MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
GEOGRAFICKÝ ÚSTAV

Analýza a kartografická vizualizace statistik o vícečetných zhoubných nádorech

Diplomová práce

Bc. Jana Havlíčková

Mgr. Radim Štampach, PhD. **Brno 2016**

Bibliografický záznam

Autor:	Bc. Jana Havlíčková Přírodovědecká fakulta, Masarykova univerzita Geografický ústav
Název práce:	Analýza a kartografická vizualizace statistik o vícečetných zhoubných nádorech
Studijní program:	Geografie a kartografie
Studijní obor:	Geografická kartografie a geoinformatika
Vedoucí práce:	Mgr. Radim Štampach, PhD.
Akademický rok:	2015/2016
Počet stran:	88+8
Klíčová slova:	Zdravotní kartografie; nádorová epidemiologie; onkologické klasifikace; klinická stádia; neznámé klinické stádium; prostorové shlukování; analýza hlavních komponent; vícečetné zhoubné novotvary

Bibliographic Entry

Author	Bc. Jana Havlíčková Faculty of Science, Masaryk University Department of Geography
Title of Thesis:	Analysis and cartographic visualization of statistics about multiple malignant neoplasms
Degree programme:	Geography and cartography
Field of Study:	Geographical cartography and geoinformatics
Supervisor:	Mgr. Radim Štampach, PhD.
Academic Year:	2015/2015
Number of Pages:	88+8
Keywords:	Health cartography; cancer epidemiology; classification in onkology; cancer staging; unknown cancer stage; spatial clustering; principal component analysis; multiple malignant neoplasms

Abstrakt

Tato diplomová práce se zabývá analýzou existujících studií vícečetných zhoubných novotvarů a použitých nástrojů statistické analýzy. Dále podrobně vysvětluje význam vícečetných zhoubných novotvarů a hodnotí jejich vývoj a výskyt v České republice v letech 1976-2010. Práce dále zahrnuje výčet metod statistické analýzy prostorových dat využitelných pro analýzu onkologických dat. Stěžejní částí práce je aplikace zvolených analytických metod na konkrétní onkologická data a prezentace zjištěných výsledků pomocí kartografických výstupů.

Abstract

This thesis deals with analysis of existing studies of multiple malignant neoplasms and tools of used statistical analysis tools. It explains meaning of multiple malignant neoplasms in detail and evaluate their development and occurrence in the Czech Republic from 1976 to 2010. The thesis also includes a list of methods of statistical analysis of spatial data, usable for analysis of oncological data. A key part of this thesis is application of chosen analytical methods to the particular oncological data and presentation of discovered results with the aid of cartographical outputs.

Masarykova univerzita

Přírodovědecká fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student:	**Jana Havlíčková**
Studijní program:	**Geografie a kartografie**
Studijní obor:	**Geografická kartografie a geoinformatika**

Ředitel Geografického ústavu PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s tématem:

Analýza a kartografická vizualizace statistik o vícečetných zhoubných nádorech

Analysis and cartographic visualization of statistics about multiple malignant neoplasms

Zásady pro vypracování:

1. Přehled možného využití, popis využívaných metod a výsledků statistické analýzy zdravotních dat o vícečetných zhoubných nádorech - v domácí i zahraniční literatuře.
2. Exploratorní analýza onkologických dat zaměřená na nalezení souvislostí mezi různými charakteristikami pacientů s diagnostikovaným následným nádorem.
3. Návrh a realizace vhodných kartografických vizualizací pro publikaci o vícečetných onkologických diagnózách.

Rozsah grafických prací: podle potřeby

Rozsah průvodní zprávy: cca 60 až 80 stran

Seznam odborné literatury:

KONEČNÝ, M., GERYK, E., KUBÍČEK, P., ŠTAMPACH, R., KOZEL, J., STACHOŇ, Z., MICHÁLEK, J., ODEHNAL, J., DÍTĚ, P., KOŠKA, P., KRAUS, R., HOLUB, J. Prevalence nádorů v České republice 1989 – 2005 – 2015. 1. vydání, Přírodovědecká fakulta Masarykovy univerzity a Občanské sdružení podpory zdraví a onkologické prevence, Brno, 2008, ISBN 978-80-903255-2-4.
GERYK, E., KOLCOVÁ, V., MARŠÍK, V., ŠIKULA, P., ŠIROKÝ, P., ZACHOVAL, J. Atlas zhoubných nádorů v České republice. Kartuziánské nakladatelství, Brno, 1995. 85 s. ISBN 80-901943-0-3. 1995.
MARŠÍK, V. Atlas výskytu zhoubných nádorů v České republice. 1. vyd., Masarykův onkologický ústav, Brno, 1998. 47 s. ISBN 80-238-1718-3.
PICKLE, L. W. Usability Testing of Map Designs. In Proceedings of the Symposium on the Interface. Salt Lake City, Utah, March 12-15, 2003. Dostupné na internetu: http://www.galaxy.gmu.edu/interface/I03/I2003Proceedings/PickleLinda/PickleLinda.paper.pdf\>ISBN 1-886658-09-9. 2003.
Závěrečné bakalářské, magisterské a disertační práce s podobnou tematikou řešené na Geografickém ústavu.

Jazyk závěrečné práce: čeština

Vedoucí diplomové práce: Mgr. Radim Štampach, Ph.D.

Podpis vedoucího práce: .....

Datum zadání diplomové práce: listopad 2014
Datum odevzdání diplomové práce: do 5. května 2016

V. Herber

RNDr. Vladimír Herber, CSc.
pedagogický zástupce ředitele ústavu

Se zadáním diplomové práce souhlasím, jsem si vědom(a), že zadání práce je závazné.

Zadání práce převzal(a): Havlíčková *dne* 17. 2. 2015

Poděkování

Na tomto místě bych chtěla mnohokrát poděkovat vedoucímu této práce Mgr. Radimu Štampachovi Ph.D. za cenné rady, postřehy, ochotu, vstřícný přístup a četné odborné konzultace při vypracování této diplomové práce.

V neposlední řadě bych chtěla ze srdce poděkovat své rodině a příteli a přátelům za neutuchající podporu, motivaci a toleranci v průběhu celého studia.

Prohlášení

Prohlašuji, že jsem svoji diplomovou práci vypracovala samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno 27. dubna 2016 ……

Bc. Jana Havlíčková

OBSAH

1 ÚVOD

Sledování zdravotního stavu lidské populace je nezbytným podkladem pro hodnocení potenciálních zdravotních rizik a následné rozhodování v oblasti řízení zdravotní péče. Monitorování zdravotního stavu se následně stává základem pro praktická opatření, preventivní aktivity a výchovu obyvatel ke zdraví.

Oblast onkologických onemocnění vyžaduje stejně jako jiné oblasti veřejného zdraví soustavné a pečlivé monitorování, jelikož riziko nádorového onemocnění paradoxně narůstá s prodlužující se délkou života. Speciálním případem jsou vícečetné zhoubné novotvary, které jsou předmětem výzkumu této práce, a které patří mezi obávané komplikace léčby Dostatečná informovanost lidí o prevenci vzniku nádorových onemocnění, možnostech včasné diagnózy a možných způsobech léčby, jakož i o možnosti vzniku následných malignit je základním předpokladem pro snížení incidence zhoubných novotvarů.

Četné informace pro hodnocení zdravotního stavu mohou poskytovat zdravotnické statistiky na státní úrovni, jejichž zřizovatelem bývá stát a správnost informací tak bývá garantována. Problémem takových dat však může být omezená možnost identifikace a hodnocení příčin vzniku sledovaných onemocnění.

Poznání hlubších souvislostí v datech je pak možné až na základě epidemiologických studií zdravotního stavu. A právě zde může k řešení problematiky přispívat kartografie, respektive její explorační forma. Nedílnou součástí tohoto přístupu jsou také dostupné geografické informační systémy (GIS), jež umožňují aplikovat pokročilé nástroje explorační kartografie na data o zdravotním stavu, potažmo o onkologických onemocněních. Výsledky kartografické explorace lze pak využít pro následnou vizualizaci, jež může přispět k distribuci důležitých informací o zdravotním stavu obyvatel a být podporou pro rozhodování politických aktérů.

1.1 Cíl práce

Tato práce shrnuje současné možnosti využití metod prostorové analýzy zdravotních dat a zmiňuje významný příspěvek vědní disciplíny nádorové epidemiologie, která napomáhá k pochopení příčin rakoviny a hodnocení zavedených preventivních opatření.

Práce se v rešeršní části popisuje existující práce, které se věnují tématice výskytu vícečetných zhoubných novotvarů. Po deskripci existujících monografií o vícečetných zhoubných novotvarech se práce věnuje čtyřem různým nástrojům, které prezentují různé typy zdravotních dat.

Významnou část práce tvoří kapitola třetí, která shrnuje problematiku vícečetných zhoubných novotvarů, popisující jejich vývoj v čase od roku 1976 do roku 2010 a jejich diferenciaci v rámci krajů České republiky. Důraz je kladen zejména na rozdíl mezi muži a ženami a na zastoupení jednotlivých klinických stádií primárních a následných diagnóz

v krajích České republiky. Kapitola dále poskytuje přehled možných onkologických klasifikací, včetně mezinárodní klasifikace nemocí, která je využívána i v rámci této práce. Následně práce zkoumá vícečetné novotvary z hlediska určeného klinického stádia a blíže se zaměřuje na otázky, týkající se výskytu neznámého klinického stádia u vícečetných diagnóz.

Dílčím úkolem práce byl popis metod využitelných pro prostorovou analýzu dat, které jsou z důvodu jejich citlivé povahy agregovány. Teoretická část, kterou shrnuje kapitola druhá, se tedy zabývá metodami, které mají napomoci identifikaci prostorového vzoru v datovém zdroji, a metodami, napomáhajícími redukci multidimenzionálních dat.

Stěžejní část práce zahrnutá v kapitole šesté pak aplikuje popsané metody na data o vícečetných zhoubných novotvarech, vyskytujících se v okresech České republiky v letech 1976-2010. Přínos této práce spočívá v analýze dosud neanalyzovaných dat o výskytu neznámého klinického stádia. Data jsou analyzována různými metodami v pořadí od nejjednodušších (identifikace odlehlých hodnot, korelace) až po složitější (prostorová autokorelace, PCA analýza) a je tak poskytnut velmi podrobný vhled do problematiky neznámých klinických stádií rakoviny v prostoru. Pro zmíněné analýzy byly použity volně dostupné programy – GeoDa a R. Pro vizualizaci výsledků byl pak využit program ArcMap 10.3.1.

1.2 Základní pojmy a definice

Následující pojmy jsou využívány v celé diplomové práci a pochopení jejich významu je nutné k pochopení všech souvislostí.

Vícečetné zhoubné novotvary – skupina novotvarů (nádorů) postihující téhož nemocného během jeho života mající odlišnou histologickou povahu a topografické uložení (GERYK a kol., 2008).

Primární novotvar – je nádor vzniklý prvotně v určitém orgánu (Velký lékařský slovník, 2008).

Následný novotvar - histologicky odlišný novotvar, rozvíjející se nejméně dva měsíce po ukončení léčby primárního novotvaru. Novotvar, který je diagnostikován v pořadí druhý, se označuje jako sekundární, třetí v pořadí jako terciární atd. (Subsequent neoplasms [on-line], (2016)).

Incidence – je ukazatelem, který kvantifikuje výskyt nové vzniklých onemocnění (s danou diagnózou) v dané populaci ve zvoleném časovém intervalu. Rozlišujeme např. roční incidenci, specifickou incidenci či hrubou incidenci (BENCKO et al., 2003).

Hrubá incidence – je definována jako podíl počtu nové zjištěných případů onemocnění v dané populaci v daném období a počtu osob v dané populaci v daném období (často přepočteno na 100 tisíc obyvatel), (ŠIROKÝ, 1999).

Morbidita – je pojem označující číselný údaj vztažený pro danou nemoc k určitému časovému úseku a počtu obyvatel nemocných za rok na 100 tisíc obyvatel (Velký lékařský slovník, 2008).

Mortalita – je termín označující úmrtnost na určitou nemoc nebo celková úmrtnost (Velký lékařský slovník, 2008).

Karcinom in situ (neinvazivní karcinom) – karcinom lokalizovaný v místě svého vzniku, např. ve sliznici daného orgánu, bez přesahu do dalších vrstev. Při detekci tohoto stavu může včasná léčba přinést plnou úzdravu (Velký lékařský slovník, 2008).

Prostorová epidemiologie (spatial epidemiology) – je popis a analýza prostorového uspořádání zdravotních dat s ohledem na demografické, environmentální, behaviorální, socioekonomické, genetické a infekční faktory. (ELLIOTT a WARTENBERG, 2004).

Nádorová epidemiologie (cancer epidemiology) – je obor epidemiologie zabývající se šířením nákazy rakoviny v populaci. Jejím konečným cílem identifikovat rizikové faktory, které mohou vést k včasnému zavedení preventivních opatření (DOS SANTOS SILVA, 1999).

2 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY

Potřeba zjišťovat informace o výskytu nemocí a prezentovat výsledky pomocí mapových výstupů má dlouhou historii. Její počátky sahají až do starověkého Řecka. Již na přelomu 5. a 4. století před naším letopočtem si Hippokrates všimnul, že existuje jistá souvislost mezi výskytem nemocí a environmentálními podmínkami. Tuto myšlenku o mnoho století později, v roce 1854, doložil skutečným výzkumem John Snow se svou nejznámější mapou výskytu cholery v Londýně. Tímto položil základní kámen pro moderní epidemiologii (MEADE a EMCH, 2010).

Právě tato skutečnost ukázala obrovský potenciál, který spočívá ve využití geografických analýz pro popis distribuce nemocí v prostoru (případně i v čase). Bez následné kartografické prezentace výsledků by však proces nebyl úplný. Je tedy třeba zdůraznit nutné provázání obou těchto disciplín.

Zde také vzniká nová vědní disciplína – prostorová epidemiologie, jejíž nejjednodušší formu představuje použití a vyhodnocování map, na kterých je zakresleno rozmístění případů nemoci. Zahrnuta by měla být taktéž statistická analýza mapovaných dat (ŠTĚPÁNOVÁ, 2011). Tématem prostorové distribuce rakovinných onemocnění se pak zabývá tzv. nádorová epidemiologie.

Nádorová epidemiologie je relativně nová věda, která vznikala v průběhu druhé poloviny 20. století. Přesto však stihla již významně přispět k našemu pochopení příčin různých typů rakoviny a hodnocení preventivních opatření (DOS SANTOS SILVA, 1999).

Tématu nádorové epidemiologie se podrobně věnuje publikace s názvem *Cancer Epidemiology: Principles and Methods* (DOS SANTOS SILVA, 1999) vydaná pod záštitou International Agency for Research on Cancer.

Vedle prostorové epidemiologie se v literatuře také velmi často setkáváme použitím prostorové analýzy zdravotních dat. Právě metody prostorové analýzy zdravotních dat nám pomáhají hodnotit rozdíly konkrétních zdravotních charakteristik, pozorované v různých zeměpisných oblastech, umožňují oddělit prostorový vzor (*spatial pattern*) od náhodného šumu, identifikovat shluky nemocí a posoudit významnost potenciálních rizik (WALLER a GOTAWAY, 2004). Díky rychlému technologickému pokroku v geoinformačních vědách dnes existuje široká škála nástrojů prostorové analýzy dat. Ne všechny jsou však bez omezení použitelné také pro data o veřejném zdraví. Jejich použití v oblasti veřejného zdraví není příliš běžné a mohou představovat určitá analytická omezení. Například všechny zdravotní události (narození, nakažení, nemoc, smrt, atd.) se projevují u osob. Tito jedinci nejsou v prostoru distribuováni náhodně. Z toho vyplývá, že při každé analýze musí být brána v potaz prostorová distribuce obyvatel. (BARCELLOS, 2001).

Z hlediska epidemiologických studií zahrnujících prostorová data i prostorové analýzy je využívána velká šíře prostorových metod. Mohou jimi být prosté techniky

mapování výskytu nemoci v administrativních jednotkách nebo její lokalizace s využitím adres pacientů nebo geografických souřadnic. Setkáváme se však také s pokročilejšími metodami predikce rozšíření choroby, analýzou prostorových vzorů, rizik, či s vizualizací v kombinaci s daty dálkového průzkumu Země.

Za pomoci dostupných nástrojů a analytických metod je možné zkoumat mnoho aspektů veřejného zdraví. V praxi se můžeme setkat s analýzami dostupnosti zdravotní péče, demografické struktury pacientů a jejich umístění nebo s komplexní analýzou konkrétní nemoci a jejího výskytu, vlivu faktorů životního prostředí nebo socioekonomických faktorů (MAREK, 2015).

V přehledech české a světové literatury nebyla dosud uspokojivě popsána problematika výskytu vícečetných zhoubných novotvarů. Dílčí sdělení uvádí ve svém článku GERYK a kol., (2009), Autoři v článku shrnují stručný teoretický základ o potenciálních příčinách vzniku vícečetných zhoubných nádorů (dále VZN) a zejména popisují výskyt primárních a následných novotvarů v české populaci. Dále také hodnotí převažující diagnostické skupiny u nemocných s primárními respektive následnými novotvary či věkové zastoupení mužů a žen. Zastoupení klinických stádií u nemocných s VZN pak hodnotí opět GERYK a kol., (2010) v publikaci *Klinická stadia u nemocných s vícečetnými novotvary*. Vývoj VZN v čase, prostoru a potřebné náklady zdravotní péče v onkologii v souvislosti se nárůstem počtu VZN diskutuje opět GERYK a kol., (2008) v článku *Vícečetné zhoubné novotvary - Ukazatel zdraví a nákladů péče v onkologii*. Zde se vyskytuje také prezentace počtu VZN v krajích v mapách prevalence.

Možnostmi využití vícerozměrných statistických analýz, zejména analýzy hlavních komponent v epidemiologických prognózách, se ve své práci zabývá BÁČOVÁ, (2012). Autorka se zaměřuje na specifickou skupinu vícečetných zhoubných novotvarů (bilaterální karcinom prsu). Potenciální možnosti časoprostorové vizualizace dat o vícečetných novotvarech v České republice hodnotí TOTUŠEK, (2011). Porovnáním incidence vícečetných novotvarů v krajích České republiky se zabývá ZETKOVÁ, (2012). Autorka v práci analyzuje především diagnózy, které se v populaci objevují nejčastěji, a které zaznamenávají nejvyšší míru úmrtnosti.

