7,90
Praha	4828	600	13	1	26,94
severní Čechy	694	100	9	3	0,44
severní Morava	1214	100	14	5	1,36
severovýchodní Čechy	3985	100	27	1	19,15
Slezsko	3052	1200	16	1	6,07
Slovácko	3472	100	28	4	13,12
střední Čechy - východ	1276	100	3	0	1,56
střední Čechy - západ	2010	2000	7	0	3,53
Valašsko	3528	100	14	1	13,57
Vysočina	711	100	6	0	0,52
západ Moravy	2088	50	19	7	5,10
celkem	50032		267	52	175,19

Vyhodnocování jednotlivých bodových zdrojů následně probíhá na základě tzv. hitscore, které udává velikost oblasti, kterou je potřeba prohledat k nalezení zdroje. Pomocí geoprofilování bylo jako potenciální zdroje lokální nákazy identifikováno 52 mlékomatů (19,5 %) z celkového počtu. Po vizuální kontrole však bylo zjištěno, že metoda DPM v některých případech nadhodnocuje hitskóre u mlékomatů, v jejichž blízkosti se vyskytuje jen malé množství výskytů onemocnění. Z tohoto důvodu byly kolem všech mlékomatů vytvořeny obalové zóny o poloměru 2 km jako dvojnásobek vzdálenosti ve funkci oslabení. Důvodem byl předpoklad nejsilnějšího významu mlékomatu pro nejbližší případy. Dalším kritériem posilující jistotu rozhodnutí u identifikace možných zdrojů byla zvolena hranice 30 případů nacházejících se v jednotlivých obalových zónách. Takto byl počet mlékomatů zvažovaných za zdroje nákazy zredukován na 24 případů (9,0 %).

Výsledek geoprofilování je možné zhlédnout na Obrázku 36, kde jsou od sebe mlékomaty barevně odlišeny právě v závislosti na konfidenci jejich odhadu jako potenciálních zdrojů. Bíle jsou vyobrazeny mlékomaty, které nebyly vyhodnoceny jako potenciální zdroj nákazy. Růžově mlékomaty, které sice byly identifikovány jako potenciální zdroj nákazy, ale v jejich blízkosti se vyskytovalo pouze menší množství případů (do 30). Červeně jsou vyobrazeny mlékomaty vyhodnocené jako potenciální zdroje, v jejichž nejbližším okolí se během studovaného období vyskytlo více než 30 případů onemocnění kampylobakteriózou. Na

mapě jsou patrná také různá nastavení thinningu, které se projevuje jako rozdílné shlazení výsledného povrchu. Kromě toho je v mapě zohledněna i oblast do 15 km od mlékomatů, která je sytá, zatímco oblasti vzdálenější jsou maskovány pomocí průhlednosti. Zde je potřeba také zdůraznit, že výsledný povrch vznikl zkombinováním 18 dílčích geoprofilů a díky tomu obsahuje i místa, která původně nebyla hodnocena. Jednotlivé povrchy společně s identifikovanými mlékomaty je možné dále zkoumat jako KML soubory vytvořené jako součást DC5 (viz volná příloha – adresář KML).



Obr. 36 Geografický profil České republiky zohledňující polohu mlékomatů

7.3 Shrnutí

Pomocí metody geografického profilování byla retrospektivně zkoumána souvislost mezi přítomností mlékomatů a zvýšenou četností výskytu onemocnění kampylobakteriózou v jejich okolí. Původně bylo zjištěno, že v období let 2008—2012 mohlo být zdrojem nákazy až 20 % mlékomatů. Po korekci byla tato hodnota snížena na 9 %. Z mapy se jeví, že největší množství identifikovaných mlékomatů se nachází na pomezí Čech a Moravy na Vysočině a v oblasti severně od ní. Vyšší množství mlékomatů mohlo být zdrojem nákazy také na Šumpersku a Českobudějovicku. Právě situace na počátku roku 2010 v Českých Budějovicích byla jedním ze spouštěčů vyostřené diskuze týkající se dopadů pití čerstvého mléka na zdraví obyvatelstva. V DC2 byl na Českobudějovicku identifikován časově omezený shluk zvýšeného relativního rizika v době odpovídající vydání ziskové zprávy a počátku diskuze. Přímo v Českých Budějovicích však žádný rizikový mlékomat identifikován nebyl. Ovšem potenciální zdroj byl nalezen v nedaleké Třeboni a také v Lišově, kde byl ale dle záznamů veterinární správy mlékomat schválen až v srpnu 2010. Ověřit, jak tomu ve skutečnosti bylo, už sice je jen stěží možné, ale geoprofilování se ukázalo jako využitelný nástroj, který je možné použít pro identifikaci mnohonásobných zdrojů nákazy v případě infekčních onemocnění.

8 GEOVISUAL ANALYTICS: GOOGLE EARTH JAKO NÁSTROJ PRO PREZENTACI A PRŮZKUM DAT V ČASE I PROSTORU [DC5]

Zkoumání a modelování prostorového rozložení onemocnění a jeho prostorových vzorů tak, jak bylo představeno v jednotlivých dílčích cílech disertační práce, je relevantním tématem pro geovědní i medicínské obory. Díky tomu je možné lépe porozumět nejen samotnému rozšíření vybraných onemocnění, ale také faktorům podmiňujícím výskyt onemocnění. Rostoucí množství (geo)dat a také jejich komplexnost si vyžádaly vznik stejně komplexních nástrojů, které umožní tato data zpracovávat a zobrazovat způsobem, který umožní usnadnit rozhodování. Zmíněnými postupy se zabývá vizuální analytika. Jde o vědní obor a teorii popisující komplexní a dynamickou povahu jevů kombinováním analytických postupů a interaktivních vizualizací, které společně umožňují efektivní porozumění, uvažování a rozhodování o daném jevu na základě rozsáhlých datových sad (Keim et al., 2010). Velké množství sbíraných dat je možné polohově určit. Díky tomu vznikla také geovizuální analytika - odnož vizuální analytiky, která zohledňuje prostorové vlastnosti jevu. Cílem geovizuální analytiky je nejen samotná analýza dat a interpretace výsledků, ale také šíření a komunikace výsledků ve vhodné formě, aby mohly být v rozhodovacích procesech využity i neodborníky (Tomaszewski, 2009). V samotném procesu hodnocení a rozhodování hraje často důležitou roli kromě prostoru také čas.

Na rozdíl od předchozích DC nebylo hlavním účelem DC5 provést analýzu ve smyslu přímé kvantifikace dat. Bylo jím však využití výsledků předchozích DC a jejich transformace do podoby vhodné k další interaktivní exploraci v prostoru i čase právě s ohledem na principy geovizuální analytiky. Pro řešení tohoto dílčího cíle byl zvolen populární nástroj Google Earth, který zde nahrazuje komplexní nástroje pro geovizuální analýzu a který zjednodušuje přenositelnost výsledků.

8.1 Google Earth a Keyhole Markup Language

Společnost Google, jeden ze současných lídrů na poli informačních technologií, poskytuje také nástroje umožňující prohlížení, vykreslování, mapování a také vizuální průzkum dat (Marek et al., 2015). Jedním z těch nejznámějších je i aplikace Google Earth. Googl Earth je populárním 3D virtuálním glóbem, který umožnuje zobrazovat prostorová data a interaktivně je prozkoumávat nejen v prostoru, ale i v čase. Ačkoliv není Google Earth plně funkční platformou pro geovizuální analýzu v pravém slova smyslu, tak je považován za vhodnou náhradu z pohledu plnění jejích hlavních cílů – prozkoumávání neznámých prostorových vzorů v datech, diseminace výsledků a srozumitelné zprostředkování analytických interpretací výsledků. Hlavní důvody pro použití Google Earth jsou tyto:

- software je dostupný zdarma a to včetně verze Google Earth Pro;
- jde o velmi rozšířený produkt a pravděpodobně vůbec nejrozšířenější prohlížečku prostorových dat (Hengl, 2007), který je ve všeobecném povědomí laické i odborné veřejnosti;
- jeho používání je považováno za jednoduché a intuitivní;

- poskytuje kvalitní podkladová data (administrativní i letecké snímky a další informace);
- je možné využít a zobrazovat i vlastní geodata (KML soubory) v prostoru i čase.

Google Earth umožňuje interaktivní zobrazování a prohlížení prostorových a časoprostorových dat včetně přibližování, změny pohledu, dotazování, přidávání vrstev nebo animací. Je však potřeba si uvědomit, že hlavním účelem tohoto programu je prohlížení dat, nikoliv jejich vytváření. To ale může vyhovovat laickým uživatelům, kteří jsou často schopni ocenit spíše maximální jednoduchost a přímočarost programu před jeho analytickou funkcionalitou. Využitelnost programu ve výzkumu týkajícím se zdravotnických dat dokládají Eisen a Lozano-Fuentes (2009), Kamadjeu (2009) nebo Bergquist (2011). Jeho srovnání s dostupnými GIS programy pak poskytuje Lozano-Fuentes et al. (2008).

Volně dostupná verze Google Earth umožňuje zobrazit pouze omezené množství formátů. V lednu roku 2015 však byla uvolněna verze Google Earth Pro, která jich umí zobrazit výrazně více (např. shapefile, GeoTIFF či Microstation dgn). Hlavním formátem využívaným v aplikaci však je v obou případech Keyhole Markup Language (KML/KMZ). Jde o formát souborů používaný pro zobrazování a ukládání prostorových dat, který je založen na tagové struktuře prvků a atributů vycházející ze standardu XML (Google, 2009). Formát byl navíc schválen OGC[42] jako standard pro výměnu prostorových dat a je tak podporován nejen v Google Earth, ale i ve většině GIS software. V KML je specifikována sada standardních prvků pro zobrazení dat (umístění, značka místa, obrázek, polygon, linie, 3D model nebo časové značky), pro které je poloha určena pomocí souřadnic v souřadnicovém systému WGS 84. Komprimovaný formát KML je označován jako KMZ.

8.2 Geovizualizace pomocí KML

Silnou stránkou Google Earth je vizualizace prostorových a časoprostorových dat. Jeho nevýhodou je ovšem nutná příprava dat mimo jeho prostředí. V disertační práci byly pro vytvoření KML souborů z vybraných výsledků analýz použity extenze *Export to KML* pro ArcMap 10.1 nebo opět statistického software **R** s využitím vhodných balíků *raster* (Hijmans, 2014) a zejména *plotKML* (Hengl et al., 2014). Konkrétně byly v podobě KML resp. KMZ souborů prezentovány následující geovizualizace a výsledky prostorových analýz:

- nálezová data agregovaná v týdenních intervalech během let 2008—2012 do pravidelné sítě o hraně 2 km vizualizovaná formou tzv. *bubble chart* tedy bublinového kartodiagramu;
- roční incidence v obcích České republiky;
- spojitý povrch týdenní hrubé incidence kampylobakteriózy vzniklý pomocí časoprostorového krigingu;
- časoprostorové shlukování obcí České republiky z pohledu relativního rizika výskytu kampylobakteriózy;
- geografické profily a mlékomaty jako potenciální zdroj lokálních rozšíření kampylobakteriózy.

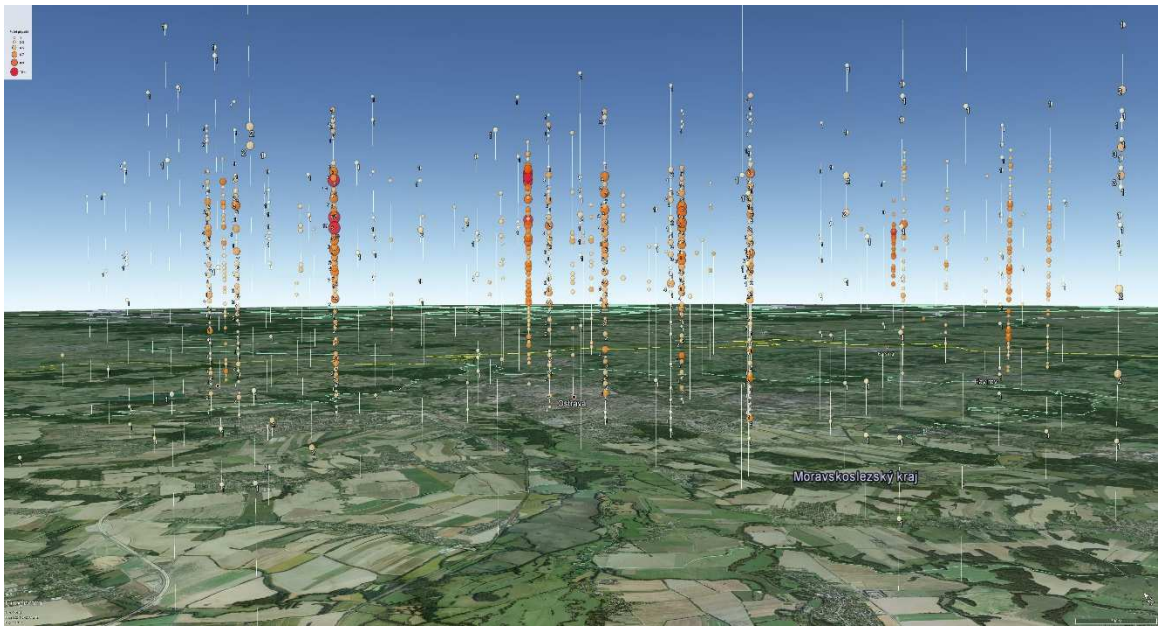
Výsledné KML soubory jsou dostupné jako součást volné přílohy na přiloženém DVD.

[42] Open Geospatial Consortium

Zobrazení výskytu kampylobakteriózy agregovaných v čase i prostoru

Důvěrná povaha dat neumožňuje prezentovat data jednotlivě přesně v místech, kde se pacient nakazil kampylobakteriózou, proto byla data nejdříve agregována. Studované území bylo překryto pravidelným čtvercovým gridem o velikosti buňky 2×2 km, do kterého byly agregovány záznamy o jednotlivých případech onemocnění v týdenních intervalech. Vzniklo tak 261 časových úseků, které je možné zobrazit. Aby bylo možné mezi sebou přímo vizuálně srovnat vývoj v čase, byla data vizualizována formou bublinového kartodiagramu. Velikost a barva bublin znázorňují kvůli snadnějšímu vzájemnému odlišení počet případů vztažený ke středu agregované jednotky. Časová složka je v případě této vizualizace zastoupena dvakrát. Poprvé formou časové značky umožňující pohyb v čase pomocí jezdce v Google Earth a podruhé je časová složka suplována výškou nad terénem, ve které jsou bubliny znázorňovány, v závislosti na pořadí týdne od začátku roku 2008. Díky těmto charakteristikám je možné postupně zkoumat časový průběh vývoje onemocnění na jednom místě nebo na více místech v časovém řezu. Jednotlivé bubliny doprovází číslo označující přesné množství případů onemocnění v jednotce. Červená barva a větší bublina znamená vyšší četnost případů, menší a světlejší bublina pak nižší počet případů. Bublina zcela chybí, pokud v jednotce nebyl v daném období zaznamenán ani jeden případ. Bubliny jsou pro snazší odhad místa doprovázeny vodicí linkou, která se vztahuje ke středu buňky pravidelné sítě. Nevýhodou této vizualizace je náročnost výsledného KML na paměť počítače z důvodu vykreslování velkého množství bodů (bublin) současně.

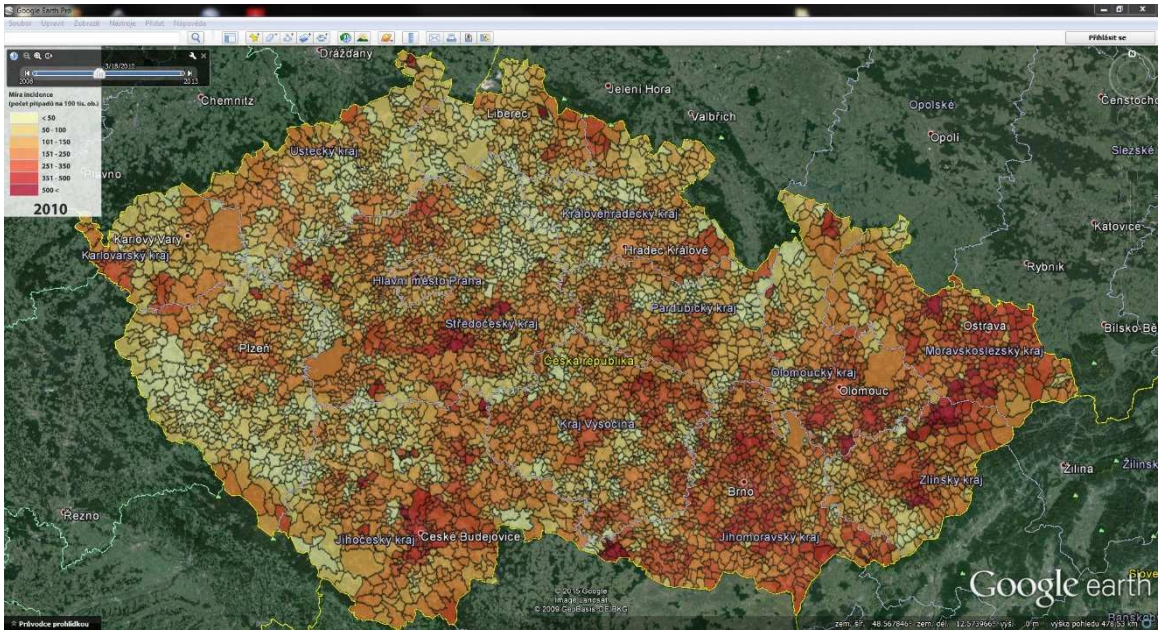
Zvolenou metodu vizualizace je možné považovat za variaci na známý model časoprostorové krychle (Kraak, 2003). Ukázka výsledného KML z oblasti Ostravska, která je dle DC2 územím nejnáchylnějším k onemocnění, je zobrazena na Obrázku 37. Geovizualizace byla vytvořena pomocí balíku *plotKML*, který slouží především k tvorbě KML z velkého množství typů dat. Výsledné KML je možné najít na přiloženém DVD nosiči.



Obr. 37 Ukázka vizualizace četností výskytu onemocnění v agregovaných jednotkách na Ostravsku v prostředí Google Earth (období 1.1.2008—3.2.2009)

Roční incidence kampylobakteriózy v obcích České republiky

Druhou vizualizací bylo zobrazení roční incidence kampylobakteriózy v obcích na 100 tisích obyvatel. V případě velkých měst byla plocha města rozdělena do městských částí. Tato vizualizace nedosahuje z pohledu časového ani prostorového rozlišení podrobnosti bublinového kartodiagramu, ale je vhodná pro sledování a srovnávání meziročních změn incidence vybraných území a hodnocení jejího vývoje. Územní jednotky jsou od sebe rozlišitelné barvou přidělené na základě incidence. Na vzniklé KML se není možno dotazovat, protože byla polygonová vrstva transformována na rastrovou. Důvodem byla velikost výsledného KML v případě exportování vektorových dat. KML bylo vytvořeno exportem z ArcGIS 10.1 využitím extenze *Export to KML*, do KML byla dále manuálně dodána legenda. Ukázka KML incidence je vidět na Obrázku 38. Výsledné KML je možné najít na přiloženém DVD. Při sledování průběhu změny incidence je nejvíce patrný vrchol během roku 2010, který byl zmiňován i v souvislosti s analýzou časových řad kampylobakteriózy v DC1 (kapitola 4.2).



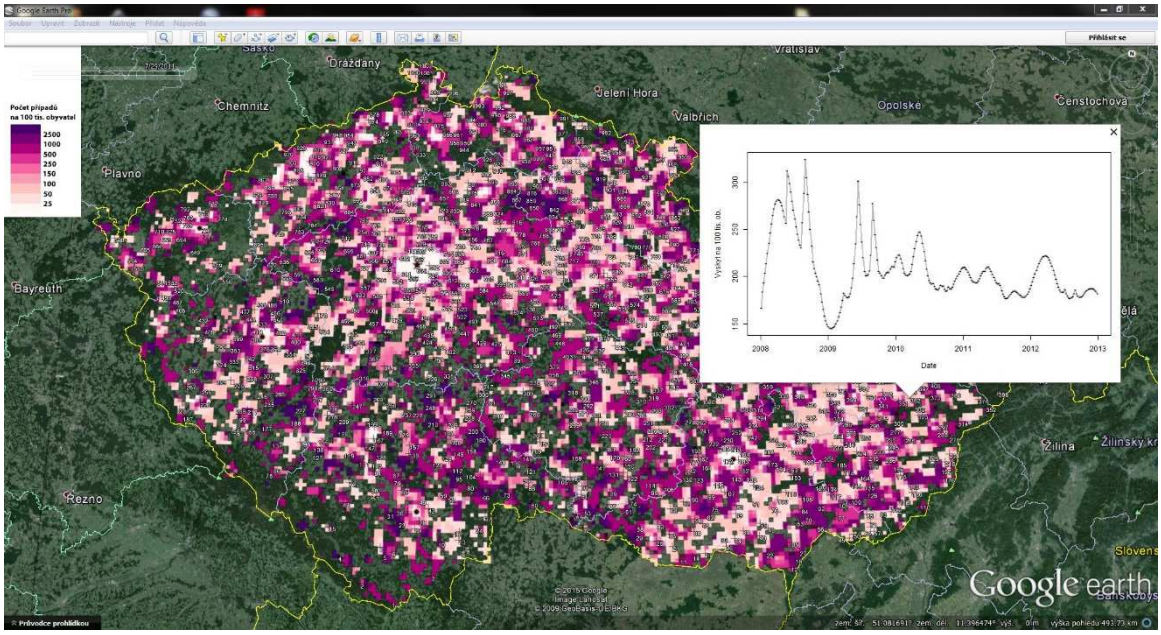
Obr. 38 Roční incidence kampylobakteriózy v obcích/městských částech

Spojitý povrchu týdenní incidence

Geovizualizací dalšího z výsledků DC1 (kapitola 4.4) je spojitý povrch vyjadřující hrubé týdenní incidence na osídlených místech České republiky, který vznikl pomocí časoprostorového krigování agregovaných dat. Tento typ výsledku je, stejně jako agregované počty, velmi vhodný pro zkoumání v prostoru i čase. Jde totiž o 261 časových řezů incidence sestrojených pro celou Českou republiku. Výsledné odhady hrubé incidence pro každý týden v letech 2008—2012 byly před vizualizací rozděleny do deseti kategorií (<25; 25—50;51—100; 101—150; 151—250; 251—500; 501—1,000; 1,001—2,500; >2,500 případů na 100 tisíc obyvatel). Následně byly týdenní časové řezy maskovány osídlenými oblastmi, které pocházely z dat CORINE Land Cover 2006. Takto upravené datové sady byly pomocí **R** balíku *raster* převedeny do podoby tzv. *raster brick* (rastrové kostky) a pomocí *plotKML* vyexportovány do KML. Nevýhodou samotného rastru byla nevhodnost sledování vývoje na jednom místě, zatímco srovnatelnost v rámci jednoho časového řezu bylo možné díky shodné barevnosti. Proto byla do KML doplněna série tisíce náhodně vybraných bodů, které po dotázání zobrazí

jednoduchý graf průběhu (nekategorizované) hrubé incidence v daném místě. Díky tomu byla umožněna geovizuální identifikace časoprostorových vzorů v rámci jednoho časového řezu i skrz více časových řezů a současně jejich srovnání s okolními místy.

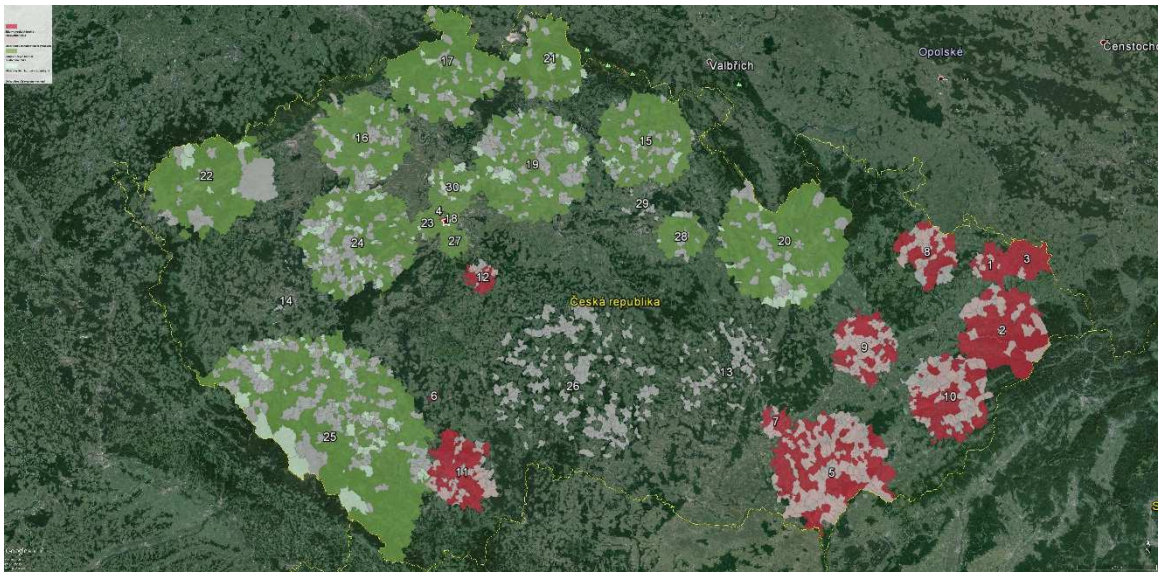
Geovizuální zkoumání vytvořeného povrchu pomohlo k odhalení několika zjištění. Některá z nich potvrdila obecně známá fakta o kampylobakterióze – například jde její o sezónnost s vrcholem výskytu během letních měsíců, která je ovšem méně patrná ve městech (hustěji osídlených obcích). Naopak více patrná je ve venkovských oblastech a také v blízkém okolí měst, které často slouží jako rekreační oblasti pro městské obyvatelstvo. Tento jev je patrný např. na Jičínsku, Bruntálsku či Hodonínsku. Inverzní vrchol onemocnění byl vysledován v podhorských oblastech Krkonoš a Jeseníků, kde je pravděpodobně spojen se zimní turistickou sezónou.



Obr. 39 Roční incidence kampylobakteriózy v obcích/městských částech

Časoprostorového shlukování

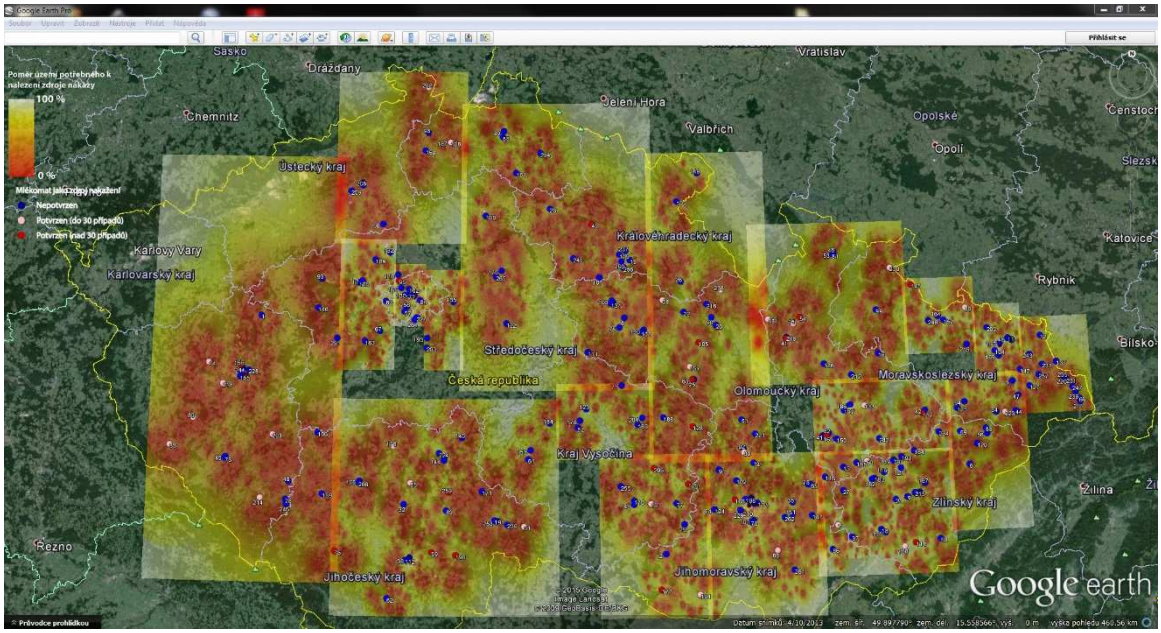
Dalším výsledkem vhodným k zobrazení a zkoumání v čase jsou výsledky DC2. Z nich šlo zejména o výstupy z časoprostorového skenování (kapitola 5.2.1). Vizualizovány byly shluky, které se ve studovaném území nacházely buď po celou sledovanou dobu, nebo měly časově omezenou dobu trvání. Obrázek 23 sice zobrazuje výsledky časoprostorového skenování v mapě, ale nezohledňuje dobu trvání jednotlivých shluků, jak je popsána v Tabulce 8. Doba trvání shluků byla výsledkům časoprostorového skenování vložena do atributové tabulky, a proto byl pomocí extenze *Export to KML* pro ArcGIS 10.1 vygenerován KML soubor, který umožnil zohlednit zobrazení shluků v čase. Vzniklé KML obsahuje polygony hodnocených územních jednotek zbarvené podle příslušnosti k typům shluků (barevnost a kategorie odpovídají mapám v kapitole 5). Kromě toho je možné i dotazování (identifikace) každého polygonu, který obsahuje další informace týkající se morbidity v obci – název obce, počet obyvatel, příslušnost ke shluku, významnost shluku (významné shluky < 0,05), počet pozorovaných případů onemocnění, očekávaný počet onemocnění, relativní riziko (SIR) v územní jednotce i shluku a dobu trvání shluku. Do KML byla manuálně dodána legenda. Ukázku výsledného KML v prostředí Google Earth je na Obrázku 40.



Obr. 40 Časoprostorové shluky kampylobakteriózy v prostředí Google Earth, ve střední části jsou patrné heterogenní části shluků, které se ale ve zobrazovaném období nevyskytovaly

Geoprofily a mlékomaty jako (ne)identifikované potenciální zdroje nákazy

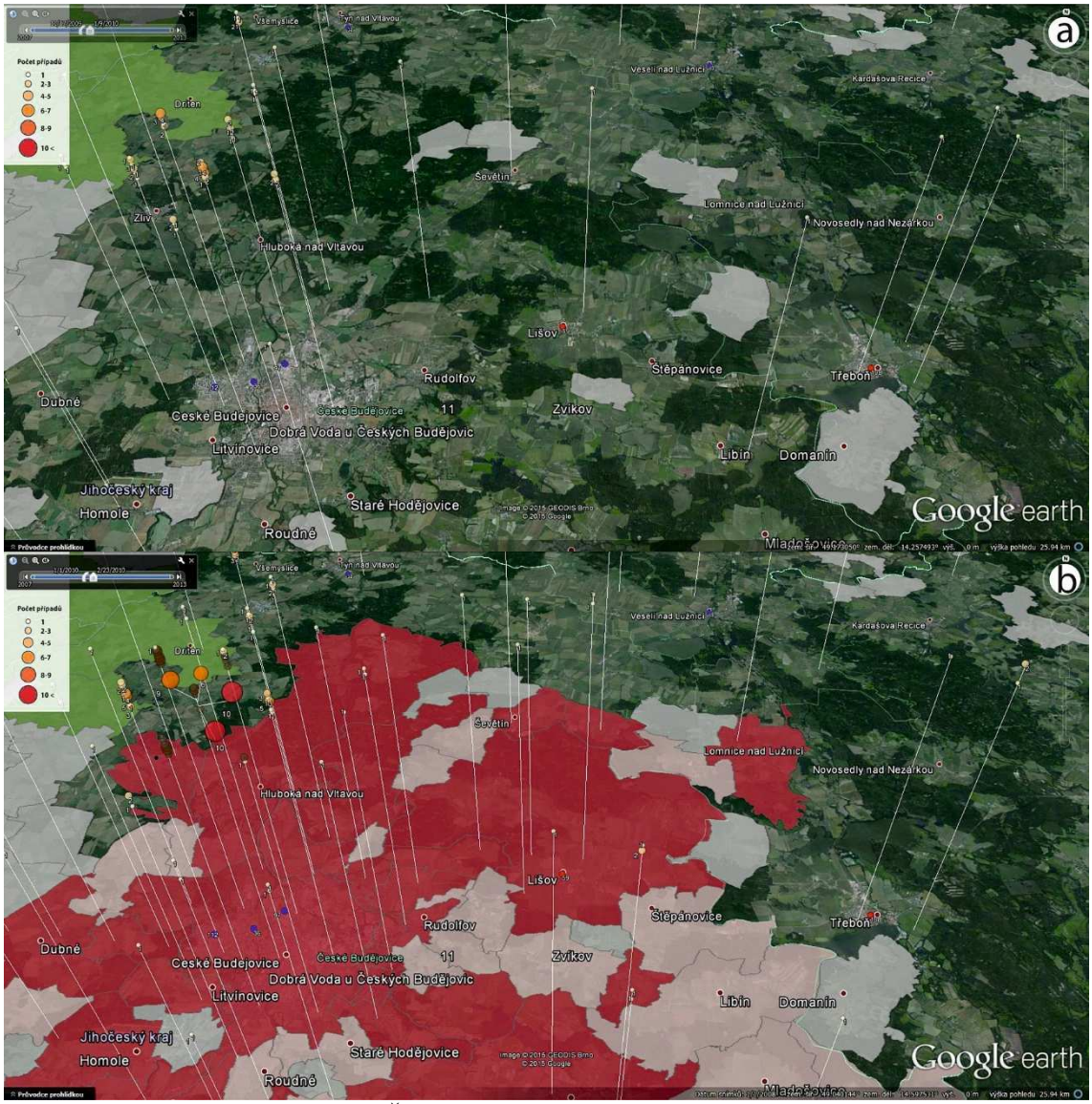
Poslední geovizualizací transformovanou do KML jsou výstupy DC4, kterými jsou geoprofily 18 oblastí České republiky sestavené podle množství případů a přítomnosti mlékomatů na jejich území (viz kapitola 7.2). Geoprofilování je sice metodou prostorovou, která nezohledňuje časovou složku jevu, ale zobrazení jeho výsledků v interaktivním prostředí umožnilo jeho detailnější zkoumání a hodnocení situace v okolí mlékomatů. Tvorba geoprofilů probíhala v prostředí **R**, stejně jako převedení jejich výsledků do podoby KML. Později byla do KML manuálně dodána legenda vysvětlující význam jednotlivých vrstev.



Obr. 41 Geoprofily a mlékomaty jako potenciální zdroje nákazy kampylobakteriózou

Geoprofily jsou zde doplněny mlékomaty, které byly barevně odlišeny podle jejich ohodnocení jako možného zdroje nákazy (Obrázek 41). Tmavě modře jsou zobrazeny mlékomaty, které nebyly vyhodnoceny jako možné zdroje, růžově mlékomaty vyhodnocené

jako potenciální zdroje s počtem případů v okolí do 30 a konečně červeně jsou zobrazeny mlékomaty ohodnocené jako potenciální zdroje nákazy s více než 30 případy v jejich okolí. I když výsledné KML postrádá časovou složku, stále umožňuje kombinaci s dalšími vrstvami a přibližování, a tedy i bližší pohled na vyhodnocovanou situaci. Ve výsledném KML jsou patrné jednotlivé geoprofily a je možné si také všimnou různé úrovně jejich shlazení, které bylo dáno velikostí thinningu během jejich tvorby (detail viz Tabulka 17).



Obr. 42 Zobrazení situace v okolí Českých Budějovic na konci roku 2009 (a) a začátku roku 2010 (b)

Příklad kombinace více vrstev v KML při geovizuálním hodnocení v Google Earth je na Obrázku 42. Hodnocena byla situace v okolí Českých Budějovic během ledna a února 2010, kdy zde došlo ke zvýšenému výskytu kampylobakteriózy, která byla dávána do souvislosti s v té době umisťovanými automaty na čerstvé mléko. Obě části obrázku obsahují stejné vrstvy – mlékomaty identifikované jako potenciální zdroj, počty případů v pravidelné síti a výsledky časoprostorového skenování. Horní část Obrázku 42 (a) představuje situaci v listopadu a prosinci 2009, dolní část (b) popisuje stav v lednu a únoru 2010. Na první pohled je patrný rozdíl. V lednu a únoru byl na Českobudějovicku identifikován shluk obcí potýkajících se se zvýšeným relativním rizikem onemocnění (červené plochy). Rozdíl je vidět také na

bublinovém kartodiagramu představujícím agregované četnosti zaznamenaných případů onemocnění, kde bylo zejména v oblasti Českých Budějovic zaznamenán zvýšený výskyt onemocnění. Pohledem na mlékomaty však nebylo zjištěno, že by mohly být potenciálními zdroji ty českobudějovické, ale paradoxně jimi mohly být mlékomaty v nedalekých obcích Třeboni a Lišově, kde byl ovšem mlékomat instalován až později než vypukla diskuze, která je motivací DC4.