V kontextu zahraniční literatury se tématu VZN značně detailně věnuje monografie *New malignancies among cancer survivors* (CURTIS, 2006). Monografie popisuje a kvantifikuje riziko vzniku nových malignit mezi více než dvěma miliony přeživších rakoviny v období let 1973 a 2000 v USA. K hodnocení autoři použili dat amerického Surveillance, Epidemiology and End Results (SEER) programu, který spadá pod záštitu National Cancer Institute. Jednotlivé kapitoly členěné dle původního umístění rakovinných buněk poskytují údaje o nebezpečí následných malignit. Velmi detailně bylo zkoumáno riziko následných malignit na základě pohlaví, věku v době diagnózy prvního nálezu a doby uplynulé od první diagnózy. Zvláštní pozornost je také věnována typu léčby původního nádoru a histologickému typu některých druhů rakoviny. Každá kapitola srovnává vzory vícečetných novotvarů s výsledky z jiných studií a porovnává výsledky s ohledem na potenciální rizikové faktory a mechanismy.

Při hodnocení literatury byla zjištěna úplná absence mapového zobrazení problematiky vícečetných zhoubných novotvarů. Přesto je ale zobrazení zdravotních dat (nejen onkologických) v mapě stále častější a je využíváno v mnoha oborech. Můžeme se setkat s aplikacemi určenými pro specialisty, ale také s jednoduššími nástroji, které může využívat i široká laická veřejnost. Níže jsou uvedeny čtyři zástupci různých typů aplikací, které využívají mapu pro prezentaci zdravotních dat.

Zvýšení výpočetního výkonu a kapacity paměti počítače v kombinaci s rostoucí dostupností geokódovaných dat způsobilo nárůst počtu „geograficko-zdravotnických" studií, jejichž cílem je zkoumat a objevovat prostorové vzory v datových souborech. Hlavním problémem při analýze zdravotních dat je fakt, že prostorové vzory velmi často odrážejí vliv komplexní konstelace demografických, sociálních, ekonomických, kulturních a environmentálních aspektů, které se mohou měnit s místem a časem. Proto je nutno na data nahlížet a zpracovávat je v různém časovém a prostorovém měřítku. To přináší širokou škálu statistických a vizualizačních technik ve většině studií zdravotního stavu (GOOVAERTS, 2010).

V dnešní „digitální" době roste význam výpočetních technologií, webových a mobilních aplikací, ale i sociálních sítí. Prostřednictvím těchto nástrojů je možné šířit důležité informace o zdravotním stavu mezi širokou veřejnost a tak v krátkém čase informovat mnoho lidí o případných zdravotních hrozbách a včas prosadit vhodná preventivní opatření. Právě tyto technologie by měly být v současnosti stěžejním bodem zájmu výzkumu v oblasti veřejného zdraví, epidemiologie a prevence.

V následujících podkapitolách jsou blíže specifikovány čtyři vybrané nástroje pro prezentaci zdravotních dat. Vybrané nástroje byly do práce začleněny jako zástupci několika různých typů aplikací. **National Cancer Institute GIS Portal** zastupuje aplikaci provozovanou národní institucí a založenou na oficiálních a ověřených datech. Aplikaci **Health Map** lze označit za jistý typ sociální sítě. Její fungování je založeno na automatickém shromažďování informací z nejrůznějších zdrojů a jejich následná publikace probíhá formou mapových výstupů. Využití **Google Earth** pro publikaci zdravotních dat má zcela jiný charakter než přechozí dvě aplikace. Jedná se sice o volně dostupnou aplikaci, jejíž použití je velmi uživatelsky přívětivé, ale data nejsou shromažďována automaticky. Naopak data musí být shromážděna a zpracována uživatelem a následně publikována ve formátech .*kml nebo *kmz. Výhoda této aplikace spočívá právě v její jednoduchosti, takže může být využívána i laiky. Pravým opakem je poslední zmíněná aplikace **EpiInfo**, která je zástupcem sofistikovaných programů pro specialisty z řad lékařů, epidemiologů a dalších. Její přínos spočívá v tom, že umožňuje zadávání vlastních dat, která mohou být dále analyzována a vizualizována v mapě.

2.1 National Cancer Institute GIS Portal

http://gis.cancer.gov/portal/

NCI GIS Portal je webová báze pro interaktivní mapování a vizualizaci onkologických geoprostorových dat. Portál je provozován americkým National Cancer Institute. Portál v sobě spojuje GIS a vědecké principy a nástroje pro harmonizaci velkého objemu vícerozměrných datasetů. Jedná se o spojení dat o rakovině a behaviorálních, environmentálních, klinických, sociálně ekonomických a politických údajů na úrovni států a okresů. Data jsou zobrazována na podkladu ESRI. Primárním účelem GIS Portálu je informovat, vzdělávat a inspirovat uživatele k vytváření a hodnocení nových výzkumných hypotéz.

Geoprostorové nástroje jsou v NCI používány pro řadu aplikací (National Cancer Institute [on-line]):

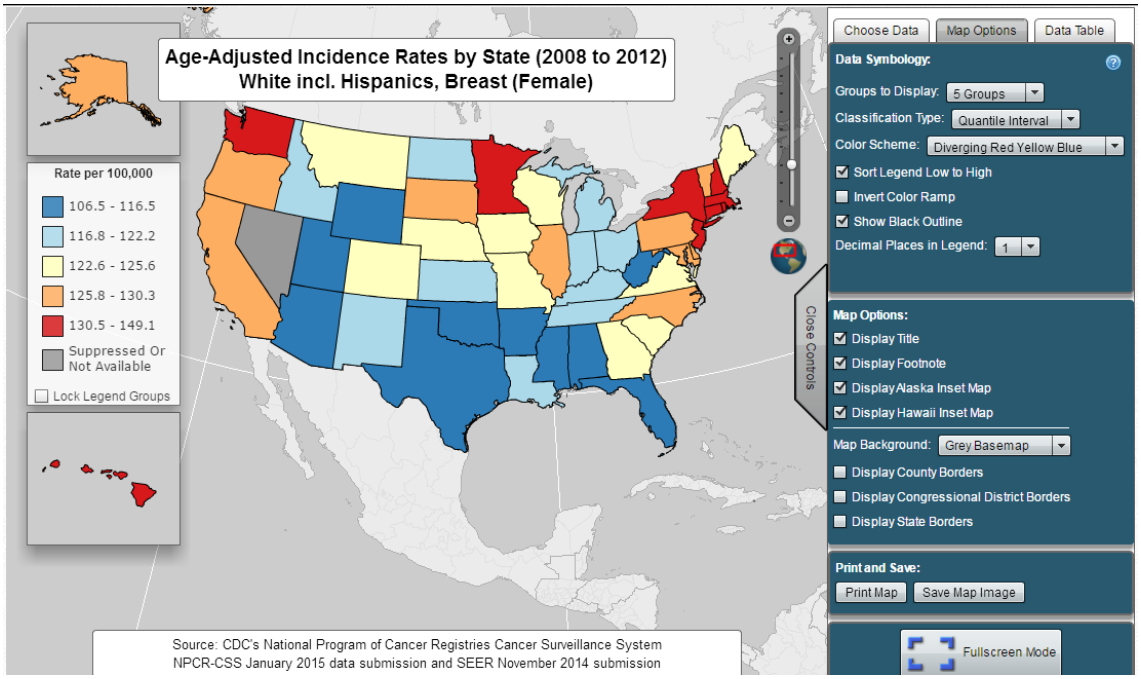
- Identifikace a zobrazení prostorových vzorů výskytu rakoviny a úmrtnosti ve Spojených státech amerických a jejich změny v čase
- Tvorba komplexních databází pro studium screeningu rakoviny, diagnóz a přežití na úrovni společenství
- Posuzování expozice životního prostředí prostřednictvím satelitního snímkování
- Prostorové statistické modely pro odhad incidence rakoviny, prevalence a přežití pro každý stát USA
- Sdělení o rakovině na místní úrovni veřejnosti a veřejným zdravotním odborníkům prostřednictvím interaktivních internetových nástrojů
- Vývoj nových metod zobrazování geoprostorových dat pro jasnou komunikaci s veřejností a pro zkoumání komplexních vícerozměrných dat

K těmto účelům portál nabízí tři různé mapovací nástroje:

NCI GeoViewer je nástroj, který uživateli umožňuje v mapě kombinovat rakovinné a demografické statistiky a rizikové faktory. Uživatel si nejprve vybere data z jedné ze šesti tematických kategorií: Demografie, Historická mortalita, Incidence, Mortalita, Prevalence a Screening a rizikové faktory. Aplikace pro zobrazení dat využívá jednoduchý homogenní kartogram. Data jsou přepočítána na 100 000 obyvatel. Uživatel si může nadefinovat počet výsledných intervalů (3-10), typ klasifikace (quantile, equal interval) a barevné schéma. Výslednou mapu je pak možné vytisknout nebo vyexportovat ve formátu .png. Aplikace také nabízí možnost zobrazení dat v tabelární fromě. Ukázka prostředí NCI GeoViewer je na Obr. 1.

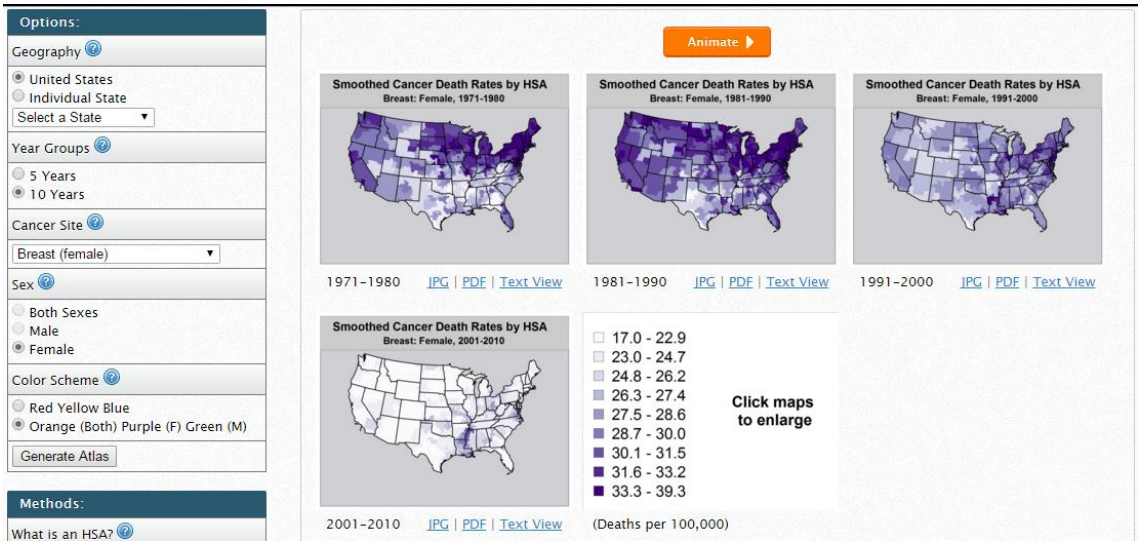
NCI Map Stories je další nástroj ze sady NCI GIS Portal, který poskytuje sérii map na dané téma včetně textového popisu jednotlivých map. Prozatím jsou Map Stories dostupné pouze pro rakovinu prsu, plic, tlustého střeva, prostaty a rakovinu děložního čípku.

Animated Historical Cancer Atlas je poslední ze sady nástrojů, který umožňuje uživatelům vytvářet animované mapy věkově standardizovaných shlazených měr úmrtnosti (v období 1971-2010) na vybraný typ rakoviny. Animaci je možno vytvořit na úrovni celých Spojených států nebo pouze za vybraný stát a může zobrazovat 5 nebo 10leté intervaly úmrtnosti. Uživatel si opět může zvolit barevnou škálu a případně pohlaví, pro které jsou výsledky prezentovány. Aplikace následně vygeneruje sérii čtyř respektive osmi map, které jsou pak vstupem výsledné animace. Každou z map je možné jednotlivě stáhnout ve formátu *.jpg, *pdf nebo ji zobrazit v textové formě, kde jsou slovně popsány hodnoty vyskytující se v jednotlivých regionech. Ukázka prostředí Animated Historical Cancer Atlas ukazuje Obr. 2.



Obr. 1: Ukázka uživatelského prostředí NCI GeoViewer na příkladu věkově standardizované incidence rakoviny prsu ve státech USA v letech 2008-2012

(Zdroj: https://gis.cancer.gov)



Obr. 2: Ukázka uživatelského prostředí Animated Historical Cancer Atlas na příkladu úmrtnosti na rakovinu prsu ve státech USA v letech 1971-2010

(Zdroj: https://gis.cancer.gov)

2.2 HealthMap

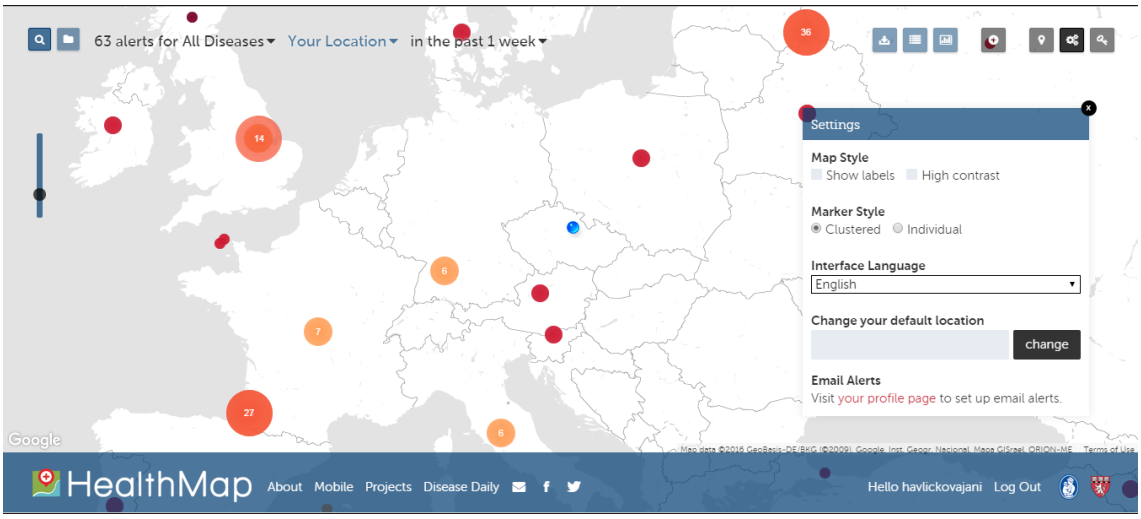
http://www.healthmap.org/en/

HealthMap je Linuxová/Apache/MySQL/PHP aplikace založená na Google Maps. Aplikace HealthMap je projektem týmu výzkumníků, epidemiologů a softwarových vývojářů Bostonské dětské nemocnice. Projekt byl založen v roce 2006 doktorem Johnem Brownsteinem a jeho týmem. Aplikace je založena na využívání on-line neformálních zdrojů pro sledování ohniska nákazy a real-time dohledu nad vznikajícím ohrožením veřejného zdraví. Volně dostupná webová stránka healthmap.org a mobilní aplikace „Outbreaks Near Me“ dodává v reálném čase zpravodajské informace o širokém spektru nově se objevujících infekčních chorob. HealthMap spojuje díky automatickým procesům různorodé zpravodajské zdroje – on-line zpravodajské agregátory, zprávy očitých svědků, odborné diskuse i ověřené oficiální zprávy. Díky tomu vzniká jednotný a detailní pohled na aktuální globální stav distribuce infekčních chorob a jejich vliv na zdraví lidí a zvířat. Prostřednictvím automatizovaného procesu, nepřetržité aktualizace a systémových kontrol organizuje, integruje, filtruje, vizualizuje a šíří on-line informace o nových chorobách v sedmi světových jazycích včetně angličtiny, španělštiny, francouzštiny, portugalštiny, ruštiny, čínštiny a arabštiny. To usnadňuje včasnou detekci globálních hrozeb pro veřejné zdraví.

Úspěch projektu spočívá v tom, že produkuje velmi srozumitelné výstupy – mapy celého světa s bodovým znázorněním jednotlivých míst výskytu nemocí. V nastavení mapy je také možné zvolit si zobrazení míst shlukově, viz Obr. 3. Barva jednotlivých bodových značek je určena algoritmem, který je zohledňuje časové hledisko jednotlivých hrozeb, počet hrozeb v oblasti a počet zdrojů poskytujících informaci o hrozbě. Velikost

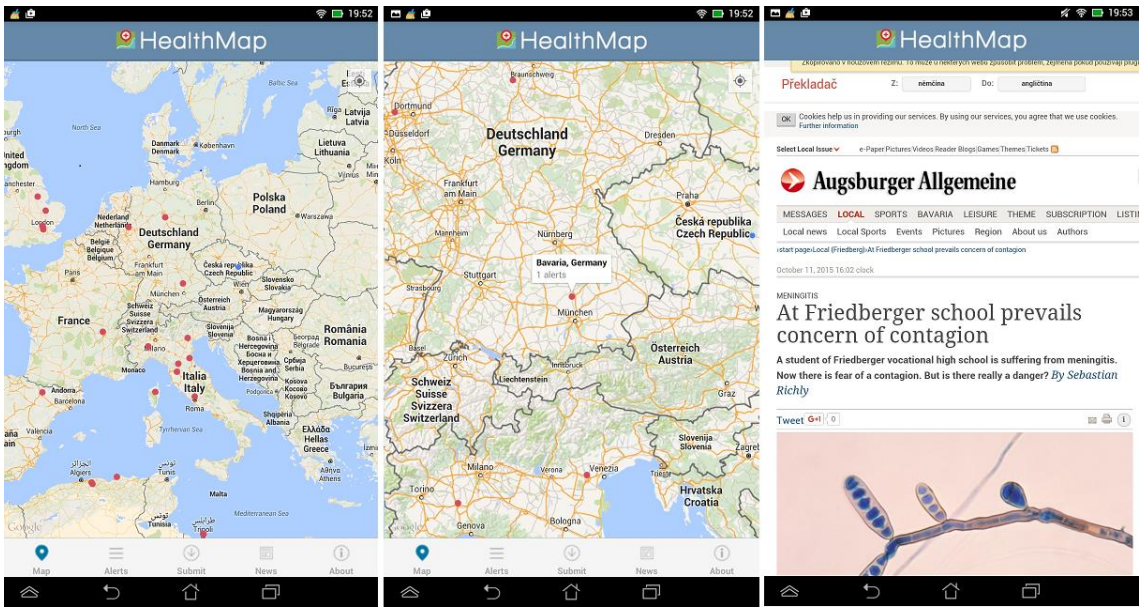
značek pak definuje, pro jaké území hrozba platí. Velký kruh označuje hrozbu na úrovni jednotlivých zemí, zatímco malý kroužek značí hrozbu na úrovni krajů, spolkových zemí, provincií atd.

Jednotlivé body jsou klikatelné a po rozkliknutí jsou zobrazeny jednotlivé relevantní zdroje s možností přesměrování na adresu konkrétní zprávy. Na stejném principu funguje taktéž mobilní aplikace Outbreaks Near Me (viz Obr. 4), jejíž funkčnost byla ověřena na mobilním zařízení a tabletu s operačním systémem Android, verze 4.2.2.



Obr. 3: Ukázka mapového výstupu aplikace HealthMap se shlukovým zobrazením výskytu rizik

(Převzato z: http://www.healthmap.org/en/)



Obr. 4: Ukázka výstupů z mobilní aplikace Outbreaks Near Me

(Převzato z: http://www.healthmap.org/en/)

HealthMap umožňuje také registraci nových uživatel. Registrovaní uživatelé mohou přidávat nová, editovat stávající, nebo změnit pořadí existujících uložených vyhledávání (včetně nastavení výchozího zobrazení) pro rychlý a snadný přístup k zájmovým vyhledáváním. Dále si mohou nastavit, ze kterých internetových stránek chtějí dostávat upozornění o nových nemocech ve zvolené geografické oblasti. Registrovaným je dále umožněno hodnotit již existující publikované hrozby a komentovat je nebo stahovat rozsáhlé tabulky s daty ve formátu *.csv, které obsahují veškeré možné informace o jednotlivých zprávách.

Aplikace také nabízí pokročilé vyhledávání, které uživateli umožňuje vyhledávat záznamy přesně dle jím zvolených kritérií. Je možno vyhledávat výsledky pouze pro zvolené choroby, konkrétní místa, dle zdroje ze kterého byla informace získána, dle živočišného druhu, pro který je ohrožení relevantní a dle data, kdy byly hrozby zaznamenány.

Přihlášení uživatelé si mohou pokročilá vyhledávání uložit do sekce *„QuickView“*, kde je možné výslednou mapu pojmenovat a stručně popsat. Výsledek tohoto vyhledávání se pak uloží jako nová domovská stránka, která se automaticky načte po přihlášení.

Webovou stránku navštíví více než jeden milion návštěvníků za rok. Specifickými uživateli HealthMap jsou vládní agentury (CDC, HHS, DOD, WHO, ECDC), úředníci z oblasti veřejného zdraví, knihovny či mezinárodní cestovatelé.

Projekt HealthMap používá algoritmy, pomocí kterých analyzuje desítky tisíc zpráv na webu. Jeho obsah je tvořen agregováním volně dostupných on-line zdrojů organizací ProfiMED Mail, WHO, OIE – World Organization for Animal Health, FAO, Eurosurveillance, Google News, Moreover, Wildlife Data Integration Network, Baidu News, SOSO Info (HealthMap [on-line], 2015).

Nedávným významným úspěchem celého projektu bylo odhalení výskytu eboly v roce 2014 dříve než jej odhalila Světová zdravotnická organizace. Algoritmus přečetl několik článků na blozích doktorů ze státu Guinea. Díky tomu byl problém odhalen již 14. března 2014 přičemž oficiální prohlášení WHO vyšlo až 23. března 2014 (Publichealthwatch [on-line], 2014).

Aplikace představuje významný nástroj pro včasné varování před nebezpečnými chorobami na každém místě na světě pomocí jednoduchých mapových výstupů, které jsou k dispozici a zároveň jednoduše pochopitelné pro širokou veřejnost.

2.3 Epi Info ™

http://wwwn.cdc.gov/epiinfo/7/

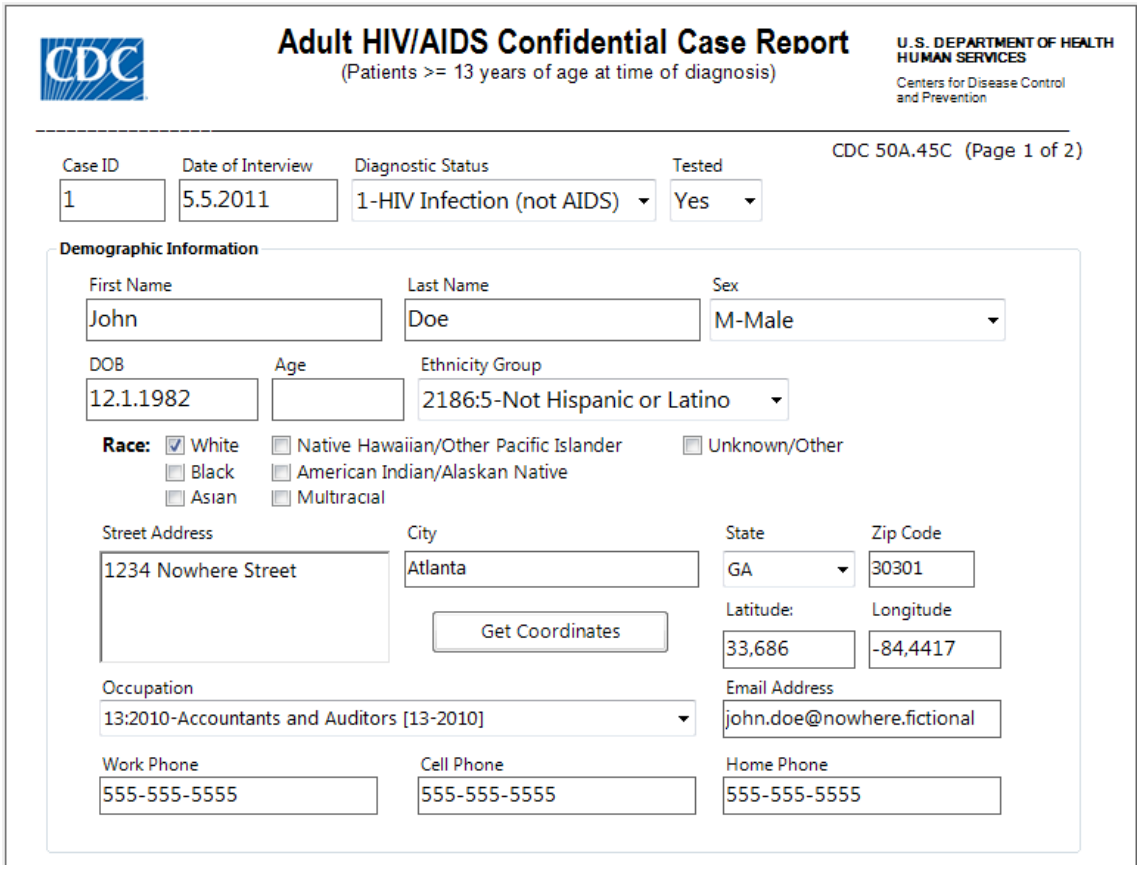
Epi Info je základním volně dosupným statistickým softwarem pro epidemiologii vyvinutý Centers for Disease Control and Prevention (CDC) ve Spojených státech Amerických v Atlantě. Epi Info je příkladem aplikace, která je primárně určena pro

odborníky epidemiologických oborů. Ti mohou pomocí aplikace sbírat, analyzovat a následně prezentovat informace o zdraví svých pacientů.

Epi Info poskytuje sadu softwarových nástrojů pro sběr dat, analýzu, následnou vizualizaci a podávání přehledových zpráv za pomoci epidemiologických metod. Je použitelný i bez připojení k internetu. Program umožňuje sběr dat, pokročilé statistické analýzy a tvorbu map.

První verze programu (Epi Info 1) byla implementována na operační systém MS-DOS počítačovým vědcem Jeffem Deanem na 5,25 palcové disketě a byla vydána v roce 1985. Nejnovější verzí je Epi Info 7.5.1, která byla vydána 19. března 2015 a je kompatibilní s operačními systémy Windows XP, Vista, 7 a 8. Za dobu své existence registruje Epi Info uživatele již ze 181 zemí světa a byl přeložen do 13 světových jazyků. Epi Info je používán celosvětově pro rychlé posouzení ohnisek nákazy.

Program je rozdělen na 4 základní moduly. Modul návrhu formulářů (*Form Designer module*) umožňuje uživatelům vytvářet dotazníky a formuláře pro sběr dat, které budou následně v programu zpracovány. Modul vstupu dat (*Enter Modul*) automaticky z předem připraveného dotazníku (vytvořeného v předchozím modulu) vytvoří databázi (viz Obr. 5). Uživatelé zde mohou zadávat údaje, upravovat již existující údaje nebo vyhledávat záznamy. Modul analýzy (*Analysis module*) je užíván ke čtení a analyzování dat zadaných pomocí modulu vstupu dat nebo získané importem dat z 24 možných datových formátů. Modul nabízí tvorbu epidemiologických statistik, tabulek, grafů. V rámci modulu je možno využít t-test, ANOVA, neparametrické testy, kontingenční tabulky, logistickou regresi (podmíněnou i nepodmíněnou), analýzu přežití a mnoho dalších (viz Obr. 6).



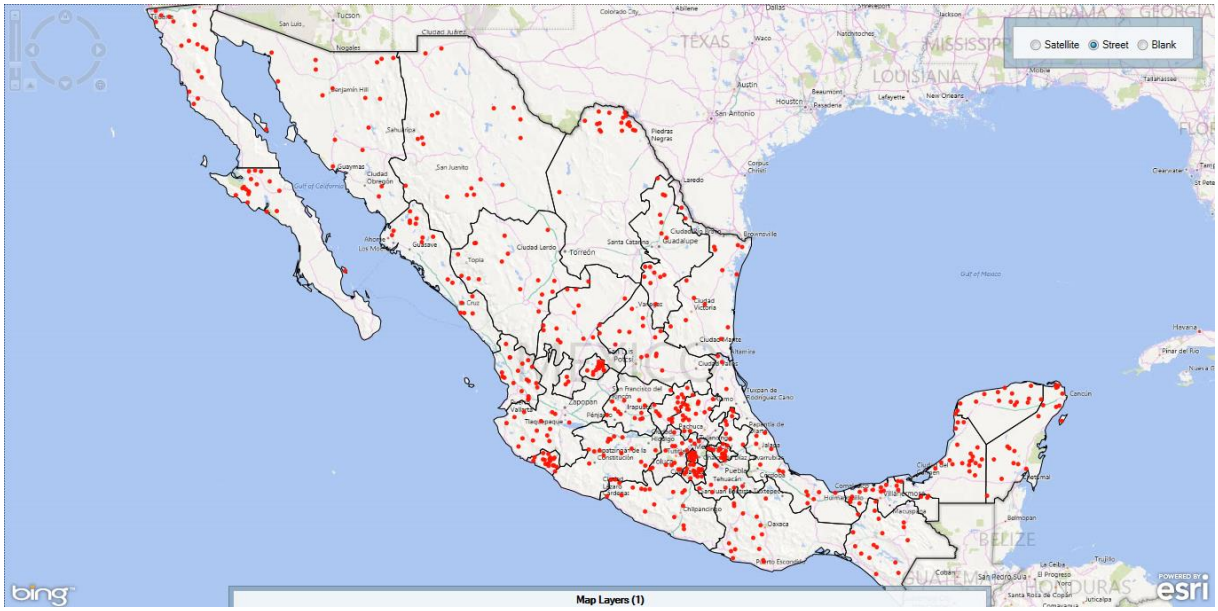
Obr. 5: Ukázka prostředí modulu Enter Data - vyplněný vzorový formulář na příkladu dat o HIV

Obr. 6: Ukázka prostředí modulu Analysis na příkladu vzorových dat o nákaze bakterií E-Coli

Modul mapy (*Epi Map Module*) zobrazuje mapy s daty z Epi Info. Epi Map je postaven na ESRI MapObject softwaru[1]. Modul využívá podkladové mapy Bing Maps. Mapy zobrazují *.shp obsahující geografické hranice států společně s vrstvou datových výstupů z modulu analýz. Uživatelé si tak mohou vytvořit vlastní mapu obsahující data o veřejném zdraví dle vlastních potřeb. Modul nabízí základní funkcionalitu jako zoom, posouvání či rotování mapy a volbu několika typů podkladů – základní, satelitní či možnost vložit vlastní mapový podklad. Dále je možné do mapy vkládat vlastní body včetně popisků a tzv. zóny, vymezující kruhovou oblast o zadaném poloměru kolem daného místa. To může být vhodné například pro vymezení spádové oblasti zdravotnických center, dosahu nákazy apod.

Nástroj umožňuje vytvářet také vlastní vrstvy: zájmové body, tzv. case clusters (vrstva shluků výskytu choroby), kartogramy (*choropleth*), tečkové mapy (*dot density*) (viz Obr. 7). Dále je možné přidat další referenční vrstvy ve formátech *.kml, *.shp nebo načíst z volitelného mapového serveru. Do jednoho mapového okna je možno nahrát více dat z různých databází, což umožňuje uživateli pozorovat vzájemné vztahy mezi daty.

Výsledné mapy je také možné ukládat jednak ve formátu *.png a dále také jako mapový soubor s koncovkou *.map7, který uloží mapu včetně jejích podkladových datových vrstev (Epi Info [on-line], 2015).



Obr. 7: Ukázka modulu Epi Map se zobrazením tečkové mapy na vzorových datech o porodech mladistvých v jednotlivých státech Mexika

V srpnu 2012 byla vydána také stejnojmenná verze pro mobilní telefony a tablety s operačním systémem Android, která je dostupná volně ke stažení (viz Obr. 8). Aplikace

[1] ESRI MapObject je soubor mapových softwarových komponent, které umožňovaly vkládat mapy do aplikací. Nyní již ESRI MapObject nenabízí.

nabízí oproti počítačovému software pouze omezenou funkcionalitu. Její výhodou je však okamžitá dostupnost chytrých mobilních telefonů a cenově dostupných tabletů, pomocí kterých mohou epidemiologové v podstatě okamžitě provádět výpočty epidemiologických statistik, nalézat ohniska nákazy, reagovat na mimořádné situace nebo provádět průzkumy veřejného zdraví v místech, kde chybí infrastruktura informačních technologií. Aplikace navíc umožňuje ukládat data na cloud, což umožní týmu epidemiologů efektivněji shromažďovat data. Data mohou být dále sbírána offline a následně po připojení k internetu mohou být synchronizována mezi jednotlivými členy týmu. V případě neustálého připojení k internetu jsou data na cloudu synchronizována okamžitě a průběžně. To umožňuje analyzovat data v reálném čase.

V roce 2009 se Epi Info stalo open source programem (od verze 7.0.5.) a na stránkách Epi Info™ - Community Edition (Epi Info [on-line], 2014) je zveřejněn zdrojový kód, včetně celé dokumentace a výukových a podpůrných materiálů. Tyto mohou být volně kopírovány, šířeny a překládány. Zároveň se vznikem open source verze vznikla Epi Info™ Community Edition, který usiluje o zdokonalení a reprodukci sady nástrojů Epi Info™ v jazyce C #, s cílem rozvoje sběru dat a analytických systému pro veřejné zdraví. Epi Info™ Community Edition tedy neustále vyvíjí nové nástroje a testuje jejich funkčnost. Tyto nástroje jsou v případě osvědčení implementovány na stabilní verzi, která je k dispozici na webových stránkách CDC Epi Info™.

Program Epi Info se stal mocným nástrojem pro včasnou detekci potenciálního ohrožení veřejného zdraví. Díky široké škále statistických a analytických nástrojů, které program nabízí, je možné analyzovat nejen rozsah nemoci, ale také například kapacitu lůžek a zdravotnického personálu, což umožňuje rychlou reakci na vzniklé situace. Významným pozitivem programu je možnost zobrazit výsledky analýz v mapě a srozumitelně je tak prezentovat široké, nejen odborné, ale i laické veřejnosti. Neboť jak uvádí PICKLE, (2003): „Obraz může vydat za tisíc slov, ale mapa může reprezentovat miliony datových bodů.“

Obr. 8: Ukázka prostředí mobilní aplikace Epi Info na tabletu – vlevo úvodní stránka, vpravo modul analýzy s přidaným mapovým oknem

2.4 Digital Earth, Google Earth a veřejné zdraví

Koncept Digital Earth je oproti mapování veřejného zdraví, jehož původ sahá až do 18. století, relativně nový koncept. Jeho původcem, který poprvé vyslovil myšlenku „digitální Země“ byl tehdejší viceprezident Spojených států Amerických Al Gore. Ve své řeči pro California Science Center v Los Angeles v roce 1998 nastínil ideu digitální, trojrozměrné, rotující zeměkoule a jejího propojení s vědeckými, kulturními či sociálními daty, které by pomohlo porozumět planetě Zemi a procesům, které na ní probíhají (GORE, 1998).

Dle KONEČNÉHO, (2015) je Digital Earth konceptem, jehož cílem je propojit mapy a data do bezešvého geoprostorového systému, který bude celosvětově dostupný.

Příkladem, který představuje částečné naplnění konceptu Digital Earth jsou tzv. *virtual globe geo-browsers*. Tím je například nejznámější aplikace Google Earth, který je celosvětově známým a oblíbeným nástrojem, který umožňuje nejen zobrazovat prostorová data, ale také je interaktivně prohlížet a analyzovat jejich změny v prostoru

a čase. Síla tohoto nástroje spočívá v jednoduchém a velmi intuitivním ovládání a je ještě více umocněna skutečností, že se jedná o zdarma dostupný software.

Ačkoli je Google Earth primárně zacílen na širokou veřejnost a obecně je nejčastěji využíván jako vyhledávací nástroj, v poslední době se mu také dostává pozornosti velké komunity uživatel, kteří oceňují široký potenciál jeho využití. Odborníci z řad lékařů a epidemiologů nejsou výjimkou. Google Earth je plně využitelný pro zkoumání prostorových vzorů v datech, poskytování a šíření výsledků mezi širokou odbornou i laickou veřejnost a srozumitelné zprostředkování analytických interpretací výsledků (MAREK, 2015).

Jak uvádí BRANDUSESCU, (2011) se vznikem Digital Earth a zejména Google Earth platformy se významně zvýšila četnost použití geovizualizace dat z různých oborů lidské činnosti, včetně oblasti veřejného zdraví. S růstem jejich popularity, vznikají nové potenciální způsoby reprezentace dat. Výhoda použití geovizualizace skrze Digital Earths oproti tradičnímu GIS spočívá v tom, že jej mohou jednoduše využívat nejen GIS specialisté, ale mohou do nich přispívat také laikové. Ve zdravotních projektech nalézají tzv. virtuální glóby stále větší uplatnění, jakožto mocný nástroj pro monitorování a prezentaci výsledků spojených s veřejným zdravím (SAARNAK et al., 2012).