8.3 Shrnutí

Analýza časoprostorových dat často probíhá nezávisle, nejdříve v čase a pak v prostoru (nebo naopak). Představené možnosti vizualizace však integrovaly oba dva důležité aspekty geodat a umožnily jejich vzájemnou (geovizální) interakci. Využitím Google Earth společně s výsledky ve formě KML dodalo analýzám přidanou hodnotu interaktivního prohlížení a jejich vzájemné kombinace. Ačkoliv Google Earth není komplexní platformou schopnou pokrýt všechny kroky nutné pro geovizuální analýzu dat včetně analytických postupů, tak je použitelný pro samotnou geovizuální analýzu a komunikaci výsledků. I přes neoddiskutovatelné výhody však je potřeba kombinovat ho s dalšími nástroji, které připraví zpracovávaná data do podoby vhodné k jejich zobrazení v tomto programu. Jeho významnou výhodou oproti desktopovým a webovým GIS aplikacím je intuitivnost, rozšířenost i mezi neodbornou veřejnost a samozřejmě také možnost interagovat nejen s prostorovou, ale i s časovou složkou dat a tím realizovat jednu z hlavních myšlenek geovizuální analytiky: *„Detekovat čekané a objevovat nečekané“* (Thomas a Cook, 2005; Kraak, 2013). Další výhodou je zobrazování KML, formátu založeném na XML, který je standardem pro skladování a výměnu prostorových dat umožňujícím velkou variabilitou typů vizualizací. V rámci kapitoly bylo pro jeho tvorbu využito jak R, tak i GIS software. Díky využití KML jako základního nástroje geovizualizace bylo dosaženo výsledků, které mohou být podrobeny dalšímu hodnocení. Výsledky DC5 přináší doplňující pohled na problematiku rozšíření onemocnění, která je současně samovysvětlující a atraktivní. A proto je také vhodná nejen pro odborníky, ale i jako forma prezentace výsledků pro veřejnost.

V rámci DC5 bylo vytvořeno pět geovizualizací vycházejících z výsledků předchozích dílčích cílů. Konkrétně jde o (1) nálezová data agregovaná v týdenních intervalech během let 2008—2012 do pravidelné sítě o hraně 2 km vizualizovaná formou tzv. *bubble chart* tedy bublinového kartodiagramu, (2) roční incidence v obcích České republiky, (3) spojitý povrch týdenní hrubé incidence kampylobakteriózy vzniklý pomocí časoprostorového krigingu, (4) časoprostorové shlukování obcí České republiky z pohledu relativního rizika výskytu kampylobakteriózy a (5) geografické profily a mlékomaty jako potenciální zdroj lokálních rozšíření kampylobakteriózy. Všechny výstupy je možné použít buď zvlášť, nebo je možné je společně kombinovat.

9 VÝSLEDKY

Disertační práce představuje příspěvek do jedné z dynamicky se rozvíjejících oblastí výzkumu v prostředí geovědních oborů – geografie zdraví či prostorové epidemiologie. Jak už je z názvu (oboru i samotné práce) patrné jde o interdisciplinární obor, který využívá a kombinuje poznatky, data a metody ze zdravotnictví, geografie i prostorové statistiky. Díky tomuto spojení je disertační práce komplexním souhrnem k tématu prostorové epidemiologie, který se postupně zabývá všemi jejími hlavními směry. Zvolené metody a postupy jsou prezentovány na výzkumné studii, která se týká rozšíření infekčního onemocnění kampylobakteriózy v České republice v letech 2008—2012. Toto onemocnění, ač není mezi veřejností příliš známé, je vůbec nejrozšířenější bakteriální střevní onemocnění nejen v České republice, ale i v celém rozvinutém světě (Weisent et al., 2011).

Hlavním cílem disertační práce bylo provedení komplexní prostorové analýzy epidemiologických dat s využitím v současnosti dostupných technologií z oblasti geoinformatiky a prostorové statistiky. Veškeré analýzy byly provedeny v souladu se standardními metodami prostorové epidemiologie a principy geografických informačních systémů. V rámci hlavního cíle bylo definováno pět na sebe navazujících dílčích cílů, které dohromady umožnily komplexně hodnotit prostorovou distribuci kampylobakteriózy v České republice a její možné podmiňující faktory. Dílčí cíle (DC) byly definovány jako:

- DC1 – mapování a popis charakteristik výskytu kampylobakteriózy v České republice v letech 2008—2012;
- DC2 – průzkum, kvantifikace a vizualizace prostorových a časoprostorových vzorů ve výskytu kampylobakteriózy v České republice v letech 2008—2012 a jejích vlastnostech;
- DC3 – identifikace a analýza možných vztahů mezi výskytem onemocnění a vnějšími environmentálními, demografickými či socioekonomickými faktory a také klasifikace územních jednotek do skupin na základě podobných vlastností a atributových vzorů souvisejících s výskytem onemocnění;
- DC4 – zhodnocení přítomnosti automatů na čerstvé mléko jako potenciálních bodových zdrojů nákazy kampylobakteriózou v jejich okolí;
- DC5 – převedení vybraných výsledků jednotlivých DC1–DC4 do podoby vhodné k další interaktivní exploraci v prostoru i čase.

DC1 – Mapování a popis charakteristik výskytu kampylobakteriózy v České republice v letech 2008—2012

Kompletním řešením prvního dílčího cíle se zabývala kapitola 4, která postupně představila základní úlohy spojené s mapováním nemocí. Základní datovou sadu pocházející z databáze EPIDAT bylo nutné pročistit a zejména geokódovat (kapitola 4.1.1). Ke geokódování byl sestaven skript, který bez omezení umožňuje lokalizaci záznamů na základě API od Mapy.cz. Geokódováno bylo 98,5 tisíc záznamů do úrovně uliční sítě (bez adresních bodů).

Neprostorovým hodnocením průběhu onemocnění a jeho charakteristikami jako celku z pohledu základní statistiky se zabývala kapitola 4.2. V té byla nejdříve zjišťována průměrná

incidence podle věku a pohlaví pacientů. Jednoznačně nejvíce ohroženou skupinou obyvatelstva jsou děti do věku 4 let (chlapci o něco více než dívky), zvýšená incidence onemocnění je patrná i u osob do 30 let nezávisle na pohlaví. S tímto faktem souvisí i nejčastější profese pacientů (dítě/žák/student – dohromady 58,3 % případů). Nejčastější nakaženi jsou zaměstnanci v potravinářství (2,3 %), u 22 % případů však není stav zjištěn. U více než 42 % případů nakažení zůstává neurčen zdroj nákazy, což může výrazně ovlivnit další prostorové i neprostorové analýzy. Nejčastějším identifikovaným zdrojem nákazy je maso – kuřecí (38,5 %), uzeniny (6,22 %), vepřové maso (3,5 %). Kromě toho je možné se nakazit od domácích mazlíčků, z mléka, vody nebo syrové zeleniny a ovoce. Četnost onemocnění (i incidence) stoupala pomalu od roku 2008 až k přelomu roku 2010 a 2011, kdy měla nemoc vrchol a od té doby počet případů klesá. V průběhu roku jsou u všech věkových kategorií nejrizikovějšími měsíci červen—září.

Samotným mapováním kampylobakteriózy se zabývala kapitola 4.3, ve které byly mapovány četnosti, průměrná hrubá a standardizovaná/vyhlazená incidence a relativní riziko (SIR – nepřímo standardizovaná incidence). Nejdříve byly pomocí tečkové metody, anamorfózy a pseudokartogramů sestaveny mapy pro agregované četnosti v hexagonové síti nahrazující územní celky a obce ČR (kapitola 4.3.1). Dále bylo využito pro mapování průměrné míry incidence a relativního rizika (SIR) metod bayesovského shlazování (kapitola 4.3.2), které umožnilo větší srovnatelnost hodnot v územních jednotkách a lepší vnímání prostorového vzoru. Doporučit lze zejména lokální bayesovské vyhlazení v případě incidence, naopak méně lze doporučit vyhlazování v případě SIR. Poslední část mapování nemocí využívá k časoprostorovému mapování nemocí časoprostorový kriging (kapitola 4.4), díky kterému byl vytvořen spojitý povrch hrubé týdenní incidence pro osídlená místa České republiky a pomocí kterého lze sledovat změnu průběhu incidence v průběhu roku a vnímat časoprostorové vzory.

Při sledování prostorové distribuce na základě kartogramů je vizuálně patrná asociace hustoty zalidnění a hustoty případů onemocnění. Výrazně více postižená je oblast Moravy a Slezska (především severovýchod Moravy, Slezsko a Brněnsko), v Čechách jde potom o jižní Čechy, Plzeňsko a oblast jihovýchodně od Prahy (Benešovsko). Při současném sledování času a prostoru s využitím interpolovaného spojitého povrchu je potvrzena sezónnost, která je u tohoto onemocnění běžná. Sezónní změny jsou pak méně patrné ve městech a více patrné v zázemí měst a na venkově, tedy oblastech, které mohu sloužit k rekreaci městského obyvatelstva během letní sezóny. V podhorských oblastech (především Jeseníků a Krkonoš) je možné pozorovat zvýšení incidence i během zimní sezóny příp. jarních prázdnin.

DC2 – Podobnosti výskytu onemocnění v čase a prostoru

Řešením druhého dílčího cíle se zabývala kapitola 5. Zde bylo nejprve pomocí metod analýzy globální a lokální prostorové autokorelace (kapitola 5.1.1) hodnocen čistě prostorový vzor relativního rizika kampylobakteriózy v jednotlivých obcích a městských částech. Tímto postupem byla potvrzena na základě map měr morbidity vizuálně stanovená hypotéza o více nakažených oblastech na východě a severovýchodě Moravy, ve Slezsku a na Českobudějovicku. Kromě toho byla vyhodnocena jako shluk vysokých hodnot i oblast severně od Prahy a Benešovsko. Ve srovnání s mapami není jako významný shluk identifikováno Plzeňsko. Celkem bylo do shluků vysokých hodnot zařazeno 8 %

obcí/městských částí, do shluku nízkých incidencí to bylo 19 %. I když je téměř 2,5× více obcí a městských částí v oblastech shluků nízkých hodnot než v oblastech shluků vysokých hodnot, tak populačně je situace srovnatelná. V oblastech shluků nízkých hodnot žije odhadem 2,27 mil. osob (21,8 % populace) a v oblastech shluků s vysokou incidencí je to dokonce 2,44 mil. osob (23,4 % populace).

Navazujícím postupem bylo hodnocení časoprostorového vzoru pomocí metody časoprostorového skenování (kapitola 5.2.1). Cílem časoprostorového skenování bylo identifikovat a vyhodnotit shluky vysokých a nízkých hodnot relativního rizika, tzn. shluky ohrožených a zdravých obcí jak v prostoru, tak i v čase. Vstupní data se sestávala z případů onemocnění kampylobakteriózou rozdělených dle pohlaví a věku agregovaných prostorově dle jejich příslušnosti k obci/městské a časově v týdenním intervalu. Relativní riziko bylo zjišťováno díky známé demografické struktuře územních jednotek. Jeden shluk zahrnoval maximálně 3 % populace při délce trvání maximálně 50 % či po celé sudované období. Pomocí časoprostorového skenování bylo identifikováno celkem 30 shluků (14 shluků zvýšeného rizika onemocnění kampylobakteriózou a 16 rizika nižšího). Primární shluk byl umístěn do oblasti Ostravska a bezprostředně na něj navazovaly i další shluky zvýšeného rizika. Diverzifikace Čech a Moravy je patrná ještě více než v případě použití prostorového shlukování. Většinu ze shluků je možné pozorovat na místě po celé studované období, ačkoliv několik jich trvalo pouze po omezenou dobu, což je případ Plzeňska, Českých Budějovic, Blanenska, jižní Moravy a Valašska, respektive Královéhradecka a Vysočiny. Celkem žije v identifikovaných shlucích vysokých hodnot až 2,6 mil. osob. Přibližně 25 % obyvatelstva ČR tak žije v oblastech se zvýšeným rizikem nakažení kampylobakteriózou, zatímco 3,9 mil. obyvatel (37 % populace ČR) žije v oblastech, kde je relativní riziko nakažení nižší.

DC3 – Analýza vztahů mezi onemocněním a vnějšími faktory prostředí

Identifikací a analýzou možných vztahů mezi výskytem onemocnění a vnějšími environmentálními, demografickými či socioekonomickými faktory, stejně jako i klasifikací územních jednotek do skupin na základě podobných charakteristik území a charakteristik morbidity v území, se postupně zabývá kapitola 6. V této kapitole bylo nejdříve pomocí korelační analýzy z množství charakteristik území vybráno jedenáct typických charakteristik, které buď mohou podmiňovat prostorovou distribuci onemocnění, nebo mohou vhodně reprezentovat území (kapitola 6.2). Tyto charakteristiky byly podrobeny analýze lokální prostorové korelace s relativním rizikem onemocnění (SIR) a také analýze lokální autokorelace pro dvě proměnné pro zjištění oblastí, kde jsou oba jevy vzájemně asociovány. Nejsilnější vazby byly vyhodnoceny v souvislosti s hustotou zalidnění a průměrnou teplotou vzduchu, dále pak s rozsahem záplavových území a ekonomickými subjekty v oblasti zemědělství a zpracování masa. Pomocí analýzy hlavních komponent (PCA) byla redukována dimenze dat na pět hlavních komponent, zatímco pomocí geograficky vážené PCA byla zkoumána proměnlivost největších zátěží v hlavních komponentách (kapitola 6.3.1). Při nastavení doporučeného adaptivního kernelu byly nejvýznamnějšími zátěžemi hlavních komponent hustota zalidnění, masozpracující ekonomické subjekty, zemědělské subjekty a relativní změna počtu obyvatel.

Možnost predikce kampylobakteriózy, resp. skupin obcí podle relativního rizika onemocnění v obci, byla testována s využitím generalizovaných lineárních modelů a jejich

modifikací (kapitola 6.3.2). Predikční úloha byla díky transformaci predikované proměnné přeformulována na úlohu klasifikační, ke které byly využity diskriminační analýza, metody strojového učení, data miningu a neuronových sítí (kapitola 6.3.3). Pro predikci absolutních četností případů onemocnění v obcích se ukázal jako nejvhodnější negativní binomický model s nadbytečnými nulami. Nejlépe hodnocenými klasifikačními postupy byly metody lokální ordinální regrese, geograficky vážené diskriminační analýzy z metod jednodušších. Z komplexnějších postupů šlo o Random Forest, neuronovou síť a Support Vector Machine s radiální funkcí báze. Klasifikací i regresními postupy se potvrdila skutečnost, že téměř polovina případů kampylobakteriózy u nás i ve světě zůstává nevysvětlena. Koncept lokální ordinální regrese je využíván zcela sporadicky a byl sestaven pro účely disertační práce.

Kromě predikce a klasifikace byla cílem modelování především explorace proměnlivosti vztahů mezi vybranými charakteristikami a výskytem a intenzitou onemocnění. Identifikována byla souvislost mezi hustotou zalidnění, průměrnou teplotou vzduchu a distribucí onemocnění. Lokálně byl také předpokládán vliv socioekonomické determinace a nízkého vzdělání obyvatelstva. Nepříliš výrazný se ukázal efekt obyvatelstva do 15 let.

Posledním úkonem třetího dílčího cíle byla klasifikace obcí České republiky do skupin podle charakteristik prostředí a charakteristik nemocnosti. K tomuto kroku byla využita shluková analýza, která v kombinaci se samoorganizačními mapami identifikovala na území ČR sedm skupin obcí. Tyto skupiny byly popsány a vizualizovány v mapě na Obrázku 34. V pěti ze sedmi skupin jde o pozvolnou proměnu charakteristik a nemocnosti v obcích, ale dvě velmi malé skupiny jsou výrazně odlišné. Potvrzeny byly zjištění z předchozích dílčích cílů, kdy se na severovýchodní Moravě a ve Slezsku vyskytuje více ohrožených obcí než ve zbytku republiky.

DC4 – Zhodnocení přítomnosti automatů na čerstvé mléko jako potenciálních bodových zdrojů nákazy kampylobakteriózou

Čtvrtý dílčí cíl disertační práce měl jako hlavní motiv identifikaci možných bodových zdrojů nákazy obyvatelstva kampylobakteriózou (kapitola 7). Začátkem roku 2010 se spolu s rostoucím počtem automatů na přímý prodej čerstvého mléka rozvinula i diskuze o možných zdravotních problémech, které jeho konzumace může způsobit. Původní velmi kladné přijetí mlékomatů obyvatelstvem bylo utlumeno vyjádřením tehdejšího hlavního hygienika upozorňujícího na tyto problémy (Ministerstvo zdravotnictví ČR, 2010), které mělo dopad na prodej čerstvého mléka z automatů a v důsledku toho i zpomalení růstu množství automatů a u některých i ukončení jejich činnosti (Andrlová, 2011). Hlavním cílem této části disertační práce byla analýza, zda v letech 2008—2012 opravdu mohly některé z mlékomatů být lokálním zdrojem nákazy i přes to, že je kvalita mléka v automatech i u jeho producentů pravidelně kontrolována.

Pro identifikaci mlékomatů jako potenciálního zdroje nákazy byla zvolena metoda geografického profilování (kapitola 7.2), která byla původně využívána zejména v kriminalistice, ale v posledních letech je úspěšně aplikována také v biologii a především v prostorové epidemiologii jako nástroj sloužící k vyhledávání a hodnocení mnohonásobných zdrojů nákazy (Le Comber et al., 2011; Verity et al., 2014). Celkem bylo testováno 267 automatů na prodej čerstvého mléka, které byly umístěny v letech 2008—2012. Samotné vyhodnocování je založeno na Dirichletově modelu pro smíšené procesy a jde tedy o bayesovské modelování

shluků s pomocí Markovových řetězců. Kvůli výpočetní náročnosti bylo území České republiky rozděleno do 18 oblastí podle množství případů kampylobakteriózy v okolí do vzdálenosti 15 km od mlékomatu. Pomocí geografického profilování bylo zjištěno, že až 52 mlékomatů (19,5 %) mohlo být v průběhu svého fungování zdrojem nákazy kampylobakteriózou. Současně byl ovšem zjištěn i nedostatek modelu, kdy byly i některé oblasti s nízkým výskytem případů onemocnění nadhodnocovány jako rizikové. Reálný odhad je tedy nižší a pohybuje se kolem 10 % zkoumaných mlékomatů. Zde je důležité zmínit, že geografické profilování neslouží k potvrzení zdrojů nákazy, ale pouze se snaží o jejich vytipování na základě podobného prostorového chování zkoumaného jevu. To, že byl mlékomat vybrán jako potenciální zdroj lokální nákazy, ještě neznamená, že tímto zdrojem opravdu je. Daná situace samozřejmě může platit i naopak.

Zajímavá byla situace na Českobudějovicku, která byla původcem celé diskuze o možných zdravotních dopadech konzumace čerstvého mléka na obyvatelstvo. Pomocí časoprostorové analýzy vzorů bylo Českobudějovicko opravdu identifikováno jako riziková oblast v období 12. ledna — 22. února 2010, ale přímo v Českých Budějovicích nebyl identifkován žádný mlékomat jako potenciální zdroj nákazy. Jako potenciálně podezřelé ovšem byly identifikovány mlékomaty v blízkém Lišově či Třeboni.

Kromě samotné analýzy byla také modifikována vizualizační část původního skriptu, která nyní lépe reflektuje pravděpodobnostní povrch. Kromě vizualizace výsledků geografického profilování byly vytvořeny také jednoduché mapy hustoty (*heat maps*) výskytu případů kampylobakteriózy a mlékomatů a prodejců čerstvého mléka umožňující rychlé vizuální srovnání intenzity jevu v prostoru (kapitola 7.2).

DC5 – Převedení vybraných výsledků jednotlivých DC1–DC4 do podoby vhodné k další interaktivní exploraci v prostoru i čase

Hlavním úkolem prvního dílčího cíle bylo mapování kampylobakteriózy, zatímco hlavními úkoly DC2—DC4 byla především analýza dat. Pátým dílčím cílem, který z těch předchozích výrazně čerpal, bylo převedení vybraných výsledků jednotlivých dílčích cílů do podoby vhodné k další interaktivní exploraci v prostoru i čase (kapitola 8). Současně byla zvolena forma a podoba prezentace, která umožní další geovizuální zkoumání uživatelům, kteří nemají zkušenosti nebo možnost využít geografických informačních systémů. Vybrané výsledky jednotlivých dílčích cílů byly vizualizovány prostřednictvím formátu KML, který je standardem OGC pro výměnu prostorových dat. Výsledná data je možné bez problémů dále využívat v prostředí GIS nebo pouze využít možnosti zobrazování např. v některém z virtuálních glóbů. Pravděpodobně nejrozšířenější prohlížečkou prostorových dat i dat časoprostorových, která je využívána laiky i odborníky a je využita i v případě této disertační práce, je program Google Earth od společnosti Google. Jeho výhodou je možnost snadno prozkoumávat prostorová data, která obsahují i časovou složku, což bylo předmětem studia několika analýz provedených v rámci disertační práce.

Konkrétně byly v podobě KML resp. KMZ souborů prezentovány tyto geovizualizace a výsledky prostorových analýz:

- nálezová data agregovaná v týdenních intervalech během let 2008—2012 do pravidelné sítě o hraně 2 km vizualizovaná formou tzv. *bubble chart* tedy bublinového kartodiagramu;
- roční incidence v obcích České republiky;
- spojitý povrch týdenní hrubé incidence kampylobakteriózy vzniklý pomocí časoprostorového krigingu;
- časoprostorové shlukování obcí České republiky z pohledu relativního rizika výskytu kampylobakteriózy;
- geografické profily a mlékomaty jako potenciální zdroj lokálních rozšíření kampylobakteriózy.

Vytvořené vizualizace je možné zobrazit ve virtuálním glóbu Google Earth a vizuálně hodnotit prostorovou distribuci rozšíření kampylobakteriózy ve vybraném období. Dále je možné pomocí časového posuvníku či animací srovnávat vývoj během různých časových období a intervalů. Vrstvy je možné prohlížet samostatně nebo je kombinovat pro lepší pochopení jevu.

Google Earth i Google Earth Pro nedávno uvolněný do bezplatné distribuce, byly vyhodnoceny jako použitelné nástroje pro geovizuální analýzu a komunikaci výsledků. I přes nesčetné výhody však je potřeba kombinovat ho s dalšími nástroji, které připraví zpracovávaná data do podoby vhodné k zobrazení a dalšímu geovizuálnímu hodnocení.

10 DISKUZE

Prostorové analýzy epidemiologických dat i obecně dat týkajících se zdraví mají ve světě i u nás dlouhou tradici sahající nejen k anglickému lékaři 19. století Johnu Snowovi, ale i dále. I přesto, že je velké množství informací přehledně zobrazováno formou map a geovizualizací, tak bohužel může mezi odborníky panovat možná nedůvěra k samotným prostorovým analýzám často pramenící jednak z oprávněné opatrnosti při poskytování (individuálních) dat, tak možná i z neznalosti výhod geografie a geoinformatiky a jejich metod pro zpracování prostorových dat.

Poskytování a přístup k dostatečně podrobným záznamům, a nemusí se nutně jednat o záznamy individuální, je pravděpodobně vůbec nejcitlivějším tématem kvůli ochraně osobních údajů, zachování anonymity uživatele a zabránění jeho zpětné identifikace na základě poskytnutých dat. V případě prostorových i neprostorových analýz dat se zdravotní či medicínskou tématikou, a zejména pak u dat týkajících se konkrétních onemocnění, existuje téměř vždy riziko způsobení často neúmyslného poplachu či dokonce zneužití objektivní studie k prosazení osobních zájmů v případě vytržení jednotlivých informací z kontextu celé studie. V současnosti je reidentifikace jedinců z poskytnutých anonymizovaných dat stále větším tématem v souvislosti se státními registry, registry soukromých marketingových společností i např. osobními informacemi, které o sobě uživatelé šíří sami na internetu. Už v roce 2000 bylo ve Spojených státech amerických možné identifikovat jedince na základě data narození, pohlaví a poštovního směrovacího čísla (Sweeney, 2000). Stejná autorka o několik let později dokázala identifikovat jedince včetně jejich záznamů na základě veřejně dostupných profilů (Sweeney et al., 2013).

Časté tvrzení o nedostatku kvalitních a dostatečně podrobných dat z oblasti zdraví nemusí být vždy pravdivé. Otázkou by nemělo být pouze, zda taková data existují, ale především zda a za jakých podmínek jsou data dostupná. Případně, jaká omezení pro jejich využití existují a jak užitečné mohou být výsledky analýz, které z dat vycházejí. Samotnému získání podrobných dat, a zejména pak dat na lokální úrovni, která mohou obsahovat informaci o adresních místech či ulici, předchází často vyjednávání a přesvědčování zainteresovaných institucí o užitečnosti geoinformačních technologií v oblasti zdraví.

V případě, že se povede data získat, tak téměř nikdy nejde o data prostorová, a proto je pro jejich prostorovou analýzu nezbytné, aby byla polohově určena v odpovídající podrobnosti. Požadovaná přesnost polohového určení je také důležitou otázkou a odvíjí se od měřítka prováděné studie. Pokud jde o studii na úrovni okresů či obcí, pak je snadné data přiřadit na základě identifikátorů územních jednotek. Pokud jde o podrobnější měřítko, jako často v případě této disertační práce, pak je nutné data umístit nejčastěji na základě adresy. Data o kampylobakterióze byla umisťována do úrovně uliční sítě bez udání adresních bodů a bylo ke geokódování bylo využita API Mapy.cz. Bylo tedy nutné spoléhat na jeho přesnost a vhodnost umístění bodů. Při vizuální kontrole bylo nalezeno a poupraveno několik nepřesností. Jiným možným způsobem by bylo například srovnávání s adresními body z Registru územní identifikace, adres a nemovitostí.

S polohovým určením dat a jejich agregací souvisí i pojem ekologická chyba. Výsledky vzniklé v určitém měřítku agregace dat často není možné zcela generalizovat na jiná měřítka

či až na individuální úroveň. Výběr prostorové (a časové) jednotky studie je tak jedním z klíčových prvků. Mělo by být voleno měřítko, které je kompromisem mezi využitelností analýzy v lokálním měřítku a ochranou soukromí zúčastněných subjektů. Současně by v případě korelačních studií měly v odpovídajícím měřítku existovat i další doplňková data.

Na měřítku jsou silně závislé výsledky většiny představených analýz. Pro ně bylo nejčastěji zvoleno prostorové rozlišení na úrovni obce či její části. Lze však předpokládat, že stejné analýzy nad stejnými daty, ale v jiném měřítku podá odlišné výsledky. Tento fakt se týká všech dílčích cílů.

Jednotlivé dílčí cíle a s nimi spojené metody jsou sice prezentovány v oddělených kapitolách, v rámci kterých jsou i diskutovány některé jejich aspekty, ale ve skutečnosti jde o sousled navazujících aktivit. Výsledky jedné části (např. DC1) představovaly pracovní hypotézu pro navazující dílčí cíl (např. DC2). Metody byly tedy voleny tak, aby se vzájemně doplňovaly. Zajímavé například může být srovnání hodnocení prostorového shlukování (DC2) a geografického profilování (DC4) v ohrožených lokalitách.

Mapování kampylobakteriózy a vizualizace výsledků

Mapováním výskytu nemocí se nezabýval pouze DC1 (v kapitole 4), ale také DC5 (v kapitole 8). Nejčastěji mapovanými charakteristikami onemocnění bývají obvykle incidence (hrubá či standardizovaná), SIR jako relativní riziko onemocnění a samozřejmě konkrétní četnosti výskytu onemocnění vyjádřené tečkovou metodou případě agregované v územní jednotce nebo pravidelné síti. Tečková metoda sice může pomoci k vizuálnímu vnímání prostorové distribuce onemocnění, ale stejně jako počty případů agregované v územních jednotkách často výrazně kopírují hustotu zalidnění. Proto je všeobecně doporučováno využít relativních měr (např. incidence), které vyjadřují počet případů na množství obyvatel. Rovněž se kvůli srovnávání míst s různou strukturou populace doporučuje využívat standardizovaných měr. Má-li být využita standardizace, opět vyvstává na povrch otázka dat, a to v podobě standardu, který je vhodné využít. Nejčastěji zmiňovaným pojmem v práci je kromě incidence také SIR – index nepřímo standardizované incidence, který může být brán také za ukazatel relativního rizika. Ukazuje srovnání mezi konkrétní morbiditou v oblasti s morbiditou, která by odpovídala standardní populaci zjištěné z celé studované oblasti. Míra je tak snadno pochopitelná. Ideálně by tedy měly být mapovány standardizované míry, ačkoliv některé studie týkající se SIR ukázaly, že srovnání hodnot mezi geografickými jednotkami bude zavádějící pouze v případě extrémně odlišných populací, což se např. v rámci jednoho státu či oblasti děje v praxi velmi zřídka (Goldman a Brender, 2000; Jarup, 2004).

Dalším významným úkonem při mapování a prezentaci měr nemocnosti je vyhlazování dat. Tento proces, zejména v případě lokálního vyhlazování, přispívá ke srovnatelnosti mezi jednotlivými územními jednotkami a také umožňuje odhalit a zvýraznit trend ve sledovaném území představovaný změnami ve střední hodnotě (Horák, 2011). V kapitole 4.3.2 byly představeny tři způsoby zobrazení měr morbidity, ze kterých lze doporučit lokální bayesovské vyhlazování. S pojmem lokální je ovšem spojena i prostorová nejistota ve smyslu zvolení vhodného sousedství tak, aby byla zachována lokální proměnlivost jevu a současně byly hodnoty shlazeny. Při využití bayesovského vyhlazování je potřeba mít na paměti, že tyto metody mají tendenci posouvat hodnoty blíže k lokálnímu či globálnímu průměru. Míry

v oblastech s větším množstvím informace (např. stabilnější odhady v hustě zalidněných regionech) jsou tak shlazeny méně než v oblastech s vyšším rozptylem hodnot (a menším množstvím vzorků - méně zalidněné oblasti) (Richardson et al., 2004). Ačkoliv se mohou mapy hrubých měr zdát příliš fragmentované a náročné k interpretaci, tak na druhou stranu vyhlazené míry mohou být příliš homogenní, a proto maskovat skutečnou distribuci jevu (Beale et al., 2008).

Využití krigingu (kapitola 4.4.2) v disertační práci nebylo z důvodu predikce, tak jak je to u této metody běžné, ale představuje alternativu k algoritmům dasymetrických metod. Vzhledem k tomu, že kriging zde slouží jako prostředek k vytvoření povrchu sloužícího spíše k vizuálnímu zkoumání než k dalším analýzám, tak nebylo příliš nutné představovat hodnocení metody z pohledu přesnosti a nejistoty predikce. K vytvoření spojitého povrchu týdenní hrubé incidence v osídlených místech České republiky byla zvolena jedna z nejnovějších implementací krigingu, a to časoprostorový kriging, který dokázal odhadovat vztahy mezi jednotlivými místy v prostoru i čase. Ústupkem za tuto výhodu však byla výpočetní a časová náročnost metody, kdy výpočet variogramu a průběh interpolace trval přibližně 50 hodin.

K metodám mapování kampylobakteriózy lze přiřadit také KML výstupy, které vznikly jako součást DC5 (kapitola 8). Zde byl program Google Earth zvolen jako médium pro další zkoumání výsledných analýz a také jako prostředek ke komunikaci výsledků. I přes četné výhody tohoto programu (rozšířenost, intuitivita, kvalitní podkladová data) je potřeba zdůraznit, že jde především o prohlížečku geografických dat a všechna data tak musela být vhodně připravena dříve, než byla postoupena geovizuální analýze. Soubory ve formátu KML se jeví jako vhodné nosné médium, jejich problémem však může být velikost souborů především v případě, kdy tyto obsahují velké množství vektorových dat, což má za následek pomalé načítání. Tvorba KML také často není tak přímočará, jak by si uživatel představoval. Vytvořené soubory vzniklé díky transformaci v některém z GIS programů či **R**, bylo nutné ještě externě doladit např. přidáním legendy, úpravou kódování apod.

Prostorové a časoprostorové vzory kampylobakteriózy

Analýza prostorových a časoprostorových vzorů onemocnění, která je obsahem DC2 a v podstatě také DC5, je jedním ze základních výzkumných témat prostorové epidemiologie. Existuje totiž všeobecný předpoklad, že je často možné nalézt prostorový vzor přírodních i socioekonomických jevů, který se přirozeně v objevuje. Někdy však nemusí být snadné tento vzor nalézt, a tak je potřeba využít sofistikovanějších metod. Je třeba přiznat, že vzor v prostoru i čase často opravdu existuje, je ale také potřeba si uvědomit, že některé metody ho mohou odhalit i tam, kde existovat nemůže, např. v náhodně vygenerovaných prostorových datech.

Pro exploraci prostorové distribuce kampylobakteriózy v ČR v letech 2008—2012 byly použity metody průzkumu prostorové autokorelace. Nejdříve byla hodnocena prostorová složka pomocí metod LISA (kapitola 5.1) a poté kombinovaný časoprostorový vzor pomocí časoprostorového skenování (kapitola 5.2). Pro úspěšné provedení obou metod je potřeba vhodně zvolit matici prostorových vztahů/sousedství pro jednotlivé územní jednotky. Současně nebyly do výpočtu zahrnuty ani další související charakteristiky mimo demografickou strukturu populace (věk/pohlaví). Tato nevýhoda teoreticky může být

odstraněna pomocí zahrnutí prostorových regresních bayesovských odhadů, které mohou zvýšit robustnost odhadů (Li et al., 2012). Pro lepší odhad inferencí byly metody prostorového shlukování randomizovány.

V případě časoprostorového skenování je kromě prostorových parametrů nutná také volba parametrů časových a maximálních velikostí shluku. Kromě v disertační práci prezentovaného nastavení byly testovány i další vstupní kombinace parametrů populace (3 %, 5 %, 10 % a 50 %), délky trvání shluků (30 dní, 105 dní and 50 % časového období) a také zpracování časových trendů. Výsledky se ovšem významně nelišily. Pouze se zvyšující se maximální populací shluků se snižoval jejich počet. Primární shluk byl vždy identifikován na stejném místě. Hodnoty tak byly zvoleny empiricky s přihlédnutím k zahraničním studiím a obecným doporučením.

K metodám prostorového shlukování lze přiřadit geografické profilování (geoprofiling). Jde o metodu, která byla původně využívána zejména v kriminalistice a představuje geografickou obdobu psychologických profilů. Předpokládá, že osoba či zdroj nákazy se drží podobného geografického chování, které tedy lze mapovat. Na základě geoprofilingu tak je možné ze sady kandidátních míst vytipovat možný zdroj lokální nákazy. Do metody nejdříve vstupují lokality událostí, na základě Dirichletova modelu pro smíšené procesy je sestaven pravděpodobnostní povrch, na základě kterého jsou hodnoceny potenciální zdroje. Metoda pracuje pouze s pozicí případů onemocnění a nikoliv vlastnostmi nakažené osoby nebo dokonce s časem. Možné rozdílnosti ve výsledcích mohou být způsobeny změnami v nastavení simulace nebo definování vzdálenosti, kde dochází k prostorové autokorelaci. Ta je totiž definována empiricky na základě grafu simulací a může docházek k značně subjektivnímu odhadu. Chybným předpokladem může teoreticky být i předpoklad popsatelného prostorového chování obyvatelstva. Současně je vhodné připomenout prostorovou nejistotu spojenou s geokódováním záznamů, kdy je známa pouze ulice a nikoliv přímo adresní bod nakažené osoby. Samotné téma mlékomatů bylo kontroverzním zejména v roce 2010, kdy bylo zdrojem sporů mezi Agrární komorou ČR a hlavním hygienikem. Geoprofiling identifikoval téměř 20 % mlékomatů jako potenciálních zdrojů nákazy. Po dalším zhodnocení výsledků je však možné počet snížit na 9 %, protože metoda zvýhodňovala některé oblasti s velmi nízkým počtem nakažení. To, že byl mlékomat vybrán jako potenciální zdroj lokální nákazy, však ještě neznamená, že tímto zdrojem opravdu je. Samozřejmě může dojít i k opačné situaci, kdy není zdroj nákazy identifikován.