2.4.1 Mapování rozšíření viru ptačí chřipky

https://www.google.com/earth/outreach/stories/showcase.html#

Aplikace Google Earth bývá pro vizualizaci zdravotních dat využívána stále častěji. Na výše uvedených webových stránkách v oddílu Public Health je výčet několika reálných příkladů. Významným příkladem je aplikace mapování rozšíření viru ptačí chřipky. Tato Google Earth prezentace je významným počinem Declana Butlera, reportéra časopisu Nature. Asociace on-line vydavatelů Butlera v roce 2006 ocenila cenou *Use of the new digital platform award*.

Butler v projektu kompiluje data od vypuknutí viru ptačí chřipky u ptáků v roce 2003. Zahrnuty jsou také potvrzené případy nákazy u lidí (SAARNAK et al. 2012). Celkově se jedná o více než 1 800 ohnisek nákazy ptačí chřipky po celém světě zobrazené na podkladu Google Earth. Přidruženy jsou navíc další dvě doplňkové vrstvy: hustota zalidnění (zdroj: CIESIN) a tzv. hustota drůbeže (*poultry density*, zdroj: FAO). Výsledný *.kml soubor je pak volně dostupný ke stažení přímo na stránkách Google Earth jako jeden z mnoha ukázkových příkladů aplikací Google Earth

(https://www.google.com/earth/outreach/stories/showcase.html#kml=Watch_the_Spread_of_Bird_Flu).

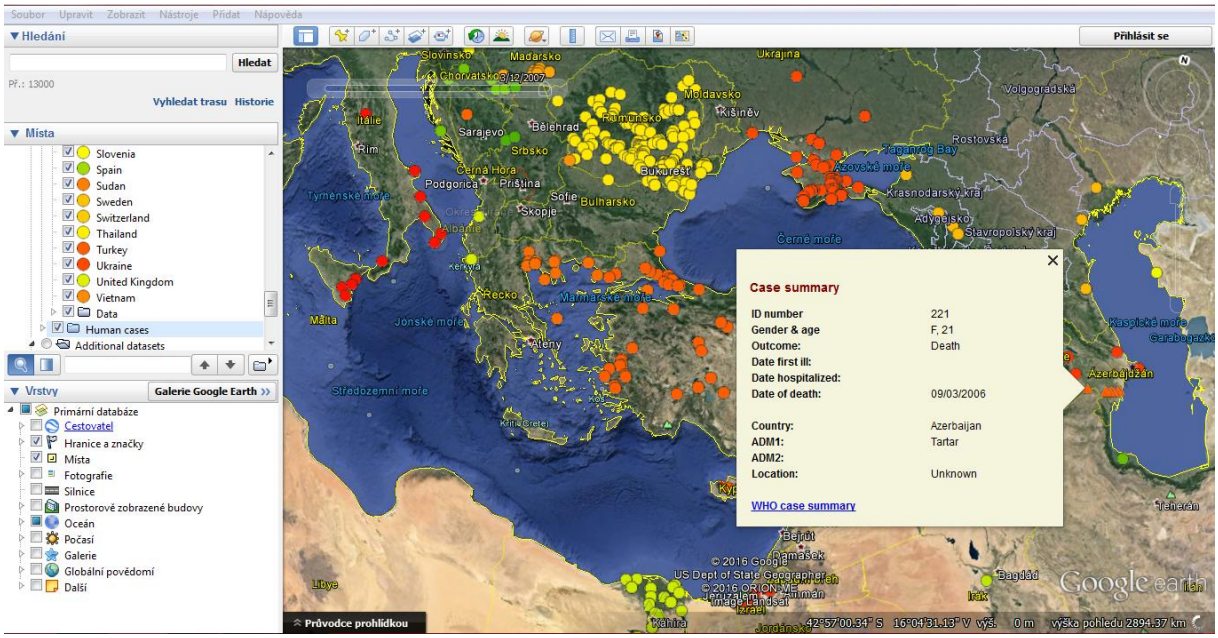
V *.kml souboru jsou kromě podkladových dat obsaženy také autorovy komentáře a charakteristika dat. Velká část dat o výskytu a rozšíření nákazy u ptáků byla sestavena z dat FAO, OIE a různých vládních zdrojů. Tato data pak byla doplněna informacemi převzatými z týdenních zpráv OIE. Data o výskytu ptačí chřipky u lidí pak byla převzata

z bulletinů WHO a různých vědeckých prací. Lokalizace potvrzených případů je definována pouze na administrativní úrovni okresů.

BUTLER, (2006) v charakteristice dat uvádí, že nejobtížnější částí práce byl úvodní preprocessing dat FAO. Důvodem byla absence zeměpisných souřadnic v původním datasetu. Data byla původně lokalizována na základě systému OSN pro definování geografických jednotek názvem místa, provincie a okresu. Geografické souřadnice tedy musely být vypočteny jednotlivě pro každé ohnisko.

Data jsou rozčleněna do dvou samostatných databází – „zvířecí ohniska“, které jsou v mapě reprezentovány symbolem kruhu a „případy u lidí“, reprezentované symbolem trojúhelníku. Funkcionalita je u všech bodů stejná – při přejetí myší je zobrazen název místa, při dvojitém poklepání na bod se zobrazí okno s doplňkovými informacemi (viz Obr. 9). Navíc je možné zobrazit vývoj ptačí chřipky v čase díky přidané časové složce. Po spuštění animace je možné sledovat vývoj od 25. 11. 2003 do 13. 3. 2010.

Tato aplikace představuje významný potenciální nástroj pro hodnocení prostorového a časového šíření chorob, pro hledání jeho prostorových vzorů a předpověď budoucího vývoje. Díky relativně přesné deskripci jednotlivých případů je také možné určit potenciálně ohroženou skupinu a zavést patřičná preventivní opatření.



Obr. 9: Ukázka mapování výskytu ptačí chřipky v prostředí Google Earth se zobrazením okna s doplňkovými informacemi pro vybraný bod

3 VÍCEČETNÉ ZHOUBNÉ NOVOTVARY

Vícečetné zhoubné novotvary (VZN) postihující téhož nemocného patří mezi obávané komplikace léčby. V anglické terminologii jsou vícečetné zhoubné novotvary označovány termínem „*second primary cancers*“. Tento termín je definován jako v pořadí druhý (respektive třetí, čtvrtý atd.) výskyt rakoviny postihující téhož nemocného během jeho života s odlišnou histologickou povahou a topografickým uložením (FELLER a LEMMER, 2012), GERYK a kol. (2008). Nejedná se však o oživení ani metastázi primární rakoviny.

Vznik následného novotvaru může mít stejnou příčinu jako primární. KOUBKOVÁ a kol. (2013) uvádí, že druhé nádory mohou vznikat také jako důsledek předchozí cytostatické léčby a radioterapie prvotního novotvaru, vrozené predispozice, životního stylu či faktorů prostředí. Pravděpodobnost vzniku nádorové duplicity resp. multiplicity vzrůstá taktéž s delším přežíváním onkologicky nemocných díky zlepšující se zdravotní péči.

3.1 Vývoj a výskyt VZN v krajích České republiky v letech 1976-2010

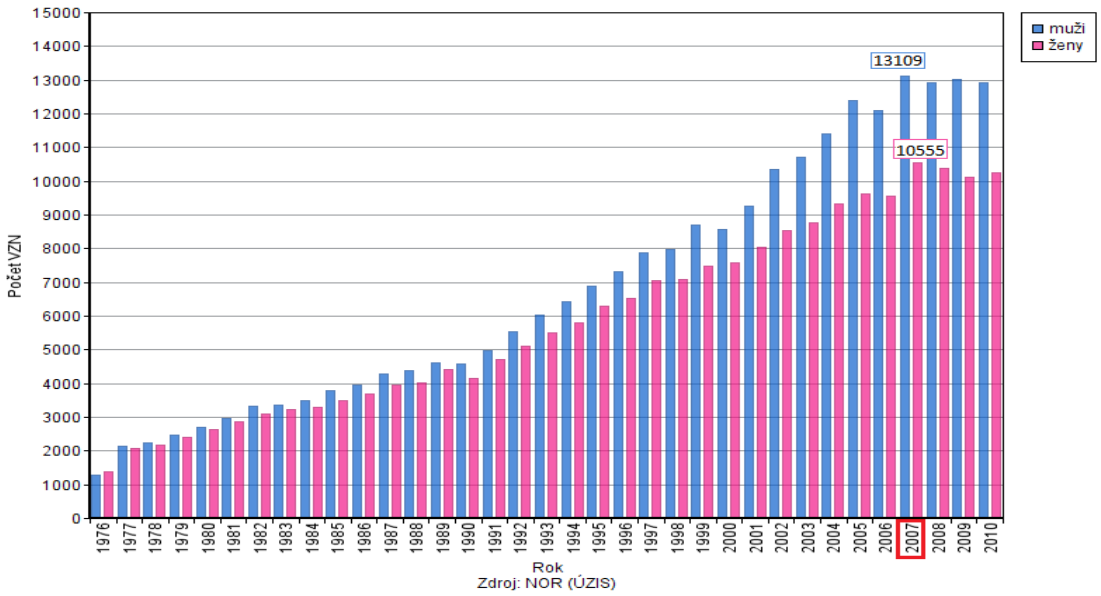
Dříve byly druhé respektive následné nádory považovány spíše za kuriozitu, dnes se řadí k třetí nejběžnější skupině malignit (KOUBKOVÁ a kol., 2013). Rozvoj více zhoubných nádorů vyskytujících se u jednoho jedince je stále častější v souvislosti s pokroky v léčbě rakoviny a prodlužováním života pacientů. GERYK a kol., (2009) uvádí, že z 1 486 984 novotvarů dg. C00-D48, evidovaných v letech 1976-2005 v registru nádorů, se u 125 262 (8,4 %) onkologicky nemocných vyskytlo 165 050 (11,1 %) následných novotvarů. Průměrná doba intervalu mezi primárním a následným onemocněním byla 6 let u mužů a 6,6 roku u žen. Do jednoho roku po zjištění primárního novotvaru se u mužů vyskytlo 15 602 (17,4 %) a u žen 11 689 (15,5 %) následných novotvarů. Z hlediska celkového počtu VZN ve věku 0-85+ let převažovali muži nad ženami. Absolutní a relativní zastoupení VZN v celkové prevalenci případů onemocnění nádory mezi roky 1989-2005 se ve věku 0-85+ let zvýšilo.

V návaznosti na zvyšující se objem onkologické péče je vhodné v prevalenčních počtech (charakterizujících počty přežívajících pacientů) odlišit zastoupení VZN. V případě evidence VZN je u nemocného postupně nahlašován výskyt každé další histologicky a topograficky odlišné malignity. „Evidované VZN u téhož pacienta zahrnují prevalenční počty jednotlivých diagnóz a celkové počty žijících případů nádorů, v jejichž statistice je nemocný uveden tolikrát, kolik zhoubných novotvarů bylo u něj nahlášeno. Oproti tomu prevalenční data osob s nádory zahrnují nemocné jen podle prvního nahlášeného novotvaru bez dalších VZN u téže osoby.“ (GERYK a kol., 2008).

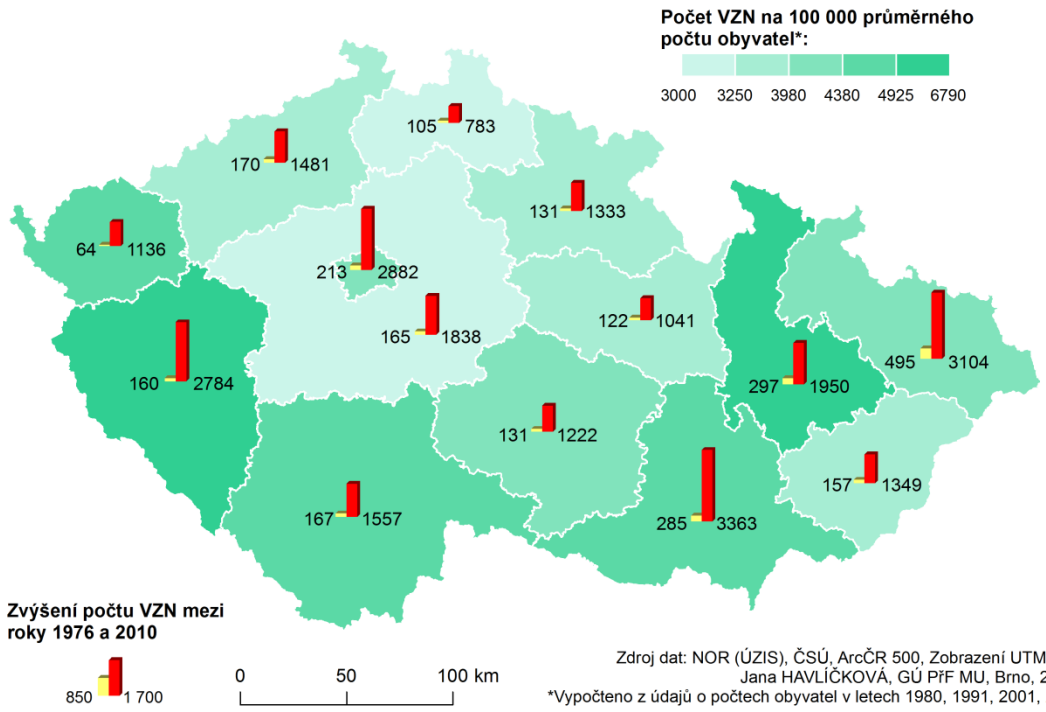
Zaměříme-li se na časový vývoj počtu vícečetných zhoubných novotvarů, který ilustruje Obr. 10, můžeme identifikovat zcela zřejmý vzrůstající trend u obou pohlaví. Počet VZN s drobnými výkyvy soustavně rostl od roku 1976 do roku 2007, kdy dosáhl jejich počet svého maxima. Tedy 13 109 diagnóz VZN u mužů (více než desetinásobek

minima v roce 1976) a 10 555 diagnóz VZN u žen (téměř osminásobek minima v roce 1976), (pozn.: BÁČOVÁ, (2012) uvádí, že data pro rok 1976 jsou evidována až od května toho roku). Relativně nízké počty VZN na počátku sledovaného období můžeme odůvodnit tak, že lze předpokládat určitou časovou prodlevu mezi diagnostikováním následného novotvaru u pacientů registrovaných s primární diagnózou. Zhruba od roku 1990 se začal zvyšovat rozdíl v počtech VZN mezi muži a ženami. Z toho pak vyplývá významný rozdíl procentuálního nárůstu počtu nově hlášených VZN na konci sledovaného období.

Porovnáme-li počet VZN v prvním a posledním pětiletém období, můžeme pozorovat 5,9násobný nárůst u mužů a 4,8násobný nárůst počtu VZN u žen. Tento alarmující nárůst by měl upozornit na významně rostoucí trend výskytu následných novotvarů a zdůraznit potřebu vzniku screeningových programů, šíření osvěty a dostatečně dlouhé následné péče.



Obr. 10: Vývoj počtu vícečetných zhoubných novotvarů u žen a mužů ve věku 0-85+ let mezi roky 1976-2010 (Zdroj dat: NOR (ÚZIS))



Obr. 11: Vícečetné zhoubné novotvary u obou pohlaví v letech 1976-2010 v krajích České republiky (Zdroj dat: NOR (ÚZIS), ČSÚ, ArcČR 500)

Prostorovou diferenciaci počtu případů VZN přepočteného na 100 000 průměrného počtu obyvatel demonstruje jednoduchý homogenní kartogram na

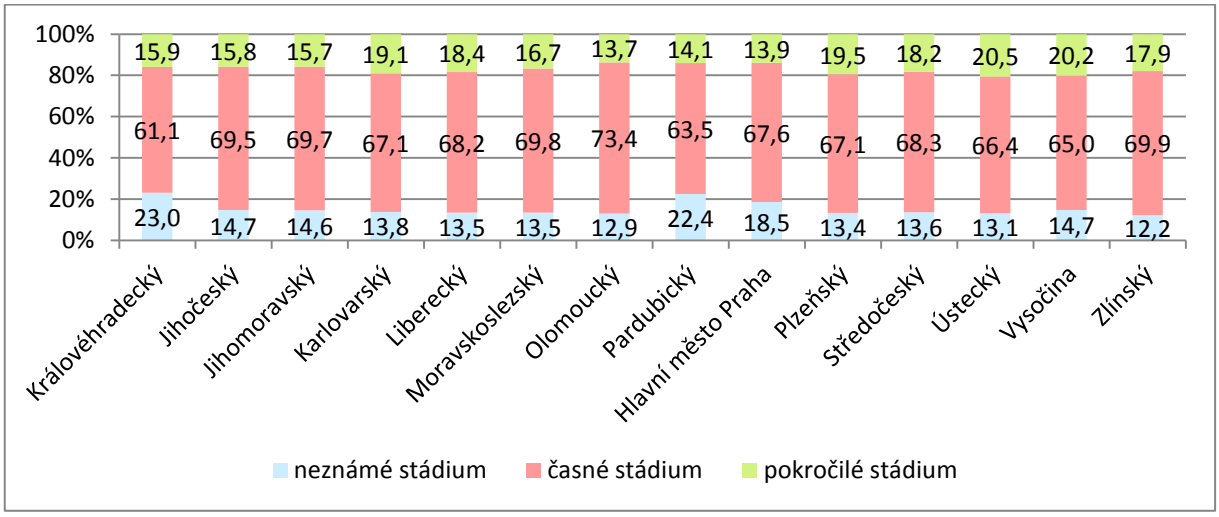
. Změna barevné intenzity znázorňuje celkový počet nově diagnostikovaných VZN od roku 1976 do roku 2010 přepočtený na průměrný počet obyvatel (výpočet je blíže specifikován v podkapitole *3.2 Charakteristika zpracovávaného datového souboru*). Výsledné hodnoty jsou rozděleny do pěti intervalů na základě výpočtu kvantilů. Tato metoda klasifikace dat definuje v každém intervalu stejný počet prvků, díky čemuž nevznikají prázdné třídy nebo třídy s příliš málo nebo mnoho hodnotami. Nejvyšších hodnot počtu VZN přepočtených na 100 000 průměrného počtu obyvatel dosahují kraje Plzeňský a Olomoucký. Vysokých hodnot dosahují také kraje Karlovarský, Jihočeský a Jihomoravský a Hlavní město Praha. Nejpříznivější situace je ve Středočeském a Libereckém kraji.

Pomocí kartodiagramu jsou v mapě vyjádřeny absolutní hodnoty počtu VZN na počátku a na konci sledovaného období, tedy v letech 1979 a 2010. Největší relativní nárůst byl pozorován v Karlovarském kraji. Zde byl počet VZN v roce 2010 17,75krát vyšší než v roce 1976. V Plzeňském kraji byl počet VZN na konci období 17,4krát vyšší než na počátku. Nejnižší relativní nárůst byl pak zaznamenán v Moravskoslezském kraji.

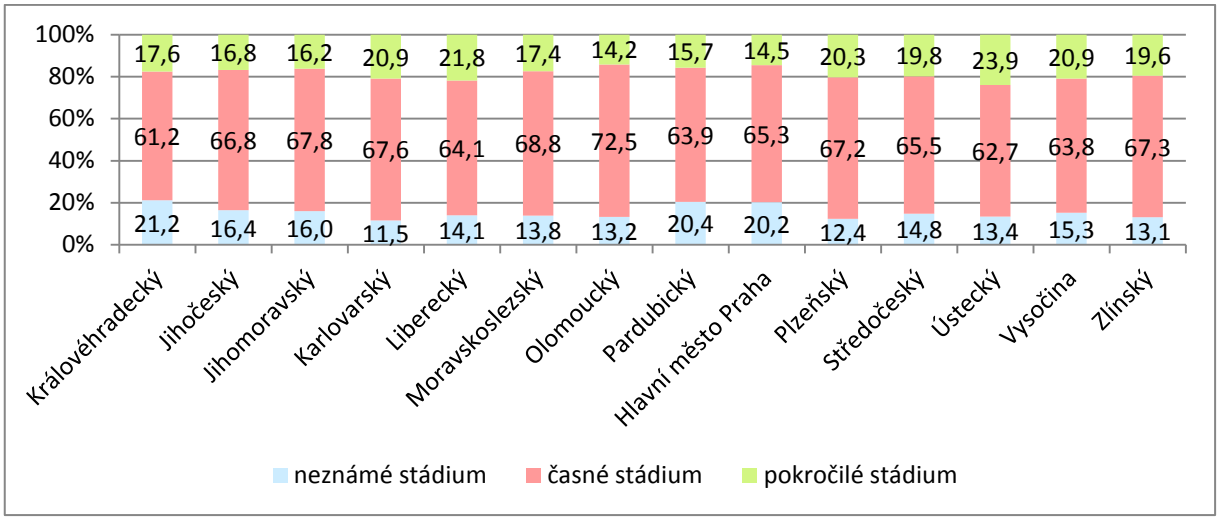
Významnou charakteristikou popisující výskyt VZN je zastoupení klinických stádií u primárních, respektive následných novotvarů, které pro jejich léčbu rozhodující. Způsob označení klinického stádia v databázi je blíže popsán v podkapitole *3.3.3 Klinické*

stádium nádorového onemocnění. Na Obr. 12 a Obr. 13 je zobrazeno zastoupení jednotlivých klinických stádií u primárních resp. následných novotvarů. Z obrázků je patrné, že největší část primárních novotvarů je diagnostikována v časném stádiu. Avšak zde zůstává poměrně značné procento případů, které jsou diagnostikovány až v pokročilém stádiu. Situace může být značně zkreslena v Královéhradeckém a Pardubickém kraji, kde je u více než 20 % diagnostikovaných případů klinické stádium označeno jako neznámé. Podobně se situace jeví u následných diagnóz. Přesto lze v porovnání s primární diagnózou pozorovat mírný nárůst podílu případů diagnostikovaných v pokročilém stádiu. Nejvyšší podíl případů v pokročilém stádiu byl diagnostikován v Ústeckém kraji. Naopak nejlépe bilance vychází pro kraj Olomoucký a Hlavní město Praha. I v případě následných diagnóz jsou zaznamenány případy v neznámém stádiu, přičemž nejvíce jich bylo pozorováno opět v kraji Královéhradeckém a Pardubickém a navíc také v Hlavním městě Praha. Ve všech případech podíl neznámých stádií přesahuje 20 %.

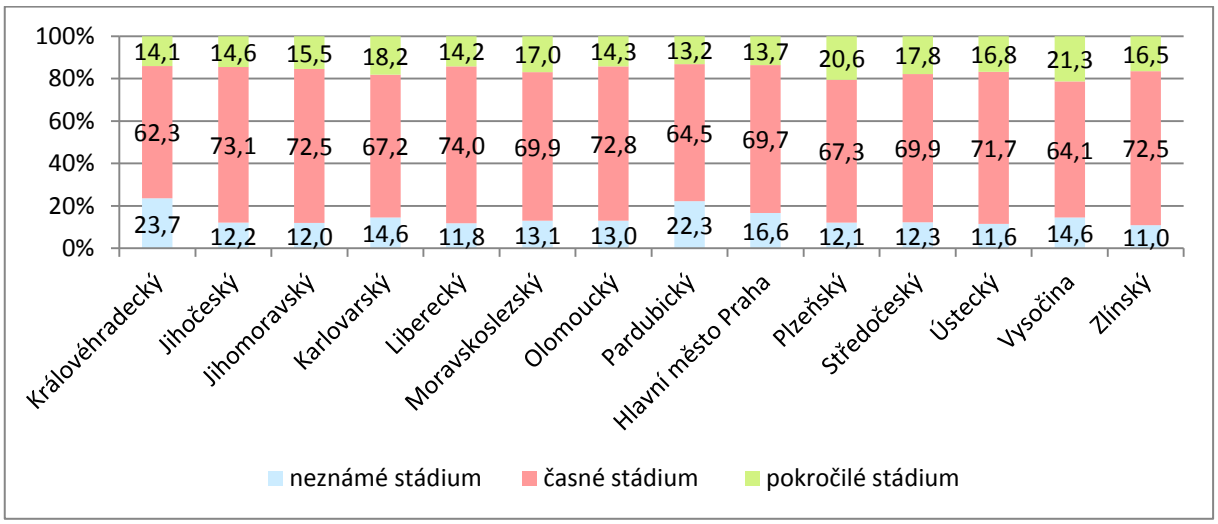
Obr. 12 - Obr. 17 pak ukazují zastoupení jednotlivých klinických stádií u primárních a následných diagnóz u mužů a u žen. Ve všech krajích je v případě primárních i neznámých diagnóz vyšší zastoupení pokročilých stádií u mužů oproti ženám. To znamená, že ženám jsou novotvary diagnostikovány obecně dříve než mužům. Nejistotu však do výsledků přináší relativně velký podíl neznámých stádií. Za zmínku stojí zejména téměř 30 % podíl neznámých stádií primárních diagnóz u žen v Královéhradeckém a Pardubickém kraji. Tématu neznámých stádií se pak blíže věnuje podkapitola *3.3.3 Klinické stádium nádorového onemocnění.*



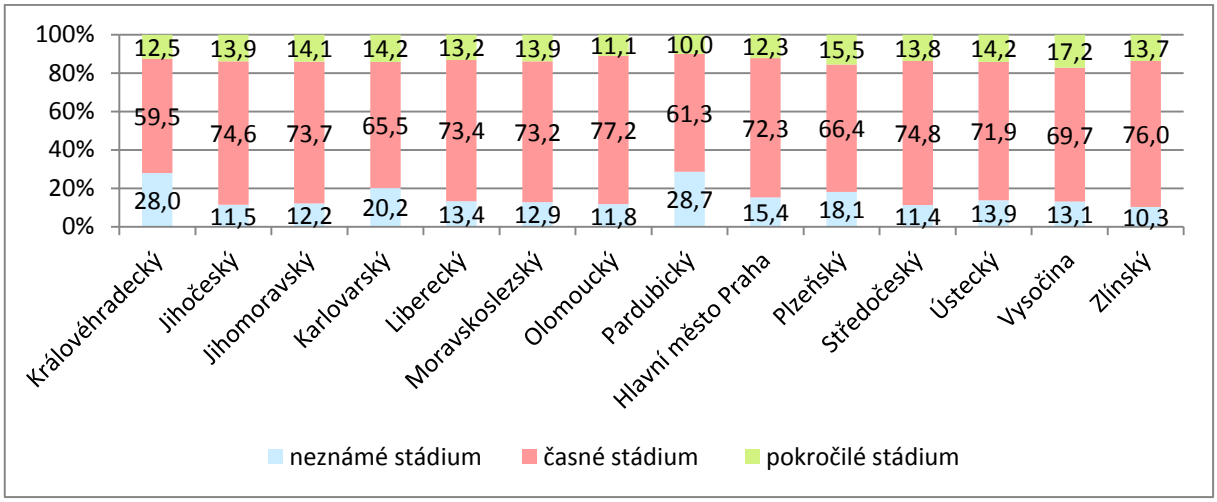
Obr. 12: Zastoupení jednotlivých stádií primární diagnózy VZN dg. C00-D48 u obou pohlaví v krajích ČR v letech 1976-2010 (Zdroj dat: NOR, ÚZIS)



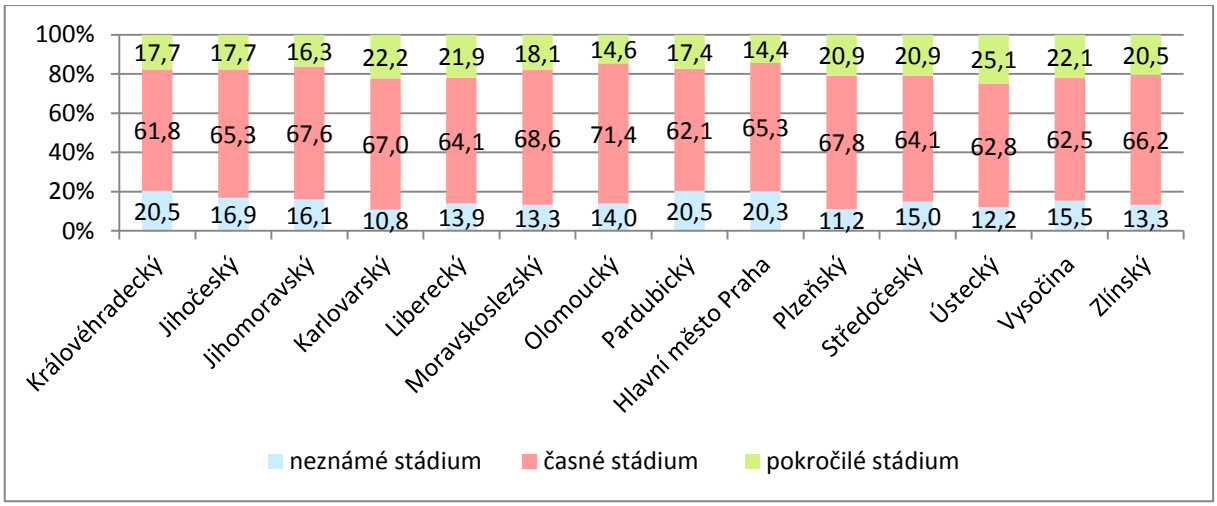
Obr. 13: Zastoupení jednotlivých stádií následných diagnóz VZN dg. C00-D48 u obou pohlaví v krajích ČR v letech 1976-2010 (Zdroj dat: NOR, ÚZIS)



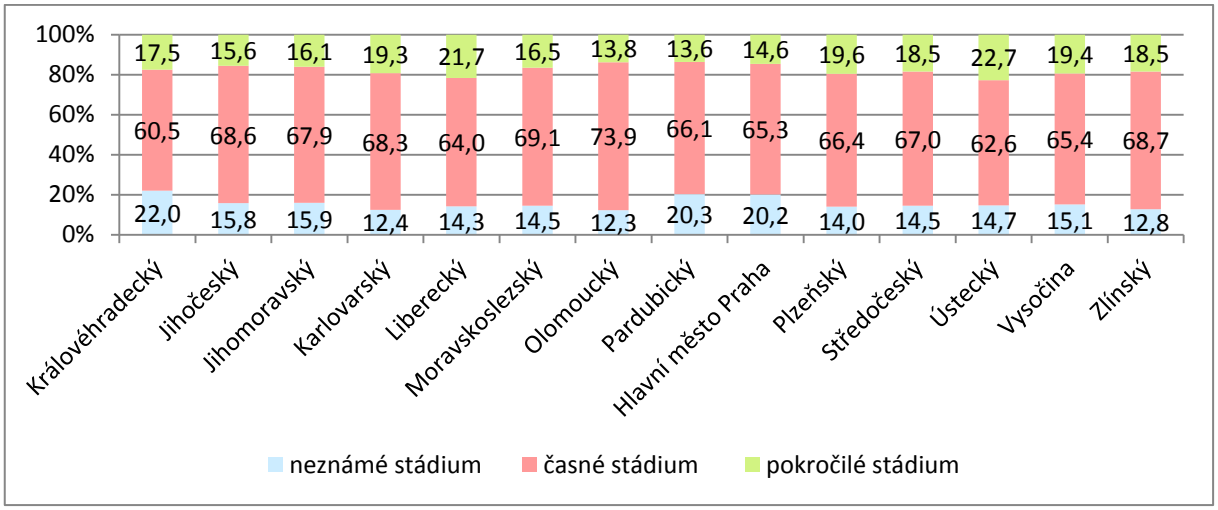
Obr. 14: Zastoupení jednotlivých stádií primárních diagnóz VZN dg. C00-D48 u mužů v krajích ČR v letech 1976-2010 (Zdroj dat: NOR, ÚZIS)



Obr. 15: Zastoupení jednotlivých stádií primárních diagnóz VZN dg. C00-D48 u žen v krajích ČR v letech 1976-2010 (Zdroj dat: NOR, ÚZIS)



Obr. 16: Zastoupení jednotlivých stádií následných diagnóz VZN dg. C00-D48 u mužů v krajích ČR v letech 1976-2010 (Zdroj dat: NOR, ÚZIS)



Obr. 17: Zastoupení jednotlivých stádií následných diagnóz VZN dg. C00-D48 u žen v krajích ČR v letech 1976-2010 (Zdroj dat: NOR, ÚZIS)

Ačkoli vícečetné zhoubné novotvary představují závažné komplikace léčby a zkracují případnou délku přežití, neexistuje příliš mnoho výzkumů zabývající se touto problematikou. Americký National Cancer Institute jako jeden z mála zaznamenává v rámci Surveilance, Epidemiology, and End Result (SEER) pogramu vícečetné malignity. SEER registr (About SEER (on-line), 2016) začal evidovat údaje o rakovině k 1. lednu 1973, tedy o více než tři roky dříve než český NOR. Data však nejsou evidována za celé území Spojených států amerických. Jedná se spíše o sběr dat z vybraných signifikantních území (tak, aby v registru byly zastoupeny všechny věkové, rasové a další skupiny obyvatel). Tato data jsou pak vztažena na území celého státu.

SEER je zaměřen na vyhodnocování vzniku následných nádorů od roku 1985. Při hodnocení je pak brána v potaz histopatologická povaha, rozsah prvotní léčby, rasové a etnické rozdíly. Ve srovnání s českým populačním NOR evidují SEER registr navíc detaily o individuální radiační zátěži a objemu chemoterapie při předchozí léčbě (GERYK, 2009).

3.2 Charakteristika zpracovávaného datového souboru

Předmětem zpracování diplomové práce byla data z Národního onkologického registru (ÚZIS). Data charakterizují vícečetné zhoubné novotvary spojené s výskytem dalších novotvarů nahlášených do NOR České republiky od května roku 1976 do roku 2010. Počty byly evidované k 31. 12. každého roku a aktualizované a verifikované Ústavem zdravotnických informací a statistiky České republiky v říjnu 2012. Údaje o počtech diagnostikovaných pacientů byly poskytnuty na administrativní úrovni okresů podle místa bydliště pacienta v okamžik nálezu první diagnózy.

Ačkoli jsou data do registru zadávána na úrovni jednotlivců, včetně velmi podrobných údajů o konkrétním pacientovi, pro zpracování práce byla data podrobena

anonymizaci, pro zajištění ochrany dat v souladu se zákonem č. 101/2000 Sb. o ochraně osobních údajů. Anonymizace proběhla formou agregace dat podle místa bydliště pacienta v okamžik nálezu první diagnózy.

Počty diagnostikovaných pacientů jsou normalizovány na průměrný počet mužů a žen žijících v daném okrese. Průměrný počet byl vypočten jako prostý aritmetický průměr z údajů ze Sčítání lidu, domů a bytů v letech 1980, 1991, 2001, 2011. Počty mužů a žen v okresech odpovídají převodům sídel mezi kraji podle změn územního členění České republiky. Data o počtech obyvatel (mužů, žen) byla čerpána z Českého statistického úřadu.

Přesto že jsou poskytnutá data v databázi agregována na úroveň okresů, objevují se zde také data, kterým jsou přiřazeny pouze kraje, bez bližšího určení okresu. Jedná se o celkem 31 diagnostikovaných případů, z toho 5 případů s neznámým označením klinického stádia, 22 případů v časném stádiu a 4 případy ve stádiu pokročilém. Vzhledem k tomu, že k databázi nebyly poskytnuty další doplňující informace, není možné zcela uspokojivě vysvětlit důvod, proč není u těchto případů definován okres bydliště pacienta. Lze se pouze domnívat, že se jednalo o cizince nebo osoby bez trvalého bydliště. Tyto případy nebyly do analýz započítávány, aby do dat nebyla zanášena chyba.

Data z NOR byla poskytnuta v databázi PosgreSQL ve dvou tabulkových prostorech: *vicecet_pacient* a *vicecet_diagnoza.* Potřebná data byla z databáze extrahována pomocí open source grafického administračního rozhraní skrze jednotlivé SQL dotazy. Soubory dat byly převedeny do kategoriálních dat a pomocí kontingenčních tabulek byla vyfiltrována potřebná data. Pro pre-processing datových souborů a tvorbu vstupních souborů pro následné zpracování a tvorbu výsledné kartografické prezentace byl využit geografický informační systém ArcMap, verze 10.3.1.

3.3 Klasifikace v onkologii

3.3.1 TNM klasifikace

Dle Linkos [on-line], 2015 slouží TNM klasifikace k jednoduchému popisu rozsahu nádoru a určení stádia onemocnění. Stádium onemocnění je pak jedním z klíčových kritérií, které lékař hodnotí při volbě léčby.