Modelování, klasifikace a analýza vztahů mezi onemocněním a faktory prostředí

V rámci DC2 byla identifikována místa s vyšší incidencí/relativním rizikem kampylobakteriózy, která se vyskytovala především v oblasti severovýchodní Moravy a Slezska. Stejně tak byla odhalena i místa, která nejsou tímto onemocněním tolik postižená. Pokud bylo možné nalézt ve studovaném období prostorový a časoprostorový vzor, pak existovala i možnost nalezení faktoru podmiňujícího tento výskyt. To je možné hodnotit pomocí metod prostorové korelace, regresních modelů a případně i klasifikátorů, což je kolem DC3 v kapitole 6.

První nejsitotou, která mohla ovlivnit predikční výkon odhadu chování kampylobakteriózy pomocí zmíněných postupů, byla volba prediktorů. Sice byla vytvořena datová sada o velkém množství socioekonomických, demografických i environmentálních

charakteristik obcí, ale z nich bylo vybráno pouze jedenáct vlastností reprezentujících všechny ostatní. Tyto vlastnosti byly dále zredukovány pomocí analýzy hlavních komponent. Většinu z dat bylo možné agregovat přímo do obce, ale například množství drůbeže bylo odhadnuto z hodnot okresů na základě počtu zemědělských podniků. V datech tak zůstává množství nejistoty.

Otázkou však je, zda by se při volbě jiných modelů či jiných proměnných by ukázala výrazná asociace neodhalená vybranými charakteristikami nebo zda je výskyt kampylobakteriózy dán především hustotou osídlení a demografickou strukturou spíše než charakteristikami území. Modelování a klasifikace na základě vybraných charakteristik nebylo výrazně úspěšné z pohledu predikce, ale bylo užitečným postupem pro exploraci a inference jednotlivých charakteristik.

Odhad přesné hodnoty průměrné incidence se už ze začátku zdál jako nereálný. Z toho důvodu byla regresní úloha transformována na úlohu klasifikační zařazující obce rozdělené do skupin dle SIR. Ačkoliv byly testovány různé prostorové i neprostorové techniky, tak stále existuje množství metod a přístupů, které by bylo možné vyzkoušet. Zmínit je možné například prostorové bayesovské modely, smíšené modely, modifikace geograficky vážených modelů nebo zahrnutí interakcí mezi jednotlivými charakteristikami obcí. Možností by bylo také zkoumání charakteristik jednotlivých demografických skupin obyvatelstva či ročních období. Některé využité modely nevyužívaly ordinality původních dat a jde tak spíše o multinomickou regresi. Je vhodné také zmínit, že prostorové metody často nemají výrazné globální nároky na data (týkající se např. homogenity či normality dat), protože je možné, že lokálně se tyto vlastnosti výrazně proměňují. Nároky na stacionaritu dat jsou často mírně přehlíženy či eliminovány vhodným interakčním schématem (sousedstvím).

Pro účely DC3 byl sestaven fungující koncept lokálního ordinálního modelu, který byl testován pro klasifikaci a ve srovnání s tradičními metodami prokazoval dobrý výkon. Jeho nevýhodou je ale komplikace hodnocení regresních koeficientů a dalších charakteristik vzhledem k faktu, že nejde o jeden model, ale o souhrn velkého množství modelů – každá obec má svůj model s vlastními koeficienty vycházející z definice prostorových vztahů mezi obcemi.

Při shlukování obcí na základě podobných socioekonomických a environmentálních charakteristik je využito, stejně jako v případě dalších vícerozměrných metod, subjektivního stanovování a intepretací. Počet skupin byl sice zvolen na základě vnitroskupinového rozptylu a simulace shlukování, ale většinou je možné využít i jiné kritérium často s jiným výsledkem. Pět ze sedmi skupin je pozvolně se měnících, dvě nejmenší jsou velmi výrazné a objevovaly se v hodnocení i při jiných nastaveních, kde se velké skupiny buď dále dělily, nebo spojovaly. Tradiční postup shlukování je obohacen o využití samoorganizačních struktur, které by měly přispět k lepšímu klasifikačnímu výkonu a snadnější interpretaci výsledných skupin. Matice podobnosti obcí nebyla záměrně prostorově vážena, aby nebyla zakryta lokální proměnlivost. Skupiny byly hodnoceny na základě jejich středních hodnot a směrodatných odchylek a v jednotlivých skupinách se tak může vyskytovat velký rozptyl hodnot.

11 ZÁVĚR

Hlavním cílem disertační práce sice provedení komplexní prostorové analýzy epidemiologických dat s využitím v současnosti dostupných technologií z oblasti geoinformatiky a prostorové statistiky. Vedlejším cílem disertační práce však bylo poskytnout ucelený materiál, který pomůže dalším podobným studiím v orientaci v tématu prostorové epidemiologie a jejích metod. Disertační práce provede zájemce hlavními postupy, se kterými se v rámci studia prostorové distribuce epidemiologických dat může setkat. Téma disertační práce a její interdisciplinární povaha předpokládaly, že pro její provedení bylo nutné kombinovat poznatky nabyté studiem geoinformatiky spolu s poznatky dalších oborů jako prostorová a aplikovaná statistika či epidemiologie.

Hlavní cíl disertační práce byl pro účely zpracování rozdělen do pěti na sebe navazujících dílčích cílů. Ty společně pokrývaly ústřední témata, jimiž se zabývá prostorová epidemiologie - mapování nemocí, identifikace prostorových shluků a geografické korelační studie. Mapováním a vizualizací výsledků analýz se zabývaly první a pátý dílčí cíl. V prvním případě byla mapována prostorová distribuce kampylobakteriózy – její četnost, incidence a také relativní riziko. Kromě toho byly doporučeny i metody vyhlazování, které v jednotlivých případech využít. Využitím metody časoprostorového krigingu jako nástroje časoprostorového dasymetrického mapování bylo docíleno možnosti sledování průběhu onemocnění v prostoru i čase současně formou spojitého povrchu. Právě využití časoprostorového krigingu je jednou z inovací, kterou disertační práce vnáší do tématu mapování morbidity. Úspěšně bylo také využito geovizualizací ve formě KML souborů možných zobrazit v prostředí geografických informačních systémů i ve virtuálním glóbu Google Earth. Google Earth v rámci práce představuje platformu pro geovizuální časoprostorové hodnocení předem připravených témat. Kromě toho se díky rozšíření tohoto programu jedná o vhodný způsob další komunikace výsledků laické i odborné veřejnosti.

Druhým ústředním tématem bylo zkoumání prostorových a časoprostorových vzorů s důrazem na identifikaci shluků v rozložení kampylobakteriózy v České republice během let 2008—2012. Nejdříve byly v rámci druhého dílčího cíle hodnoceny předpoklady stanovené v prvním dílčím cíli na základě vizuálního hodnocení. Bylo potvrzeno, že nejvíce onemocněním ohrožená je oblast Ostravska a dále severovýchodní Moravy a na Benešovsku, kde byly identifikovány shluky vysokých měr morbidity v prostoru i v čase. Z dalších míst tomu tak bylo po omezené časové období na Plzeňsku a Českobudějovicku. Hodnocením prostorového vzoru konkrétních případů se zabývá ve spojitosti s identifikací možných bodových zdrojů nákazy čtvrtý dílčí cíl. Zde byl pomocí metody geografického profilování hodnocen vztah mezi umístěním automatů na prodej čerstvého mléka a výskytem případů kampylobakteriózy v jejich okolí. Toto téma bylo velmi aktuální během roku 2010, kdy vznikl spor mezi hlavním hygienikem ČR a zástupci Agrární komory ČR, který vznikl na základě upozornění na množství případů v kampylobakteriózy v Českých Budějovicích. Na základě časoprostorového skenování byl opravdu identifikován shluk zvýšeného relativního rizika onemocnění na Českobudějovicku v lednu a únoru 2010. Nicméně jako potenciální zdroj nákazy byly ohodnoceny mlékomaty v nedalekých obcích a nikoliv přímo v Českých Budějovicích identifikován žádný z přítomných automatů jako potenciální zdroj lokální epidemie. Kromě toho bylo označeno za potenciální zdroje nákazy téměř 20 % mlékomatů.

V úvahu však musí být bráno nadhodnocování významu některých případů a množství tedy může být redukováno na zhruba 10 %.

Třetím hlavním tématem byla analýza možných vztahů mezi výskytem onemocnění a vnějšími faktory prostředí. Vzhledem k tomu, že byly identifikovány a popsány prostorové shluky onemocnění, tak existoval předpoklad existence možných podmiňujících faktorů prostředí. Tento předpoklad se ovšem potvrdil pouze částečně, kdy byly jako hlavní faktory ohodnoceny zejména demografie obyvatelstva s možnými lokálními vlivy teploty vzduchu, socioekonomické deprivace a zemědělství. Predikční schopnost sestavených modelů sice nepřekračuje 50 %, ale je třeba konstatovat, že zdroj a způsob nakažení pouze zhruba poloviny případů u nás i ve světě je uspokojivě identifikováno. Na použité postupy a modely může být nahlíženo z pohledu zkoumání lokálních inferencí, kde prokázaly svou užitečnost. V rámci tohoto tématu byl sestaven koncept lokální ordinální logistické regrese jako modifikace ordinální logistické regrese pro využití s prostorovými daty.

Pokud by měly být zmíněny hlavní přínosy práce, pak je to využití časoprostorového krigingu v mapování onemocnění, časoprostorového skenování a geografického profilování při zkoumání prostorových vzorů a využití geograficky vážených metod pro analýzu asociací mezi výskytem onemocnění a lokálními charakteristikami. Kromě samotných metod je to i měřítko případové studie – kdy šlo o obce či jejich části na území celé České republiky.

Zkoumání zdraví a faktorů zdraví ovlivňujících je jedním z velmi aktuálních témat nejen v prostředí geověd, ale také v rámci širokého spektra dalších vědních oborů od medicíny po fyzikální vědy. V České republice je však často vidět disparita mezi lékařskými individuálními studiemi a studiemi v dalších prostorových podrobnostech, která může být způsobena jednak ochranou osobních údajů a s ní spojeným malým množstvím veřejně dostupných dat v lokálním měřítku, a pak také možnou nedůvěrou ve schopnosti a přínos geografických a prostorových metod a analýz. Je však nutno zmínit, že situace se v současnosti zlepšuje díky aktivitám výzkumných týmů na univerzitních pracovištích i aktivitám zainteresovaných státních institucí.

POUŽITÁ LITERATURA A INFORMAČNÍ ZDROJE

AAMODT, G. et al. (2006): A simulation study of three methods for detecting disease clusters. International journal of health geographics, 5, s. 15.

AGRÁRNÍ KOMORA ČR (2010): Strašák zvaný automat na čerstvé mléko [online]. Dostupné z: http://www.apic-ak.cz/strasak-zvany-automat-na-cerstve-mleko.php

ALEXANDER, G. L. et al. (2003): Marginalization and health geomatics. Journal of Biomedical Informatics, 36, č. 4-5, s. 400–407.

AMBROŽOVÁ, H. (2011): Letní průjmy. Medicína pro praxi, 8, č. 5, s. 7–9.

ANDRLOVÁ, L. (2011): Perspektivy prodeje (bio)mléka v prodejních automatech. Diplomová práce. Jihočeská univerzita v Českých Budějovicích, 107 s.

ANSELIN, L. (1988): Spatial econometrics: methods and models. Kluwer Academic, Dordrecht, 284 s.

ANSELIN, L. (1994): Exploratory spatial data analysis and geographic information systems. New tools for spatial analysis.

ANSELIN, L. (1995): Local indicators of spatial association—LISA. Geographical analysis, 27, č. 2.

ANSELIN, L. (1996): The Moran scatterplot as an ESDA tool to assess local instability in spatial association. Spatial analytical perspectives on GIS.

ANSELIN, L. et al. (2002): Visualizing multivariate spatial correlation with dynamically linked windows. In: Anselin, L. a Rey, S. (ed.): New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting. Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara, s. 1–20.

ANSELIN, L. (2003): GeoDa™ 0.9 User's Guide.

ARAGON, T. J. (2012): epitools: Epidemiology Tools. Verze 0.5-7. Dostupné z: http://cran.r-project.org/package=epitools

ARCDATA PRAHA (2013): Geoportál Koordinačního střediska pro resortní zdravotnické informační systémy [online]. Dostupné z: http://www.arcdata.cz/digitalAssets/323152_case_study_ksrzis.pdf

ARENTZE, T. A. (2009): Spatial Data Mining, Cluster and Pattern Recognition. In: Kitchin, R. a Thrift, N. (ed.): International Encyclopedia of Human Geography. Elsevier, Oxford, s. 325–331.

ARMITAGE, P. et al. (2008): Statistical methods in medical research. Blackwell Science Ltd., Oxford, UK.

ARMSTRONG, M. P. et al. (1999): Geographically masking health data to preserve confidentiality. Statistics in medicine, 18, č. 5, s. 497–525.

ARSENAULT, J. (2010): Épidémiologie spatiale de la campylobactériose au Québec. Diplomová práce. Université de Montréal,

ARSENAULT, J. et al. (2012): Environmental and demographic risk factors for campylobacteriosis: do various geographical scales tell the same story? BMC infectious diseases, 12, s. 318.

ARSENAULT, J. et al. (2013): How to choose geographical units in ecological studies: Proposal and application to campylobacteriosis. Spatial and spatio-temporal epidemiology, 7, s. 11–24.

ASSUNCAO, R., REIS, E. (1999): A new proposal to adjust Moran's I for population density. Statistics in medicine, 2162, č. November 1998, s. 2147–2162.

BADDELEY, A. (2010): Analysing spatial point patterns in R.

BAILEY, T. (2001): Spatial statistical methods in health. Cadernos de Saúde Pública, 17, č. 5, s. 1083–1098.

BAILEY, T. C., GATRELL, A. C. (1995): Interactive spatial data analysis. Longman Scientific & Technical, Essex.

BARDON, J. et al. (2009): Prevalence of Campylobacter jejuni and its resistance to antibiotics in poultry in the Czech Republic. Zoonoses and Public Health, 56, č. 3, s. 111–116.

BEALE, L. et al. (2008): Methodologic issues and approaches to spatial epidemiology. Environmental health perspectives, 116, č. 8, s. 1105–10.

BELL, B. S. et al. (2006): Current practices in spatial analysis of cancer data: mapping health statistics to inform policymakers and the public. International journal of health geographics, 5, s. 49.

BENCKO, V. et al. (2003): Statistické metody v epidemiologii. Nakladatelství Karolinum, Praha, 505 s.

BENEŠ, V. et al. (2005): A case study on point process modelling in disease mapping. 24, s. 159–168.

BERGQUIST, R. (2011): New tools for epidemiology: a space odyssey. Memórias do Instituto Oswaldo Cruz, 106, č. 7, s. 892–900.

BERNARD, J. et al. (2014): Existují prostorové kontextové vlivy na volební chování i v relativně nacionalizovaném stranickém systému? Příklad Česka. Geografie - sborník České geografické společnosti, 119, č. 3, s. 240–258.

BIVAND, R. S. (2007): Analysing Spatial Data in R : Worked example : point patterns.

BIVAND, R. S. et al. (2008): Applied Spatial Data Analysis with R. Springer New York, New York, NY.

BIVAND, R. S. et al. (2014): rgdal: Bindings for the Geospatial Data Abstraction Library. Verze 0.9-1. Dostupné z: http://cran.r-project.org/package=rgdal

BLAŽEK, J., NETRDOVA, P. (2012): Aktuální tendence lokální diferenciace vybraných socioekonomických jevů v Česku: Směřuje vývoj k větší mozaikovitosti uspořádání? Geografie - sborník České geografické společnosti, 117, č. 3, s. 266–288.

BOBAK, M. et al. (1998): Socioeconomic factors, perceived control and self-reported health in Russia. A cross-sectional survey. Social Science and Medicine, 47, č. 2, s. 269–279.

BREIMAN, L. (1984): Classification and regression trees. CRC Press.

BRUNSDON, C. et al. (2007): Geographically weighted discriminant analysis. Geographical Analysis, 39, s. 376–396.

BUSCEMA, M. et al. (2009): Outbreaks source: A new mathematical approach to identify their possible location. Physica A: Statistical Mechanics and its Applications, 388, č. 22, s. 4736–4762.

CANNON, A. J. (2012): monmlp: Monotone multi-layer perceptron neural network. Verze 1.1.2. Dostupné z: http://cran.r-project.org/package=monmlp

CARVALHO, A. et al. (2009): Spatial Hierarchical clustering. Rev. Bras. Biom., 27, č. 3, s. 411–442.

CASAS, I. (2009): Neural networks. In: Kitchin, R. a Thrift, N. (ed.): International Encyclopedia of Human Geography. Elsevier, Oxford, s. 419–422.

CÍCHA, V. (2013): Správa, analýza a prezentace zdravotnických prostorových dat pomocí R. Diplomová práce. Univerzita Palackého v Olomouci, 76 s.

CLAYTON, D., BERNARDINELLI, L. (1996): Bayesian methods for mapping disease risk. In: Elliott, P. et al. (ed.): Geographical and Environmental Epidemiology: Methods for Small Area Studies. Oxford University Press, Oxford.

CLIFF, A. D., ORD, K. J. (1973): Spatial autocorrelation. Pion Ltd, London, London, 178 s.

CRESSIE, N. A. C. (1993): Statistics for Spatial Data. John Wiley & Sons, New York.

CROMLEY, E. K. (2003): GIS and disease. Annual review of public health, 24, s. 7–24.

ČESKÁ REPUBLIKA (2015): Zákon č. 89/1995 Sb., o státní statistické službě, ve znění pozdějších předpisů, s účinností do 1. 1. 2015. Sbírka zákonů České republiky, s. 40.

ČESKÝ ROZHLAS (2010): Agrární komora zvažuje trestní oznámení na hlavního hygienika [online]. Dostupné z: http://www.rozhlas.cz/zpravy/politika/_zprava/728204

ČSÚ (2015): Český statistický úřad [online]. Dostupné z: https://www.czso.cz/

DA SILVA, A. R., RODRIGUES, T. C. V. (2014): Geographically Weighted Negative Binomial Regression-incorporating overdispersion. Statistics and Computing, 24, s. 769–783.

DAVENHALL, B. (2012): Geomedicine: Geography and Personal Health. Esri, Redlands, 33 s.

DELMELLE, E. et al. (2010): H.E.L.P: A GIS-based Health Exploratory AnaLysis Tool for Practitioners. Applied Spatial Analysis and Policy, 4, č. 2, s. 113–137.

DEMOGRAFIE (2014): Ukazatele nemocnosti [online]. Dostupné z: http://www.demografie.info/?cz_nemocnostukazatele=

DOMASOVÁ, I. (2014): Kampylobakterióza. Praha, 2 s.

DORLING, D. (1998): Mapping disease patterns. In: Encyclopedia of Biostatistics.

DVORSKÝ, J., DRAŽILOVÁ, P. (2011): Neuronové sítě. In: Voženílek, V. et al. (ed.): Metody umělé inteligence v geoinformatice. Univerzita Palackého v Olomouci, Olomouc, s. 5–17.

DZÚROVÁ, D. et al. (2006): Demographic and Social Correlates of Suicide in the Czech Republic. Sociologický časopis, 42, č. 3, s. 557–571.

DZÚROVÁ, D. et al. (2010): Social inequalities in alcohol consumption in the Czech Republic: a multilevel analysis. Health & place, 16, č. 3, s. 590–7.

EISEN, L., LOZANO-FUENTES, S. (2009): Use of mapping and spatial and space-time modeling approaches in operational control of Aedes aegypti and dengue. PLoS neglected tropical diseases, 3, č. 4, s. e411.

EKDAHL, K. et al. (2005): Could flies explain the elusive epidemiology of campylobacteriosis? BMC infectious diseases, 5, s. 11.

ELLIOTT, P. et al. (2000): Spatial Epidemiology: Methods and Applications. Oxford University Press, 504 s.

ELLIOTT, P., BEST, N. (1998): Geographic patterns of disease. Encyclopedia of biostatistics.

ELLIOTT, P., WARTENBERG, D. (2004): Spatial Epidemiology: Current Approaches and Future Challenges. Environmental Health Perspectives, 112, č. 9, s. 998–1006.

ESRI (2012): ArcGIS Desktop: Release 10.1. Environmental Systems Research Institute., Redlands, CA.

ESTER, M. et al. (1997): Spatial data mining: A database approach. In: Proceedings of the Fifth Int. Symposium on Large Spatial Databases.

EUROSTAT (2009): Health Statistics: Atlas on Mortality in the European Union.

EUROSTAT (2014): Eurostat Regional Statistics Illustrated [online]. Dostupné z: http://epp.eurostat.ec.europa.eu/cache/RSI/#?vis=nuts2.health

EVERITT, B., HOTHORN, T. (2011): An Introduction to Applied Multivariate Analysis with R. Springer, Heidelberg.

F. DORMANN, C. et al. (2007): Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, 30, č. 5, s. 609–628.

FAWCETT, T. (2006): An introduction to ROC analysis. Pattern Recognition Letters, 27, č. 8, s. 861–874.

FENG, C. (2011): Models and Methods For Spatial Data: Applications in Epidemiological, Environmental and Ecological Studies. Diplomová práce. Simon Frase University, 160 s.

FOTHERINGHAM, A. S. et al. (2000): Quantitative geography: Perspectives on spatial data analysis. Sage, London, 270 s.

FOTHERINGHAM, A. S. et al. (2002): Geographically Weighted Regression: the analysis of spatially varying relationships. John Wiley & Sons.

FRIEDMAN, J. H. (2001): Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29, č. 5, s. 1189–1232.

FRIEDMAN, J. H. et al. (2010): Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 36, č. 1, s. 1–22.

GABRIEL, E. et al. (2010): Spatio-temporal epidemiology of Campylobacter jejuni enteritis, in an area of Northwest England, 2000-2002. Epidemiology and infection, 138, č. 10, s. 1384–90.

GASTNER, M. T., NEWMAN, M. E. J. (2004): From The Cover: Diffusion-based method for producing density-equalizing maps. Proceedings of the National Academy of Sciences of the United States of America, 101, č. 20, s. 7499–7504.

GATRELL, A. A. C. et al. (1996): Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology. Transactions of the Institute of British Geographers, 21, č. 1, s. 256.

GELFAND, A. et al. (2010): Handbook of spatial statistics. CRC Press.

GETHING, P. et al. (2007): A local space-time kriging approach applied to a national outpatient malaria dataset. Computers & geosciences, 33, č. 10, s. 1337–1350.

GETHING, P. W. et al. (2006): Improving imperfect data from health management information systems in Africa using space-time geostatistics. PLoS medicine, 3, č. 6, s. e271.

GILLESPIE, I. et al. (2008): Demographic determinants for Campylobacter infection in England and Wales: implications for future epidemiological studies. Epidemiology and infection, 136, č. 12, s. 1717–25.

GOLDMAN, D., BRENDER, J. (2000): Are standardized mortality ratios valid for public health data analysis? Statistics in medicine, 19, č. 8, s. 1081–8.

GOLLINI, I. et al. (2015): GWmodel : an R package for exploring spatial heterogeneity. Journal of Statistical Software, 63, č. 17, s. 1–50.

GOOGLE (2009): Keyhole Markup Language [online]. Dostupné z: https://developers.google.com/kml/documentation/kml_tut

GOOVAERTS, P. (2005): Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. International journal of health geographics, 4, s. 31.

GOOVAERTS, P. (2006): Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using Poisson kriging and p-field simulation. International journal of health geographics, 5, s. 7.

GRÄLER, B. (2012): Chapter 1: Different concepts of spatio-temporal kriging. In: Summer School GEOSTAT 2012: Spatio-Temporal Geostatistics. s. 1–39.

GRÄLER, B. et al. (2012): Spatio-temporal analysis and interpolation of PM10 measurements in Europe. 37 s.

GRÄLER, B. (2013): Different concepts of spatio-temporal kriging. München, 1-39 s.

GREEN, C. G. et al. (2006): Spatial analysis of campylobacter infection in the Canadian province of Manitoba. International journal of health geographics, 5, č. 2, s. 14.

GRIFFITH, D., ARBIA, G. (2010): Detecting negative spatial autocorrelation in georeferenced random variables. International Journal of Geographical Information Science, 24, č. 3, s. 417–437.

GRUEBNER, O. et al. (2011): A spatial epidemiological analysis of self-rated mental health in the slums of Dhaka. International journal of health geographics, 10, č. 1, s. 36.

HAINING, R. (1998): Spatial statistics and analysis of health data. In: GIS and health. Taylor and Francis, London, s. 29 – 47.

HAINING, R. (2004): Spatial Data Analysis: Theory and Practice. Cambridge University Press.

HASHIZUME, M. et al. (2008): Factors determining vulnerability to diarrhoea during and after severe floods in Bangladesh. Journal of Water and Health, 6, č. 3, s. 323–332.

HAVELAAR, A. et al. (2013): Estimating the true incidence of campylobacteriosis and salmonellosis in the European Union, 2009. Epidemiology and infection, 141, č. 2, s. 293–302.

HEBÁK, P. et al. (2004): Vícerozměrné statistické metody 1. Informatorium, Praha, 240 s.

HEBÁK, P. et al. (2005a): Vícerozměrné statistické metody 2. Informatorium, Praha, 240 s.

HEBÁK, P. et al. (2005b): Vícerozměrné statistické metody 3. Informatorium, Praha, 256 s.

HENGL, T. (2007): A practical guide to geostatistical mapping of environmental variables. 143 s.

HENGL, T. (2009): A Practical Guide to Geostatistical Mapping. Office for Official Publications of the European Communities, Luxembourg.

HENGL, T. et al. (2012): Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. Theoretical and Applied Climatology, 107, č. 1-2, s. 265–277.

HENGL, T. et al. (2014): plotKML: Scientific visualization of spatio-temporal data. Journal of Statistical Software July, 58, č. II, s. 24.

HEUVELINK, G. B. M., GRIFFITH, D. a. (2010): Space-time geostatistics for geography: A case study of radiation monitoring across parts of Germany. Geographical Analysis, 42, č. 2, s. 161–179.

HIJMANS, R. J. (2014): raster: raster: Geographic data analysis and modeling. Verze 2.3-12. Dostupné z: http://cran.r-project.org/package=raster

HÖNIG, V. et al. (2011): Klíšťata a jimi přenášená onemocnění v Jihočeském kraji a Bavorsku.

HORÁK, J. (2009): Zpracování dat v GIS. VŠB-TU Ostrava, HGF, Institut geoinformatiky, Ostrava, 199 s.

HORÁK, J. (2011): Prostorové analýzy dat. VŠB-TU Ostrava, HGF, Institut geoinformatiky, Ostrava, 170 s.

HORÁK, J. et al. (2012): Prostorové hierarchické shlukování. In: GIS Ostrava 2012. VŠB-TU Ostrava, HGF, Institut geoinformatiky, s. 10.

HREJSEMNOU, O. (2009): Koncepce vybraných tématických map zdravotního stavu a zdravotní péče v ČR: analýza a interpretace současného stavu. Diplomová práce. Masarykova univerzita v Brně, 76 s.

CHARLTON, M., FOTHERINGHAM, A. S. (2009): Geographically weighted regression: White paper.

CHEN, C. et al. (2014): SpatialEpi: Methods and Data for Spatial Epidemiology. Verze 1.2.1. Dostupné z: http://cran.r-project.org/package=SpatialEpi

CHEN, J. et al. (2008): Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. International journal of health geographics, 7, s. 57.

INSPIRE: THEMATIC WORKING GROUP HUMAN HEALTH AND SAFETY (2013): INSPIRE: D2.8.III.5 - Data Specification on Human Health and Safety – Technical Guidelines.

IVAN, I., TVRDÝ, L. (2007): Změny v prostorovém pohybu obyvatelstva Moravskoslezského kraje. In: Sborník Území, znalosti a rozvoj v rámci konference Zvyšování konkurenceschopnosti aneb nové výzvy pro rozvoj regionů, států a mezinárodních trhů. VŠB – Technická univerzita Ostrava, Ostrava, s. 20.

JAGAI, J. S. et al. (2007): The Use of Köppen Climate Classification System for Public Health Research. Epidemiology, 18, č. 5, s. 7–212.

JARUP, L. (2004): Health and Environment Information Systems for Exposure and Disease Mapping, and Risk Assessment. Environmental Health Perspectives, 112, č. 9, s. 995–997.

JELINSKI, D. E., WU, J. (1996): The modifiable areal unit problem and implications for landscape ecology. Landscape Ecology, 11, č. 3, s. 129–140.

KALOGIROU, S. (2011): Testing local versions of correlation coefficients. Jahrbuch für Regionalwissenschaft, 32, č. 1, s. 45–61.

KALOGIROU, S. (2015): lctools: Local Correlation, Spatial Inequalities and Other Tools. Verze 0,2. Dostupné z: http://cran.r-project.org/package=lctools

KAMADJEU, R. (2009): Tracking the polio virus down the Congo River: a case study on the use of Google Earth in public health planning and mapping. International journal of health geographics, 8, s. 4.

KANEVSKI, M. et al. (2009): Machine learning for spatial envi- ronmental data: Theory, applications and software. EPFL Press, Lausanne, 377 s.

KAŇOK, J. (1999): Klasifikace stupnic a zásady jejich tvorby pro kartogram a kartodiagram. Kartografické listy, č. 7, s. 75–86.

KARATZOGLOU, A. et al. (2004): kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software, 11, č. 9, s. 1–20.

KASAL, P. et al. (2011): České zdravotnické registry - současný stav a perspektivy. In: Medsoft. Creative Connections s. r. o., Praha, s. 65–72.

KAUKO, T., GOETGELUK, R. (2005): Spatial and multidimensional analysis of the Dutch housing market using the Kohonen Map and GIS. ERSA conference papers.

KEIM, D. et al.ed. (2010): Mastering The Information Age - Solving Problems with Visual Analytics. Eurographics Association, Goslar, Germany, 168 s.

KLASCHKA, J. (2011): Klasifikační metody založené na rozhodovacích stromech. In: Voženílek, V. et al. (ed.): Metody umělé inteligence v geoinformatice. Univerzita Palackého v Olomouci, Olomouc, s. 31–39.

KLUFOVÁ, R. (2009): Využití nástrojů GIS při analýze vztahů socio-ekonomických faktorů a úrovně sociální péče. In: Sborník GIS Ostrava. VŠB-TU Ostrava, HGF, Institut geoinformatiky, Ostrava, s. 6.

KOHONEN, T. (1982): Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, č. 1, s. 59–69.

KOHONEN, T. (2001): Self-Organizing Maps. Springer Science & Business Media, 502 s.

KOCH, T. (2005): Cartographies of Disease: Maps, Mapping and Medicine. ESRI Press, Redlands, CA, 412 s.

KOCH, T. (2009): Disease Mapping. In: Kitchin, R. a Thrift, N. (ed.): International Encyclopedia of Human Geography. s. 234–241.

KOCH, T., DENIKE, K. (2004): Medical mapping: The revolution in teaching—and using—maps for the analysis of medical issues. Journal of Geography.

KOPERSKI, K. et al. (1996): Spatial Data Mining: Progress and Challenges Survey paper. SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96), SIGMOD'96, s. 1–2.

KRAAK, M. (2003): The space-time cube revisited from a geovisualization perspective. In: Proc. 21st International Cartographic Conference. Document Transformation Technologies, Durban, RSA, s. 1988–1996.

KRAAK, M.-J. (2013): From Cartography to Geographic Information Science The map and Geographic Information Science. Twente, NL, 8 s.

KSRZIS (2010): Koordinační středisko pro rezortní zdravotnické informační systémy [online]. Dostupné z: http://www.ksrzis.cz/

KULLDORFF, M. (1999): Spatial scan statistics: models, calculations, and applications. In: Scan statistics and applications. Birkhäuser, Boston, s. 303 – 322.

KULLDORFF, M. et al. (2005): A space-time permutation scan statistic for disease outbreak detection. PLoS medicine, 2, č. 3, s. e59.

KULLDORFF, M., INFORMATION MANAGEMENT SERVICES INC (2009): SaTScan v9.3: Software for the spatial and space-time scan statistics. s. 109.

KULLDORFF, M., NAGARWALLA, N. (1995): Spatial disease clusters: detection and inference. Statistics in medicine, 14, č. 8, s. 799–810.

KYRIAKIDIS, P. C., JOURNEL, A. G. (1999): Geostatistical space-time models: A review. Mathematical Geology, 31, č. 6, s. 651–684.

LAROSE, D. T., LAROSE, C. D. (2014): Discovering knowledge in data: An introduction to Data Mining. John Wiley & Sons, Inc., Hoboken, NJ, USA, 336 s.

LAST, J. M., ABRAMSON, J. (2001): A Dictionary of Epidemiology. Oxford University Press, USA, 196 s.

LAWSON, A. B. et al. (1999): Disease mapping and its uses. In: Lawson, A. B. et al. (ed.): Disease mapping and risk assessment for public health. John Wiley & Sons, Ltd, s. 3–13.

LAWSON, A. B. (2002): Spatial Cluster Modelling. CRC, Boca Raton, 284 s.

LAWSON, A. B. et al. (2003): Disease Mapping with WinBUGS and MLwiN. John Wiley & Sons, Ltd, Chichester.

LAWSON, A. B. (2009): Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. CRC Press.

LE COMBER, S. C. et al. (2011): Geographic profiling as a novel spatial tool for targeting infectious disease control. International journal of health geographics, 10, č. 1, s. 35.

LE COMBER, S. C., STEVENSON, M. D. (2012): From Jack the Ripper to epidemiology and ecology. Trends in Ecology and Evolution, 27, s. 307–308.

LEXOVÁ, P. et al. (2013): Výskyt infekčních onemocnění přenášených potravinami a vodou v ČR – rok 2012 a trendy nemocnosti. Zprávy CEM, 22, č. 7, s. 233–239.

LI, H. et al. (2012): One-step estimation of spatial dependence parameters: Properties and extensions of the APLE statistic. Journal of Multivariate Analysis, 105, č. 1, s. 68–84.

LIAW, A., WIENER, M. (2002): Classification and Regression by randomForest. R news, 2/3, č. December, s. 18–22.

LLOYD, C. D. (2010): Analysing population characteristics using geographically weighted principal components analysis: A case study of Northern Ireland in 2001. Computers, Environment and Urban Systems, 34, č. 5, s. 389–399.

LOH, W.-Y. (2011): Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1, č. 1, s. 14–23.

LOZANO-FUENTES, S. et al. (2008): Use of Google Earth to strengthen public health capacity and facilitate management of vector-borne diseases in resource-poor environments. Bulletin of the World Health Organization, 86, č. 9, s. 718–725.

LU, B. et al. (2014): The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. Geo-spatial Information Science, 17, č. 2, s. 85–101.

LUNARDON, N. et al. (2014): ROSE: A Package for Binary Imbalanced Learning. The R Journal, 6, č. 1, s. 79–89.

MANITZ, J., HÖHLE, M. (2013): Bayesian outbreak detection algorithm for monitoring reported cases of campylobacteriosis in Germany. Biometrical journal. Biometrische Zeitschrift, 55, č. 4, s. 509–26.

MAREK, L. et al. (2012): Spatial Analyses of Epidemiological Data: Case Study In Olomouc Region. In: 12th International Multidisciplinary Scientific GeoConference SGEM: SGEM 2012, Proceedings Volume II. STEF92 Technology Ltd, Sofia, Bulgaria, s. 1155 – 1162.

MAREK, L. et al. (2013a): Health Datasets in Spatial Analyses: What We Want, What We Get and What We Can Use. In: Recent Advances in Geodesy and Geomatics Engineering. WSEAS Press, Antalya, s. 7.

MAREK, L. et al. (2013b): On Estimation of the Spatial Clustering: Case Study of Epidemiological Data In Olomouc Region, Czech Republic. VŠB – Technická univerzita Ostrava, Ostrava.

MAREK, L. et al. (2014): Spatial Clustering of Disease Events Using Bayesian Methods. In: DATESO 2014. s. 10.

MAREK, L. et al. (2015): Using geovisual analytics in Google Earth to understand disease distribution: a case study of campylobacteriosis in the Czech Republic. International journal of health geographics, 14, č. 7, s. 1–13.

MARTIN, R. et al. (2009): Hunting patterns and geographic profiling of white shark predation. Journal of Zoology, 279, č. 2, s. 111–118.

MAŠKARINEC, P. (2013): Prostorová analýza prezidentských voleb v České republice v roce 2013. Sociológia-Slovak Sociological Review, 45, č. 5, s. 435–469.

MCBRIDE, G. B., MITTINTY, M. N. (2007): Explaining Differential Timing of Peaks of a Pathogen Versus a Faecal Indicator During Flood Events. In: MODSIM 2007: International Congress on Modelling and Simulation: Land, Water and Environmental Management: Integrated Systems for Sustainability. Modelling & Simulation Soc Australia & New Zealand Inc, Christchurch, NZ, s. 2417–2423.

MCLAFFERTY, S. L. (2003): GIS and health care. Annual review of public health, 24, s. 25–42.

MEADE, M. S., EMCH, M. (2010): Medical geography. The Guilford Press, New York, NY, 498 s.