TNM systém pro klasifikaci zhoubných nádorů vyvinul mezi roky 1943 a 1952 Francouz Pierre Denoix. Jelikož byly do pravidel klasifikace určitých anatomických lokalizací v průběhu let zaváděny odchylky, dohodly se národní výbory pro TNM v roce 1982 na formulování jednotné TNM klasifikace. Za účelem sjednocení a aktualizace stávající klasifikace bylo uskutečněno několik setkání. Výsledkem bylo 4. vydání klasifikace TNM. V roce 1993 pak byl vydán Doplněk TNM (TNM Supplement). Tento dokument měl poskytnout především podrobný výklad pravidel TNM klasifikace

s praktickými příklady. (Upraveno dle TNM Klasifikace zhoubných novotvarů, 7. Vydání (originál 2011) [on-line], 2013).

Dle Linkos [on-line], (2015), TNM systém není univerzální. Pro každou nádorovou lokalizace je vypracován vlastní systém. TNM se určuje na základě klinického vyšetření, které zahrnuje vyšetření onkologem a zobrazovací vyšetření. Tato klasifikace se může lišit, je-li prováděna na žijícím či zemřelém pacientovi.

Písmeno T znamená „Tumor“ a popisuje rozsah primárního nádoru, a to buď jeho velikost, nebo vztah k okolním strukturám. Zde je možno také definovat, že nádor není přítomen, nebo že jej nelze klasifikovat.

Písmeno N znamená „Node“ a popisuje postižení regionálních lymfatických uzlin a rozsah takového postižení.

Písmeno M pak popisuje přítomnost či nepřítomnost vzdálených metastáz.

3.3.2 MKN – 10: Mezinárodní klasifikace nemocí a přidružených zdravotních problémů

MKN označuje mezinárodní statistickou klasifikaci nemocí a přidružených zdravotních problémů. MKN lze definovat jako soustavu kategorií, do kterých jsou zařazovány chorobné jevy podle zavedených kritérií (MKN-10 [on-line], 2008).

První vydání Mezinárodní klasifikace nemocí je datováno do roku 1893. Vznikla formalizací tzv. Bertollonovy klasifikace neboli Mezinárodního seznamu příčin smrti. Nyní je v platnosti již 10. revize, která byla zpracovávána od roku 1983 a v platnost vešla k 1. 1. 2009.

Zdravotní problémy jsou zde klasifikovány způsobem, který je považován za nejvhodnější pro obecné epidemiologické účely a pro hodnocení zdravotní péče. Klasifikace má podobu číselníku s diagnostickými popisy a výkladem u jednotlivých položek a kapitol. Hlavním klíčem je znakový kód nemoci.

Jednotlivé druhy nemocí jsou označeny alfanumerickými kódy. První znak je písmeno latinské abecedy, které udává hlavní kategorii. Numerické znaky na druhém a třetím místě určují hlavní skupinu diagnóz. Na čtvrtém, respektive dalším místě jsou numerické znaky pro podrobnější členění (MKN-10 [on-line], 2008).

Onkologická onemocnění jsou označena kódy začínajícími písmenem C (*C00-C97 - zhoubné novotvary*) a D (*D00-D09 – novotvary in situ, D10-D39 – nezhoubné novotvary, D37-D48 – novotvary nejistého nebo neznámého chování*). Přehled klasifikace onkologických onemocnění dle klasifikace MKN shrnuje Tab. 1.

Tab. 1: Klasifikace onkologických onemocnění dle klasifikace MKN – 10. revize

Kód	Popis
C00-C14	Zhoubné novotvary rtu, dutiny ústní a hltanu
C15-C26	Zhoubné novotvary trávicího ústrojí
C30-C39	Zhoubné novotvary dýchací soustavy a nitrohrudních orgánů
C40-C41	Zhoubné novotvary kosti a kloubní chrupavky
C43-C44	Melanom a jiné zhoubné novotvary kůže
C45-C49	Zhoubné novotvary mezotelové a měkké tkáně
C50	Zhoubné novotvary prsu
C51-C58	Zhoubné novotvary ženských pohlavních orgánů
C60-C63	Zhoubné novotvary mužských pohlavních orgánů
C64-C68	Zhoubné novotvary močového ústrojí
C69-C72	Zhoubné novotvary oka, mozku a jiných částí centrální nervové soustavy
C73-C75	Zhoubné novotvary štítné žlázy a jiných žláz s vnitřní sekrecí
C76-C80	Zhoubné novotvary nepřesně určených sekundárních a neurčených lokalizací
C81-C96	Zhoubné novotvary mízní, krvetvorné a příbuzné tkáně
C97	Zhoubné novotvary mnohočetných samostatných (primárních) lokalizací
D00-D09	Novotvary in-situ
D10-D36	Nezhoubné novotvary
D37-D48	Novotvary nejistého nebo neznámého chování

(*Zdroj: MKN-10, [on-line], 2008)*

3.3.3 Klinické stádium nádorového onemocnění

Jedním z klíčových ukazatelů onkologických dat je *klinické stádium* onemocnění. Dělení klinických stádií vychází z TNM klasifikace. Nádorová onemocnění jsou dle celkového rozsahu dělena do čtyř stádií (I-IV), u některých stádií jsou pak ještě dále definovány podskupiny (A, B, C). Do databáze NOR jsou klinická stádia udávána v podobě arabských číslic, jejichž význam objasňuje Tab. 2. Onemocnění lze na základě znalosti stádia označit jako *časné* (neboli lokalizované), které svým charakterem odpovídá I. a II. stádiu, dále *pokročilé*, které odpovídá III. a IV. stádiu nebo jako *neznámo* (Závazné pokyny NZIS, 2013). Čím vyšší stádium, tím je onemocnění závažnější a pravděpodobnost metastazování nádoru vyšší.

Tab. 2: Význam typů stádií onkologických onemocnění zadávaných do NOR

Číslo	Význam
0	Stadium 0 (novotvar in situ, neboli zatím neinvazivní)
1	Stádium I.
2	Stádium II.
3	Stádium III.
4	Stádium IV.
6	metastázy u nádoru neznámé primární lokalizace
7	neuvádí se
9	neznámo

(Zdroj: Závazné pokyny NZIS, 2013)

4 METODY PRO ANALÝZU AGREGOVANÝCH DAT

Analýza a vizualizace prostorově agregovaných dat může být negativně ovlivňována rozdíly ve velikosti a tvaru administrativních jednotek, různou velikostí datové základny a z toho vyplývajících výkyvů sledovaného jevu, který nemusí odrážet jeho variabilitu ale právě rozdíly v mapovaných jednotkách (ŠTĚPÁNOVÁ, 2011).

4.1 Korelační analýza a korelační diagram

Korelace obecně označuje míru stupně závislosti dvou proměnných. Dvě proměnné jsou korelované, jestliže určité hodnoty jedné proměnné mají tendenci se vyskytovat společně s určitými hodnotami druhé sledované proměnné. Míra korelace může být různá od neexistence korelace až po absolutní korelaci (HOŠKOVÁ, 2006).

Nejjednodušším způsobem pro rozpoznání, zda je mezi hodnotami dvou náhodných veličin nějaká závislost je vykreslení bodového grafu – korelačního diagramu (*scatter plot*). Ten zobrazuje, jak hodnoty jedné veličiny rostou nebo klesají v závislosti na druhé veličině. Každý bod v diagramu odpovídá jednomu páru měření, tzv. *korelační dvojici* (x_i, y_i). Pomocí korelačního diagramu se snažíme zjistit, zda se statistická závislost mezi sledovanými jevy blíží některé funkční závislosti. Jsou-li body v něm rozloženy více méně rovnoměrně, je to důkaz, že posuzované veličiny vykazují nulovou korelaci. Blíží-li se rozložení bodů v grafu přímce, jedná se o tzv. *lineární závislost* (BEDÁŇOVÁ, VEČEREK, 2007).

Nevýhodou korelačního diagramu je však absence kvantifikace funkčního vztahu sledovaných veličin. Pro kvantifikaci lineárního vztahu byl zaveden tzv. *Pearsonův korelační koeficient (Pearson correlation coefficient),* (HOLČÍK, KOMENDA a kol., 2015).

Pearsonův korelační koeficient měří sílu lineární závislosti mezi dvěma veličinami. Korelační koeficient r je počítán z tzv. kovariance s_{xy} a směrodatných odchylek jednotlivých proměnných s_x a s_y. Pearsonův korelační koeficient je dán následujícím vztahem:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{cov\,(x,y)}{s_x s_y}, kde\ s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} \qquad [4.1]$$

Výsledek Pearsonova korelačního koeficientu může nabývat všech hodnot z intervalu ⟨-1,1⟩. Kladných hotnot koeficient r nabývá, pokud vyšší hodnoty první proměnné souvisí s vyššími hodnotami druhé proměnné. Záporných hodnot r nabývá pokud nižší hodnoty první proměnné souvisí s vyššími hodnotami druhé proměnné. Hodnotu 1 resp. -1 dostaneme v případě, že lze vztah obou proměnných vyjádřit přímkou

(HOŠKOVÁ, 2006). Přibližnou interpretaci možných výsledků korelačního koeficientu *r* shrnuje Tab. 3.

HOŠKOVÁ, (2006) však uvádí, že interpretace korelačního koeficientu není tak přímočará. Proto se doporučuje dopočítat další charakteristiky, jako například parametry proložené přímky nebo směrodatná odchylka odhadu při regresi. Dále je také vhodné doplnit intervalem spolehlivosti, která nám dá informace o variabilitě tohoto odhadu a následně také doplnit testem hypotézy o nulové korelaci dvou náhodných veličin. Blíže se této problematice věnuje HOLČÍK, KOMENDA a kol., 2015.

Tab. 3: Přibližná interpretace hodnot korelačního koeficientu

Koeficient korelace	Interpretace
|r| = 1	Naprostá závislost (funkční závislost)
1,0 > |r| ≥ 0,9	Velmi vysoká závislost
0,9 > |r| ≥ 0,7	Vysoká závislost
0,7 > |r| ≥ 0,4	Střední závislost
0,4 > |r| ≥ 0,2	Nízká závislost
0,2 > |r| ≥ 0,0	Slabá závislost
|r| = 0	Naprostá nezávislost

(Zdroj: CHRÁSKA, M., 2000)

Druhou mocninou korelačního koeficietnu je *koeficient determinace*, který se značí R^2, který udává kolik procent celkové variability může být vysvětlitelných zvoleným regresním modelem. Tzn. informuje o těstnosi závislosti mezi zvolenými proměnnými. Koeficient determinace nabývá hodnot z intervalu ⟨0,1⟩. Čím vyšší hodnoty nabývá, tím lépe model vysvětluje zkoumaná dat. Koeficient determinace je vyjádřen jako poměr vysvětlené variability k celkové variabilitě:

$$R^2 = \frac{\sum(y'_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} \qquad [4.2]$$

Výslednou hodnotu lze taktéž interpretovat v procentuálním zastoupení. Takový údaj pak udává z kolika procent jsou změny ve vysvětlované proměnné závislé na změnách vysvětlujících proměnných (MALÝ, 2015). Tab. 4 podává přibližnou interpretaci procentuálních hodnot koeficientu determinace.

Tab. 4: Stupnice těsnosti závislosti dle hodnocení koeficientu determinace

Koeficient determinace	Interpretace
R^2 = < 10 %	Těsnost nízká
10 % ≤ R^2 < 25 %	Těsnost mírná
25 % ≤ R^2 < 50 %	Těsnost význačná
50 % ≤ R^2 < 80 %	Těsnost velká
80 % ≤ R^2	Těsnost velmi vysoká

(Zdroj: HOŠKOVÁ, P., 2006)

4.2 Prostorové shlukování (Spatial clustering)

Při identifikace prostorového rozložení dat, v tomto případě onkologických, která jsou prostorově agregována, je podstatné zaměřit se také na lokální proměnlivost výskytu nemocnosti než pouze zkoumat její charakter v globálním měřítku. Při zkoumání prostorové diferenciace výskytu nemoci či její konkrétní charakteristiky jsou využitelné identifikátory prostorového shlukování. Prostorového uspořádání incidence či mortality nemoci je předmětem zkoumání prostorové epidemiologie (ŠTĚPÁNOVÁ, 2011).

Prostorová autokorelace je jedním z nejvýznamnějších způsobů zkoumání a vyjádření existence a míry prostorového shlukování. V kontextu zdravotních dat byla prostorová autokorelace zmíněna již ve 40. letech 20. století, kdy Cruickshank upozornil na přítomnost pozitivní prostorové autokorelace v relativní míře úmrtnosti na rakovinu v Anglii a Walesu (SPURNÁ, 2008). Zkoumání prostorové autokorelace a zejména identifikace prostorových shluků regionů se statisticky vyššími a nižšími úrovněmi studovaného jevu umožňuje zaměření výzkumu na konkrétní oblasti a odhalení lokálních faktorů, které mohou přispívat k vyšší či nižší úrovni jevu. Identifikace prostorových vzorů může přispět k porozumění procesům, které tyto prostorové vzory utvářejí.

Prostorová autokorelace je korelací mezi hodnotami jedné proměnné, který přímo odpovídá jejich vzájemné relativní poloze v rovině a představuje prostorovou obdobu tradičního statistického předpokladu odchylek od nezávislého pozorování (GRIFFITH, ARBIA, 2010 cit. podle MAREK, 2015). Jinými slovy prostorová autokorelace udává do jaké míry hodnoty atributu v určitém bodě souvisí či nesouvisí s hodnotami v bodech okolních. Jde tedy v podstatě o aplikaci prvního zákona geografie *„Všechno souvisí se vším, ale blízké věci spolu souvisí více než věci vzdálené.“* (TOBLER, 1970). Můžeme rozlišit pozitivní a negativní prostorovou autokorelaci. Pozitivní autokorelace reflektuje stav, kdy sousední či blízké regiony nesou podobné hodnoty studovaného jevu. Naopak negativní prostorová autokorelace popisuje situaci, kdy jsou hodnoty jevu sousedních nebo blízkých regionů velmi odlišné. Rozlišujeme globální a lokální míru prostorové autokorelace. Při výpočtu globální autokorelace dostáváme pouze jednu hodnotu pro celou studovanou oblast, která popisuje celkový prostorový vzor převládající ve studovaném území. Globální míry tedy udávají, zda dochází ve studované oblasti ke shlukování, není však možné přesně shluky lokalizovat. Pro tyto účely pak slouží lokální míry prostorové autokorelace, které identifikují konkrétní polohu a rozsah shluků (WALLER, GOTAWAY, 2004).

Podobně jako při dalších statistických metodách, mezi které se identifikace prostorového shlukování řadí, musí být na počátku stanovena nulová hypotéza (H_0), vůči které jsou data srovnávána. Nulová hypotéza tvrdí, že hodnoty ve studovaném území jsou rozmístěny náhodně a neexistuje zde tedy žádný prostorový vzor ani tendence ke shlukování. Tato hypotéza je následně testována oproti tzv. hypotéze alternativní (H_a), tedy že v prostoru existuje prostorový vzor, který není výsledkem náhodného procesu. Prezentace výsledku by vždy měla být doprovázena odhadem statistické významnosti identifikovaných shluků (p-value). Statistická významnost vypočtených hodnot, která zamítá nulovou hypotézu o neexistenci prostorové autokorelace může být testována například pomocí permutační procedury.

Permutace neboli podmíněná randomziace vytváří empirické hladiny významnosti. Randomizace je podmíněná v tom smyslu, že hodnota y_i je v bodě i fixní, zatímco zbývající proměnné jsou náhodně permutovány skrze geografické jednotky v datech. Jednoduchost metody spočívá ve snadnosti provedení, protože pro každou lokaci je třeba převzorkovat pouze tolik hodnot, kolik je v souboru sousedících jednotek (ANSELIN, 1995).

Jedny z nejčastěji používaných metod pro hodnocení globální i lokální autokorelace jsou *Moranovo I kritérium a Gearyho C koeficient*. Tyto nástroje jsou mimo jiné implementovány v programu GeoDa, který byl použit pro účely této práce. Pro hodnocení prostorového vzoru v poskytnutých datech bylo použito *Moranovo I kritérium* v jeho globální i lokální verzi. Více viz kapitola *6.3 Identifikace shluků v prostoru.*

4.2.1 Moranovo I kritérium (Morans‘ I)

Moranovo I kritérium je vhodným východiskem pro posouzení a vizualizaci geografických vzorů sledovaného jevu. Moranovo I kritérium je podobné Pearsonovu korelačnímu koeficientu, který měří statistickou závislost lineárních dat.

Moranovo I kritérium je charakterizováno následujícím vztahem:

$$I = \frac{n \sum_i \sum_j w_{ij}(z_i - \overline{z})(z_j - \overline{z})}{\sum_i \sum_j w_{ij} \sum_k (z_k - \overline{z})^2} \qquad [4.3]$$

n … $celkový\ počet\ oblastí$

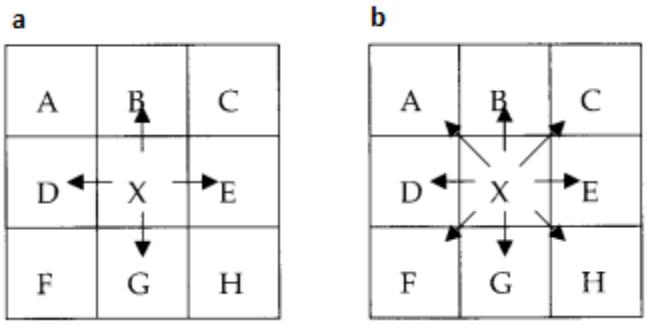
$z_{i,j}$ … $residuum(rozdíl\ mezi\ pozorovanou\ a\ očekávanou\ hodnotou)\ nebo\ hodnota\ proměnné\ v\ ploše\ i$, j

$\bar{z}$ … $průměrná\ hodnota\ proměnné$

w_{ij} … $míra\ blízkosti\ mezi\ místy\ i\ a\ j\ (prostorová\ vážící\ funkce)$

k … $celkový\ počet\ studovaných\ míst$

Nejvýznamnějším úkolem při výpočtu Moranova I kritéria je volba prostorové vážící funkce w_{ij}, respektive konstrukce matice prostorových vah *W* (*spatial weight matrix*). Matice sestává z řádků a sloupců, jež reprezentují každou prostorovou jednotku. Konkrétní hodnota buněk matice pak představuje prostorový vztah mezi studovanými prostorovými jednotkami. Tato hodnota přímo závisí na způsobu definování sousedství, viz Obr. 18 (WALLER, GOTWAY, 2004).