MELOUN, M. (2011): Počítačová analýza vícerozměrných dat v oborech přírodních , technických a společenských věd.

MELOUN, M., MILITKÝ, J. (2004): Přednosti analýzy shluků ve vícerozměrné statistické analýze. In: Zajištění kvality analytických výsledků: sborník přednášek ze semináře. Univerzita Pardubice, Pardubice, s. 18.

METZ, C. E. (1978): Basic principles of ROC analysis. Semin Nucl Med, 8, č. 4, s. 283–298.

MEYER, D. et al. (2014): e1071: Misc Functions of the Department of Statistics (e1071). Verze 1.6-4. Dostupné z: http://cran.r-project.org/package=e1071

MIDDEL, A. (2007): A Framework for Visualizing Multivariate Geodata. In: Hagen, H. et al. (ed.): Visualization of Large and Unstructured Data Sets. Department of Informatics, Bonn, Kaiserslautern, s. 11.

MILLER, H. J., HAN, J.ed. (2009): Geographic Data Mining and Knowledge Discovery. CRC Press, Boca Raton, 443 s.

MILLER, J., FRANKLIN, J. (2002): Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecological Modelling, 157, č. 2-3, s. 227–247.

MINAMI, M. et al. (2007): Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. Fisheries Research, 84, č. 2, s. 210–221.

MINISTERSTVO ZDRAVOTNICTVÍ ČR (2010): Ministerstvo zdravotnictví ČR upozorňuje na možná zdravotní rizika způsobená konzumací mlékárensky neupraveného mléka. 1 s.

MOORE, D. a, CARPENTER, T. E. (1999): Spatial analytical methods and geographic information systems: use in health research and epidemiology. Epidemiologic reviews, 21, č. 2, s. 143–61.

MORAN, P. A. P. (1950): Notes on continuous stochastic phenomena. Biometrika, 37, č. 1, s. 17–23.

MULLNER, P. et al. (2010): Molecular and spatial epidemiology of human campylobacteriosis: source association and genotype-related risk factors. Epidemiology and infection, 138, č. 10, s. 1372–1383.

MYERS, D. E. (2004): Estimating and Modeling Space-time variograms. Proceedings of TIES Spatial Accuracy.

NAISH, S. et al. (2011): Spatio-temporal patterns of Barmah Forest virus disease in Queensland, Australia. PloS one, 6, č. 10, s. e25688.

NETRDOVÁ, P., NOSEK, V. (2009): Přístupy k měření významu geografického rozměru spoločenských nerovnoměrností. Geografie, 114, č. 1, s. 52–65.

NUR AIDI, M., PURWANINGSIH, T. (2013): Modeling Spatial Ordinal Logistic Regression and The Principal Component to Predict Poverty Status of Districts in Java Island. International Journal of Statistics and Applications, 3, č. 1, s. 1–8.

NYGÅRD, K. et al. (2004): Association between environmental risk factors and campylobacter infections in Sweden. Epidemiology and infection, 132, s. 317–325.

NYLEN, G. et al. (2002): The seasonal distribution of campylobacter infection in nine European countries and New Zealand. Epidemiology and infection, 128, č. 3, s. 383–390.

O'LEARY, M. (2010): Implementing a Bayesian approach to criminal geographic profiling. In: Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application - COM.Geo '10. s. 8.

OAKES, J. M. (2004): The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. Social science & medicine, 58, č. 10, s. 1929–52.

OPEN GEOSPATIAL CONSORTIUM (2008): OGC KML 2.2.0. Open Geospatial Consortium, 252 s.

OPENSHAW, S. (1984a): Ecological fallacies and the analysis of areal census data. Environment and Planning A, 1, č. 16, s. 17 – 31.

OPENSHAW, S. (1984b): The Modifiable Areal Unit Problem. In: Geo Books. Norwich.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (2014): OECD Regional eXplorer [online]. Dostupné z: http://stats.oecd.org/OECDregionalstatistics/#story=0

OSEI, F. B. (2014): Current Statistical Methods for Spatial Epidemiology: A Review. Austin Biometrics and Biostatistics, 1, č. 2, s. 7.

OSTFELD, R. S. et al. (2005): Spatial epidemiology: an emerging (or re-emerging) discipline. Trends in ecology & evolution, 20, č. 6, s. 328–36.

PÁSZTO, V. et al. (2014): Using a fuzzy inference system to delimit rural and urban municipalities in the Czech republic in 2010. Journal of Maps, 11, č. 2, s. 231–239.

PEBESMA, E. (2012): spacetime: Spatio-temporal data in r. Journal of Statistical Software, 51, č. 7, s. 30.

PEBESMA, E., GRÄLER, B. (2014): Spatio-temporal geostatistics using gstat. Münster, DE, 1-11 s.

PECÁKOVÁ, I. (2007): Logistická regrese s vícekategoriální vysvětlovanou proměnnou. Acta Oeconomica Pragensia, 15, č. 1, s. 86–96.

PEKÁR, S., BRABEC, M. (2009): Moderní analýza biologických dat: Zobecněné lineární modely v prostředí R. Scientia, Praha, 225 s.

PENG, R. D., DOMINICI, F. (2008): Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health. Springer New York, New York, NY.

PFEIFFER, D. et al. (2008): Spatial analysis in epidemiology. Oxford University Press.

R CORE TEAM (2014): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RAINE, N. E. et al. (2009): Geographic profiling applied to testing models of bumble-bee foraging. Journal of the Royal Society, Interface / the Royal Society, 6, č. 32, s. 307–319.

REZAEIAN, M. et al. (2007): Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. Journal of epidemiology and community health, 61, č. 2, s. 98–102.

RICHARDSON, S. et al. (2004): Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies. Environmental Health Perspectives, 112, č. 9, s. 1016–1025.

ROSSMO, D. K. (1995a): Geographic profiling: Target patterns of serial murderers. Diplomová práce. 581-581 p. s.

ROSSMO, D. K. (1995b): Place, Space, and Police Investigations: Hunting Serial Violent Criminals. In: Weisburd, D. a Eck, J. E. (ed.): Crime and place. NY: Criminal Justice Press, Monsey, s. 217 – 235.

ROSSMO, D. K. (1999): Geographic Profiling. CRC Press, Boca Raton, 376 s.

ROSSMO, D. K. (2000): Geographic profiling. CRC Press, New York, 376 s.

ROUHANI, S., MYERS, D. E. (1990): Problems in space-time kriging of geohydrological data. Mathematical Geology, 22, č. 5, s. 611–623.

RUSHTON, G. (2003): Public health, GIS, and spatial analytic tools. Annual review of public health, 24, s. 43–56.

RYTKÖNEN, M. J. (2004): Not all maps are equal: GIS and spatial analysis in epidemiology. International journal of circumpolar health, 63, č. 1, s. 9–24.

SARI KOVATS, R. et al. (2005): Climate variability and campylobacter infection: An international study. International Journal of Biometeorology, 49, č. 4, s. 207–214.

SCOTT, L. M., JANIKAS, M. V (2010): Spatial Statistics in ArcGIS. In: Fischer, M. M. a Getis, A. (ed.): Handbook of Applied Spatial Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, s. 27–42.

SHEKHAR, S. et al. (2011): Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1, č. 3, s. 193–214.

SCHLIEP, K., HECHENBICHLER, K. (2014): kknn: Weighted k-Nearest Neighbors. Verze 1.2-5. Dostupné z: http://cran.r-project.org/package=kknn

SMOLA, A. J. et al. (1998): The connection between regularization operators and support vector kernels. Neural Networks, 11, č. 4, s. 637–649.

SNOOK, B. et al. (2005): Commentary: Shortcuts to Geographic profiling success: A reply to Rossmo (2005). Applied Cognitive Psychology, 19, č. 5, s. 655–661.

SPENCER, S. et al. (2012): The spatial and temporal determinants of campylobacteriosis notifications in New Zealand, 2001-2007. Epidemiology and infection, 140, č. 9, s. 1663–77.

SPEYBROECK, N. (2012): Classification and regression trees. International Journal of Public Health, 57, č. 1, s. 243–246.

SPIELMAN, S. E., THILL, J.-C. (2008): Social area analysis, data mining, and GIS. Computers, Environment and Urban Systems, 32, č. 2, s. 110–122.

SPILKOVA, J. et al. (2014): Perception of neighborhood environment and health risk behaviors in Prague's teenagers: a pilot study in a post-communist city. International Journal of Health Geographics, 13, č. 1, s. 41.

SPILKOVÁ, J. et al. (2011): Inequalities in smoking in the Czech Republic: Societal or individual effects? Health and Place, 17, č. 1, s. 215–221.

SPURNÁ, P. (2008): Prostorová autokorelace – všudypřítomný jev při analýze prostorových dat? Sociologický časopis/Czech Sociological Review, 44, č. 4, s. 767–787.

STAESSEN, J. et al. (1999): Environmental exposure to cadmium, forearm bone density, and risk of fractures: prospective population study*. The Lancet.

STEVENSON, M. D. et al. (2012): Geographic profiling as a novel spatial tool for targeting the control of invasive species. Ecography, 35, č. December 2011, s. 704–715.

STEVENSON, M. D. (2014): User Guide to Rgeoprofile (version 1.2). 11 s.

STEVENSON, M. D. (2015): epiR: Tools for the Analysis of Epidemiological Data. Verze 0.9-62. Dostupné z: http://cran.r-project.org/package=epiR

STEVENSON, M. D., VERITY, R. (2014): Rgeoprofile: Geographic Profiling in R. Verze 1.3. Dostupné z: http://evolve.sbcs.qmul.ac.uk/lecomber/sample-page/geographic-profiling/

SWEENEY, L. (2000): Simple demographics often identify people uniquely. Pittsburgh, 1-34 s.

SWEENEY, L. et al. (2013): Identifying Participants in the Personal Genome Project by Name. SSRN Electronic Journal, s. 1–4.

SZÚ ČR (2010): Zdravotní nezávadnost mléka prodávaného prostřednictvím automatů [online]. Dostupné z: http://apps.szu.cz/svi/hygiena/show.php?kat=novinky_vse#100118c

SZÚ ČR (2014): Státní zdravotní ústav [online]. Dostupné z: http://www.szu.cz/

ŠIROKÝ, P. (1999): Výpočet a odhad měr incidence, prevalence a mortality. Klinická onkologie, 12, č. 23-24, s. 2.

ŠTAMPACH, R. (2010): Explorační geografická analýza zdravotních dat a jejich kartografická prezentace. Diplomová práce. Masarykova Univerzita v Brně,

ŠTAMPACH, R. (2013): Mezinárodní zdroje dat a map se zdravotní tématikou: srovnání a potenciál. In: 7. kartografický den2. Olomouc, s. 40.

ŠTAMPACH, R., GERYK, E. (2012): Health statistics in international databases and their cartographic visualization. Quaestiones Geographicae, 31, č. 3, s. 77–88.

TATEM, A. J. et al. (2004): Terra and Aqua: new data for epidemiology and public health. International Journal of Applied Earth Observation and Geoinformation, 6, č. 1, s. 33–46.

TATEM, A. J. et al. (2012): Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation. Population health metrics, 10, č. 1, s. 8.

THE CENTER FOR FOOD SECURITY & PUBLIC HEALTH (2013): Campylobacteriosis. Ames, 7 s.

THERNEAU, T. et al. (2014): rpart: Recursive Partitioning and Regression Trees. Verze R package version 4.1-8. Dostupné z: http://cran.r-project.org/package=rpart

THOMAS, J. J., COOK, K. A.ed. (2005): Illuminating the path: The research and Development Agenda for visual analytics. IEEE Computer Society Press, Chicago, USA, 184 s.

TIMM, N. H. (2002): Applied Multivariate Analysis. Springer, New York, 693 s.

TOBLER, W. R. (1970): A Computer Movie Simulation Urban Growth in the Detroit Region. Economic Geography, 46, č. 332, s. 234–240.

TOMASZEWSKI, B. (2009): Emerging Applications and Challenges for Geovisual Analytics Research. 43 s.

TVRDÍK, J. (2003): Analýza vícerozměrných dat.

ÚSTAV ZDRAVOTNICKÝCH INFORMACÍ A STATISTIKY ČESKÉ REPUBLIKY (2013): Infekční nemoci 2013. 60 s.

ÚZIS ČR (2014): Ústav zdravotnických informací a statistiky ČR [online]. Dostupné z: http://www.uzis.cz/

VELKÝ LÉKAŘSKÝ SLOVNÍK (2008): Velký lékařský slovník [online]. Dostupné z: http://lekarske.slovniky.cz/pojem/morbidita

VENABLES, W. N., RIPLEY, B. D. (2002): Modern Applied Statistics with S. Springer, New York, 498 s.

VERITY, R. et al. (2014): Spatial targeting of infectious disease control: identifying multiple, unknown sources. Methods in Ecology and Evolution, 5, č. 7, s. 647–655.

VOŽENÍLEK, V. et al. (2011): Metody tematické kartografie: vizualizace prostorových jevů. Univerzita Palackého v Olomouci, Olomouc, 216 s.

WALESIAK, M., DUDEK, A. (2007): Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych–oprogramowanie komputerowe i wyniki badaÅ„. Zeszyty Naukowe Uniwersytetu Szczecinskiego, č. 450, s. 635–646.

WALLER, L. (2005): Bayesian thinking in spatial statistics. Handbook of Statistics, 25, s. 589–618.

WALLER, L. (2009): Detection of clustering in spatial data. The SAGE handbook of spatial analysis, s. 34.

WALLER, L. A., GOTWAY, C. A. (2004): Applied Spatial Statistics for Public Health Data. John Wiley & Sons.

WANG, F. (2009): Factor Analysis and Principal-Components Analysis. In: Kitchin, R. a Thrift, N. (ed.): International Encyclopedia of Human Geography. Elsevier, Oxford, s. 1–7.

WEHRENS, R., BUYDENS, L. M. C. (2007): Self-and super-organizing maps in R: the Kohonen package. Journal Of Statistical Software, 21, č. 5, s. 19.

WEISENT, J. et al. (2011): Detection of high risk campylobacteriosis clusters at three geographic levels. Geospatial health, 6, č. 1, s. 65–76.

WEISENT, J. et al. (2012): Socioeconomic determinants of geographic disparities in campylobacteriosis risk: a comparison of global and local modeling approaches. International Journal of Health Geographics, 11, č. 1, s. 45.

WIKLE, B. C. K., CRESSIE, N. (1999): A dimension-reduced approach to space-time Kalman filtering. s. 815–829.

WILLIAMS, G. (2011): Data mining with Rattle and R: the art of excavating data for knowledge discovery.

WU, X. et al. (2007): Top 10 algorithms in data mining. 1-37 s.

XIE, X. (2008): A review of recent advances in surface defect detection using texture analysis techniques. Electronic Letters on Computer Vision and Image Analysis, 7, č. 3, s. 1–22.

ZAFARANI, R. et al. (2014): Social Media Mining: An Introduction. Cambridge University Press, Cambridge, 380 s.

ZEILEIS, A. et al. (2008): Regression models for count data in R. Journal of Statistical Software, 27, č. 8, s. 1–25.

ZELEŇÁKOVÁ, L. et al. (2012): Application of epidemiological information system (EPIS) in the Slovak republic within the surveillance of salmonellosis and campylobacteriosis outbreaks in the European Union (2001-2010). Acta Universitatis Agriculturae et Silvicluturae Mendelianae Brunensis, LX, č. 1, s. 189–200.

ZHAO, Y. (2013): R and Data Mining: Examples and Case Studies. Elsevier, 160 s.

SUMMARY

The main objective of the dissertation thesis was to carry out the complex spatial analysis of the epidemiological data with the usage of recent technologies from fields of geoinformatics and spatial statistics. The practical part of the dissertation examined an infectious disease called campylobacteriosis during 2008—2012 in the Czech Republic. Besides the primary objective, there was also the aim to provide the coherent work that may provide the overview of usable spatial epidemiology methods and provide the support and inspiration for other similar studies. The main theme of the thesis and its interdisciplinary nature required to combine the knowledge acquired during the GIS studies with the findings of other disciplines such as spatial and applied statistics or spatial epidemiology.

During the processing of the dissertation, its main objective was split to five consecutive partial objectives that together covered main topics addressed in the field of spatial epidemiology. These are – a disease mapping, an identification of spatial patterns and geographical correlation studies. The first and fifth partial objectives were examining mapping and geovisualization issues. The first partial objective mapped the spatial distribution of the campylobacteriosis – its frequency, incidence and relative risk. In addition, the smoothing methods were applied, and recommendations were provided. The spatiotemporal mapping was realised by the employment of the spatiotemporal kriging that supplied methods of dasymetric mapping. Using the spatiotemporal kriging, the continuous surface of weekly raw incidence was created so the examination of the disease could have been monitored in space and time simultaneously. The engagement of the spatiotemporal kriging in the exploration of the morbidity is one of the dissertation's highlights. Maps and geovisualisations were also transformed into KML files in order to provide the transferability of the data as well as to communicate the results in Google Earth. Google Earth provided the platform for the geovisual interactive spatiotemporal analytics and, due to its intuitive interface and wide distribution, a convenient way for the spreading the results to lay and professional audience.

The second important topic in the spatial epidemiology and the second partial objective of the dissertation research are analyses of spatial and spatiotemporal patterns of the disease distribution. Initially, patterns were evaluated visually based on the outputs of the first partial objective. Then, it was confirmed that the most vulnerable areas are Ostrava and it surrounding and also the north-east Morava, Benešov and several others, where clusters of high rates of morbidity were identified in space and in time. There was also cluster around Plzeň and České Budějovice but they appeared for the limited time. The method of geoprofiling was utilized in order to evaluate the spatial pattern of particular cases and possibly to identify the likely association between small local disease outbreaks and the location of fresh milk vending machines. This was proceed in order to provide the retrospective analysis of the interesting situation that was actual in 2010, when the dispute over the health risks of the fresh milk made the controversy regarding the situation in České Budějovice and the public announcement by the head hygienist of the Czech Republic. The cluster of high relative risk was identified in České Budějovice lasting during the January and February 2010, however the geoprofiling did not identify any potential source of the outbreak located near the fresh milk vending machines, although some potential outbreak sources were found in the neighbourhood. During the analyses of all vending machines in the Czech

Republic up to 20% of fresh milk vending machines were identified as potential sources. Considering the overestimation of the significance of some cases reduced the estimation of sources to approximately 9%.

The third partial objective covered the third main theme of the spatial epidemiology, which is the geographical correlation analysis of the association between the disease distribution and environmental factors. The existence of the association was presumed based on the identification of clusters of both, high and low rates in the preceding partial objectives. The assumption itself could have been confirmed only partly. The demography was identified as the most associated factor, but also the local influence of the temperature, socioeconomic deprivation and agriculture. The prediction and classification performances of both, spatial and traditional methods were never higher than 50% of the real situation, which was not very much but it corresponds with the fact, that only half of the campylobacteriosis cases is well described in the Czech Republic and also worldwide. Although, models used in this part of the dissertation did not perform well, they were still very useful for the exploration of local inferences. Within this partial objective, the concept of the local ordinal logistic regression was introduced as the modification of ordinal regression for spatial data.

Among the others, main benefits or highlights presented in the dissertation can be summarized as the use of the spatiotemporal kriging within the disease mapping, the application of the space-time scan statistics and geoprofiling in the investigation of spatial and spatiotemporal patterns, and the utilization of geographically weighted methods for the analysing of the association between the disease occurrence and local environmental characteristics. The local scale of the case study should be also mentioned in addition to all methods.

The investigation of the health and health-related topics is one of the very favourite topics not only in the geosciences but also in a wide range of other disciplines. However, the disparity between medical studies based on individuals and studies in other scales is usually visible. On one hand, this can be caused by the protection of personal data (also connected to the small amount of publicly available data), but it can be also caused by possible distrust in the capabilities and benefits of spatial and spatiotemporal methods. Nevertheless, it should be noted that the situation is improving slowly but constantly because of activities of the research teams in universities as well as in the involved state institutions.

ANOTACE

Název práce: Prostorové a vícerozměrné statistické analýzy epidemiologických dat

Autor práce: Mgr. Lukáš Marek

Pracoviště: Katedra geoinformatiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci

Anotace: Disertační práce je příspěvkem k multidisciplinární povaze geoinformatiky, geovizualizací a geografie. Jejich využití v případě zdravotnických dat je bezpochyby jedním z trendů ve výzkumu v zahraničí a dostává se do popředí i u nás. Hlavní cíl disertační práce se zaměřuje na provedení komplexní prostorové analýzy epidemiologických dat využitím metod geoinformatiky a prostorové statistiky. Práce je tematicky rozdělena v souladu s hlavním cílem a metodologií výzkumů v prostorové epidemiologii. Hlavním spojujícím motivem zkoumání je datová sada týkající se výskytu infekční nemoci zvané kampylobakterióza v České republice v období let 2008 - 2012. Tato nemoc je nejprve hodnocena tradiční popisnou statistikou. Dále je mapována s pomocí tradičních metod kartografie a také s použitím bayesovského shlazování. V navazující části je zkoumán globální a lokální časoprostorový vzor onemocnění, na což navazuje identifikace možných socioekonomických a environmentálních faktorů a modelování rizika ohrožení nemocí pro jednotlivé obce ČR s využitím statistických vícerozměrných metod, metod strojového učení a jejich prostorových ekvivalentů. Kromě modelování jsou obce České republiky klasifikovány do skupin na základě relativního rizika ohrožení nemocí a převládajících socioekonomických, demografických a environmentálních vlastností území. V poslední části je pomocí metody geografického profilování retrospektivně zkoumán možný vliv přítomnosti automatů na prodej čerstvého mléka na lokální zvýšení počtu případů v jejich okolí.

Klíčová slova: Prostorová epidemiologie, mapování onemocnění, geovizualizace, geografická korelace, geografické profilování, prostorový vzor

Rozsah práce: 136 stran

Jazyk práce: čeština

ANNOTATION

Title: Spatial and multivariate statistical analyses of epidemiological data

Author: Mgr. Lukáš Marek

Department: Department of Geoinformatics, Faculty of Science, Palacky University, Olomouc

Abstract: The dissertation thesis is a contribution to the multidisciplinary nature of geoinformatics, geography and geovisualization. The investigation of the health and health-related topics is one of the most popular topics not only in the geosciences but also in a wide range of other disciplines. The main objective of the dissertation thesis was to carry out the complex spatial analysis of the epidemiological data with the usage of recent technologies from fields of geoinformatics and spatial statistics. Its main objective was split to five consecutive partial objectives that together covered main topics addressed in the field of spatial epidemiology. The common motive of all partial objectives was the spatial analysis and exploration of the campylobacteriosis in the Czech Republic in 2008—2012. Initially, the disease was described statistically and the spatial distribution was mapped using cartographic methods and smoothing. Then, the spatial and spatiotemporal patterns on both, global and local scale were assessed and the associations of the disease occurrence and environmental and socioeconomic factors were inspected using spatial and multivariate statistics and machine learning. In addition to the modelling, Czech municipalities were clustered to groups of similar entities based on their common characteristics of environment and morbidity. The last research question dealt with the evaluation of fresh milk vending machines as possible sources of local outbreaks utilizing the geoprofiling.

Keywords: Spatial epidemiology, disease mapping, geovisualisation, geographic correlation, spatial pattern, geoprofiling

Range: 136 pages

Language: Czech

PŘÍLOHY

SEZNAM PŘÍLOH

Vázané přílohy

Volné přílohy

1× DVD

Struktura DVD:

Složka s mapami prezentovanými v práci

Složka s KML soubory

Doplňková data

Geoprofily

Text disertační práce

PŘÍLOHY

SEZNAM PŘÍLOH

Vázané přílohy

Volné přílohy

1× DVD

Struktura DVD:

Složka s mapami prezentovanými v práci

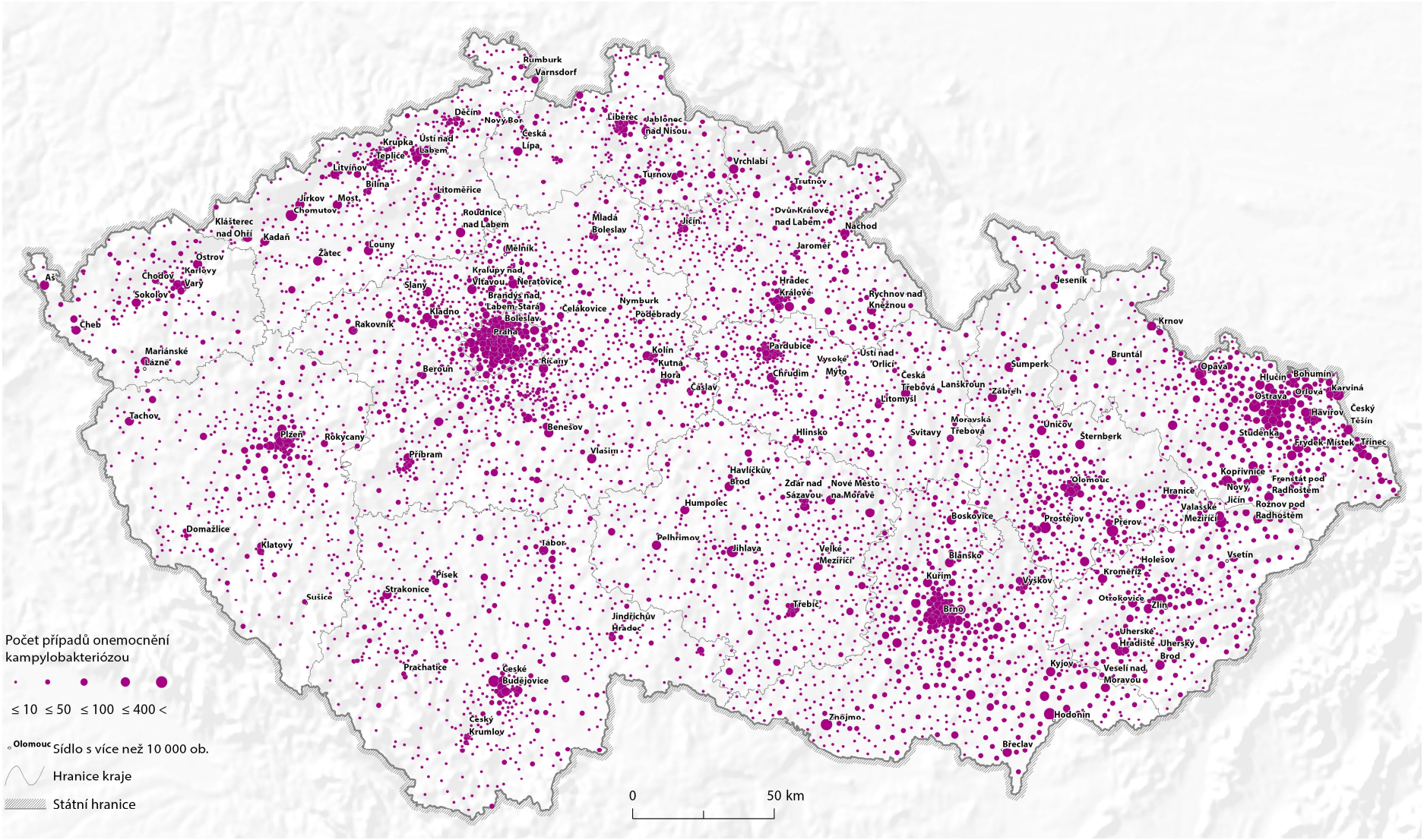
Složka s KML soubory

Složka s doplňujícími daty

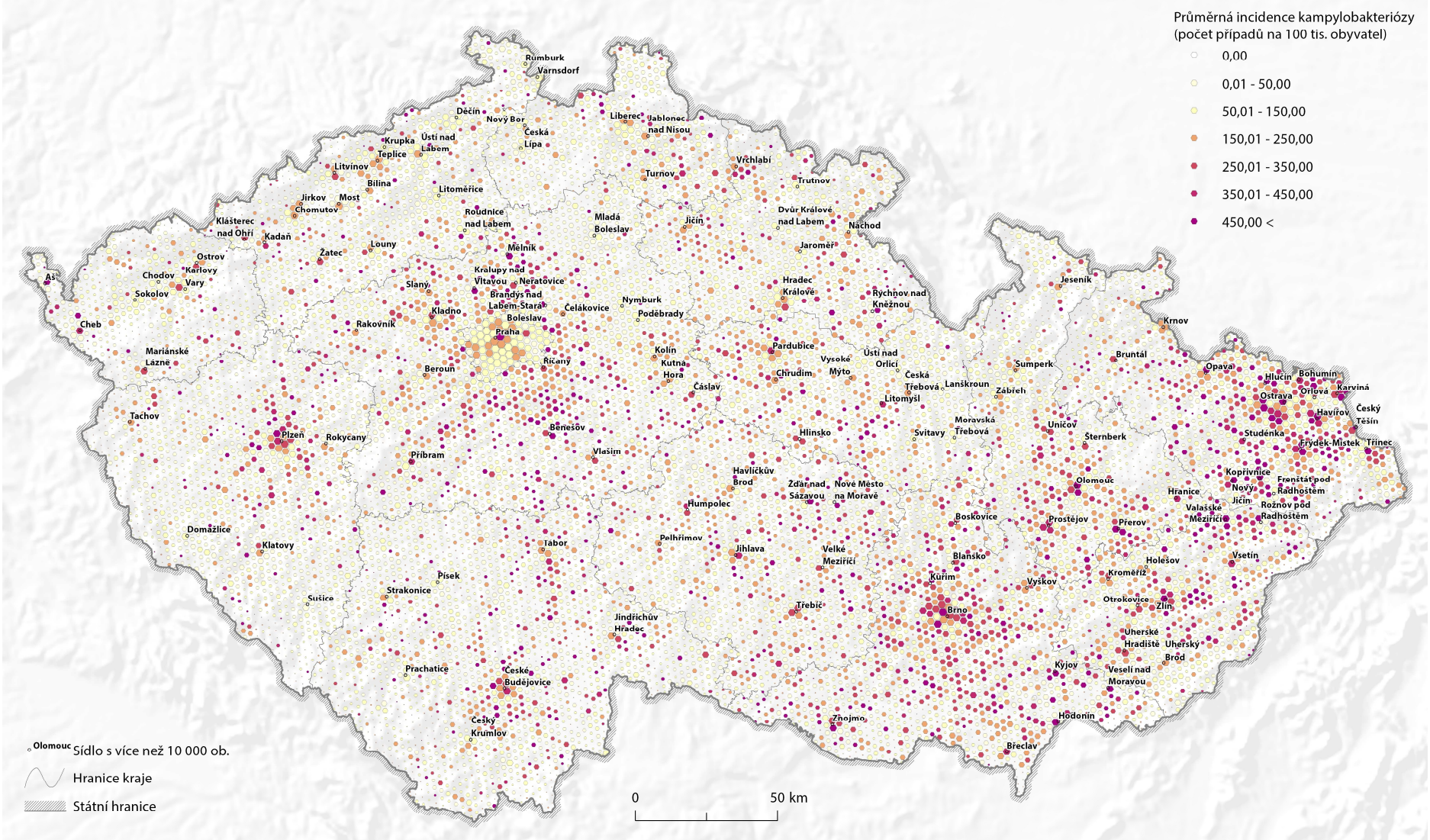
Složka s geoprofily

Text disertační práce

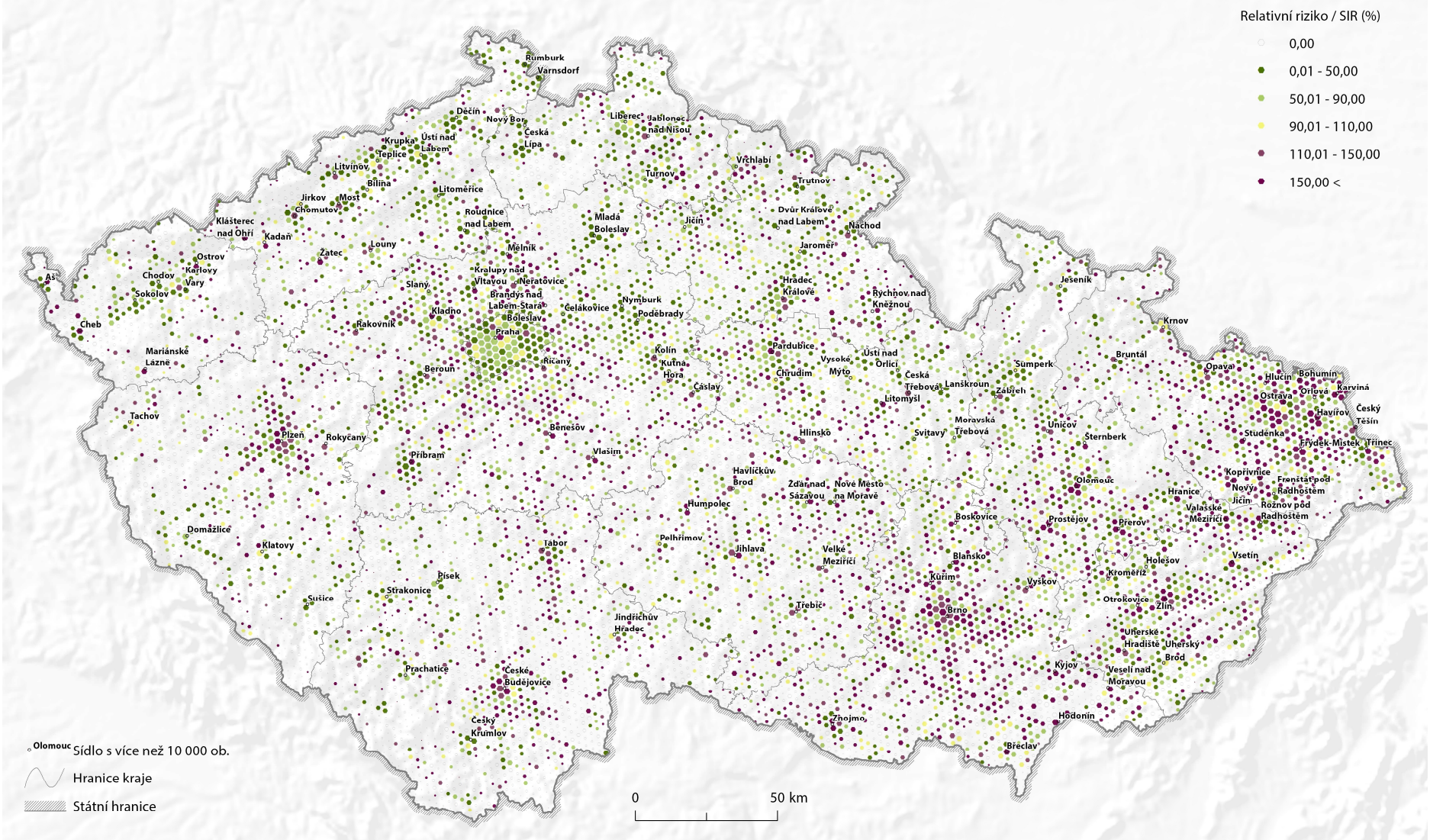
Příloha 1: Případy kampylobakteriózy zobrazené pomocí metody teček s využitím vážených teček



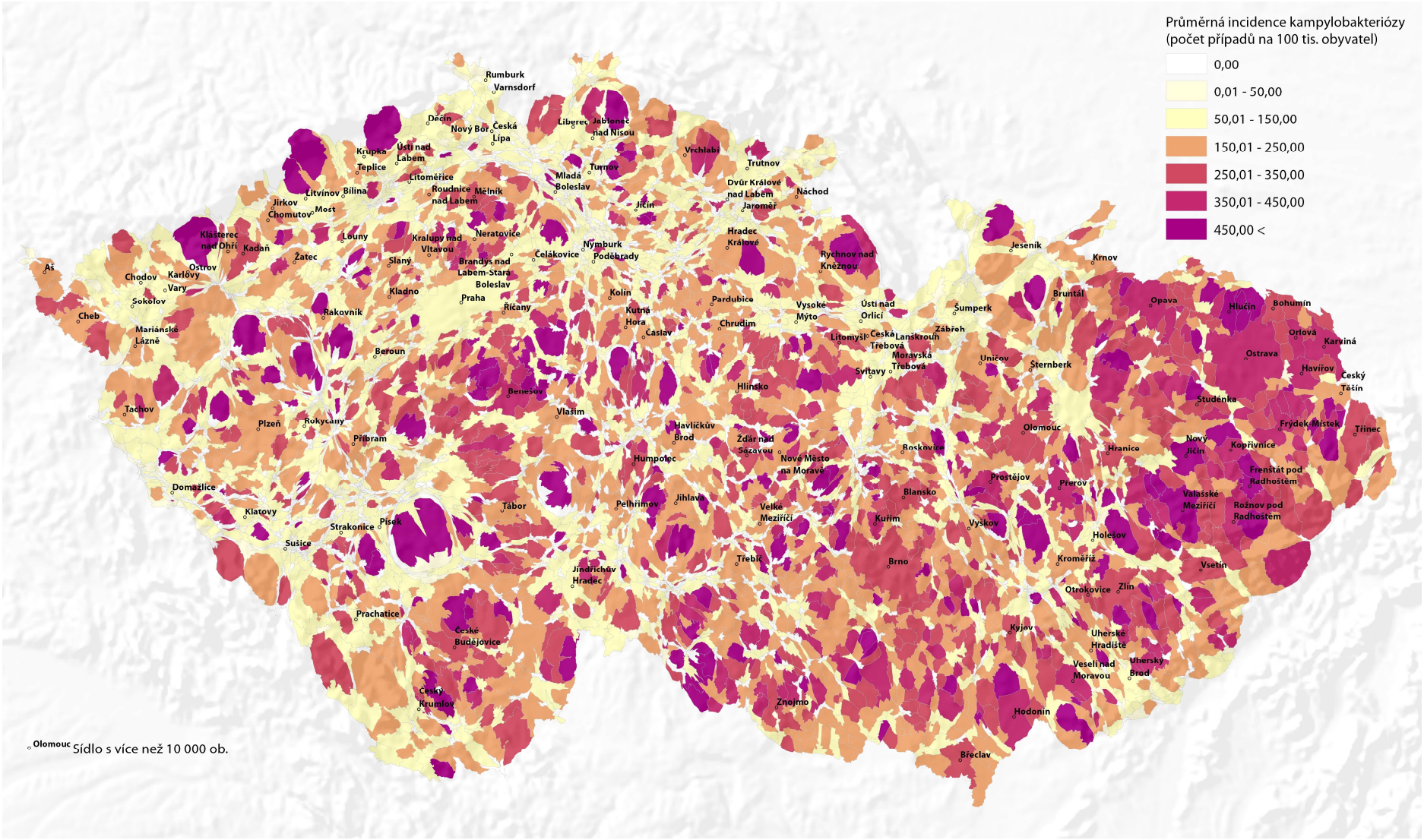
Příloha 2: Průměrná hrubá incidence kampylobakteriózy v České republice vizualizovaná prostřednictvím hexagonů



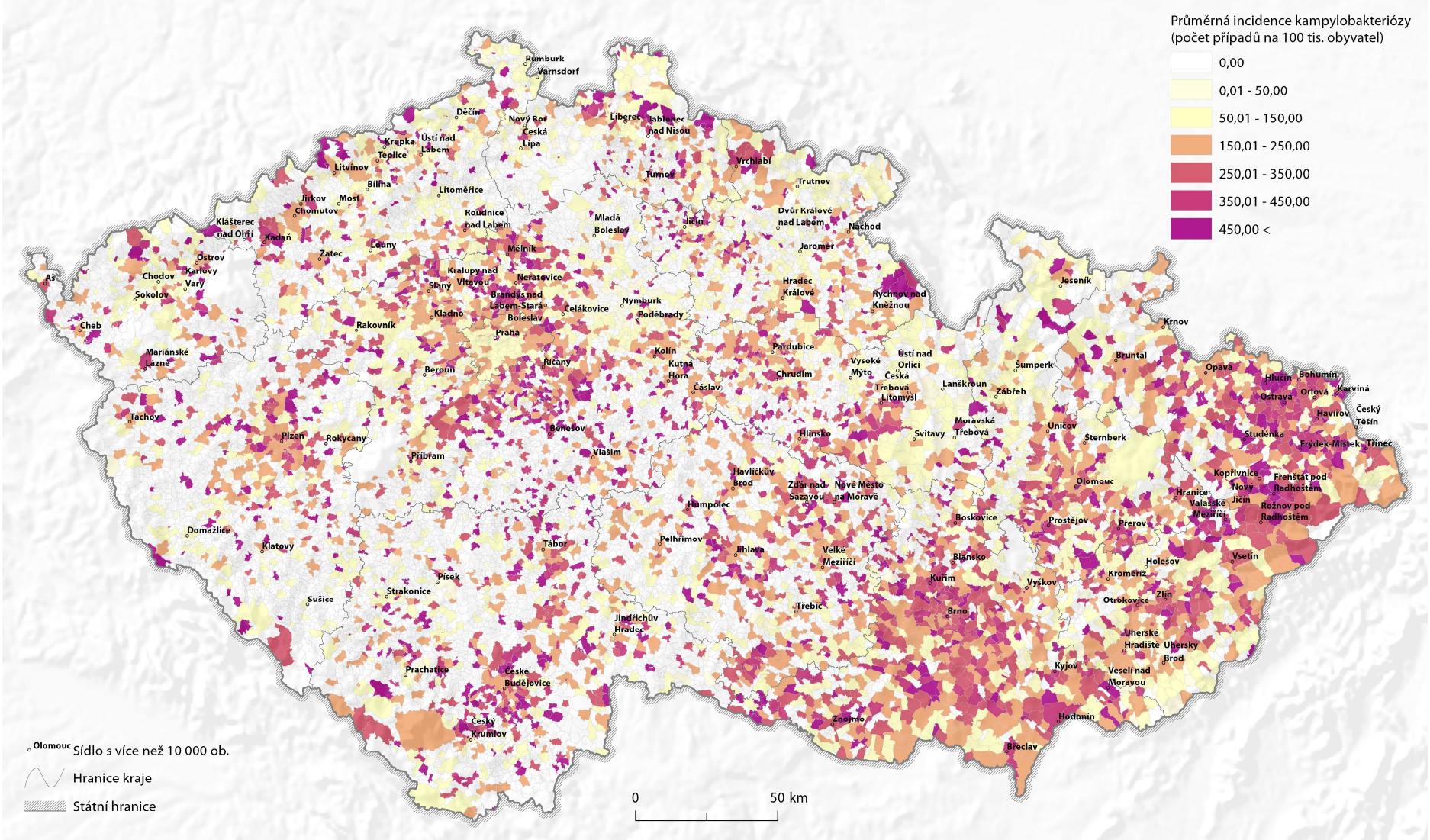
Příloha 3: Relativní riziko (SIR) vizualizované prostřednictvím hexagonů



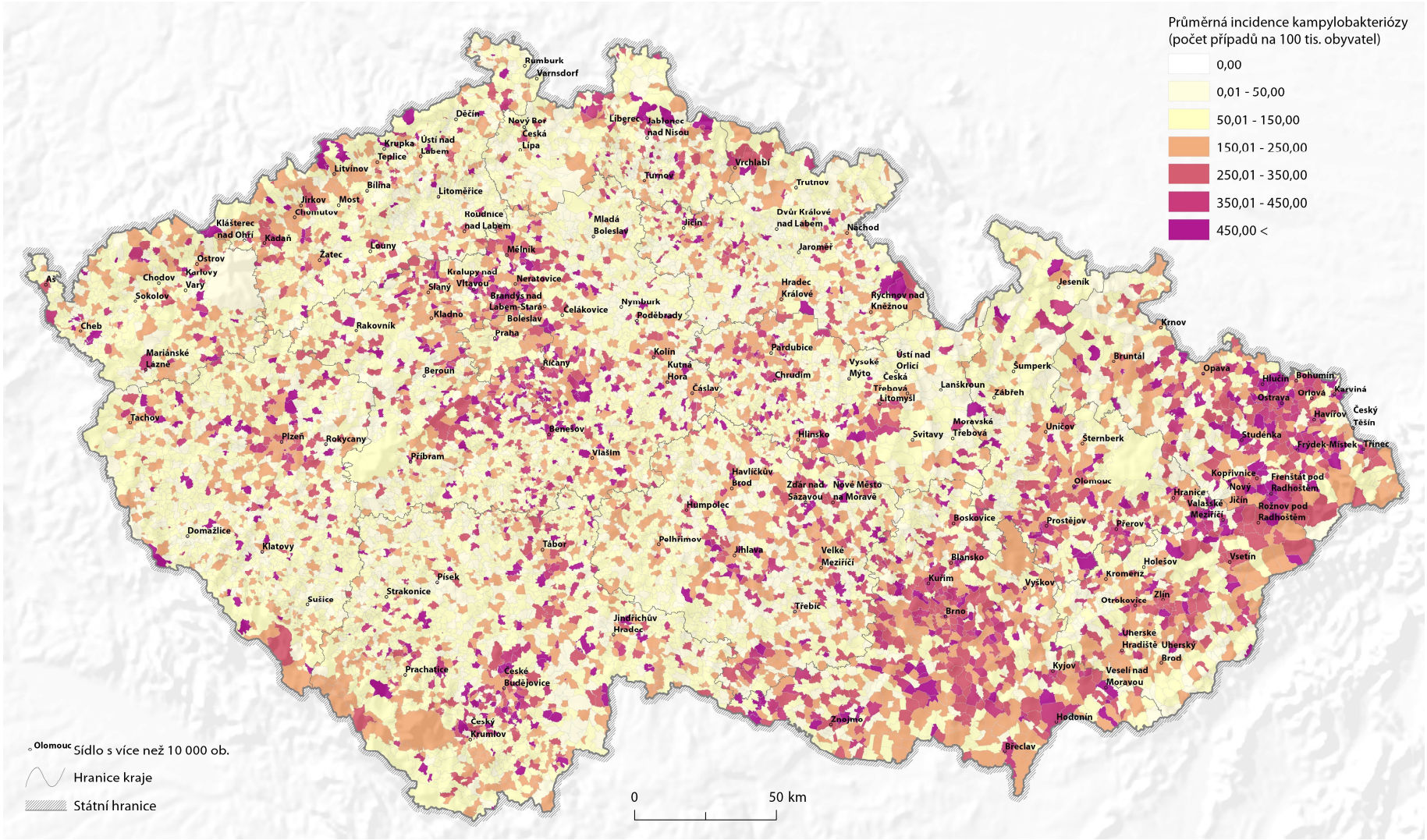
Příloha 4: Zobrazení průměrné incidence kampylobakteriózy formou kartografické anamorfózy



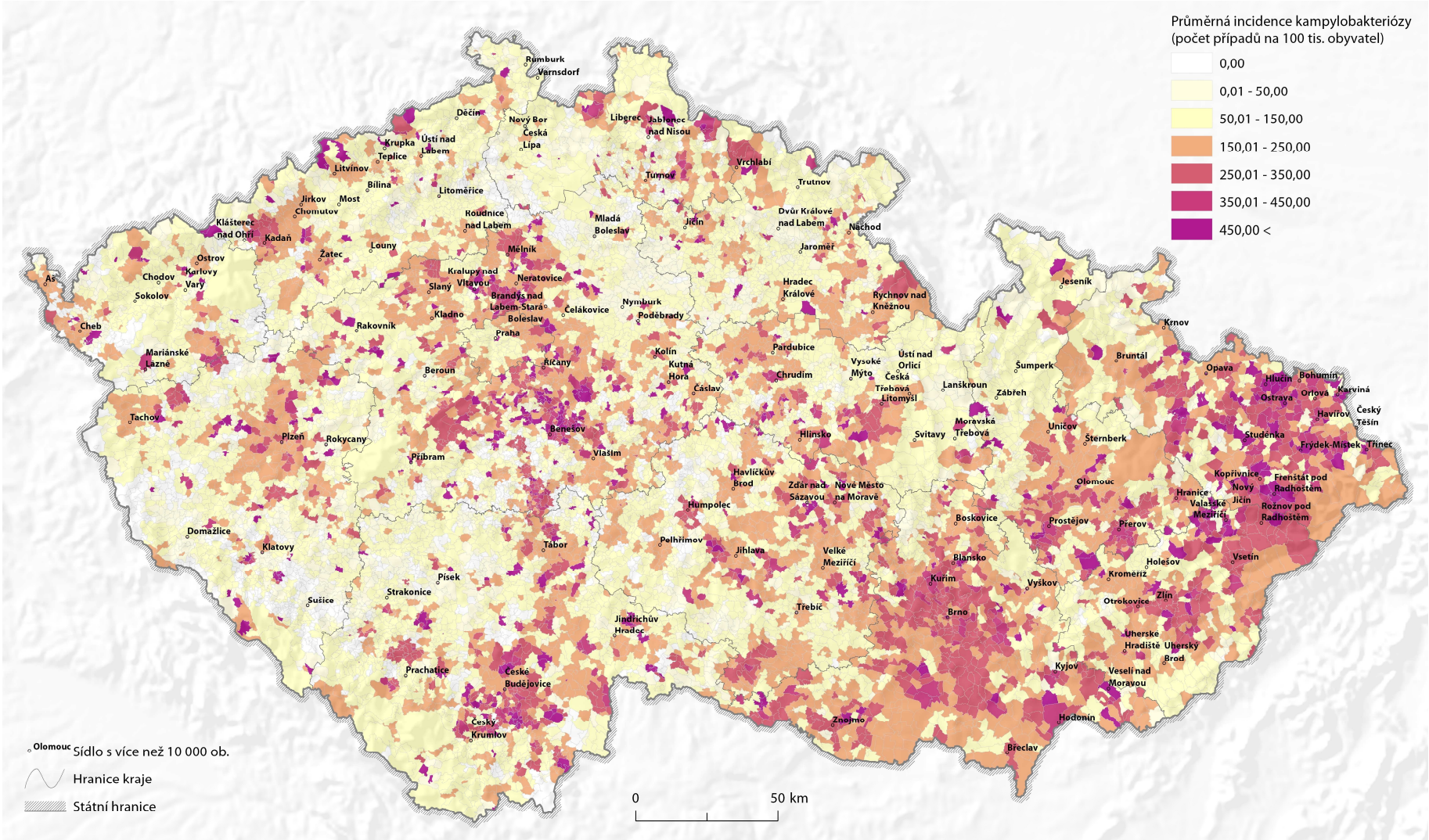
Příloha 5: Hrubá průměrná incidence kampylobakteriózy v částech obcí v ČR v letech 2008—2012 v počtech případů na 100 tisíc obyvatel



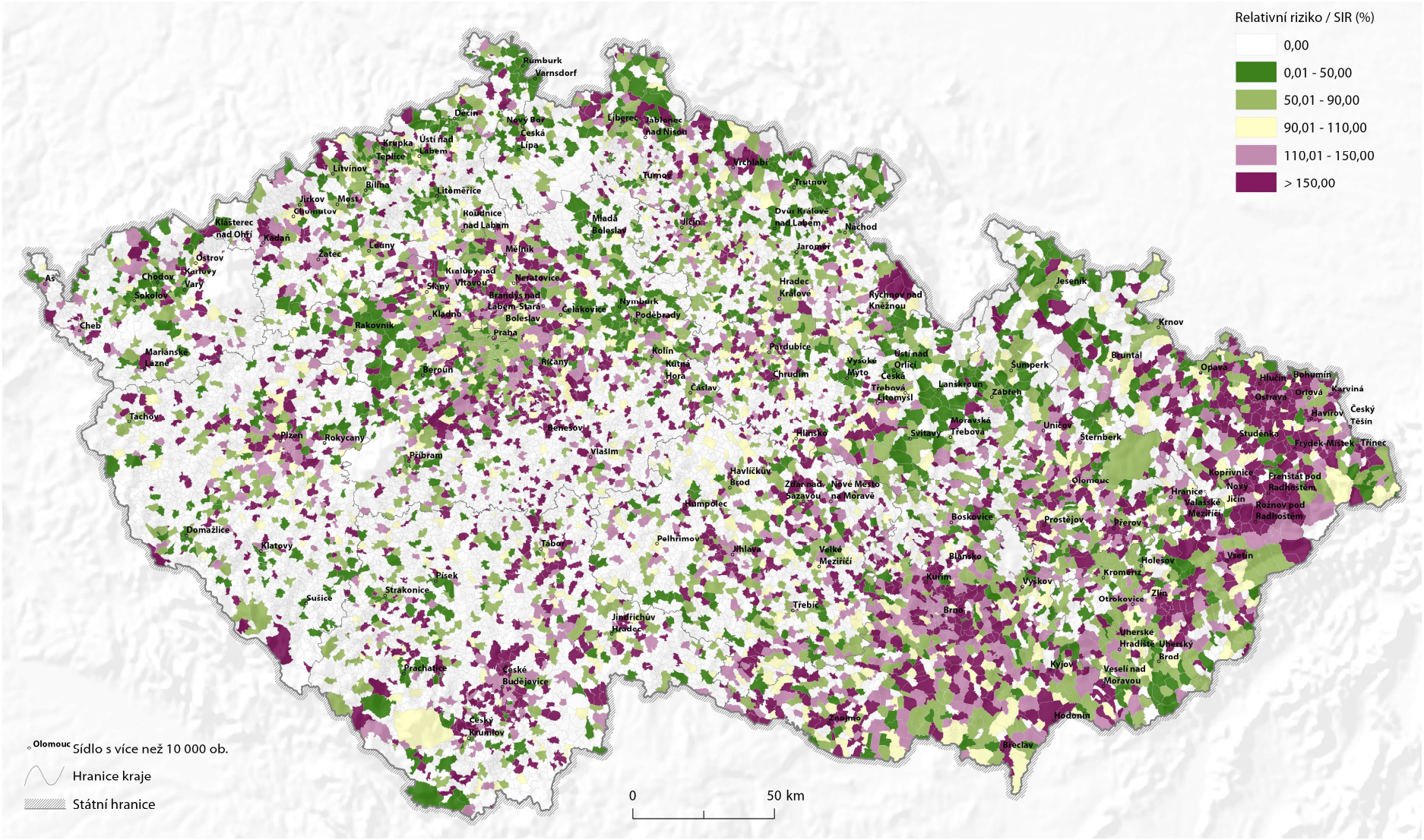
Příloha 6: Vyhlazená standardizovaná průměrná incidence kampylobakteriózy v částech obcí v ČR v letech 2008—2012 v počtech případů na 100 tisíc obyvatel, která vznikla pomocí globálního Bayesova vyhlazení založeného na negativním binomickém rozdělení



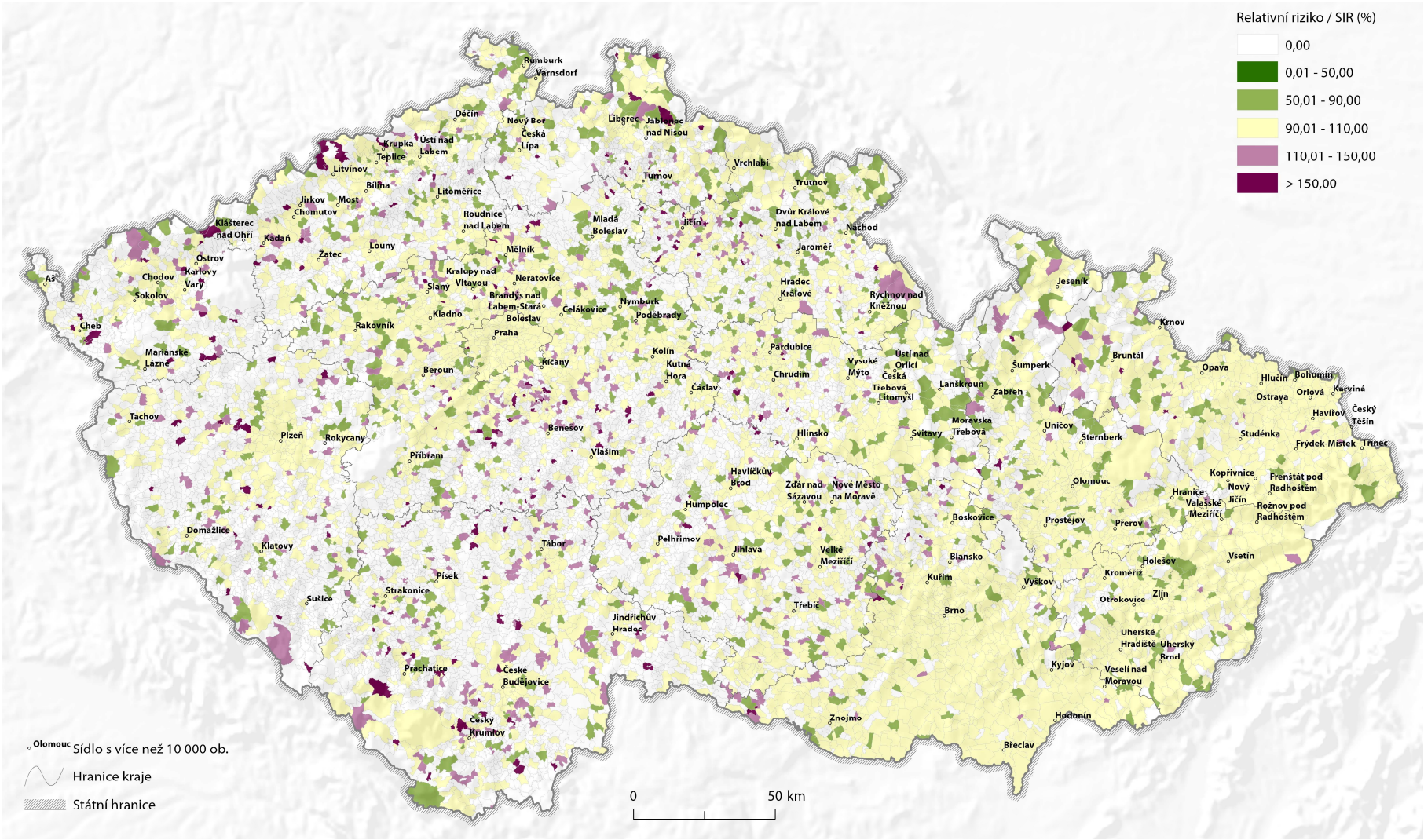
Příloha 7: Vyhlazená standardizovaná průměrná incidence kampylobakteriózy v částech obcí v ČR v letech 2008—2012 v počtech případů na 100 tisíc obyvatel, která vznikla pomocí lokálního Bayesova vyhlazení založeného na negativním binomickém rozdělení a sousedství typu královna 1. řádu



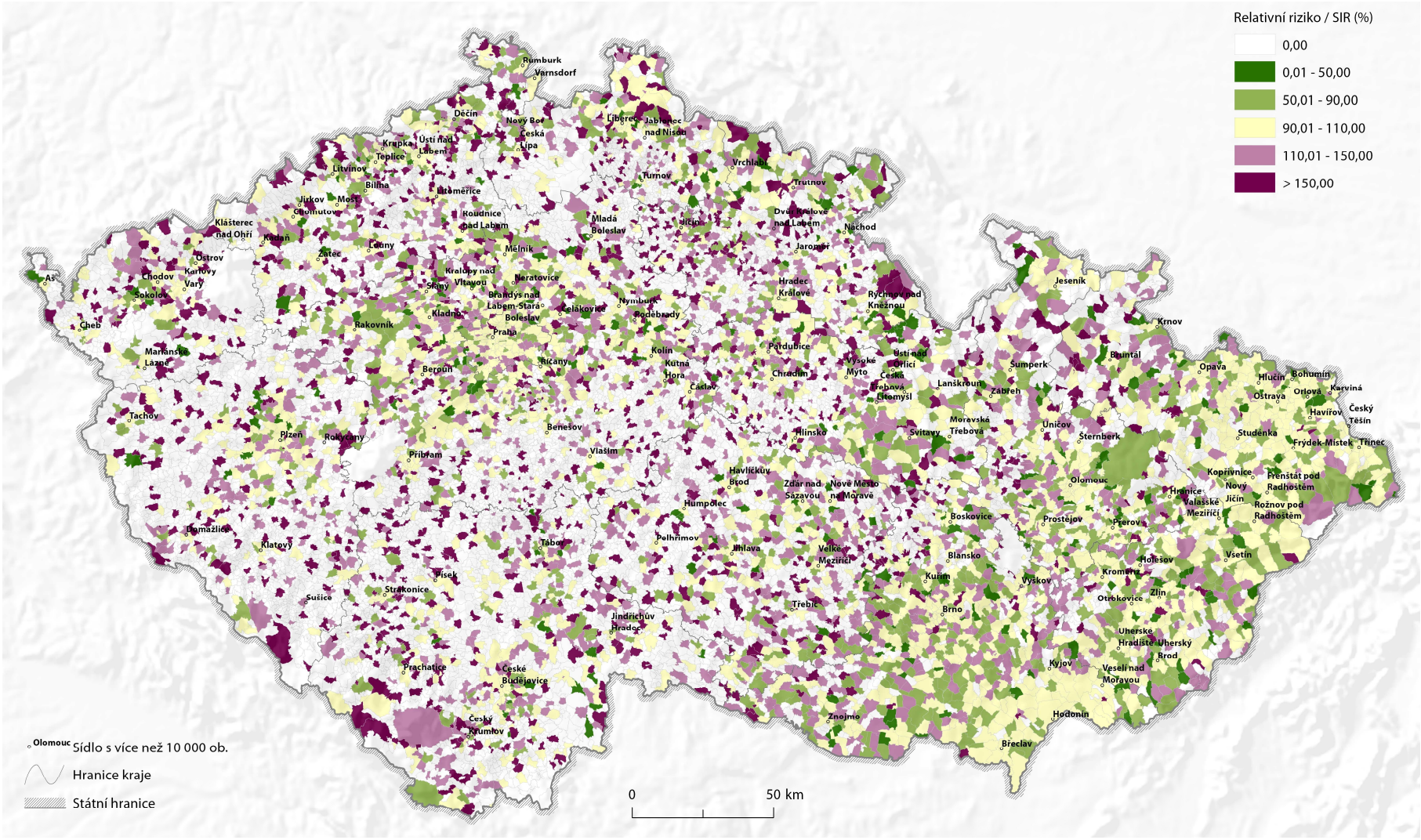
Příloha 8: Relativní riziko kampylobakteriózy (v %) v částech obcí v ČR v letech 2008—2012



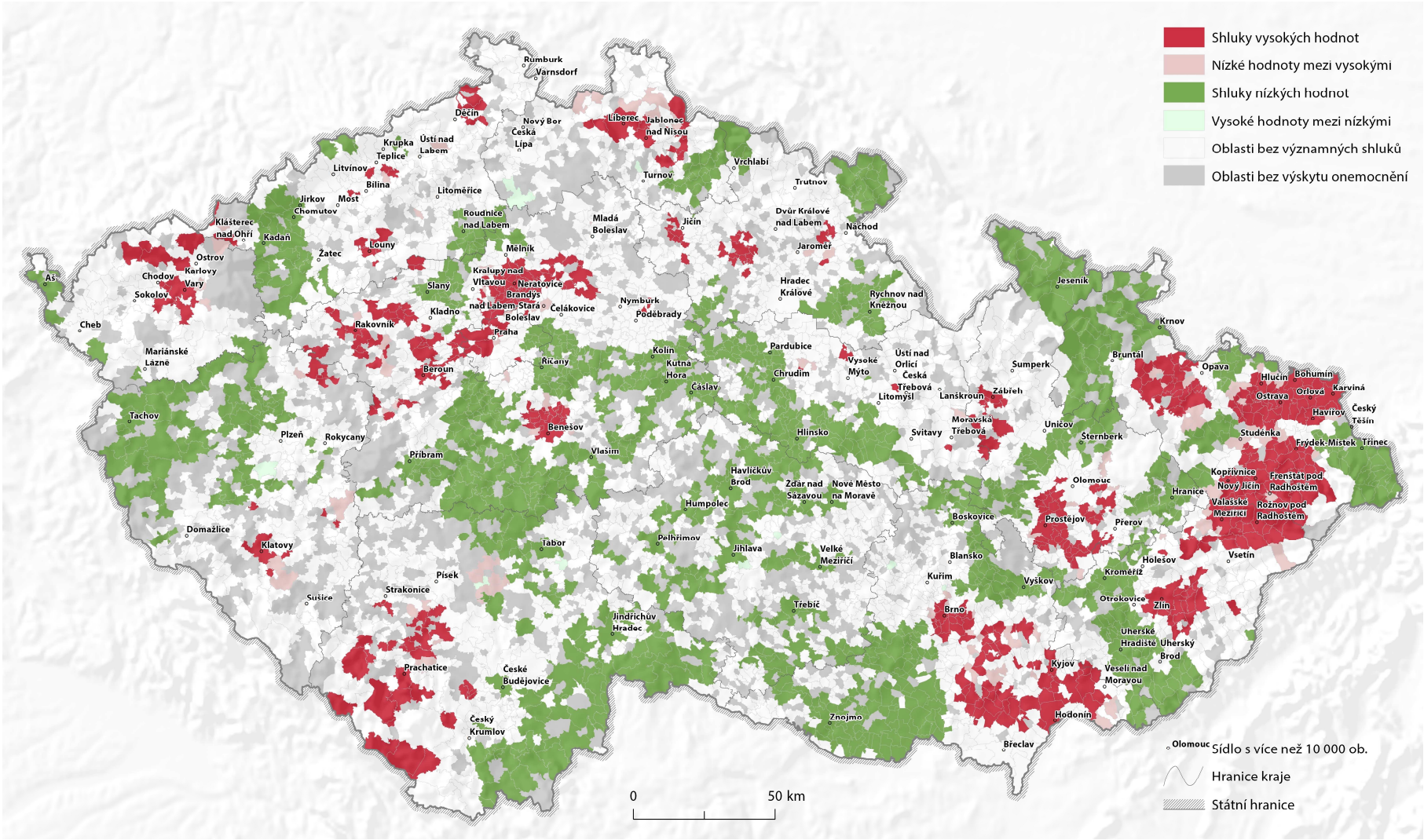
Příloha 9: Relativní riziko kampylobakteriózy (v %) v částech obcí v ČR v letech 2008—2012 získané na základě globálního Bayesova vyhlazení založeného na negativním binomickém rozdělení



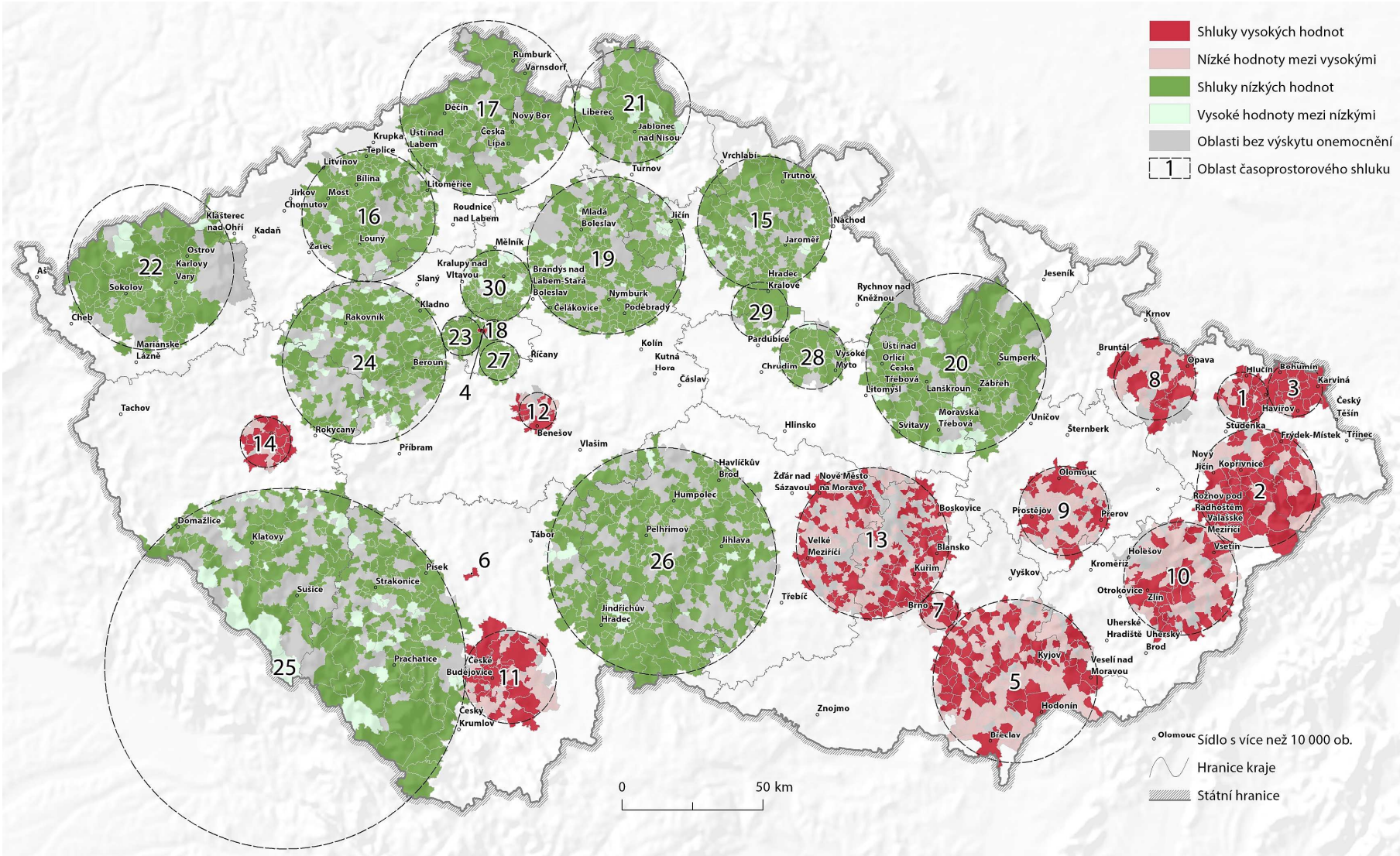
Příloha 10: Relativní riziko kampylobakteriózy (v %) v částech obcí v ČR v letech 2008—2012 získané na základě lokálního Bayesova vyhlazení založeného na negativním binomickém rozdělení a sousedství typu královna 1. řádu



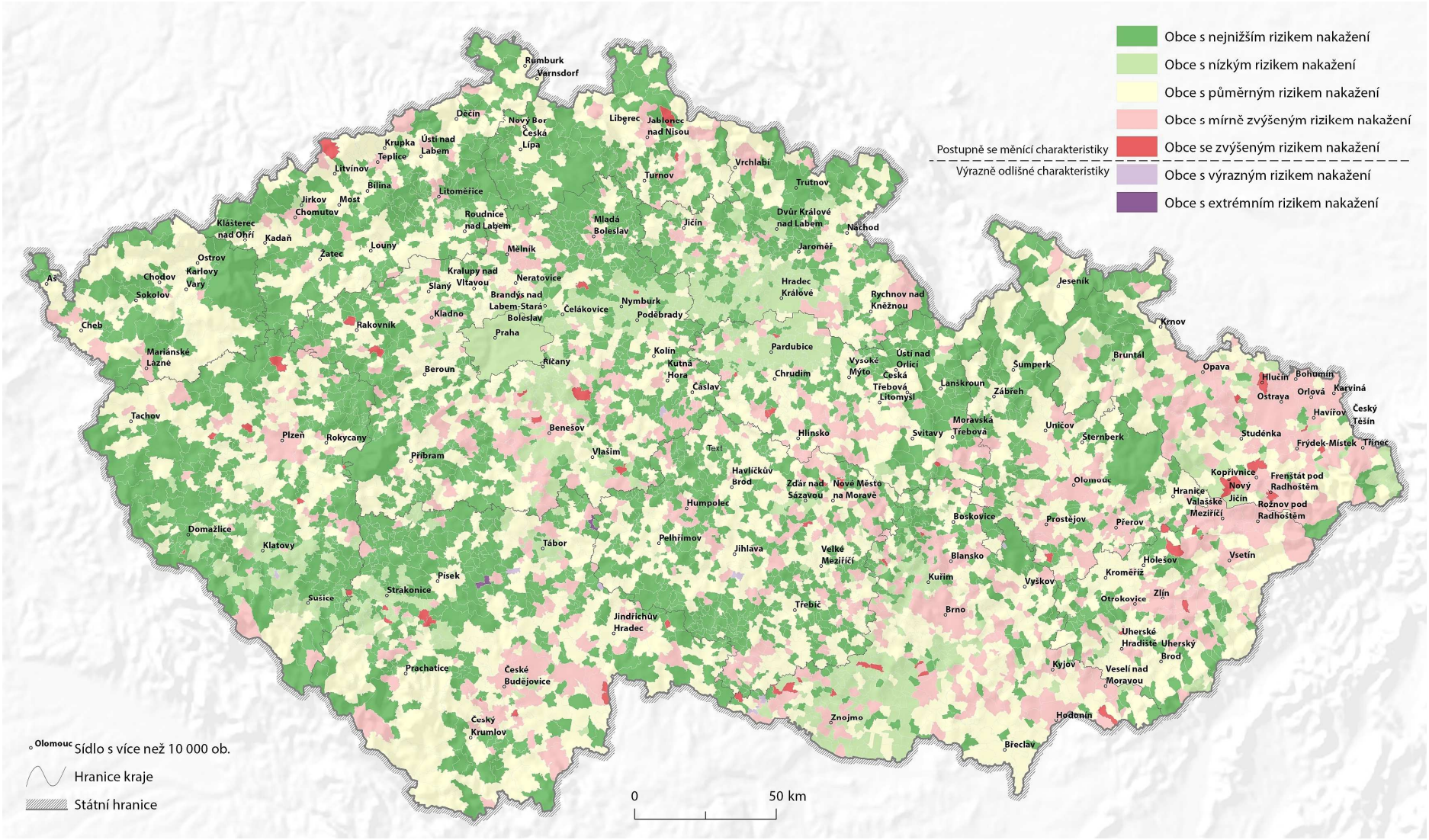
Příloha 11: Prostorové shluky vysokých a nízkých hodnot relativního rizika získané pomocí lokálního Moranova I kritéria s využitím empirického bayesovského principu a randomizace