Obr. 18: Možné způsoby definování sousedství (a: Rook's case – věž, b: Queen's case – dáma), (Převzato z: DOBROVOLNÝ, 2015)

Výsledkem Moranova I kritéria je hodnota indexu, který charakterizuje, zda v prostoru existuje tendence ke shlukování, respektive definuje převažující prostorový vzor v daném území pro studovanou charakteristiku. Hodnota indexu nabývá hodnot z intervalu ⟨-1,1⟩, přičemž hodnota -1 vypovídá o silné negativní prostorové autokorelaci, hodnota 1 značí naopak silnou pozitivní autokorelaci. Obecně platí, že je-li hodnota indexu kladná, vypovádá to o existenci jednotek s podobnými hodnotami a naopak. Jsou-li vysoké hodnoty v regionu doprovázeny vysokými hodnotami (případně regiony s nízkými hodnotami sousedí s dalšími regiony s nízkými hodnotami jevu), můžeme hovořit o pozitivní prostorové autokorelaci neboli prostorovém shlukování (*spatial clustering*). Vyjde-li hodnota indexu nulová nebo blízká nule, znamená to, že se jedná o náhodné uspořádání. Tzn. že ve studované oblasti neexistuje žádný prostorový vzor (DOBROVOLNÝ, 2015).

Výsledkem *globálního Moranova I* kritéria je jedna hodnota indexu, která předpokládá jeho platnost pro celé území bez ohledu na lokální odlišnosti. Tento předpoklad je ale většinou velmi nepravděpodobný. Významnou autokorelaci může pozorovaný jev vykazovat pouze v určité oblasti, v jiné může být prostorové uspořádání náhodné. Může nastat také situace, kdy pozitivní prostorová autokorelace přechází v negativní (SPURNÁ, 2008). Z toho důvodu je třeba se zaměřit právě na lokální rozdíly a identifikovat lokální prostorové vzory. Pro tyto účely slouží *lokální Moranovo kritérium*. Toto kritérium bývá také označováno jako lokální indikátor prostorové autokorelace (*Local indicator of Spatial Autocorrelation - LISA*). S jeho pomocí je možné odlišit shluky podobných hodnot (clusters) od skupin hodnot nepodobných (outliers). Navíc umožňuje určit přesnou lokalizaci těchto shluků na základě definované matice vah (MAREK, 2015).

Lokální Moranovo I kritérium lze vyjádřit vztahem:

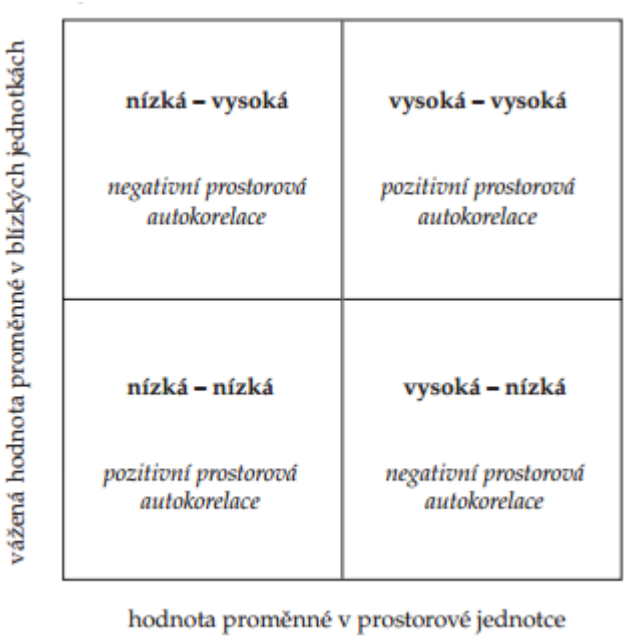
$$I_i = r_i \sum_j (w_{ij} r_j) \qquad [4.4]$$

r_i, r_j … *standardizované hodnoty veličiny z*

w_{ij} … *matice vah*

Primární výsledky analýzy prostorové autokorelace je možno vyjádřit pomocí Moranova diagramu, který je vyobrazen na Obr. 19. Tento diagram zobrazuje závislost původních hodnot proměnné (horizontální osa) na vypočtených průměrných hodnotách ze sousedních prostorových jednotek (vertikální osa). Výsledný sklon proložené regresní přímky odpovídá hodnotě Moranova I kritéria (SPURNÁ, 2008).

Výsledky výpočtu LISA je možno rozdělit do čtyř základních kategorií podle typu prostorové autokorelace. Tyto typy pak odpovídají čtyřem kvadrantům Moranova diagramu. Můžeme rozlišit prostorové shluky s výrazně vyššími nebo nižšími hodnotami sledované proměnné, které odpovídají hodnotám v okolních jednotkách. Takovéto jednotky jsou v diagramu situovány v pravém horním (vysoká-vysoká = shluky vysokých hodnot proměnné, hot spots) a levém dolním kvadrantu (nízká-nízká = shluky nízkých hodnot proměnné, cold spots). Nepodobné hodnoty neboli prostorové odchylky s výrazně vyšší/nižší hodnotou proměnné než ve svém okolí se pak v diagramu nachází v levém horním a pravém dolním kvadrantu.



Obr. 19: Moranův diagram (zdroj: SPURNÁ, 2008)

Přínos lokální varianty Moranova I kritéria spočívá v možnosti přesně lokalizovat regiony s nadprůměrnými či podprůměrnými hodnotami studované proměnné. Analýza navíc výsledky doplňuje hodnocením statistické významnosti, díky čemuž je možné eliminovat nevýznamné výsledky a předejít tak chybné interpretaci výsledků. Zároveň je díky tomu možné se zaměřit pouze na statisticky významné výsledky a pominout oblasti nevýznamné a koncentrovat se na hledání a studium příčin vzniku prostorových shluků nebo odlehlých hodnot.

Z výše uvedeného vyplývá, že analýza LISA vhodně doplňuje výpočet globálního Moranova I kritéria.

4.2.2 Hodnocení významnosti výsledků na základě p-value

Místo porovnání hodnoty testovacího kritéria s kritickými hodnotami lze pro rozhodnutí o platnosti či neplatnosti nulové hypotézy použít i tzv. *p-hodnotu (neboli p-value).*

Hodnota p-value poskytuje více informací, o výsledku statistického testování než je pouhé zamítnutí nebo nezamítnutí nulové hypotézy. Zhodnocení její velikosti říká, zda je výsledek testování významný nebo naopak.

P-value vyjadřuje pravděpodobnost za platnosti H_0 (nulová hypotéza – tedy že je rozdělení náhodné), s níž bychom získali stejnou nebo extrémnější (ještě méně pravděpodobnou) hodnotu testové statistiky. Platí tedy, že čím nižší p-value testu je, tím menší tento test indikuje pravděpodobnost, že platí nulová hypotéza. (HOLČÍK, KOMENDA a kol., 2015).

P-value může nabývat hodnot od 0 do 1 a je interpretována následujícím způsobem:

- Malá hodnota (typicky $p < 0,05$) indikuje silný důkaz pro zamítnutí nulové hypotézy, můžeme tedy H_0 odmítnout
- Vysoká hodnota ($p > 0,05$) indikuje slabý důkaz pro zamítnutí nulové hypotézy, takže není možné H_0 odmítnout
- Hodnota p-value blízká mezní hodnotě 0,05, je považována za hraniční (může být interpretována oběma způsoby), (RUMSEY, 2011).

Podstata hodnoty p-value má velmi podobný charakter jako hladina významnosti α. Hladina významnosti definuje pravděpodobnost, že bude nulová hypotéza zamítnuta i přesto že platí. Obecně mohou při testování statistických hypotéz nastat dva typy chyb:

- *Chyba 1. druhu α – nulová hypotéza je zamítnuta, přestože platí*
- *Chyba 2. druhu β – nulová hypotéza je přijata přestože neplatí*

Hladina významnosti testu, tedy pravděpodobnost chyby prvního druhu α je volena blízká nule, nejčastěji 0,05 nebo 0,01. Někdy také bývá uváděna v procentech (tedy 5 % nebo 1 % (BEDNÁŘ, 2006). Provedeme-li test na hladině významnosti 0,05,

znamená to, že pravděpodobnost, že nesprávně přijmeme alternativní hypotézu, je 5 %, (jinak řečeno: nulovou hypotézu odmítáme s jistotou 95 %), (CHRÁSKA, 2000). Výsledek testu je pak významný na zvolené hladině α. Blíže se tématu hladin významnosti a statistického testování obecně věnuje v e-learningové učebnici HOLČÍK, KOMENDA a kol., 2015.

4.3 Analýza hlavních komponent

Analýza hlavních komponent (*principal component analysis* – PCA) je jednou z nejčastějších metod redukce dimenzionality dat. Jejím cílem je vysvětlit celkový rozptyl vektoru náhodných veličin, respektive jeho podstatnou část pomocí méně veličin. Tato metoda byla poprvé zavedena Pearsonem již v roce 1901 a nezávisle Hotellingem v roce 1933 (MELOUN, 2011). BÁČOVÁ a kol., (2013) uvádí, že využití analýzy hlavních komponent je dnes nejčastější především v oblasti zpracování satelitních snímků v dálkovém průzkumu země, kde je používán pro omezení pásmové dimenzionality zpracovávaného obrazu. Další oblastí, kde je její využití známo je kriminalistika. Zde je PCA využívána pro rozpoznávání obličejů. Dále tvrdí, že je její využití vhodné také v kartografii pro detekci kauzalit mnoharozměrných dat a pro předpověď jejich vývoje. Nejpodstatnější přednost analýzy pro obor kartografie spočívá v možnosti redukce datového souboru pro následnou kartografickou vizualizaci na dvě až tři proměnné bez ztráty podstatných informací.

V analýze hlavních komponent nejsou znaky děleny na závislé a nezávislé. Vypočtené hlavní komponenty, neboli latentní proměnné vznikají lineární transformací původních znaků na nové. Každá z komponent má určitou míru variability neboli rozptyl. Hlavní komponenty jsou následně řazeny podle klesajícího rozptylu. Většina variability původních dat je soustředěna v první komponentě, nejméně variability je pak obsaženo v komponentě poslední (MELOUN, 2011).

Původní prostor o n dimenzích je definován osami, které odpovídají původním proměnným. PCA je matematicky definována jako ortogonální lineární transformace dat z původního do nového souřadnicového systému, jehož osy jsou tvořeny hlavními komponentami. Osy procházejí ve směru maximálního rozptylu dat, protože podmínka nezávislosti komponent vede ke kolmosti os (LUKÁŠOVÁ 2012).

Dle MELOUNA, (2011) je základním cílem PCA transformace původních znaků x_i, $j=1, \ldots, m$ do menšího počtu latentních proměnných y_j. Tyto proměnné mají vhodnější vlastnosti než původní dataset – je jich méně, vystihují téměř celou proměnlivost původních znaků a jsou vzájemně nekorelované.

Rozdíl mezi souřadnicemi objektů v původních znacích a v hlavních komponentách se nazývá *mírou těsnosti proložení modelu PCA* nebo také *chybou modelu PCA*. PCA transformuje původní matici do nového systému os a snižuje rozměrnost dat nahrazením několika hlavními komponentami, které vystihují strukturu v datech.

Při PCA je zdrojová matice X_C rozložena na matici komponentních skóre T ($n \times k$) a matici komponentních zátěží P_T ($k \times m$). Původní zdrojová matice X je rozdělena na část *struktury* - TP^T (první hlavní komponenty – obvykle první tři) a část *šumu* - E(ostatní hlavní komponenty). Model hlavních komponent má tedy tvar:

$$X = TP^T + E \qquad [4.5]$$

T je v tomto případě matice komponentního skóre, PT je transponovaná matice komponentních vah, E je matice reziduí, která není objasněna modelem hlavních komponent. Matice E souvisí s „těsností proložení“ a ukazuje, jak dobře jsou objekty proloženy modelem hlavních komponent. Zdrojovou matici je však třeba nejdříve vycentrovat.

Rovnici [4.5] lze dle MELOUNA, (2011) rozepsat následovně:

$$X = t_1 p_1^T + t_2 p_2^T + \cdots + t_A p_A^T + E \qquad [4.6]$$

Každý sčítanec $t_1 p_1^T$ je matice rozměru $n \times m$, která má hodnost 1. Výpočet probíhá v několika krocích:

1. Nejprve je vypočteno t_1 a p_1 z X například s využitím SVD (singular value decomposition - rozbití matice na jednodušší smysluplné části) a vyčíslí se $E_1 = X - t_1 p_1^T$.
2. Dále je vypočteno t_2 a p_2 z E_1 a vyčíslí se $E_2 = E_1 - t_2 p_2^T$.
3. Dále je vypočteno t_3 a p_3 z E_2 a vyčíslí se $E_3 = E_2 - t_2 p_3^T$.

Takto se pokračuje tak dlouho, až je vyčíslen dostatečný počet A hlavních komponent. Hledáme optimální počet hlavních komponent, abychom dosáhli nejlepšího proložení a matice byla co nejmenší. Čím nižší hodnota E tím lepšího modelu jsme dosáhli. (LUKÁŠOVÁ, 2012). Hlavní komponenty jsou vycentrované. To znamená, že mají společný počátek, který odpovídá těžišti celého shluku objektů.

Každá hlavní komponenta představuje lineární kombinaci všech m vektorů v prostoru znaků. Lineární kombinace v každé hlavní komponentě obsahuje m koeficientů p_{ka}, kde k je index m-tého znaku a a je index směru hlavní komponenty. Tyto komponenty představují *komponentní váhy*. Váhy pro všechny hlavní komponenty tvoří matici P, která je transformační maticí, která převádí původní znaky zdrojové matice X do nových latentních proměnných. Váhy podávají informaci o vztahu mezi původními znaky m a hlavními komponentami

Komponentní skóre představuje souřadnice každého objektu na osách hlavních komponent. Každý objekt má svůj soubor komponentních skóre pro každou z komponent.

Vlivem aproximace původního souboru dat dochází k určité ztrátě informace, která je představována velikostí projekční vzdálenosti e_i, neboli rezidua. Všechna rezidua jsou obsažena v matici reziduí E. Velikost reziduí je spojena s vhodností použitého modelu. Příliš velká rezidua značí, že model vhodně nepopisuje data, malá rezidua naopak značí, že je použitý model vhodný (MELOUN, 2011).

Výsledek analýzy hlavních komponent je možné mimo matematického odvození zobrazit také v grafické podobě:

a) Scree plot (Cattelův indexový graf úpatí vlastních čísel)

Scree plot je sloupcovým nebo liniovým diagramem vlastních čísel nebo reziduálního rozptylu proti stoupající hodnotě indexu, pořadového čísla A. Graf zobrazuje relativní velikost jednotlivých vlastních čísel. Cattel definuje „scree“ jako zlomové místo mezi „kolmou stěnou“ a „vodorovným dnem“. Vybrané užitečné hlavní komponenty tvoří „stěnu“ a neužitečné představují „vodorovné dno“. Užitečné komponenty jsou tak zřetelně odděleny zřetelným zlomovým místem.

b) Graf komponentních vah, zátěží (Plot Components Weights)

Graf komponentních vah zobrazí komponentní váhy pro první dvě hlavní komponenty. V grafu jsou porovnávány vzdálenosti mezi proměnnými. Krátká vzdálenost mezi dvěma proměnnými znamená korelaci. Tento graf ukazuje, jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent.

c) Rozptylový diagram komponentního skóre (Scatterplot)

Rozptylový diagram komponentního skóre zobrazuje komponentní skóre neboli hodnoty většinou prvních dvou hlavních komponent u všech objektů. Diagram se používá k identifikaci odlehlých objektů, identifikaci trendů, tříd, shluků objektů, k objasnění podobnosti objektů atd. Není možné analyzovat všechny možné rozptylové diagramy, protože jich je velmi mnoho. Obvykle se volí první hlavní komponenta (jelikož obsahuje největší míru proměnlivosti v datech) a kombinuje se s druhou, třetí, čtvrtou a dalšími komponentami.

d) Dvojný graf (Biplot)

Dvojný graf kombinuje předchozí dva grafy. Úhel mezi průvodiči dvou znaků x_j a x_k je nepřímo úměrný velikosti korelace mezi těmito dvěma znaky. Čím je menší úhel, tím je větší korelace. Každý průvodič má své souřadnice na první a na druhé hlavní komponentě. Délka této souřadnice je úměrná příspěvku původního znaku x_j do hlavní komponenty, čili je úměrná komponentní váze.

e) Graf reziduí jednotlivých objektů

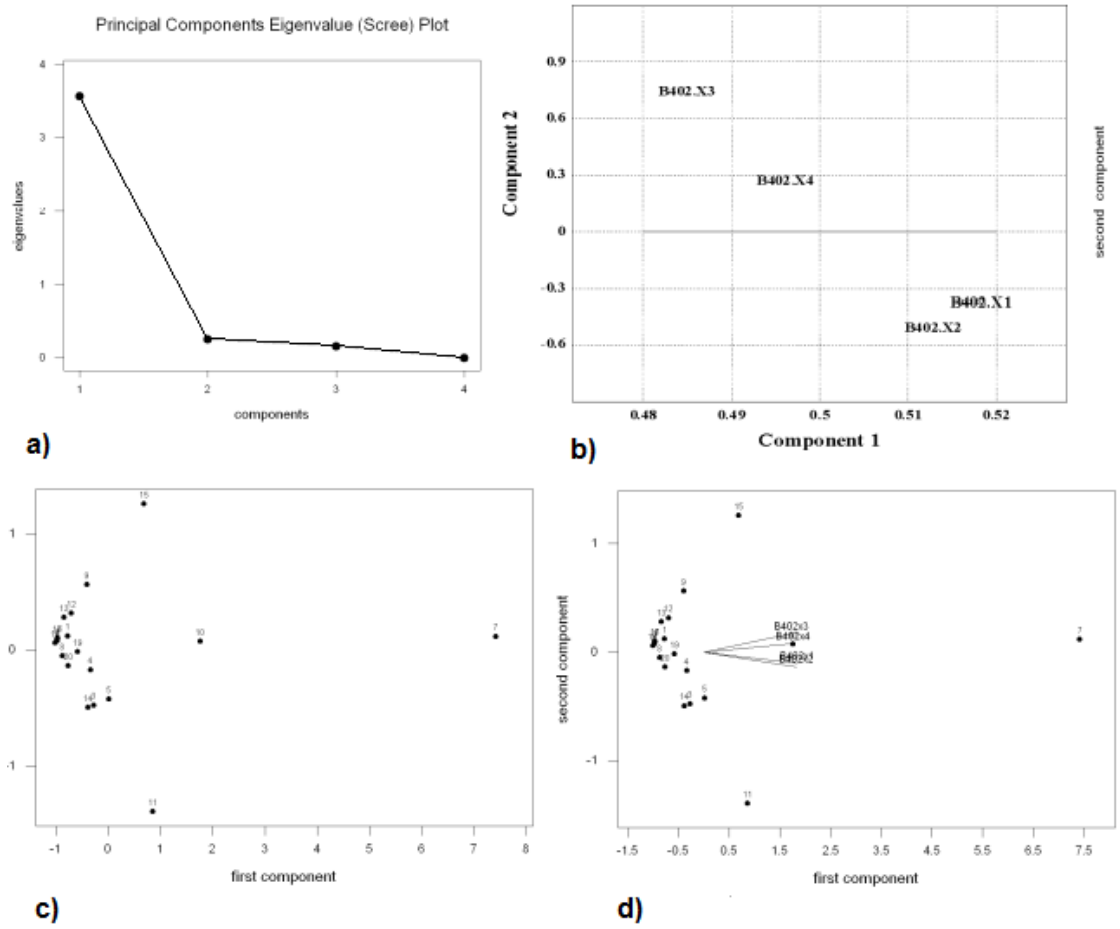
Rozptyl reziduí jednoho i-tého objektu představuje vzdálenost mezi tímto objektem a modelem. Tento graf je zvláště výhodný, potřebujeme-li porovnat rezidua jednotlivých objektů mezi sebou.

f) Graf celkového reziduálního rozptylu všech objektů

Rozptylová matice E_1 poskytne novou hodnotu celkového rozptylu reziduí $e_{tot,1}^2$, která musí být menší než $e_{tot,0}^2$. Postupným přidáváním dalších hlavních komponent se hodnota celkového rozptylu reziduí $e_{tot,i}^2$ bude zmenšovat tak, že každá další hodnota rozptylu reziduí bude menší než předešlá. Vynesením hodnot $e_{tot,i}^2$ proti počtu použitých hlavních komponent, získáme indexový graf úpatí celkového rozptylu všech objektů.