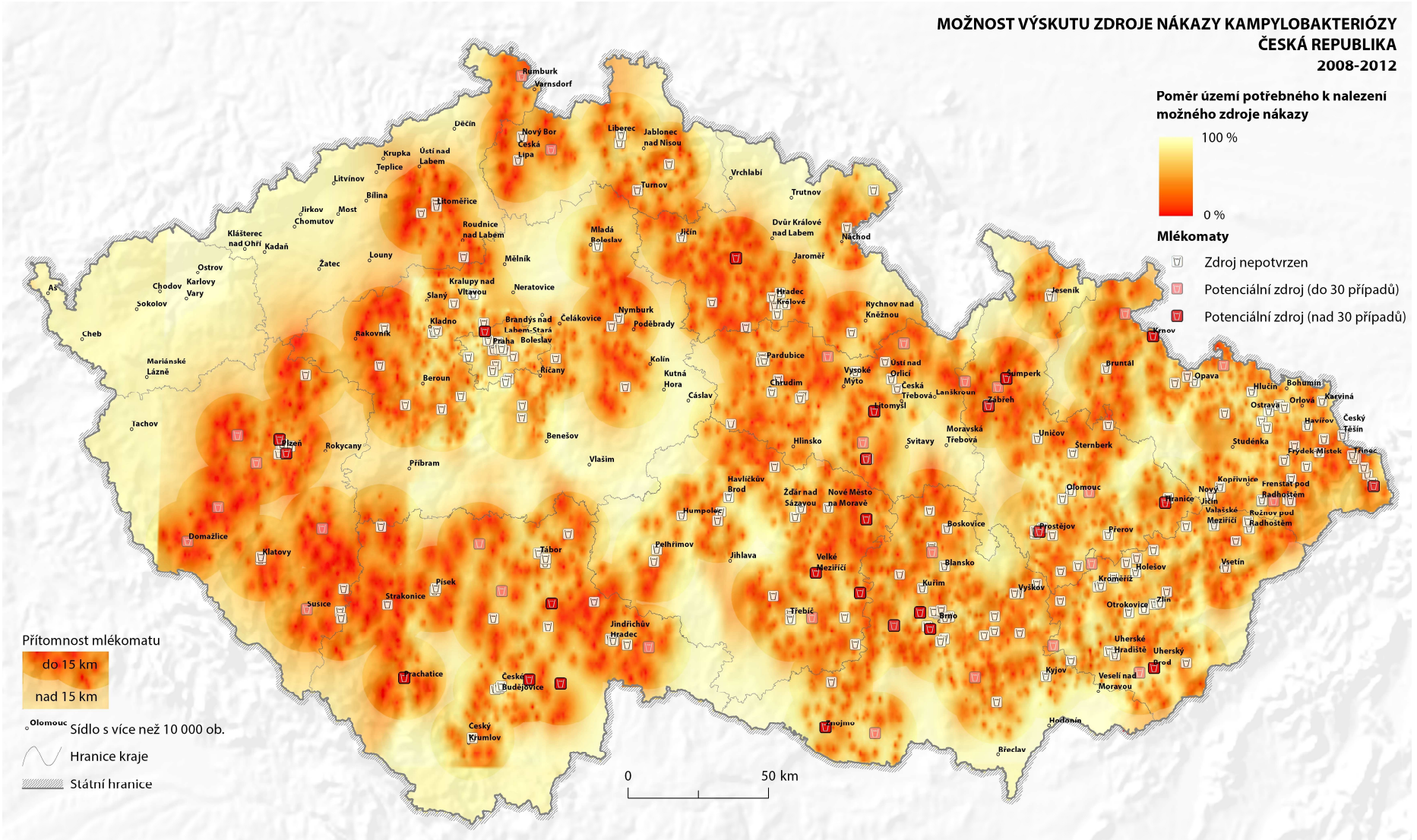
Příloha 12: Časoprostorové shluky onemocnění kampylobakteriózou v letech 2008—2012



Příloha 13: Mapa příslušnosti obcí k identifikovaným skupinám



Příloha 14: Geografický profil České republiky zohledňující polohu mlékomatů



Příloha 15: Ukázka kódu pro výpočet lokální ordinální logistické regrese

```
# popis postupu
1. nahrani SHP a vyber vstupnich promennych
2. kontrola NA a jejich odstraneni
3. vytvoreni matice sousedstvi - k-nejblizsich sousedu (musi byt relativne
   vysoke cislo)
4. postupne vypocty polr, vypocitani p-value, koeficienty a jejich odlogaritmovani
5. predikce a pseudo-R-squared
6. zapisovani koeficientu a jejich p-value do seznamu nebo matice
7. zapisovani predikce a jeji procentualni klasifikace do jednotlivych trid
7. vyhodoceni - cely model a postupne hodnoty koeficientu (min, max, mean)
8. prirazeni hodnot koeficientu a jejich odlogaritmovani

library(rgdal)
library(spdep)
library(MASS)

setwd("D:/Skola/SZU/doplnkova data")

# nacteni dat k modelu
load("D:/Skola/SZU/doplnkova data/data_model.RData")

# procisteni dat
d.rr <- data.model.r
d.rr <- na.omit(d.rr)

# nahrani a uprava  geodat
obce <- readOGR(dsn = "D:/Skola/SZU/doplnkova data/shp", "Obce_model2_bod")
#  obce  <-  spTransform(obce,  CRSobj  =  CRS("+proj=utm  +zone=33  +ellps=WGS84
+datum=WGS84 +units=m +no_defs"))
obce@data$ICOB <- as.character(obce@data$ICOB)
for(sloupec in 2:22) obce@data[,2] <- NULL
# vybere pouze obce bez NA
obce <- obce[obce$ICOB %in% row.names(d.rr),]

# funkce pocitajici pomer spravne zarazenych obci ku vsem, vstupem je tabulka
spravnosti klasifikace
# neboli confusion matrix / matice zamen
CCR <- function(x){sum(diag(x))/sum(x)*100}

# vytvoreni matice sousedstvi
ID <- obce@data$ICOB
dn <- dnearneigh(obce,0,45000, row.names = ID)
dn.ID <- dn
for(l in 1:length(dn.ID)) dn.ID[[l]] <- c(ID[l],ID[dn[[l]]])

# seznam pro nacitani jednotlivych modelu
lpolr.cat <- list()

# model postupne pocitajici hodnoty pro sousedstvi
# postupny vyber okolnich obci diky matici sousedstvi d.rr[row.names(d.rr) %in%
dn.ID[[1]],]
```

```
for(n in 1:length(dn)){
  tryCatch({
    mdata <- d.rr[row.names(d.rr) %in% dn.ID[[n]],-c(12,13,15)]
    m <- polr(Std_inc_2008_12 ~ ., data = mdata, Hess = TRUE, na.action = na.omit,
    method = "logistic")

    # predikce z modelu
    pred <- data.frame("pred" = predict(m))
    row.names(pred) <- row.names(m$model)

    # predikovana hodnota s kodem obce
    pred.val <- data.frame(predicted = pred[row.names(pred) %in% dn.ID[[n]][1],],
    row.names = dn.ID[[n]][1])

    # pridani p-value a konfidencnich intervalu k hodnotam koeficientu
    ctab <- coef(summary(m))
    p <- round(pnorm(abs(ctab[, "t value"]), lower.tail = FALSE) * 2,5)
    ci <- confint.default(m)
    ctab <- cbind(ctab, "CIl" = 0, "CIu" = 0 , "p value" = p)
    ctab[1:nrow(ci),c("CIl", "CIu")] <- ci

    # odlogaritmovana tabulka / Odd ratio
    ctab.exp <- ctab[,c(1,4,5,6)]
    ctab.exp[,1:3] <- exp(ctab[,1:3])

    # CCR
    confusion <- CCR(table(m$model$Std_inc_2008_12, pred[,1]))

    #zapsani vsech vysledku do seznamu
    l <- list(pred.value = pred.val, predict = pred, model = m, coefficients =
    ctab, odd.ratio = ctab.exp, classification = confusion)
    lpolr.cat[[n]] <- l

    # procisteni na konec cyklu
    l <- NULL
    gc(reset = T)

    # vypsani cisla cyklu
    print(n)
  }, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}

save(lpolr.cat, file = "local_polr.RR.RData")
```

KATEDRA GEOINFORMATIKY
Univerzita Palackého v Olomouci | Přírodovědecká fakulta

PROSTOROVÉ A VÍCEROZMĚRNÉ STATISTICKÉ ANALÝZY EPIDEMIOLOGICKÝCH DAT

AUTOREFERÁT DISERTAČNÍ PRÁCE

Studijní program: P1314 Geografie
Obor studia: 1302V011 Geoinformatika a kartografie
Školitel: doc. Mgr. Jiří Dvorský, Ph.D.

Mgr. Lukáš MAREK

SPATIAL AND MULTIVARIATE STATISTICAL ANALYSES OF EPIDEMIOLOGICAL DATA

Ph.D. THESIS SUMMARY

Study Programme: Geography
Specialization: Geoinformatics and Cartography
Supervisor: doc. Mgr. Jiří Dvorský, Ph.D.

Department of Geoinformatics
Faculty of Science, Palacký University Olomouc

Olomouc 2015

Disertační práce byla vypracována v prezenční formě doktorského studia na Katedře geoinformatiky Přírodovědecké fakulty Univerzity Palackého v Olomouci.
Dissertation thesis was compiled within Ph.D. study at the Department of Geoinformatics, Faculty of Science, Palacký University Olomouc.

Předkladatel / Submitter:
Mgr. Lukáš Marek

Školitel / Supervisor:
doc. Mgr. Jiří Dvorský, Ph.D.
Katedra geoinformatiky
Přírodovědecká fakulta Univerzity Palackého v Olomouci
17. listopadu 50
771 46 Olomouc

Oponenti / Reviewers:
Prof. Mgr. Jaroslav Hofierka, Ph.D. (Univerzita P.J. Šafárika, Košice)
Doc. RNDr. Dagmar Dzúrová, CSc. (Univerzita Karlova, Praha)
Doc. Mgr. Eva Fišerová, Ph.D. (Univerzita Palackého, Olomouc)

Autoreferát byl zaslán dne / Summary was posted on: _____

Obhajoba disertační práce se koná dne _____ před komisí pro obhajoby disertačních prací doktorského studia v oboru P1314 Geografie, studijním oboru 1302V011 Geoinformatika a kartografie, v prostorách Katedry geoinformatiky Přírodovědecké fakulty Univerzity Palackého v Olomouci, 17. listopadu 50, 771 46 Olomouc.
The defence of the dissertation thesis will be held on _____ at the commission for the defence of dissertation thesis of Ph.D. degree in study programme P1314 Geography, specialization Geoinformatics and cartography, in the premises of the Department of Geoinformatics, Faculty of Science, Palacký University Olomouc, 17. listopadu 50, 771 46 Olomouc.

S disertační prací je možno se seznámit na studijním oddělení Přírodovědecké fakulty Univerzity Palackého v Olomouci, 17. listopadu 12, 77 46 Olomouc.
The dissertation thesis is available at the Study Department, Faculty of Science, Palacký University in Olomouc, 17. listopadu 12, 771 46 Olomouc.

ISSN 1805-7500

ISBN 978-80-244-4547-2

Obsah

Contents

ANOTACE

Zkoumání zdraví a faktorů, které zdraví ovlivňují, jsou jedním z velmi aktuálních témat nejen v prostředí geověd, ale také v rámci širokého spektra dalších vědních oborů od medicíny po fyzikální vědy. Disertační práce z tohoto pohledu přispívá k rozšíření povědomí o možnostech prostorových analýz zdravotnických dat. Hlavní cíl disertační práce se zaměřuje na provedení komplexní prostorové analýzy rozšíření kampylobakteriózy v České republice v letech 2008—2012 s využitím metod geoinformatiky a prostorové statistiky. Pro potřeby disertační práce bylo vytyčeno pět cílů dílčích.

První dílčí cíl se zabývá jednak statistickým popisem onemocnění a jeho charakteristikou v české populaci, ale především mapováním prostorové distribuce onemocnění v ČR. Kromě sestrojení map morbidity bylo využito i časoprostorového krigingu k vytvoření spojitého povrchu týdenních hrubých incidencí České republice.

Druhý dílčí cíl se zabývá hodnocením a odhalováním prostorových a časoprostorových shluků obcí/městských částí s vysokým relativním rizikem, které jsou více náchylné k onemocnění než jejich okolí. Byl nalezen primární shluk a třináct shluků sekundárních se zvýšeným relativním rizikem v čase i prostoru.

Třetí dílčí cíl se zabývá identifikací možných faktorů podmiňujících rozšíření onemocnění či jeho útlum. Využito je regresních a klasifikačních postupů. V druhé části jsou obce ČR klasifikovány do skupin dle jejich společných charakteristických vlastností území a morbidity.

Čtvrtý dílčí cíl zkoumá vztah mezi přítomností automatů na čerstvé mléko a lokálními zvýšeními četností případů onemocnění. K tomu je využito metody geografického profilování založeného na bayesovských procesech a simulacích.

Pátý dílčí cíl využívá výsledků předchozích dílčích cílů, jejichž výstupy převádí do podoby vhodné k dalšímu interaktivnímu zkoumání v prostoru i čase.

Primárním cílem disertační práce sice bylo provedení komplexní prostorové analýzy kampylobakteriózy v České republice, ale stejně důležitý je i cíl vedlejší. Tím bylo poskytnout ucelený materiál, který pomůže dalším podobným studiím v orientaci v tématu prostorové epidemiologie a jejích metod a provede zájemce hlavními postupy, se kterými se v rámci studia prostorové distribuce epidemiologických dat může setkat.

Klíčová slova: Prostorová epidemiologie, mapování onemocnění, geovizualizace, geografická korelace, geografické profilování, prostorový vzor

1. Úvod

Je součástí přirozené lidské povahy hledat povědomé vzory i ve zdánlivě nahodilých situacích. Příkladem mohou být souhvězdí na noční obloze, mraky připomínající beránky nebo spojování teček na papíře, ze kterých vznikne smysluplný obrázek, jsou-li spojeny ve správném pořadí. Je-li tato zvídavost přenesena do vědeckého prostředí, pak formalizované metody připomínající právě zmíněné spojování teček a hledání vzorů mohou poskytnout nástroje vhodné k identifikaci a kvantifikaci prostorových vzorů reálných jevů vyskytujících se v prostředí a případně pomoci k odhalení jejich podmiňujících faktorů (Waller a Gotway, 2004).

V případě zdravotnických či konkrétně epidemiologických dat, je nejčastěji analyzována trojice „čas-osoba-místo" (Elliott a Best, 1998). V epidemiologických studiích nedávné minulosti ovšem převažuje důraz na první dva prvky z trojice, tedy čas a osobu, zatímco geografický aspekt byl dlouho v pozadí (Ostfeld et al., 2005). Situace se však změnila s rostoucí dostupností (prostorových) datových sad, programových prostředků schopných analyzovat prostorová data a samozřejmě i s výkonem současné výpočetní techniky, která umožňuje urychlit zpracování i výpočetně náročných úkolů (Marek et al., 2012). Lze hovořit o nově vzniklém interdisciplinárním vědním oboru, jehož základem je aplikovaná statistika, epidemiologie a geovědní obory, který může nést různá označení jako geografická epidemiologie, prostorová epidemiologie (Elliott et al., 2000), lékařská geografie nebo dokonce geomedicína (Davenhall, 2012).

Současně žijeme v době, kdy jsou podpora zdraví a s ní související studie, stejně jako pronikání informačních technologií do běžného života i vědy, v popředí zájmu laické i odborné veřejnosti. Díky tomu se i původně úzce specializované vědecké obory, jakým epidemiologie bezpochyby byla, stávají více interdisciplinárními. Geografické informační systémy (GIS) se v důsledku schopnosti efektivně spravovat, analyzovat a zobrazovat prostorová data staly důležitým nástrojem v geovědních oborech a také všude tam, kde je potřeba nebo možnost zpracovávat geodata. Geografické informační systémy proto nejsou vnímány pouze jako nástroje pro tvorbu jednoduchých tematických map, ale jako plnohodnotný analytický nástroj, jehož silná stránka tkví právě ve schopnosti prostorových analýz a odhalování prostorových souvislostí (Rezaeian et al., 2007). Takto se geografickým informačním systémům mimo jiné povedlo proměnit i analýzy zdravotnických dat, které jsou v současnosti jedním z nejaktuálnějších témat v geovědách, což dokazuje i množství nově vznikajících publikací (Davenhall, 2012; Pfeiffer et al., 2008), specializovaných sekcí nebo

odborných konferencí (např. Esri Health GIS Conference, GEOMED, Spatial Statistics Conference, …) nebo specializovaného software (Epi Info, SpaceStat, ClusterSeer5, SaTScan , atd.).

Předkládaná disertační práce představuje v ucelené formě možnosti aplikací prostorových a statistických analýz v souvislosti s epidemiologickými nálezovými daty. Zvolené teoretické postupy jsou prakticky využity v komplexní případové studii týkající se kampylobakteriózy, která je v současnosti nejrozšířenější bakteriální střevní infekcí v České republice (ÚZIS, 2013).

Hlavní motivací k vypracování této disertační práce byly zejména dvě skutečnosti. První skutečností je fakt, že ačkoliv je zdraví jedním z nejaktuálnějších témat současné společnosti, tak mu v geovědních disciplínách na našem území stále není věnována dostatečná pozornost. I přes existenci specializovaných pracovišť se k laické i odborné veřejnosti dostávají informace často v ne zcela vhodné či úplné podobě a prostorové analýzy a vazby často mohou být považovány za méně podstatné. Mapy či geovizualizace tak mohou být prezentovány a považovány jen za doplňující složku schopnou vyjádřit pouze jednoduché prvky reality a nikoliv za komplexní výstup, který může poskytnout komplexní popis situace a pomoci k pochopení celého procesu. Druhou skutečností a hlavním motivem případové studie je kampylobakterióza. Infekční onemocnění, které je vůbec jedno z nejčastějších v rozvinutých zemích, Českou republiku nevyjímaje, a stále je mezi veřejností spíše neznámé. Velkou výzvou bylo také samotné zpracování rozsáhlých (prostorových) dat a operace s nimi.

2. CÍLE PRÁCE

Hlavním cílem disertační práce je **provedení komplexní prostorové analýzy epidemiologických dat s využitím v současnosti dostupných technologií z oblasti geoinformatiky a prostorové statistiky**. Veškeré analýzy budou provedeny v souladu se standardními metodami prostorové epidemiologie a principy geografických informačních systémů. Analyzovaná data týkající se infekčního onemocnění kampylobakteriózy pocházejí z databáze EPIDAT, která je výstupem stejnojmenného programu sloužícího k zajištění povinného hlášení, evidence a analýzy výskytu infekčních nemocí v České republice. Prostorově se data vztahují k území České republiky, časově jsou omezena lety 2008 a 2012.

Disertační práce ve svých dílčích částech hledá odpovědi na otázky týkající se časoprostorové distribuce výskytu kampylobakteriózy, jejich vztahu k případným vnějším environmentálním rizikovým faktorům a identifikace možných zdrojů nákazy. Právě odpovědi na dotazy typu *„Je možné nalézt prostorové trendy a prostorové vzory okolo místa výskytu?“*, *„Je dále možné tyto vzory popsat a kvantifikovat pomocí exaktních metod?“* nebo *„Je riziko onemocnění stejné v celém regionu nebo se prostorově liší?“* či *„Je možné nalézt vztah mezi průměrným počtem nemocných danou chorobou a vlivem okolí?“*, poskytují v doktorské práci představené nástroje, metody a postupy.

Postup řešení, který vede k naplnění hlavního cíle práce, je s ohledem na metody a postupy prostorové epidemiologie rozdělen na následující dílčí cíle [DC]:

- Prvním dílčím cílem [DC1] je **mapování a celkový popis charakteristik výskytu kampylobakteriózy** v České republice v letech 2008—2012.
- Druhým dílčím cílem [DC2] je **průzkum, kvantifikace a vizualizace prostorových a časoprostorových vzorů** ve výskytu kampylobakteriózy v České republice v letech 2008—2012 a jejích vlastnostech.
- Třetím dílčím cílem [DC3] je **identifikace a analýza možných vztahů mezi výskytem onemocnění a vnějšími environmentálními, demografickými či socioekonomickými faktory** pomocí vícerozměrné statistiky a statistických modelů a následné hodnocení jejich přesnosti a využitelnosti pro predikci ohrožení vybraných územních jednotek kampylobakteriózou. Současně je cílem také klasifikace územních jednotek do skupin na základě jejich podobných vlastností a atributových vzorů souvisejících s výskytem onemocnění.

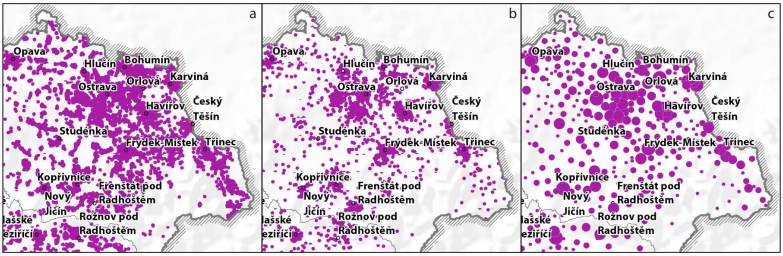
- Čtvrtým dílčím cílem [DC4] je **zhodnocení přítomnosti automatů na čerstvé mléko jako potenciálních bodových zdrojů nákazy** kampylobakteriózou v jejich okolí.
- Pátým dílčím cílem [DC5] je **převedení vybraných výsledků jednotlivých DC do podoby vhodné k další interaktivní exploraci v prostoru i čase**.

Výjimečnost disertační práce leží především ve dvou rovinách. První je mezioborový přesah studie a zapojení pokročilých metod prostorové statistiky (hodnocení prostorové autokorelace, geograficky vážené modelování, geoprofiling) a GIS v epidemiologii. Ačkoliv jsou metody a postupy prostorové epidemiologie známy a ve světě používány, tak na našem území se jejich aplikace vyskytují pouze v omezeném množství. Druhou rovinou je potom měřítko analýz, které probíhají zejména na lokální úrovni (obcí či jejich částí). Dosavadní studie provedené v ČR totiž nejčastěji zahrnují pouze úrovně okresů.

3. DATOVÁ ZÁKLADNA

Základní datovou sadou v disertační práci jsou data pocházející z databáze EPIDAT (Epidemiologická databáze), která byla poskytnuta Státním zdravotním ústavem v Praze. Databáze obsahuje kompletní záznamy o výskytu kampylobakteriózy na celém území České republiky v letech 2008—2012. Tato datová sada je anonymizovaná, což znamená, že není známo jméno pacienta s hlášenou nemocí a ani jeho přesná adresa, takže lokalizace je možná do úrovně uliční sítě. Kromě věku a pohlaví pacienta jsou v databázi obsaženy i další atributy jako datum hlášení, zaměstnání, diagnóza, národnost apod. Datová sada obsahuje téměř 100 tisíc záznamů se 78 atributy s různou úplností a kvalitou vyplnění, které je nutné zkontrolovat, vyčistit a opravit. Zkontrolovaným datům byla na základě místa nakažení a/nebo místa bydliště přiřazena poloha pomocí geokódování s využitím API poskytovaným mapovým portálem Mapy.cz (Obr. 1).

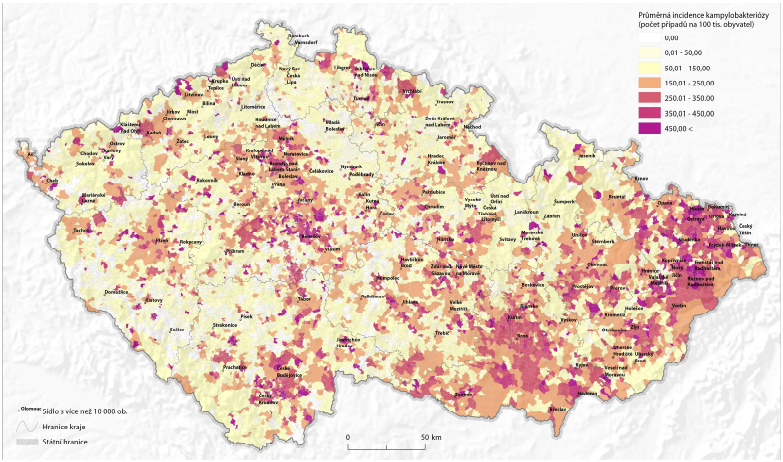
Dalšími datovými sadami, které byly využity v analýzách, jsou zejména ty týkající se demografické struktury obyvatelstva a jeho ekonomické aktivity. Příkladem jsou např. data ze Sčítání lidí, domů a bytů 2001 a 2011, Registr emisí a zdrojů znečištění ovzduší atd. Dále byla použita data fyzicko-geografická jako digitální modely terénu, klimatické jevy (průměrná teplota vzduch, atmosférické srážky), využití půdy, radonové riziko, znečištění ovzduší apod. nebo data odvozená (např. vzdálenost od vodních toků nebo silnic, apod.). Pro analýzy a vizualizace byla použita data administrativních jednotek, případně pravidelné sítě pro agregaci dat.



Obr. 1 Poloha případů kampylobakteriózy vyjádřená pomocí tečkové metody: a) topografický způsob, b) topografický způsob - vážené tečky – 1 tečka odpovídá 10 případům, c) kartogramový způsob - vážené tečky – velikost tečky odpovídá kategorii množství případů v území

4. Mapování onemocnění: Co nám mohou prozradit mapy? [DC1]

Hlavním úkolem DC1 je s pomocí základních statistických metod a analýzy časových řad, poskytnout celkový přehled o výskytu případů kampylobakteriózy a jejich základních charakteristikách na území České republiky ve zkoumaných letech. V kapitole jsou přehledně zobrazeny výčty a statistiky nemocných s ohledem na jejich věk, pohlaví, či zaměstnání a to jak v numerické, tak i v grafické podobě.

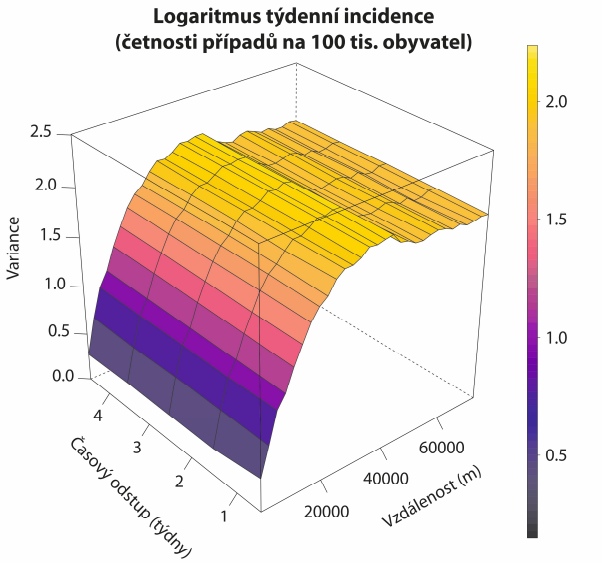


Obr. 2 Vyhlazená standardizovaná průměrná incidence kampylobakteriózy v částech obcí v ČR v letech 2008—2012 v počtech případů na 100 tis. obyvatel, která vznikla pomocí lokálního Bayesova vyhlazení založeného na negativním binomickém rozdělení a sousedství typu královna 1. řádu

Současně s tím je provedeno mapování výskytu nemoci v podobě map měr morbidity (nemocnosti). Konkrétně je mapována incidence jako relativní míra onemocnění ve zvolené územní jednotce a standardizovaná incidence, kterou je možno považovat za míru relativního rizika postižení zkoumaného území onemocněním (Bivand et al., 2008). Při výpočtu měr morbidity bylo využito především nepřímé věkové standardizace. Vzhledem k malému rozsahu dat u některých mapovaných jednotek jsou kromě map standardizovaných měr morbidity vytvořeny také jejich shlazené ekvivalenty. Pro výsledné shlazení jsou využity metody globálního empirického Bayesova odhadu, kde je pro vyhlazování měr v jednotlivých areálech aplikován konstantní průměr a rozptyl vycházející z negativně binomického rozdělení. Dále je využito také lokálního (adaptivního) empirického Bayesova vyhlazování, kde jsou průměrná hodnota

a rozptyl využité pro vyhlazování odhadu měr morbidity definovány na základě sousedství 1. řádu typu královna (Obr. 1Obr. 2).

Poslední částí DC1 je vytvoření a vizualizace souvislého povrchu incidence v týdenních intervalech, který je oproštěn od zatížení administrativních hranic obcí. Základem pro tvorbu spojitého povrchu je grid hustoty zalidnění a agregovaná data výskytu onemocnění. K vytvoření spojitého povrchu je využito časoprostorového krigingu, který na rozdíl od tradiční varianty krigingu umožňuje zohlednit pro tvorbu spojitého povrchu nejen prostorovou strukturu (varianci) jevu, ale také strukturu změny jevu v čase i v čase a prostoru současně (Pebesma a Gräler, 2014). Zmíněné vztahy jsou vyjádřeny časoprostorovým variogramem (Obr. 3), který tvoří základ pro interpolaci spojitého povrchu incidence síti o dvoukilometrové hraně buňky. Výsledný povrch je pro účely vizualizace zobrazen pouze pro obydlená území.

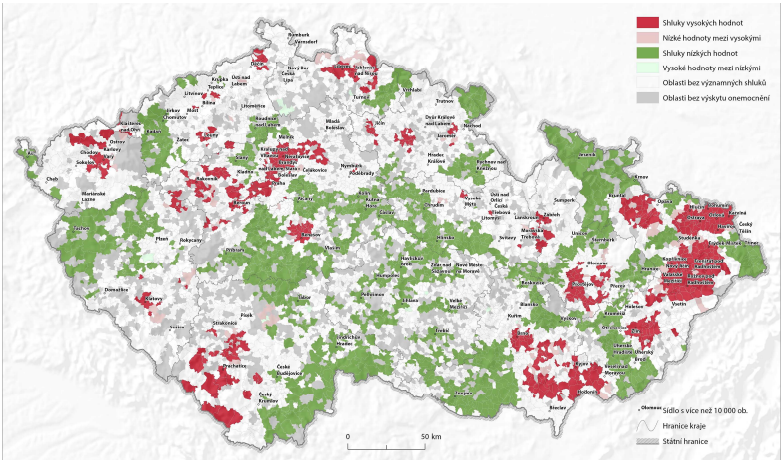


Obr. 3 Časoprostorový teoretický model variogramu použitý pro interpolaci

Výsledky statistické analýzy a mapování slouží jako základ pro stanovení hypotéz týkajících prostorových a časoprostorových vzorů v území, které jsou konkrétně kvantifikovány a vyjádřeny numericky i vizuálně v DC2.

5. PODOBNOSTI VÝSKYTU ONEMOCNĚNÍ V ČASE I PROSTORU [DC2]

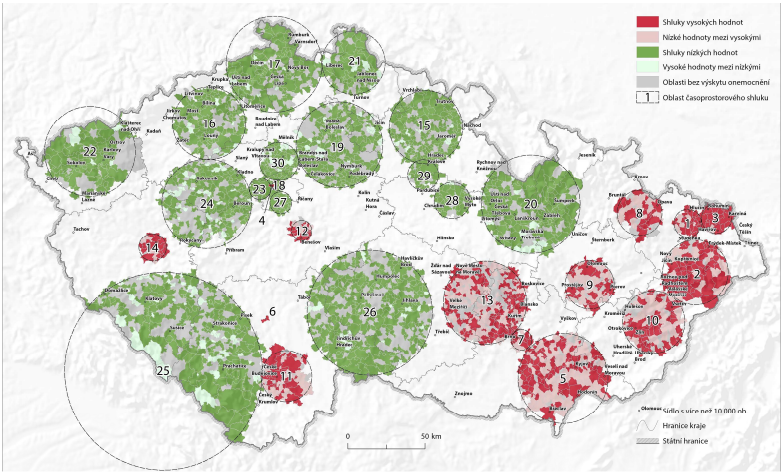
Vizuálním vyhodnocením DC1 je zjištěno, že v rámci České republiky existovaly v letech 2008—2012 spojité oblasti se zvýšeným výskytem kampylobakteriózy a současně i oblasti, které byly touto nemocí téměř neb zcela nepostihnuty. Druhý dílčí cíl (DC2) přímo navazuje na tato zjištění a snaží se tento jev kvantifikovat, kvantifikaci vizualizovat a následně i popsat. K řešení DC2 je využito několika hlavních metod a postupů prostorové a časoprostorové statistiky.



Obr. 4 Prostorové shluky vysokých a nízkých hodnot relativního rizika získané pomocí lokálního Moranova I kritéria s využitím empirického bayesovského principu a randomizace

V prvním kroku je za pomoci metod průzkumu prostorové autokorelace zkoumán pouze prostorový vzor. K tomuto je využito lokálních indikátorů prostorové asociace (LISA) a především lokálního Moranova I (Anselin, 1995), založeném na základě sousedství 1. řádu typu královna. Pro srovnání je použita jak základní varianta tohoto indexu, tak i varianta využívající k lepšímu odhadu skutečného prostorového vzoru jevu lokálního empirického Bayesova vyhlazování v kombinaci s randomizací a určením významnosti výsledků pomocí permutačních testů (Obr. 4). Vizualizovány jsou oblasti na hladině významnosti do 5 %. Výsledkem jsou mapy identifikující oblasti shluků vysokých nebo nízkých hodnot, tedy s vyšším/nižším výskytem onemocnění,

a pak tzv. „*outliers*“ neboli území, která vybočují z trendu v jejich okolí. Tato vybočující území představují buď oblasti nízkých hodnot v blízkosti více nemocí postižených oblastí, nebo naopak území s výskytem vyšším poblíž oblastí, které nejsou výrazně zasáhnuty.

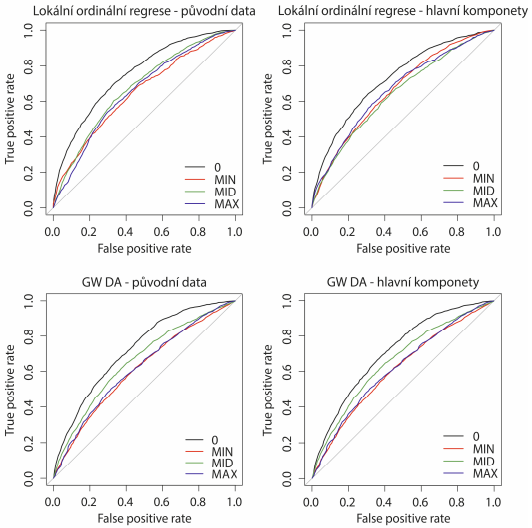


Obr. 5 Časoprostorové shluky onemocnění kampylobakteriózou v letech 2008—2012

V dalším navazujícím kroku DC2 je k průzkumu prostorového vzoru onemocnění přiřazen i rozměr času. Jednotlivé případy nákazy jsou agregovány do územní jednotky (městských částí) v týdenních intervalech a současně rozlišením věku a pohlaví. Agregovaná data jsou následně srovnávána s podobně strukturovanými demografickými údaji území (počet obyvatel, věková struktura a struktura pohlaví) s pomocí metody *spatio-temporal scan statistics* - SaTScan (Kulldorff et al., 2005). Metoda využívá k odhalení shluků s podobnou strukturou simultánního skenování území v čase i prostoru. Výsledky jsou podobné jako v případě LISA, ovšem s tím rozdílem že je současně indikována primární shluk i shluky sekundární. V případě této analýzy byla zvolena maximální velikost shluku odpovídající 3 % celkové populace a maximálním dobou trvání 50 % časového rozsahu dat, ovšem s výjimkou shluků vyskytujících se po celé období. Analýza identifikovala tři desítky shluků (14 shluků vysokých hodnot a 16 nízkých) (Obr. 5).

6. Analýza vztahů mezi onemocněním a vnějšími faktory prostředí [DC3]

Předchozí dílčí cíle DC1 a DC2 odhalily nenáhodný proces shlukování obcí s vyšší incidencí kampylobakteriózy a lze tak stanovit předpoklad, že v daných územích bude existovat i environmentální, demografický či socioekonomický faktor, který může přispívat ke zvýšené incidenci a relativnímu riziku. V první fázi třetího dílčího cíle jsou z velkého množství zdrojů nashromážděna a sjednocena data, ze kterých je vytvořena datová sada vlastností území obcí obsahující více než stovku charakteristik ke každé obci České republiky. Pro účely modelování však nejsou takto rozsáhlá data vhodná ani potřebná a postupně byla datová sada pomocí korelační analýzy, logických úvah, dostupných literárních zdrojů a konzultací redukována až na 11 spolu téměř nekorelujících charakteristik. Výsledná sada byla ještě pomocí analýzy hlavních komponent redukována na pět nově vytvořených proměnných, které slouží jako vstup do modelů. Jako závislá proměnná, případně jako klasifikační schéma je využito relativní riziko (SIR - standardized incidence ratio), podle kterého jsou obce rozděleny do čtyř skupin (nulové, minimální, průměrné a vysoké riziko).



Obr. 6 ROC křivky pro lokální ordinální regresi a GW DA

Analýza vztahů mezi onemocněním a vnějšími faktory prostředí je postupně realizována pomocí metod vícerozměrného regresního modelování, prostorových regresních modelů a metod strojového učení a data miningu. Jednotlivé metody a modely jsou hodnoceny pomocí ROC křivek a AUC (Obr. 6). Vzhledem k nízké úspěšnosti většiny modelů je pro srovnání aplikován negativní binomický regresní model pro případy s velkým množstvím nul, kterým je modelován průměrný počet případů v obcích (Tab. 1).

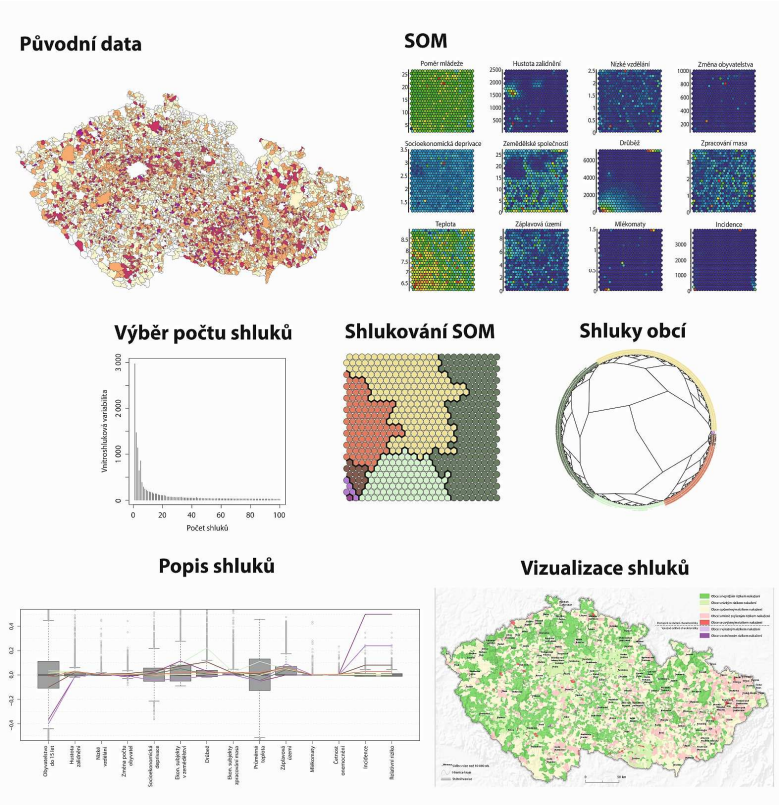
Tab. 1 Vyhodnocení klasifikačního výkonu regresních modelů a diskriminační analýzy, hodnocená jako celek i pro jednotlivé skupiny; Acc – přesnost hodnocení (%), AUC – plocha pod křivkou (%); tučně jsou označeny nejvyšší hodnoty Acc a AUC pro jednotlivé třídy i celkově

	Multinomická regrese		Ordinální regrese		Lokální ordinální regrese		Diskriminační analýza		Prostorově vážená diskriminační analýza	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
	Lineární kombinace reprezentativních charakteristik									
0	**67,9**	75,0	53,9	68,6	59,7	**75,3**	47,5	70,6	49,2	72,4
MIN	65,7	63,3	**74,5**	60,0	1,8	65,0	73,6	60,8	44,5	**65,9**
MID	10,6	64,9	0,0	55,0	**74,4**	**67,7**	13,4	63,3	42,3	66,1
MAX	0,0	57,7	0,0	51,2	0,3	**65,7**	0,0	57,1	**29,7**	63,4
Celková	45,7	65,2	42,2	58,7	**47,2**	**68,4**	43,2	63,0	45,3	67,0
Vážená	36,1		32,1		34,1		33,7		**41,5**	
	Lineární kombinace hlavních komponent									
0	45,9	68,5	36,3	68,3	**54,1**	**72,9**	45,4	68,5	53,3	71,4
MIN	74,6	60,8	**82,7**	53,3	64,3	**66,4**	75,3	60,6	60,1	61,2
MID	8,2	62,3	4,5	62,1	26,5	64,3	10,8	62,0	**32,4**	**66,0**
MAX	0,0	58,1	0,0	53,7	5,5	**66,6**	0,0	57,6	**6,4**	62,1
Celková	41,8	62,4	40,9	59,4	**45,6**	**67,6**	42,5	62,2	45,3	65,2
Vážená	32,2		30,9		37,6		32,9		**38,0**	

0 – obce bez zaznamenaného výskytu onemocnění; MIN – SIR ≤ 0,80; MID - 0,80 < SIR < 1,50; MAX - SIR ≥ 1,50; Celková – počet správně klasifikovaných obcí / průměrné AUC; Vážená – vážená průměrná přesnost určení zohledňující četnosti obcí ve skupinách

Předchozí metody prostorového shlukování představené v DC2 jsou založeny pouze na incidenci v jednotlivých částech České republiky a její geografické podobnosti s okolními územními jednotkami (prostorové autokorelaci). V poslední části DC3 je navíc zkoumána podobnost v atributové rovině. Na základě analýzy důležitosti a vhodnosti doplňkových dat pro modelování relativního rizika onemocnění v obcích jsou vybrány potenciálně nejdůležitější doplňková data a také míry morbidity. Tato data slouží jako vstupní údaje pro shlukovou analýzu. Jejím cílem je identifikovat skupiny obcí s podobnými

vlastnostmi v rámci skupiny tak, aby se od sebe jednotlivé skupiny současně co nejvíce odlišovaly (Hebák et al., 2005). Pro účely nalezení podobných datových struktur je v práci využita kombinace samoučící se neuronové sítě, konkrétně Kohonenovy samoorganizační mapy (SOM) a hierarchického shlukování. Cílem SOM je vytvořit mapu skupin co nejpodobnějších objektů, které je možné vizualizovat ve 2D prostoru a současně co nejvíce zachovat jejich topologii (Kohonen, 1982, 2001). V procesu shlukování může komplexní výstup ze SOM doplnit výpočetně jednodušší míry podobnosti, které tvoří základ hierarchického shlukování. Buňky SOM tak mohou být kategorizovány a následně ohodnoceny i územní jednotky do buněk spadající. Tímto postupem je vytvořeno, popsáno a vizualizováno 7 hlavních skupin územních jednotek podle environmentálních a socioekonomických vlastností a vlastností spojených s onemocněním kampylobakteriózou (Obr. 7).

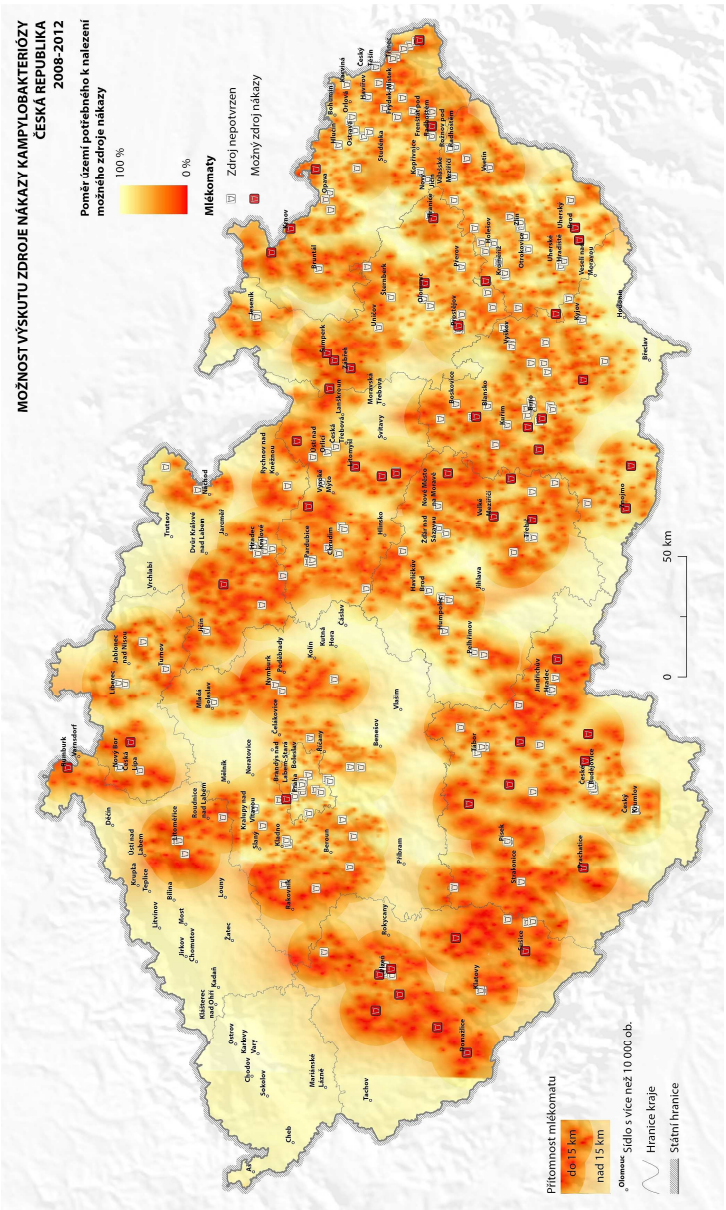


Obr. 7 Analytický postup použitý během shlukování obcí

7. GEOPROFILING: IDENTIFIKACE MOŽNÝCH ZDROJŮ INFEKCE [DC4]

Čtvrtý dílčí cíl případové studie měl jako hlavní motiv identifikaci možných zdrojů nákazy. Po konzultaci s MUDr. Michaelem Vítem, Ph.D. (bývalým hlavním hygienikem ČR a v současnosti vedoucím Centra hygieny práce a pracovního lékařství Státního zdravotního ústavu) byly za možné zdroje nákazy zvoleny automaty na prodej čerstvého mléka (tzv. mlékomaty), které se začaly objevovat právě mezi roky 2008—2012. Informace o umístění mlékomatů pochází z několika zdrojů – z portálu registrovaných subjektů Státní veterinární správy, kde ovšem nejsou veškerá historická data, ale pouze v současnosti funkční mlékomaty a dále z Venkovského fóra nebo přímo od provozovatelů, např. společnost TOKO. Celkem bylo testováno 267 automatů na prodej čerstvého mléka, které byly umístěny v letech 2008—2012.

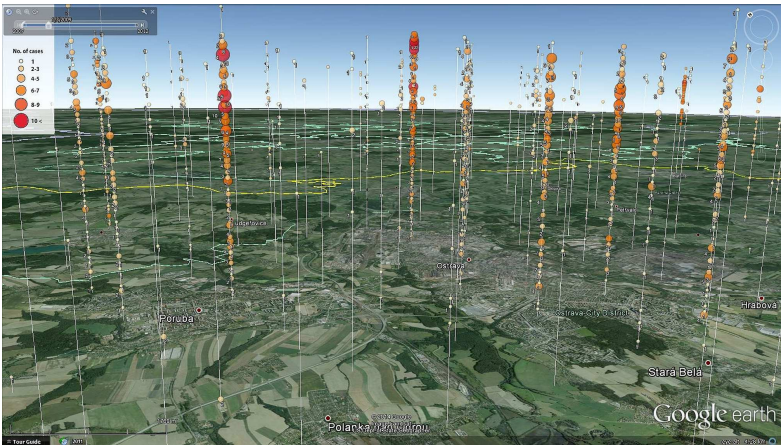
Pro identifikaci mlékomatů jako potenciálního zdroje nákazy byla zvolena metoda geografického profilování, která byla původně využívána zejména v kriminalistice, ale v posledních letech je úspěšně aplikována také v biologii a především v prostorové epidemiologii jako nástroj sloužící k vyhledávání a hodnocení mnohonásobných zdrojů nákazy (Le Comber et al., 2011; Verity et al., 2014). Vyhodnocování je založeno na Dirichletově modelu pro smíšené procesy (DPM - Dirichlet process mixture model) a jde tedy o bayesovské modelování shluků s pomocí Markovových řetězců (MCMC). Vyhodnocování jednotlivých bodových zdrojů následně probíhá na základě tzv. hitscore, které udává velikost oblasti, kterou je potřeba prohledat k nalezení zdroje. Vzhledem k výpočetní náročnosti operace bylo nutné nejdříve redukovat datovou sadu míst nákazy na okolí do 15 km od mlékomatu a dále území rozdělit do 18 oblastí. Pomocí geografického profilování bylo zjištěno, že množství mlékomatů mohlo být v průběhu svého fungování zdrojem nákazy kampylobakteriózou. Současně byl ovšem zjištěn i nedostatek modelu, kdy jsou oblasti s nízkým výskytem případů onemocnění nadhodnocovány jako rizikové. Reálný odhad je tedy nižší a pohybuje se kolem 10 % zkoumaných mlékomatů (Obr. 8).



Obr. 8 Vizualizace povrchu histscore (pravděpdobnosti) a mlékomaty identifikované jako potenciální zdroje nákazy

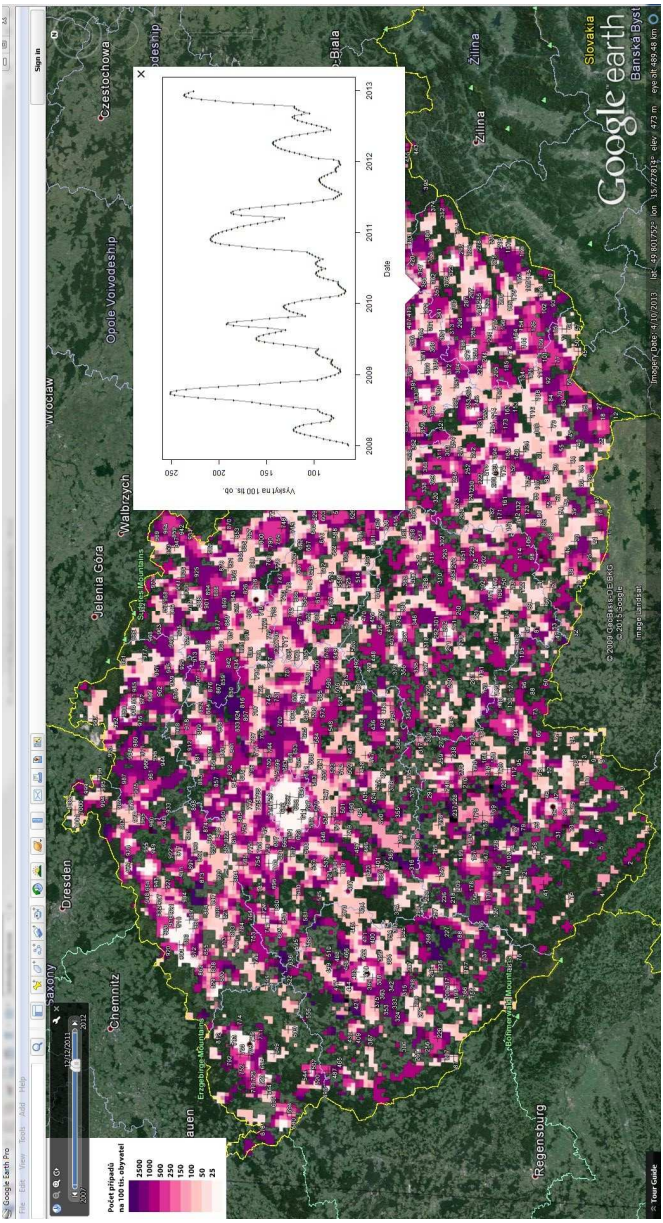
8. GEOVISUAL ANALYTICS: GOOGLE EARTH JAKO NÁSTROJ PRO PREZENTACI A PRŮZKUM DAT V ČASE I PROSTORU [DC5]

Výstupem prostorových analýz bývají nejčastěji statické mapy a geovizualizace, které sice umožnují vhodným způsobem sdělit jejich výsledky, ale problémem může být jejich další využitelnost, případně kombinace s dalšími zjištěními. Náplní posledního dílčího cíle (DC5) je zpřístupnění nejdůležitějších výsledků dříve definovaných v DC1—DC4 v podobě, která umožní jejich snadný interaktivní průzkum jak v prostoru, tak ideálně i v čase s pomocí časových značek.



Obr. 9 Bublinový kartodiagram v Google Earth

Výsledky analýz jsou upraveny do podoby vhodné k interaktivnímu zobrazení a z původních formátů exportovány do formátu KML (Google, 2009), který je standardem pro výměnu a zobrazení geografických dat (Open Geospatial Consortium, 2008). Jako nástroj vhodný k další geovizuální analýze je zvolen Google Earth, který je v současnosti vůbec nejrozšířenější prohlížečkou prostorových dat (Hengl, 2007) a jeho ovládání je natolik známé a intuitivní, že je hojně využíván jak laiky, tak i odborníky. Jeho nevýhodou je ovšem nutná příprava dat v prostředí GIS (extenze *Export to KML* pro ArcMap 10.1 nebo pluginu *GEarthView* pro QGIS), případně s využitím vhodných balíků statistického software **R** (např. *plotKML* a *raster*). Takto připravená data umožňují zmíněné komplexní interaktivní časoprostorové hodnocení pomocí geovizuální analýzy. Pro interaktivní zobrazení jsou zvoleny vizualizace agregovaných nálezových dat (Obr. 9), spojitého povrchu týdenní incidence (Obr. 10), statistik časoprostorového shlukování, roční incidence a geografického profilování.



Obr. 10 Zobrazení časoprostorového krigingu v prostředí Google Earth, který díky podpoře časové složky dat umožňuje snadné prohlížení a (geo)vizuální analýzu

9. VÝSLEDKY

Disertační práce představuje příspěvek do jedné z dynamicky se rozvíjejících oblastí výzkumu v prostředí geovědních oborů – geografie zdraví či prostorové epidemiologie. Jak už je z názvu (oboru i samotné práce) patrné jde o interdisciplinární obor, který využívá a kombinuje poznatky, data a metody ze zdravotnictví, geografie i prostorové statistiky. Díky tomuto spojení je disertační práce komplexním souhrnem k tématu prostorové epidemiologie, který se postupně zabývá všemi jejími hlavními směry. Zvolené metody a postupy jsou prezentovány na případové studii, která se týká rozšíření infekčního onemocnění kampylobakteriózy v České republice v letech 2008—2012. Toto onemocnění, ač není mezi veřejností příliš známé, je vůbec nejrozšířenější bakteriální střevní onemocnění nejen v České republice, ale i v celém rozvinutém světě (Weisent et al., 2011).

Hlavním cílem disertační práce bylo provedení komplexní prostorové analýzy epidemiologických dat s využitím v současnosti dostupných technologií z oblasti geoinformatiky a prostorové statistiky. Veškeré analýzy byly provedeny v souladu se standardními metodami prostorové epidemiologie a principy geografických informačních systémů. V rámci hlavního cíle bylo definováno pět na sebe navazujících dílčích cílů, které dohromady umožnily vytvořit si obrázek o prostorové distribuci kampylobakteriózy v České republice a jejích možných podmiňujících faktorech.

DC1 – Mapování a popis charakteristik výskytu kampylobakteriózy v České republice v letech 2008—2012

Kompletním řešením prvního dílčího cíle se zabývá kapitola **Chyba! Nenalezen zdroj odkazů.**, která postupně představila základní úlohy spojené s mapováním nemocí. Základní datovou sadu pocházející z databáze EPIDAT, která bylo nutné pročistit a zejména geokódovat. Ke geokódování byl sestaven skript, který bez omezení umožňuje lokalizaci záznamů na základě API od Mapy.cz. Geokódováno bylo 98,5 tis. záznamů do úrovně uliční sítě (bez adresních bodů).

Neprostorovým hodnocením průběhu onemocnění a jeho charakteristikami jako celku z pohledu základní statistiky se zabývala. V té byla nejdříve zjišťována průměrná incidence podle věku a pohlaví pacientů. Jednoznačně nejvíce ohroženou skupinou obyvatelstva jsou děti do věku 4 let (chlapci o něco více než dívky), zvýšená incidence onemocnění je patrná i u osob do 30 let nezávisle na pohlaví. S tímto faktem souvisí i nejčastější profese pacientů (dítě/žák/student – dohromady 58,3 % případů), nejčastější nakaženi jsou zaměstnanci v potravinářství (2,3 %), u 22 % případů však není stav zjištěn. U více než 42 %

případů nakažení zůstává neurčen zdroj nákazy, což může výrazně ovlivnit další prostorové i neprostorové analýzy. Nejčastějším identifikovaným zdrojem nákazy je pak maso – kuřecí 38,5 %, uzeniny (6,22 %), vepřové maso (3,5 %). Kromě toho je možné se nakazit od domácích mazlíčků, z mléka, vody nebo syrové zeleniny a ovoce. Četnost onemocnění (i incidence) stoupala pomalu od roku 2008 až k přelomu roku 2010 a 2011, kdy měla nemoc vrchol a od té doby počet případů klesá. V průběhu roku jsou u všech věkových kategorií nejrizikovějšími měsíci červen—září.

V DC1 byly mapovány četnosti kampylobakteriózy, průměrná hrubá a standardizovaná/vyhlazená incidence a relativní riziko (SIR – nepřímo standardizovaná incidence). Nejdříve byly pomocí tečkové metody, anamorfózy a pseudokartogramů sestaveny mapy pro agregované četnosti, hexagonovou síť nahrazující územní celky a obce ČR. Dále bylo využito pro mapování průměrné míry incidence a relativního rizika (SIR) metod bayesovského shlazování, které umožnilo větší srovnatelnost hodnot v územních jednotkách a lepší vnímání prostorového vzoru. Doporučit lze zejména lokální bayesovské vyhlazení v případě incidence, naopak méně lze vyhlazování doporučit v případě SIR. Poslední část mapování nemocí využívá k časoprostorovému mapování nemocí časoprostorový kriging, díky kterému byl vytvořen spojitý povrch hrubé týdenní incidence pro osídlená místa České republiky. Díky tomu lze sledovat změnu průběhu incidence v průběhu roku a vnímat časoprostorové vzory.

Při sledování prostorové distribuce na základě kartogramů je vizuálně patrná asociace hustoty zalidnění a hustoty případů onemocnění. Výrazně více postižená je oblast Moravy a Slezska (především severovýchod Moravy, Slezsko a Brněnsko), v Čechách jde potom o jižní Čechy, Plzeňsko a oblast jihovýchodně od Prahy (Benešovsko). Při současném sledování času a prostoru s využitím interpolovaného spojitého povrchu je potvrzena sezonalita, která je u tohoto onemocnění běžná. Sezónní změny jsou pak méně patrné ve městech a více patrné v zázemí měst a na venkově, tedy oblastech, které mohu sloužit k rekreaci městského obyvatelstva během letní sezóny. V podhorských oblastech (především Jeseníků a Krkonoš) je možné pozorovat zvýšení incidence i během zimní sezóny příp. jarních prázdnin.

DC2 – Podobnosti výskytu onemocnění v čase a prostoru

Při řešení druhého dílčího cíle byl nejprve pomocí metod analýzy globální a lokální prostorové autokorelace hodnocen čistě prostorový vzor relativního rizika kampylobakteriózy v jednotlivých obcích a městských částech. Tímto postupem byla potvrzena na základě map měr morbidity vizuálně stanovená hypotéza o více nakažených oblastech na východě a severovýchodě Moravy a ve

Slezsku a dále pak na Českobudějovicku. Kromě toho byla vyhodnocena jako shluk vysokých hodnot i oblast severně od Prahy a Benešovsko. Ve srovnání s mapami není jako významný shluk identifikováno Plzeňsko. Celkem bylo do shluků vysokých hodnot zařazeno 8 % obcí/městských částí, do shluku nízkých to bylo incidencí 19 %. I když je téměř 2,5× více obcí a městských částí v oblastech shluků nízkých hodnot než v oblastech shluků vysokých hodnot, tak populačně je situace srovnatelná. V oblastech shluků nízkých hodnot žije odhadem 2,27 mil. osob (21,8 % populace) a v oblastech shluků s vysokou incidencí je to dokonce 2,44 mil. osob (23,4 % populace).

Navazujícím postupem pak bylo hodnocení časoprostorového vzoru pomocí metody časoprostorového skenování. Cílem časoprostorového skenování bylo identifikovat a vyhodnotit shluky vysokých a nízkých hodnot relativního rizika, tzn. shluky ohrožených a zdravých obcí jak v prostoru, tak i v čase. Vstupní data se sestávala z případů rozdělených dle pohlaví a věku agregovaných prostorově dle jejich příslušnosti k obci/městské a časově v týdenním intervalu. Relativní riziko bylo zjišťováno díky známé demografické struktuře územních jednotek. Jeden shluk zahrnoval maximálně 3 % populace při délce trvání maximálně 50 % či celé sudované období. Pomocí časoprostorového skenování bylo identifikováno celkem 30 shluků (14 shluků zvýšeného rizika onemocnění kampylobakteriózou a 16 rizika nižšího). Primární shluk byl umístěn do oblasti Ostravska a bezprostředně na něj navazovaly i další shluky zvýšeného rizika. Diverzifikace Čech a Moravy je patrná ještě více než v případě pouze prostorového shlukování. Většinu ze shluků je možné pozorovat na místě po celé studované období, ačkoliv několik jich trvalo pouze po omezenou dobu, což je případ Plzeňska, Českých Budějovic, Blanenska, jižní Moravy a Valašska, resp. Královéhradecka a Vysočiny. Celkem žije v identifikovaných shlucích vysokých hodnot až 2,6 mil. osob., tzn., že přibližně 25 % obyvatelstva ČR žije v oblastech se zvýšeným rizikem nakažení kampylobakteriózou, zatímco 3,9 mil. obyvatel (37 % populace ČR) žije v oblastech, kde je relativní riziko nakažení nižší.

DC3 – Analýza vztahů mezi onemocněním a vnějšími faktory prostředí

Identifikací a analýza možných vztahů mezi výskytem onemocnění a vnějšími environmentálními, demografickými či socioekonomickými faktory a klasifikací územních jednotek do skupin na základě podobných charakteristik území a charakteristik morbidity v území se postupně zabývá DC3. V DC3 bylo nejdříve pomocí korelační analýzy z množství charakteristik území vybráno jedenáct typických charakteristik, které buď mohou podmiňovat prostorovou distribuci onemocnění, nebo mohou vhodně reprezentovat území. Tyto charakteristiky byly podrobeny analýze lokální prostorové korelace s relativním

rizikem onemocnění (SIR) a také analýze lokální autokorelace pro dvě proměnné, aby bylo možné zjistit, zda existují oblasti, kde jsou oba jevy vzájemně asociovány. Nejsilnější vazby byly vyhodnoceny v souvislosti s hustotou obyvatelstva a průměrnou teplotou, dále pak s rozsahem záplavových území a ekonomickými subjekty v oblasti zemědělství a zpracování masa. Pomocí PCA byla redukována dimenze dat na pět hlavních komponent, pomocí geograficky vážené PCA pak byla zkoumána proměnlivost největších zátěží v hlavních komponentách. Při nastavení doporučeného adaptivního kernelu byly nejvýznamnějšími zátěžemi hlavních komponent hustota obyvatelstva, masozpracující ekonomické subjekty, zemědělské subjekty a relativní změna počtu obyvatel.

Možnost predikce kampylobakteriózy, resp. skupin obcí podle relativního rizika onemocnění v obci, byla testována s využitím generalizovaných lineárních modelů a jejich modifikací. Predikční úloha byla díky transformaci predikované proměnné přeformulována na úlohu klasifikační, ke které byly využity diskriminační analýza, metody strojového učení, data miningu a neuronových sítí. Pro predikci absolutních četností případů onemocnění v obcích se ukázal jako nejvhodnější negativní binomický model s nadbytečnými nulami. Nejlépe hodnocenými klasifikačními postupy byly metody lokální ordinální regrese, geograficky vážené diskriminační analýzy z metod jednodušších a dále z komplexnějších postupů Random Forest, neuronová síť a Support Vector Machine s radiální funkcí báze. Klasifikací i regresními postupy se potvrdila skutečnost, že téměř polovina případů kampylobakteriózy u nás i ve světě zůstává nevysvětlena. Koncept lokální ordinální regrese je využíván zcela sporadicky a byl sestaven pro účely disertační práce.

Kromě predikce a klasifikace byla cílem modelování především explorace proměnlivosti vztahů mezi vybranými charakteristikami a výskytem a intenzitou onemocnění. Identifikována byla souvislost mezi hustotou zalidnění a průměrnou teplotou a distribucí onemocnění, lokálně byl také předpokládán vliv socioekonomické determinace a nízkého vzdělání obyvatelstva. Nepříliš výrazný se ukázal efekt obyvatelstva do 15 let.

Posledním úkonem tohoto dílčího cíle byla klasifikace obcí České republiky do skupin podle charakteristik prostředí a charakteristik nemocnosti. K tomuto kroku byla využita shluková analýza, která v kombinaci se samoorganizačními mapami identifikovala na území ČR sedm skupin obcí. Tyto skupiny byly popsány a vizualizovány v mapě. V pěti ze sedmi skupin jde o pozvolnou proměnu charakteristik a nemocnosti v obcích, ale dvě velmi malé skupiny jsou výrazně odlišné. Potvrzeny byly zjištění z předchozích dílčích cílů, kdy se na

severovýchodní Moravě a ve Slezsku vyskytuje více ohrožených obcí než ve zbytku republiky.

DC4 – Zhodnocení přítomnosti automatů na čerstvé mléko jako potenciálních bodových zdrojů nákazy kampylobakteriózou

Čtvrtý dílčí cíl disertační práce měl jako hlavní motiv identifikaci možných bodových zdrojů nákazy obyvatelstva kampylobakteriózou. Začátkem roku 2010 se spolu s rostoucím počtem automatů na přímý prodej čerstvého mléka rozhořela i diskuze o možných zdravotních problémech, které jeho konzumace může způsobit. Původní nadšení obyvatelstva bylo utlumeno vyjádřením tehdejšího hlavního hygienika upozorňujícího na tyto problémy (Ministerstvo zdravotnictví ČR, 2010), které mělo dopad na prodej čerstvého mléka z automatů a v důsledku toho i zpomalení růstu množství automatů a u některých i ukončení jejich činnosti (Andrlová, 2011). Hlavním cílem této části disertační práce bylo pokusit se zjistit, zda v letech 2008—2012 opravdu mohly některé z mlékomatů být lokálním zdrojem nákazy i přes to, že je kvalita mléka v automatech i u jeho producentů pravidelně kontrolována.

Pro identifikaci mlékomatů jako potenciálního zdroje nákazy byla zvolena metoda geografického profilování, která byla původně využívána zejména v kriminalistice, ale v posledních letech je úspěšně aplikována také v biologii a především v prostorové epidemiologii jako nástroj sloužící k vyhledávání a hodnocení mnohonásobných zdrojů nákazy (Le Comber et al., 2011; Verity et al., 2014). Celkem bylo testováno 267 automatů na prodej čerstvého mléka, které byly umístěny v letech 2008—2012. Vyhodnocování samotné je založeno na Dirichletově modelu pro smíšené procesy a jde tedy o bayesovské modelování shluků s pomocí Markovových řetězců. Kvůli výpočetní náročnosti bylo území České republiky rozděleno do 18 oblastí podle množství případů kampylobakteriózy v okolí do vzdálenosti 15 km od mlékomatu. Pomocí geografického profilování bylo zjištěno, že až 52 mlékomatů (19,5 %) mohlo být v průběhu svého fungování zdrojem nákazy kampylobakteriózou. Současně byl ovšem zjištěn i nedostatek modelu, kdy byly i některé oblasti s nízkým výskytem případů onemocnění nadhodnocovány jako rizikové. Reálný odhad je tedy nižší a pohybuje se kolem 10 % zkoumaných mlékomatů. Zde je důležité zmínit, že geografické profilování neslouží k potvrzení zdrojů nákazy, ale pouze se snaží o jejich vytipování na základě podobného prostorového chování zkoumaného jevu. To, že byl mlékomat vybrán jako potenciální zdroj lokální nákazy, ještě neznamená, že tímto zdrojem opravdu je. Daná situace samozřejmě může platit i naopak.

Zajímavá byla situace na Českobudějovicku, která byla původcem celé diskuze o možných zdravotních dopadech konzumace čerstvého mléka na obyvatelstvo. Pomocí časoprostorové analýzy vzorů bylo Českobudějovicko opravdu identifikováno jako riziková oblast v období 12. ledna — 22. února 2010, ale přímo v Českých Budějovicích nebyl identifkován žádný mlékomat jako potenciální zdroj nákazy. Jako potenciálně podezřelé ovšem byly identifikovány mlékomaty v blízkém Lišově či Třeboni.

Kromě samotné analýzy byla také modifikována vizualizační část původního skriptu, která nyní lépe reflektuje pravděpodobnostní povrch. Kromě vizualizace výsledků geografického profilování byly vytvořeny také jednoduché mapy hustoty (*heat maps*) výskytu případů kampylobakteriózy a mlékomatů a prodejců čerstvého mléka umožňující rychlé vizuální srovnání intenzity jevu v prostoru.

DC5 – Převedení vybraných výsledků jednotlivých DC1–DC4 do podoby vhodné k další interaktivní exploraci v prostoru i čase

Hlavním úkolem DC1 bylo mapování kampylobakteriózy, zatímco hlavními úkoly DC2—DC4 byla především analýza dat. Pátým dílčím cílem, který z těch předchozích výrazně čerpal, je převedení vybraných výsledků jednotlivých dílčích cílů do podoby vhodné k další interaktivní exploraci v prostoru i čase. Současně byla zvolena forma a podoba prezentace, která umožňuje další geovizuální zkoumání uživatelům, kteří nemají zkušenosti nebo možnost využít geografických informačních systémů. Vybrané výsledky jednotlivých dílčích cílů byly vizualizovány a tato vizualizace byla převedena do formátu KML, který je standardem OGC pro výměnu prostorových dat. Výsledná data je tak možné bez problémů dále využívat v prostředí GIS nebo pouze využít možnosti zobrazování např. v některém z virtuálních glóbů. Pravděpodobně nejrozšířenější prohlížečkou prostorových dat i dat časoprostorových, která je využívána laiky i odborníky a je využita i v případě této disertační práce, je Google Earth od společnosti Google. Jeho výhodou je možnost snadno prozkoumávat prostorová data, která obsahují i časovou složku, což je případem několika analýz provedených v rámci disertační práce.

Konkrétně byly v podobě KML resp. KMZ souborů prezentovány tyto geovizualizace a výsledky prostorových analýz:

- nálezová data agregovaná v týdenních úsecích během let 2008—2012 do pravidelné sítě o hraně 2 km vizualizovaná formou tzv. *bubble chart* tedy bublinového kartodiagramu;
- roční hrubé incidence v obcích České republiky;

- spojitý týdenní povrch hrubé incidence kampylobakteriózy vzniklý pomocí časoprostorového krigingu;
- časoprostorové shlukování obcí České republiky z pohledu relativního rizika výskytu kampylobakteriózy;
- geografické profily a mlékomaty jako potenciální zdroj lokálních rozšíření kampylobakteriózy.

Vytvořené vizualizace je možné zobrazit ve virtuálním glóbu Google Earth a vizuálně hodnotit prostorovou distribuci rozšíření kampylobakteriózy ve vybraném období nebo pomocí časového posuvníku či animací srovnávat vývoj během různých časových období a intervalů. Vrstvy je možné prohlížet samostatně nebo je kombinovat pro lepší pochopení jevu. Google Earth, či Google Earth Pro nedávno uvolněný do bezplatné distribuce, byly vyhodnoceny jako použitelné nástroje pro geovizuální analýzu a komunikaci výsledků. I přes nesčetné výhody však je potřeba kombinovat ho s dalšími nástroji, které připraví zpracovávaná data do podoby vhodné k zobrazení a dalšímu geovizuálnímu hodnocení.

10. Diskuze

Poskytování a přístup k dostatečně podrobným záznamům, a nemusí se nutně jednat o záznamy individuální, je pravděpodobně vůbec nejcitlivějším tématem kvůli ochraně osobních údajů, zachování anonymity uživatele a zabránění jeho zpětné identifikace na základě poskytnutých dat. Časté tvrzení o nedostatku kvalitních a dostatečně podrobných dat z oblasti zdraví nemusí být vždy pravdivé. Otázkou by nemělo být pouze, zda taková data existují, ale především zda a za jakých podmínek jsou data dostupná, případně jaká omezení pro jejich využití existují a jak užitečné mohou být výsledky analýz, které z dat vycházejí.

S polohovým učením dat a jejich agregací souvisí i pojem ekologická chyba. Výsledky vzniklé v určitém měřítku agregace dat často není možné zcela generalizovat na jiná měřítka či až na individuální úroveň.

Mapování kampylobakteriózy a vizualizace výsledků

Je všeobecně doporučováno využít relativních měr (např. incidence), které vyjadřují počet případů na množství obyvatel. Rovněž se kvůli srovnávání míst s různou strukturou populace doporučuje využívat standardizovaných měr, ačkoliv některé studie týkající se SIR ukázaly, že srovnání hodnot mezi geografickými jednotkami bude zavádějící pouze v případě extrémně odlišných populací, což se např. v rámci jednoho státu či oblasti děje v praxi velmi zřídka (Goldman a Brender, 2000; Jarup, 2004). Při využití bayesovského vyhlazování je potřeba mít na paměti, že tyto metody mají tendenci posouvat hodnoty blíže k lokálnímu či globálnímu průměru. Ačkoliv se mohou mapy hrubých měr zdát příliš fragmentované a náročné k interpretaci, tak na druhou stranu vyhlazené míry mohou být příliš homogenní a tak maskovat skutečnou distribuci jevu (Beale et al., 2008). Využití krigingu v disertační práci nebylo z důvodu predikce, tak jak je to u této metody běžné, ale představuje alternativu k algoritmům dasymetrických metod.

I přes četné výhody Google Earth je potřeba zdůraznit, že jde především o prohlížečku geografických dat a všechna data, tak musela být vhodně připravena dříve, než byla postoupena geovizuální analýze. KML soubory jako nosné médium se jeví jako vhodné, problémem však může být velikost souborů především v případě, kdy tyto obsahují velké množství vektorových dat, což má za následek pomalé načítání.

Prostorové a časoprostorové vzory kampylobakteriózy

Existuje všeobecný předpoklad, že je často možné nalézt prostorový vzor přírodních i socioekonomických jevů, který se přirozeně v objevuje. Je ale také

potřeba si uvědomit, že některé metody ho mohou odhalit i tam, kde existovat nemůže, např. v náhodně vygenerovaných prostorových datech jak to bylo prokázáno v případě GWR. Pro exploraci prostorové distribuce kampylobakteriózy byly použity metody průzkumu prostorové autokorelace - LISA a časoprostorové skenování. Pro úspěšné provedení obou metod je potřeba vhodně zvolit matici prostorových vztahů a také parametry analýz (populaci, kovariáty, časové okno) pro jednotlivé územní jednotky.

Možné rozdílnosti ve výsledcích geoprofilingu mohou být způsobeny změnami v nastavení simulace nebo definování vzdálenosti, kde dochází k prostorové autokorelaci, ta je totiž definována empiricky na základě grafu simulací a může tak docházek k značně subjektivnímu odhadu. Chybným předpokladem může teoreticky být i předpoklad prostorového chování obyvatelstva. Geoprofiling identifikoval téměř 20 % mlékomatů jako potenciálních zdrojů nákazy. Po dalším zhodnocením výsledků je však možné počet snížit na 10 %, protože metoda zvýhodňovala některé oblasti s velmi nízkým počtem nakažení. To, že byl mlékomat vybrán jako potenciální zdroj lokální nákazy, však ještě neznamená, že tímto zdrojem opravdu je.

Modelování, klasifikace a analýza vztahů mezi onemocněním a faktory prostředí

První otázkou, která mohla ovlivnit predikční výkon odhadu chování kampylobakteriózy pomocí zmíněných postupů, byla volba prediktorů. Sice byla vytvořena datová sada o velkém množství socioekonomických, demografických i environmentálních charakteristik obcí, ale z nich bylo vybráno pouze jedenáct vlastností reprezentujících všechny ostatní. Otázkou však je, zda by se při volbě jiných modelů či jiných proměnných by ukázala výrazná asociace neodhalená vybranými charakteristikami nebo zda je výskyt kampylobakteriózy dán především hustotou osídlení a demografickou strukturou spíše než charakteristikami území. Pro účely DC3 byl sestaven fungující koncept lokálního ordinálního modelu, který byl otestován pro klasifikaci a ve srovnání s tradičními metodami prokazoval dobrý výkon. Jeho současnou nevýhodou je ale komplikace hodnocení regresních koeficientů a dalších charakteristik vzhledem k faktu, že nejde o jeden model, ale o souhrn velkého množství modelů – každá obec má svůj model s vlastními koeficienty vycházející z definice prostorových vztahů mezi obcemi.

Matice podobnosti obcí nebyla záměrně prostorově vážena, aby nebyla zakryta lokální proměnlivost. Skupiny byly hodnoceny na základě jejich středních hodnot a směrodatných odchylek a v jednotlivých skupinách se tak může vyskytovat velký rozptyl hodnot.

11. Závěr

Hlavním cílem disertační práce sice bylo provedení komplexní prostorové analýzy epidemiologických dat s využitím v současnosti dostupných technologií z oblasti geoinformatiky a prostorové statistiky. Vedlejším cílem disertační práce však bylo také poskytnout ucelený materiál, který pomůže dalším podobným studiím v orientaci v tématu prostorové epidemiologie a jejích metod a provede zájemce hlavními postupy, se kterými se v rámci studia prostorové distribuce epidemiologických dat může setkat. Téma disertační práce a její interdisciplinární povaha předpokládaly, že pro její úspěšné provedení bylo nutné kombinovat poznatky nabyté studiem geoinformatiky spolu s poznatky dalších oborů jako prostorová a aplikovaná statistika či epidemiologie.

Hlavní cíl disertační práce byl pro účely zpracování rozdělen do pěti na sebe navazujících dílčích cílů, které společně pokrývaly ústřední témata, jimiž se zabývá prostorová epidemiologie, a to mapováním nemocí, identifikací prostorových shluků a geografickými korelačními studiemi. Mapováním a vizualizací výsledků analýz se zabývaly první a pátý dílčí cíl. V prvním případě byla mapována prostorová distribuce kampylobakteriózy – její četnost, incidence a také relativní riziko. Kromě toho byly doporučeny i metody vyhlazování, které v jednotlivých případech využít. Využitím metody časoprostorového krigingu jako nástroje časoprostorového dasymetrického mapování bylo docíleno možnosti sledování průběhu onemocnění v prostoru i čase současně formou spojitého povrchu. Právě využití časoprostorového krigingu, je jednou z inovací, kterou disertační práce vnáší do tématu mapování morbidity. Úspěšně bylo také využito geovizualizací ve formě KML souborů možných zobrazit v prostředí geografických informačních systémů i ve virtuálním glóbu Google Earth. Google Earth v rámci práce představuje platformu pro geovizuální časoprostorové hodnocení předem připravených témat a kromě toho jde díky rozšíření tohoto programu také o vhodný způsob další komunikace výsledků laické i odborné veřejnosti.

Druhým ústředním tématem bylo zkoumání prostorových a časoprostorových vzorů v rozložení kampylobakteriózy v České republice během let 2008—2012. Nejdříve byly v rámci druhého dílčího cíle hodnoceny předpoklady stanovené v prvním dílčím cíli na základě vizuálního hodnocení. Bylo potvrzeno, že nejvíce onemocněním ohrožená je oblast Ostravska a dále severovýchodní Moravy a na Benešovsku, kde byly identifikovány shluky vysokých měr morbidity v prostoru i v čase. Z dalších míst tomu tak bylo po omezené časové období na Plzeňsku a Českobudějovicku. Hodnocením prostorového vzoru konkrétních případů, se

zabývá ve spojitosti s identifikací možných bodových zdrojů nákazy čtvrtý dílčí cíl. Zde byl pomocí metody geografického profilování hodnocen vztah mezi umístěním automatů na prodej čerstvého mléka a výskytem případů kampylobakteriózy v jejich okolí. Toto téma bylo velmi aktuální během roku 2010, kdy se rozhořel spor mezi hlavním hygienikem ČR a zástupci Agrární komory ČR, který vznikl na základě upozornění na množství případů v kampylobakteriózy v Českých Budějovicích. Na základě časoprostorového skenování byl opravdu identifikován shluk zvýšeného relativního rizika onemocnění na Českobudějovicku v lednu a únoru 2010. Nicméně jako potenciální zdroj nákazy byl ohodnocen mlékomat v nedalekém Lišově a Třeboni a nikoliv přímo v Českých Budějovicích identifikován žádný z přítomných automatů jako potenciální zdroj lokální epidemie. Kromě toho bylo označeno za potenciální zdroje nákazy téměř 20 % mlékomatů. V úvahu však musí být bráno nadhodnocování významu některých případů a množství tedy může být redukováno na zhruba 10 %.

Třetím hlavním tématem byla analýza možných vztahů mezi výskytem onemocnění a vnějšími faktory prostředí. Vzhledem k tomu, že byly identifikovány a popsány prostorové shluky onemocnění, tak existoval předpoklad existence možných podmiňujících faktorů prostředí. Tento předpoklad se ovšem potvrdil pouze částečně, kdy byly jako hlavní faktory ohodnoceny zejména demografie obyvatelstva s možnými lokálními vlivy teploty, socioekonomické deprivace a zemědělství. Predikční schopnost sestavených modelů sice nepřekračuje 50 %, ale je třeba konstatovat, že zdroj a způsob nakažení pouze zhruba poloviny případů u nás i ve světě je uspokojivě identifikováno. Na použité postupy a modely může být nahlíženo z pohledu zkoumání lokálních inferencí, kde prokázaly svou užitečnost. V rámci tohoto tématu byl sestaven koncept lokální ordinální logistické regrese jako modifikace ordinální logistické regrese pro využití s prostorovými daty.

Pokud by měly být zmíněny hlavní přínosy práce, pak by to mohlo být využití časoprostorového krigingu v mapování onemocnění, časoprostorového skenování a geografického profilování při zkoumání prostorových vzorů a využití geograficky vážených metod pro analýzu asociací mezi výskytem onemocnění a lokálními charakteristikami. Kromě samotných metod je to pak měřítko případové studie – kdy šlo o obce či jejich části na území celé České republiky.

POUŽITÁ LITERATURA A INFORMAČNÍ ZDROJE

ANSELIN, L. (1995): Local indicators of spatial association—LISA. Geographical analysis, 27, č. 2.

BEALE, L. et al. (2008): Methodologic issues and approaches to spatial epidemiology. Environmental health perspectives, 116, č. 8, s. 1105–10.

BIVAND, R. S. et al. (2008): Applied Spatial Data Analysis with R. Springer New York, New York, NY.

DAVENHALL, B. (2012): Geomedicine: Geography and Personal Health. Esri, Redlands, 33 s.

ELLIOTT, P. et al. (2000): Spatial Epidemiology: Methods and Applications. Oxford University Press, 504 s.

ELLIOTT, P., BEST, N. (1998): Geographic patterns of disease. Encyclopedia of biostatistics.

GOLDMAN, D., BRENDER, J. (2000): Are standardized mortality ratios valid for public health data analysis? Statistics in medicine, 19, č. 8, s. 1081–8.

GOOGLE (2009): Keyhole Markup Language [online]. Dostupné z: https://developers.google.com/kml/documentation/kml_tut

HEBÁK, P. et al. (2005): Vícerozměrné statistické metody 3. Informatorium, Praha, 256 s.

HENGL, T. (2007): A practical guide to geostatistical mapping of environmental variables. 143 s.

HORÁK, J. (2011): Prostorové analýzy dat. VŠB-TU Ostrava, HGF, Institut geoinformatiky, Ostrava, 170 s.

JARUP, L. (2004): Health and Environment Information Systems for Exposure and Disease Mapping, and Risk Assessment. Environmental Health Perspectives, 112, č. 9, s. 995–997.

KOHONEN, T. (1982): Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, č. 1, s. 59–69.

KOHONEN, T. (2001): Self-Organizing Maps. Springer Science & Business Media, 502 s.

KULLDORFF, M. et al. (2005): A space-time permutation scan statistic for disease outbreak detection. PLoS medicine, 2, č. 3, s. e59.

LE COMBER, S. C. et al. (2011): Geographic profiling as a novel spatial tool for targeting infectious disease control. International journal of health geographics, 10, č. 1, s. 35.

LI, H. et al. (2012): One-step estimation of spatial dependence parameters: Properties and extensions of the APLE statistic. Journal of Multivariate Analysis, 105, č. 1, s. 68–84.

MAREK, L. et al. (2012): Spatial Analyses of Epidemiological Data: Case Study In Olomouc Region. In: 12th International Multidisciplinary Scientific GeoConference SGEM: SGEM 2012, Proceedings Volume II. STEF92 Technology Ltd, Sofia, Bulgaria, s. 1155 – 1162.

OPEN GEOSPATIAL CONSORTIUM (2008): OGC KML 2.2.0. Open Geospatial Consortium, 252 s.

OSTFELD, R. S. et al. (2005): Spatial epidemiology: an emerging (or re-emerging) discipline. Trends in ecology & evolution, 20, č. 6, s. 328–36.

PEBESMA, E., GRÄLER, B. (2014): Spatio-temporal geostatistics using gstat. Münster, DE, 1-11 s.

PFEIFFER, D. et al. (2008): Spatial analysis in epidemiology. Oxford University Press.

REZAEIAN, M. et al. (2007): Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. Journal of epidemiology and community health, 61, č. 2, s. 98–102.

RICHARDSON, S. et al. (2004): Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies. Environmental Health Perspectives, 112, č. 9, s. 1016–1025.

SWEENEY, L. (2000): Simple demographics often identify people uniquely. Pittsburgh, 1-34 s.

SWEENEY, L. et al. (2013): Identifying Participants in the Personal Genome Project by Name. SSRN Electronic Journal, s. 1–4.

ÚSTAV ZDRAVOTNICKÝCH INFORMACÍ A STATISTIKY ČESKÉ REPUBLIKY (2013): Infekční nemoci 2013. 60 s.

VERITY, R. et al. (2014): Spatial targeting of infectious disease control: identifying multiple, unknown sources. Methods in Ecology and Evolution, 5, č. 7, s. 647–655.

WALLER, L. A., GOTWAY, C. A. (2004): Applied Spatial Statistics for Public Health Data. John Wiley & Sons.

Odborný životopis autora / *Curriculum vitae*

OSOBNÍ ÚDAJE / *PERSONAL INFORMATION*

Jméno / *Name*	Lukáš MAREK
Bydliště / *Adress*	U Vodojemu 1229
	757 01 Valašské Meziříčí
E-mail	lukys.marek@gmail.com
Narozen / *Birth*	20. 1. 1985, Valašské Meziříčí

VZDĚLÁNÍ / *EDUCATION*

2009–dosud	Univerzita Palackého v Olomouci, **doktorské studium,** *obor Geoinformatika a kartografie* / *PhD study Geoinformatics and Cartography*
2007–2009	Univerzita Palackého v Olomouci, navazující **Mgr. studium**, obor *Geoinformatika* / *Master's degree: specialization Geoinformatics*
2004–2007	Univerzita Palackého v Olomouci, **Bc. studium**, obor *Geografie-Geoinformatika* / *Bachelor's degree: Geography–Geoinformatics*

PRAXE / *EXPERIENCE*

2012–dosud	Univerzita Palackého v Olomouci, **projekt OP VK StatGIS team** (vědecký pracovník / researcher) *Budování výzkumně-vzdělávacího týmu v oblasti modelování přírodních jevů a využití geoinformačních systémů, s vazbou na zapojení do mezinárodních sítí a programů*
2013	Cleopa GmbH, Hennigsdorf (DE) (datový analytik / data analyst)
2011–dosud	GISportal.cz (editor a spoluzakladatel / editor and co-founder)
2009	Oddělení územního rozvoje a městského plánování MěÚ Valašské Meziříčí (GIS technik / GIS technician)
2007	Ageris s.r.o, Brno (GIS analytik / GIS analyst)

VÝUKOVÉ AKTIVITY / *TEACHING*

2009–dosud	cvičení a přednášky na Katedře geoinformatiky UP: Geostatistika, Statistika, Metody geocomputation v GIS, Základy informatiky, Matematika *seminars and lectures of: Statistics, Geostatistics, Geocomputational methods in GIS, Informatics, Mathematics*

2015	University of Canterbury, Nový Zéland
2014	University of Belgrade, Srbsko
2014	University of La Laguna, Španělsko
2012	Norwegian School of Economics, Norsko
2012	Jagiellonian University Krakow, Polsko
2011	University of Belgrade, Srbsko
2010	Norwegian University of Science and Technology, Norsko

Seznam vybraných publikací autora souvisejících s disertační prací / Author's selected publications related to the dissertation

Publikace v odborném časopise v databázi Web of Science (Jimp)
Publications in a scientific journal indexed on the Web of Science (Jimp)

- **Marek, L.,** Tuček, P., Pászto, V. (2015): Using geovisual analytics in Google Earth to understand disease distribution: a case study of campylobacteriosis in the Czech Republic (2008-2012), International Journal of Health Geographics, 14:7.

Kapitola v odborné knize (C) / Chapter in a scientific book (C)

- **Marek, L.,** Pászto, V., Tuček, P. (2015): Bayesian mapping of medical data, Modern Trends in Cartography, Springer International Publishing, s. 489-505.

Publikace v konferenčním sborníku v databázi ISI Proceedings (D)
Papers in conference proceedings on the database ISI Proceedings (D)

- **Marek, L.,** Pászto, V., Tuček, P., Sádovská, P. (2013): Space-time evaluation of health data: Case of Olomouc area, Czech Republic. SGEM Conference: 13th International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 911-918.
  study in Olomouc region, SGEM Conference: 12h International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 1155-1162.

Publikace v konferenčním sborníku v databázi Scopus (D)
Papers in conference proceedings on the database Scopus (D)

- **Marek, L.**, Pászto, V., Tuček, P., Dvorský, J. (2014): Spatial Clustering of Disease Events Using Bayesian Methods, CEUR Workshop Proceedings – DATESO 2014, s. 25-34.
- **Marek, L.**, Pászto, V., Dvorský, J., Tuček, P. (2013): On Estimation of the Spatial Clustering: Case Study of Epidemiological Data In Olomouc Region, Czech Republic, CEUR Workshop Proceedings – DATESO 2013, s. 1-12.

Publikace v konferenčním sborníku domácí (ostatní)
Papers in conference proceedings – Czech (other)

- **Marek, L.,** Pászto, V., Tuček, P. (2014): Bayesian mapping of medical data. CARTOCON2014, Olomouc.
- **Marek, L.** (2014): Spatial Analyses of Health Data: From Points to Models. Third InDog Doctoral Conference, Olomouc, 2014.
- **Marek, L.** (2014): Prostorová geo-demografická klasifikaca úmrtnosti v České republice v letech 1994—2012. Second StatGIS Conference Proceedings, Olomouc, 2014.
- **Marek, L.** (2014): The exploration of spatio-temporal patterns: An example of campylobacteriosis in the Czech Republic. Second StatGIS Conference Proceedings, Olomouc, 2014.
- **Marek, L.** (2013): Health Datasets in Spatial Analyses: What We Want, What We Get and What We Can Use. First StatGIS Conference Proceedings, Olomouc, 2014.
- **Marek, L.** (2013): Prostorová analýza výskytu kampylobakteriózy v České Republice. First StatGIS Conference Proceedings, Olomouc, 2014.
- **Marek, L.**, Pászto, V. (2013): Prostorové statistiky zdravotnických dat: případová studie v Olomouckém kraji, Juniorstav 2013, VÚT Brno, 474 s.
- **Marek, L.** (2013): Spatial Clustering and Multivariate Statistics in Analysis of Infectious Diseases. Second InDog Doctoral Conference, Olomouc, 2013.
- **Marek, L.** (2012): Spatial statistics of health data: Case study in Olomouc Region, Czech Republic. First InDog Doctoral Conference, Olomouc, 2012.

Publikace v konferenčním sborníku zahraniční (ostatní)
Papers in conference proceedings – foreign (other)

- **Marek, L.** et al. (2013): Health Datasets in Spatial Analyses: What We Want, What We Get and What We Can Use, Proceedings of GENG'13 – Antalya.
- **Marek, L.** et al. (2013): Geography of Campylobacter in the Czech Republic, GEOMED 2013, Shefield.

Ostatní publikace autora / Another author's publications

Zahraniční publikace / Foreign publications

- Koukol, M., Zajíčková, L., **Marek, L.,** Tuček, P. (2015): Fuzzy Logic in Traffic Engineering – A Review on Signal Control, Mathematical Problems in Engineering, in press. (IF 1.082, WoS, Scopus)
- Pászto, V., Brychtová, A., Tuček, P., **Marek, L.**, Burian, J. (2015): Using a fuzzy inference system to delimit rural and urban municipalities in the Czech republic in 2010. Journal of Maps, 11(2), s. 231-239. (IF 0,895, WoS, Scopus)
- **Marek, L.,** Miřijovský, J., Tuček, P., Pászto, V. (2014): Surveying and analysis of the landslides using UAS remote sensing, SGEM Conference: 14th International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 825-832.

- Pászto, V., Dařena, F., **Marek, L.**, Fusková, D. (2014): Spatial analyses of Twitter data – case studies, SGEM Conference: 14th International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 785-792.
- Caha, J, **Marek, L.,** Pászto, V. (2014): Spatial Prediction Using Uncertain Variogram, DailyMeteo.org 2014, Serbia, Beograd.
- **Marek, L**., Pászto, V., Tuček, P., Sádovská, P. (2013): Space-time evaluation of health data: Case of Olomouc area, Czech Republic, SGEM Conference: 13th International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 911-918. (WoS, Scopus)
- Pászto, V., **Marek, L**., Hartmannová, S. (2013): Analyzing major cities of European Union member states using shape metrics. SGEM Conference: 13th International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 313-320. (WoS, Scopus)
- Tuček, P., Tučková, M., **Marek, L.**, Pászto, V. (2013): Comparison of regression models with constraints in geodetical measurement, SGEM Conference: 13th International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 343-354. (WoS, Scopus)
- Burian, J., **Marek, L.,** Pászto, V., Tučková, M., Tuček, P., Zajíčková, L., Chrudimská, J. (2013): Quantification of the threat of the transport infrastructure by selected natural hazards: case study in Zlinsky region, SGEM Conference: 13th International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 535-542. (WoS, Scopus)
- Brychtová, A., Pászto, V., **Marek, L.**, Pánek, J. (2013): Web-design evaluation of The Crisis Map of The Czech Republic using eye-tracking, SGEM Conference: 13th International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 1065-1072. (WoS, Scopus)
- Tuček, P., Tučková, M., **Marek, L.**, Burian, J. (2013): Statistical processing of laminar water flow data – nonlinear regression based approach, SGEM Conference: 13th International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 343-354. (WoS, Scopus)
- Pászto, V., **Marek, L.**, Tuček, P. (2013): Are Shape Metrics Useful for a Geocomputation? CORINE Land-Cover Analysis Case Study, CEUR Workshop Proceedings – DATESO 2013, s. 26-35. (Scopus)
- Pászto, V., **Marek, L.,** Tuček, P. (2013): Rural or Urban? Delimitation of The Czech Republic Municipalities Using Łukasiewicz T-norm, Recent Advances in Geodesy and Geomatics Engineering (GENG 2013), Antalya, Turkey, s. 38-44.
- Pászto, V., **Marek, L.**, Sedoník, J. (2012): Fuzzy approach of rural and urban areas delimitiation in GIS, SGEM Conference: 12h International Multidisciplinary Scientific Geoconference, Bulgaria, Albena, s. 1049-1056. (WoS, Scopus)
- Pászto, V., **Marek, L.**, Tuček, P. (2011): Fractal Dimension Calculation for CORINE LandCover Evaluation in GIS – A Case Study, CEUR Workshop Proceedings – DATESO 2011, s. 196-2015. (Scopus)
- Tuček, P., **Marek, L.,** Pászto, V., Tučková, M., Růžička, O., Voženílek, V., Schmidt, N. (2011): GPS Tourist Multimedia Guide – Urban Sightseeing Navigation System, LBS 2011 Wien, Wien.
- Pászto, V., Marek, L., Tuček, P. (2011): Perspectives of fractal geometry in GIS analyses. In Sborník – Symposium GIS Ostrava 2011, Ostrava

- Tuček, P., **Marek, L.,** Pászto, V., Janoška, Z. (2011): Fractal perspectives of GIScience based on the leaf shape analysis. Geocomputation 2011, UK, London
- **Marek, L.**, Tuček, P., Pászto, V., Marek, J. (2010): Stochastic approach for determining landslide activity, Advances in GIT, Ostrava.
- Pászto, V., Tuček, P., **Marek, L**., Kuprová, L., Burian, J. (2010): Statistical inferences - visualization possibilities and fuzzy approach computing, Advances in GIT, Ostrava.

Domácí publikace / Czech publications

- **Marek, L.,** Pánek, J., Pászto, V. (2014): Krizová mapa České republiky. In Sborník abstraktů z konference Geoinformace ve veřejné správě 2014, Praha,
- Pászto, V., Brychtová, A., Marek, L. (2014): On shape metrics in cartographic generalization - a case study of the building footprint geometry. In: CARTOCON 2014: Conference proceedings, Olomouc.
- Pánek, J., **Marek, L.,** Pászto, V. (2014): Crisis Mapping via Internet in Disaster Management of the Czech Republic. In: CARTOCON 2014: Conference proceedings, Olomouc
- Vávra, A., **Marek, L.** (2013): Utilization of Phenological Observation in the Landscape Mapping with using GIS, Proceedings of GENG'13 – Antalya
- Pászto, V., Marek, L., Tuček, P. (2013): Rural or Urban? Delimitation of The Czech Republic Municipalities Using Łukasiewicz T-norm, Proceedings of GENG'13 – Antalya
- Burian, J., Pászto, V., Tuček, P., a kol. (2013): Geoinformatika při analýzách rurálního a urbánního prostoru. Univerzita Palackého v Olomouci, Vydavatelství UP, 130 s.
- Vránová, V. et al. (2013): Struktura osídlení v období popelnicových polí na střední Moravě (Settlements structures in urnfield culture period in the central Moravia), Archaeologiae Regionalis Fontes 12, Olomouc.
- Pászto, V., a kol. (2012): Obce České republiky: příslušnost obcí České republiky k městskému a venkovskému prostoru k 31. 12. 2010. Edice M.A.P.S. Univerzita Palackého v Olomouci, 29 s.
- Pászto, V., **Marek, L.** (2012): Matematika na pozadí geografických informačních systémů. Kvaternion, 1, VUT Brno, s. 27-36.
- Marjanovič. M. et al. (2012): Využití klasifikačních algoritmů metod strojového učení pro účely prostorového modelování. (Application of Machine Learning classification algorithms in spatial modeling framework) In Ed. VOŽENÍLEK, V., DVORSKÝ, J., HÚSEK, D. Metody umělé inteligence v geoinformatice. Univerzita Palackého v Olomouci. Olomouc, 2011.
- Svobodová, J., **Marek, L.**, Tuček, P. (2012): Analysis of the relationships among error values and values of morphometric parameters derived from the DEM. Sborník symposia GIS Ostrava 2012, Ostrava, 2012, ISBN 978-80-248-2558-8.
- Burian, J. a kol. (2011): Sborník příspěvků letní školy GEOCOMPUTATION IN GIS. Univerzita Palackého v Olomouci, 96 s. ISBN: 978-80-244-2982-3
- **Marek, L.,** Pászto, V., Janoška, Z., Tuček, P. (2011): Příklad využití fraktální geometrie v geoinformatice: analýza typologie říční sítě založená na morfologii listů, Geografie a geoinformatika – výzva pro praxi a vzdělávání, Brno.

- Brus, J., Fekiač, M., Marek, L. (2011): Aspekty vyjadřování nejistoty v kartografii. Bratislava, 19. Cartographic Conference
- **Marek, L.,** Tuček, P., Pászto, V., Marek, J. (2010): Hypotheses testing in regression models and their using in landslide detection, In Sborník – Symposium GIS Ostrava 2010. VŠB – Technická univerzita Ostrava.
- Voženílek, V., Tuček, P., Pászto, V., **Marek, L**., Schmidt, N. (2010): Zkušenosti z pilotního nasazení městské multimediální navigace, Symposium GIS Ostrava 2010, Ostrava.
- Voženílek, V., Brus, J, Pászto, **Marek, L.** (2010): Perspectives of visualisation based on a geocomputation of climatological datasets. IGU 2010 Regional Meeting. Tel Aviv.

Annotation

The investigation of the health and health-related factors that may affect it is one of the popular topics not only in the geosciences but also in a wide range of other disciplines. From this perspective, the dissertation thesis contributes to spreading the awareness of possibilities of the spatial analyses of health data. The main objective of the dissertation focuses on a complex spatial analysis of the campylobacteriosis distribution in the Czech Republic in 2008—2012 utilizing the variety of methods from geosciences and spatial statistics. The main objective was split to five consecutive partial objectives that together covered main topics addressed in the field of spatial epidemiology.

The first partial objective deals with both, the statistical description of the disease and its characteristics in the Czech population, but it mainly deals with mapping of the spatial distribution of the campylobacteriosis in the Czech Republic. The continuous surface of the weekly raw incidence rates was created using the spatiotemporal kriging in addition to maps of morbidity.

The second partial objective aims to explore, identify and evaluate spatial and spatiotemporal pattern of Czech municipalities regarding the vulnerability to the disease described by increased relative risk comparing to their neighbourhood. There were one primary and thirteen secondary clusters of high rates in space and time identified in the Czech Republic.

The third partial objective focuses on the identification of possible associations of the disease distribution and factors that influence the spread of the disease or its inhibition using the (spatial) regression and classification methods. Besides the modelling, the classification based on the characteristics of morbidity and environment was created by multivariate clustering.

Fourth partial objective examines the relationship between the location of fresh milk vending machines and a local increase in the disease occurrence. This was carried out by geographical profiling, the method that incorporates Bayesian processes and Monte Carlo simulations.

Fifth partial objective uses results and outputs of preceding partial objectives in order to transform them into a form that is suitable for their clear communication as well as further interactive exploration in space and time.

Besides the primary objective, there was also the aim to create the coherent work that may provide the overview of usable spatial epidemiology methods and provide the support and inspiration for other similar studies.

Keywords: Spatial epidemiology, disease mapping, geovisualisation, geographic correlation, spatial pattern, geoprofiling

SUMMARY

The main objective of the dissertation thesis was to carry out the complex spatial analysis of the epidemiological data with the usage of recent technologies from fields of geoinformatics and spatial statistics. The practical part of the dissertation examined an infectious disease called campylobacteriosis during 2008—2012 in the Czech Republic. Besides the primary objective, there was also the aim to create the coherent work that may provide the overview of usable spatial epidemiology methods and provide the support and inspiration for other similar studies. The main theme of the thesis and its interdisciplinary nature required to combine the knowledge acquired during the GIS studies with the findings of other disciplines such as spatial and applied statistics or spatial epidemiology. During the processing of the dissertation, its main objective was split to five consecutive partial objectives that together covered main topics addressed in the field of spatial epidemiology. These are denoted as: (1) a disease mapping, (2) an identification of spatial patterns and (3) geographical correlation studies. However, they are usually closely interconnected. Partial objectives of the dissertation were following these topics and were stated as:

- The first partial objective was to map and statistically describe the distribution of the campylobacteriosis in the Czech Republic during 2008—2012.
- The second partial objective aimed to explore, quantify and geovisualise the spatial and spatiotemporal patterns in the distribution of the campylobacteriosis in the Czech Republic during 2008—2012 and its characteristics.
- The third partial objective was to identify and analyse possible associations of the disease distribution and environmental, demographic and socioeconomic factors. These were utilized using the multivariate statistics and (spatial) statistical modelling with the subsequent evaluation of their prediction and classification performance. Besides the modelling, the classification based on the characteristics of morbidity and environment was created.
- The fourth partial objective was to evaluate the fresh milk vending machines as potential sources of local campylobacteriosis outbreaks during 2008—2012.
- The fifth partial objective was to transform selected results from previous partial objectives into form that is suitable for the further geovisual analysis in both, space and time.

The first and fifth partial objectives were examining mapping and geovisualization issues. The first partial objective mapped the spatial distribution of the campylobacteriosis – its frequency, incidence and relative risk. In addition, the smoothing methods were applied, and recommendations were provided. The spatiotemporal mapping was realised by the employment of the spatiotemporal kriging that supplied methods of dasymetric mapping. Using the spatiotemporal kriging, the continuous surface of weekly raw incidence was created so the examination of the disease could have been monitored in space and time simultaneously. The engagement of the spatiotemporal kriging in the exploration of the morbidity is one of the dissertation's highlights. Maps and geovisualisations were also transformed into KML files in order to provide the transferability of the data as well as to communicate the results in Google Earth. Google Earth provided the platform for the geovisual interactive spatiotemporal analytics and, due to its intuitive interface and wide distribution, a convenient way for the spreading the results to lay and professional audience.

The second important topic in the spatial epidemiology and the second partial objective of the dissertation research are analyses of spatial and spatiotemporal patterns of the disease distribution. Initially, patterns were evaluated visually based on the outputs of the first partial objective. Then, it was confirmed that the most vulnerable areas are Ostrava and it surrounding and also the north-east Morava, Benešov and several others, where clusters of high rates of morbidity were identified in space and in time. There was also cluster around Plzeň and České Budějovice but they appeared for the limited time. The method of geoprofiling was utilized in order to evaluate the spatial pattern of particular cases and possibly to identify the likely association between small local disease outbreaks and the location of fresh milk vending machines. This was proceed in order to provide the retrospective analysis of the interesting situation that was actual in 2010, when the dispute over the health risks of the fresh milk made the controversy regarding the situation in České Budějovice and the public announcement by the head hygienist of the Czech Republic. The cluster of high relative risk was identified in České Budějovice lasting during the January and February 2010, however the geoprofiling did not identify any potential source of the outbreak located near the fresh milk vending machines, although some potential outbreak sources were found in the neighbourhood. During the analyses of all vending machines in the Czech Republic up to 20% of fresh milk vending machines were identified as potential sources. Considering the overestimation of the significance of some cases reduced the estimation of sources to approximately 10%.

The third partial objective covered the third main theme of the spatial epidemiology, which is the geographical correlation analysis of the association between the disease distribution and environmental factors. The existence of the association was presumed based on the identification of clusters of both, high and low rates in the preceding partial objectives. The assumption itself could have been confirmed only partly. The demography was identified as the most associated factor, but also the local influence of the temperature, socioeconomic deprivation and agriculture. The prediction and classification performances of both, spatial and traditional methods were never higher than 50% of the real situation, which was not very much but it corresponds with the fact, that only half of the campylobacteriosis cases is well described in the Czech Republic and also worldwide. Although, models used in this part of the dissertation did not perform well, they were still very useful for the exploration of local inferences. Within this partial objective, the concept of the local ordinal logistic regression was introduced as the modification of ordinal regression for spatial data.

Among the others, main benefits or highlights presented in the dissertation can be summarized as the use of the spatiotemporal kriging within the disease mapping, the application of the space-time scan statistics and geoprofiling in the investigation of spatial and spatiotemporal patterns, and the utilization of geographically weighted methods for the analysing of the association between the disease occurrence and local environmental characteristics. The local scale of the case study should be also mentioned in addition to all methods.

The investigation of the health and health-related topics is one of the very favourite topics not only in the geosciences but also in a wide range of other disciplines. However, the disparity between medical studies based on individuals and studies in other scales is usually visible. On one hand, this can be caused by the protection of personal data (also connected to the small amount of publicly available data), but it can be also caused by possible distrust in the capabilities and benefits of spatial and spatiotemporal methods. Nevertheless, it should be noted that the situation is improving slowly but constantly because of activities of the research teams in universities as well as in the involved state institutions.

Mgr. Lukáš Marek

PROSTOROVÉ A VÍCEROZMĚRNÉ STATISTICKÉ ANALÝZY EPIDEMIOLOGICKÝCH DAT

SPATIAL AND MULTIVARIATE STATISTICAL ANALYSES OF EPIDEMIOLOGICAL DATA

Určeno pro studenty, partnerská akademická pracoviště a veřejnost.

Výkonný redaktor prof. RNDr. Zdeněk Dvořák, DrSc. et Ph.D.
Odpovědná redaktorka Mgr. Věra Krischková
Technická redakce Mgr. Lukáš Marek

Publikace neprošla redakční jazykovou úpravou.

Vydala a vytiskla Univerzita Palackého v Olomouci
Křížkovského 8, 771 47 Olomouc
www.vydavatelstvi.upol.cz
www.e-shop.upol.cz
vup@upol.cz

1. vydání
Olomouc 2015
Edice GEOINFO-CARTO-THESIS, svazek VII.
ISSN 1805-7500
ISBN 978-80-244-4547-2

Neprodejná publikace