Graf celkového reziduálního rozptylu je obdobou grafu úpatí vlastních čísel a nalezení zlomu na sestupné křivce čili úpatí analogicky vystihuje optimální počet využitelných hlavních komponent.

Příklady grafů a) – d) jsou na Obr. 20.



Obr. 20: Možnosti grafického vyjádření výsledků analýzy hlavních komponent: a) scree plot, b) graf komponentních vah, c) rozptylový diagram komponentního skóre, d) dvojný graf (Zdroj: MELOUN, MILITKÝ, 2002)

Přesné matematické odvození analýzy hlavních komponent a další podrobnosti k analýze poskytuje monografie MELOUN, MILITKÝ, (2002).

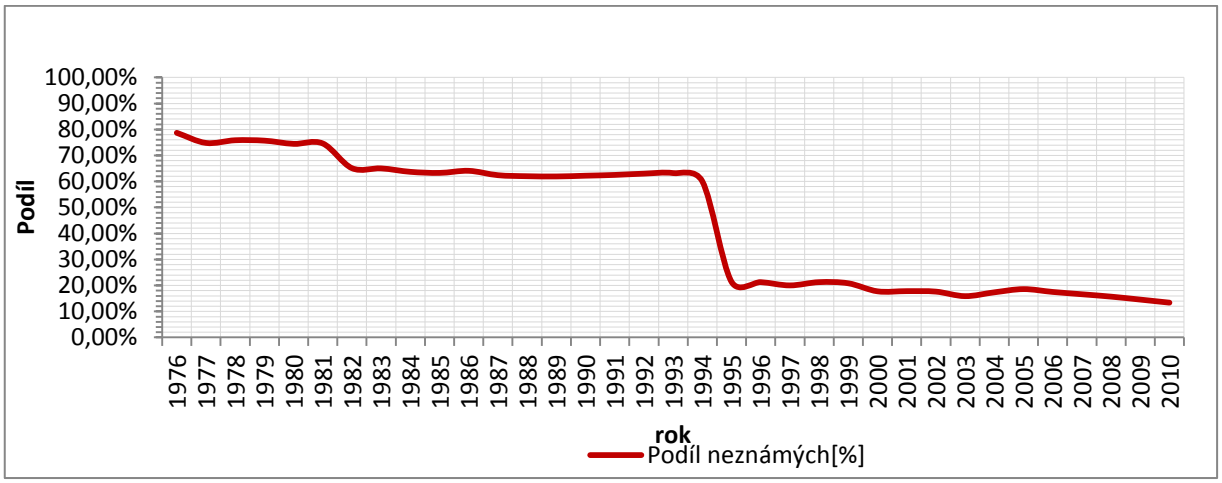
5 ANALÝZA NEZNÁMÝCH STÁDIÍ RAKOVINY

Analýze stádií rakoviny byla podrobena data z databáze NOR (ÚZIS) o počtech vícečetných novotvarů v okresech ČR v letech 1976-2010. V analyzovaných datech bylo zjištěno velké množství diagnóz v neznámém stádiu. To znamená, že lékařem nebylo definováno přesné stádium rakoviny. V datech bylo toto stádium označeno jako 0 a 9. Tato data byla pro analýzu vybrána z toho důvodu, že prozatím nebyla nijak zkoumána ani analyzována. Cílem bylo odhalit bližší souvislosti a poukázat na možné příčiny jejich četného výskytu v onkologických datech.

Byly stanoveny dvě možné teorie, které by mohly podat vysvětlení této skutečnosti. Teorie byly navrženy epidemiologem MUDr. Edvardem Gerykem z Fakultní nemocnice Brno-Bohunice.

5.1 Analýza neznámých stádií v čase

První teorií bylo, že se tento stav se výrazně zhoršil po roce 1994, kdy se změnily předpisy pro hlášení nových diagnóz. Tato teorie byla ověřena vypočtením podílů neznámých stádií ze všech diagnóz v každém roce před a po roce 1994. Výsledky dokumentuje Obr. 21.

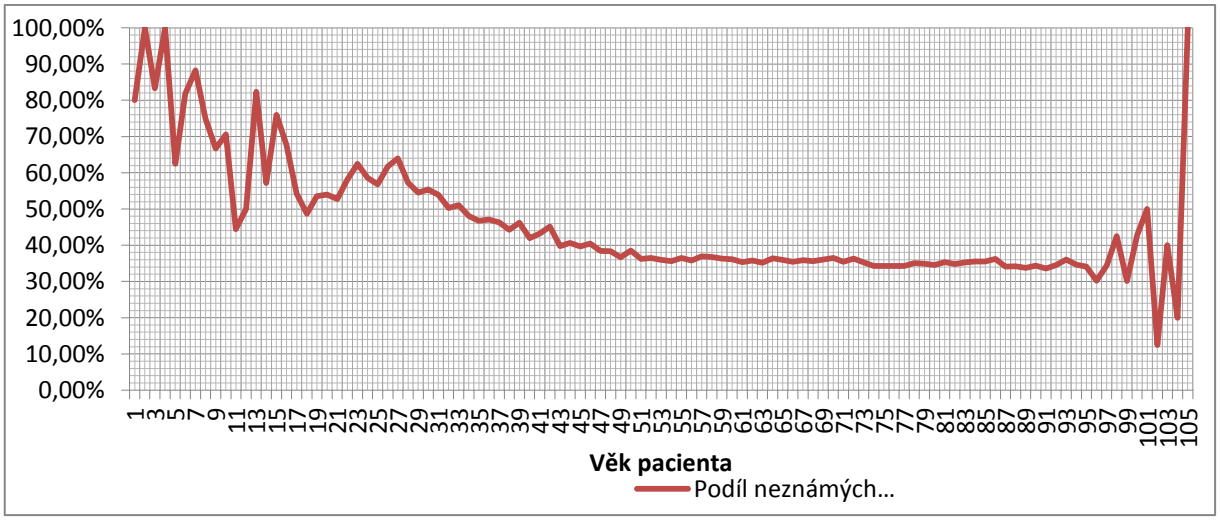


Obr. 21: Vývoj podílu neznámých stádií rakoviny v letech 1976-2010 [%]

Jak můžeme vyvodit z výše uvedeného grafu, první teorie, tedy že došlo ke zhoršení stavu (zlomovému nárůstu podílu neznámých stádií rakoviny po roce 1994) se nepotvrdila. Naopak si můžeme všimnout, že po roce 1994 došlo k výraznému poklesu podílu neznámých stádií.

Druhou teorií, která by mohla vysvětlovat vysoký podíl neznámých stádií v datech je, že neochota lékařů důkladně zjišťovat diagnózu pacienta roste s jeho věkem, "protože

již nemá význam to zjišťovat". Tato teorie byla ověřena vypočtením podílu neznámých stádií ze všech diagnóz v jednotlivých věkových kategoriích pacientů (viz Obr. 22).



Obr. 22: Podíl neznámých stádií rakoviny podle věku pacienta [%]

Ani v tomto případě nelze říci, že byla původní hypotéza správná. Přestože lze v grafu pozorovat jistý nárůst a zároveň výrazné kolísání podílu v nejvyšších věkových kategoriích, vysvětlení nalezneme spíše v mnohem nižším celkovém počtu diagnóz v daných věkových kategoriích. Podobně je tomu i v případě nejnižších věkových kategorií. Jinak je ale podíl neznámých stádií nemoci přibližně stabilní a pohybuje se přibližně kolem jedné třetiny.

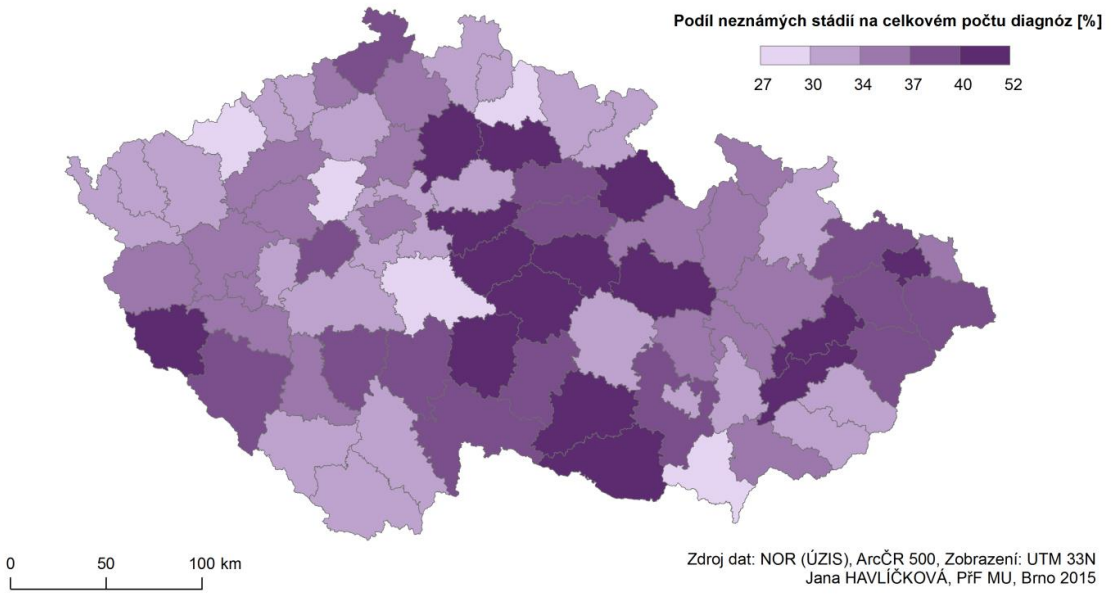
Na základě těchto výsledků lze říci, že se ani jedna z teorií nepotvrdila. Jelikož se nepodařilo získat zasvěcenou odpověď ze strany odborné lékařské veřejnosti, není možné usuzovat na možné příčiny vysokého podílu neznámých stádií rakoviny v datech.

Jak uvádí GERYK a kol., (2010) neznámá stádia převažovala nad ostatními stádii až do roku 1994. „Příčinou chybění klinického stádia může být zjištění nádoru při pitvě, náhlé úmrtí pacienta, kontraindikace léčby nebo její odmítnutí, část případů byla způsobena neúplným nahlášením onemocnění do registru.“ (GERYK a kol., 2010).

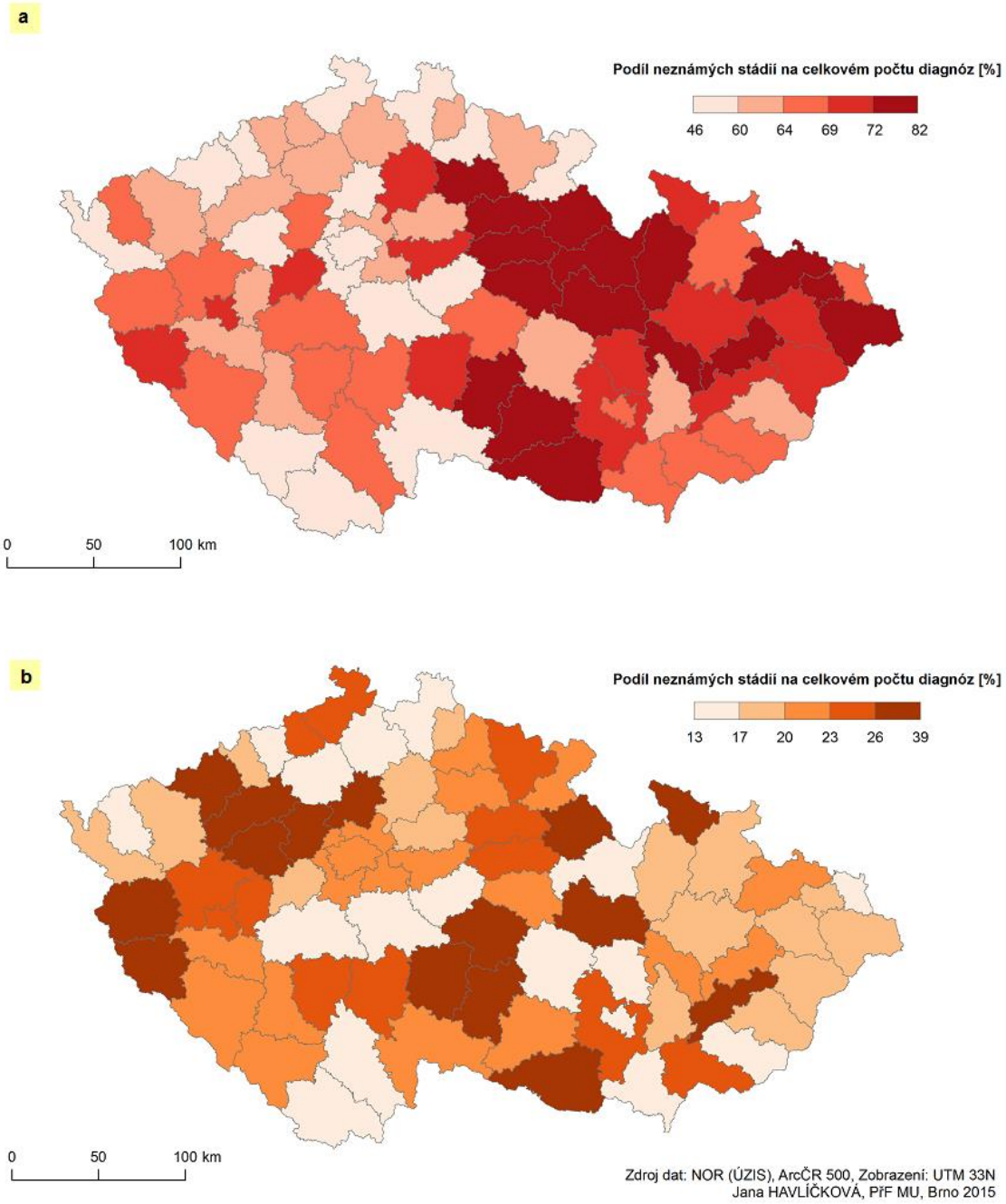
5.2 Analýza neznámých stádií v prostoru

Jelikož poskytnutá data sahají až na úroveň okresů, bylo dalším krokem zpracování zjistit, zda se liší podíl neznámých stádií rakoviny k celkovému počtu diagnóz v rámci okresů České republiky a zda se zde vyskytují okresy s výrazně vyšším podílem neznámých stádií. Dle GERYKA a kol., (2010) byl největší podíl neznámých stádií u 125 262 nemocných s primárními novotvary v letech 1976 - 2005 v Královéhradeckém kraji a Kraji Vysočina (55,6 %), dále v kraji Moravskoslezském (53,6 %), Jihomoravském (53 %) a Pardubickém (51,5 %).

Dle GERYKA a kol., (2010) zkreslovaly v letech 1976 – 1994 vysoké počty neznámých stádií diagnóz celkové počty nahlášených vícečetných diagnóz. Za zlomový rok byl tedy označen rok 1994. Pro prvotní přiblížení situace může velmi dobře sloužit nepravý kartogram, viz Obr. 23. Metodou prezentace byl i v tomto případě zvolen jednoduchý homogenní pseudokartogram. Hodnoty jsou rozděleny do pěti intervalů na základě výpočtu kvantilů. Vysoký podíl diagnóz v neznámém stádiu za celé sledované období byl pozorován v okresech na pomezí Pardubického kraje, Kraje Vysočina a Středočeského kraje (okresy Kolín, Kutná Hora, Svitavy, Chrudim, Havlíčkův Brod, Pelhřimov. Dále byl vysoký podíl zaznamenán v okrese Mladá Boleslav, Jičín, Rychnov nad Kněžnou, Ostrava-město, Přerov, Kroměříž, Znojmo, Třebíč, a Domažlice. Porovnání situace před rokem 1994 a po něm zobrazuje Obr. 24. Na základě vizuální analýzy obou mapových oken lze obecně konstatovat, že ve všech okresech došlo po roce 1994 k významnému poklesu podílu diagnóz v neznámém stádiu k celkovému počtu diagnóz. V letech 1976 – 1994 byl pozorován vysoký podíl diagnóz v neznámém stádiu v celém Pardubickém a v části Královéhradeckého kraje. Po roce 1994 právě tyto okresy zaznamenaly nejvýznamnější pokles podílu neznámých diagnóz. Obecně by se dalo říci, že před rokem 1994 byl podíl neznámých vyšší ve východní části České republiky. Zde také byla změna po roce 1994 nejvýznamnější. Méně významný pokles podílu neznámých zaznamenaly okresy ve Středočeském a Ústeckém kraji. V případě okresů Plzeňského kraje došlo k nejméně výrazné změně a v období let 1995-2010 se tak dostává na relativně nejhorší pozici.



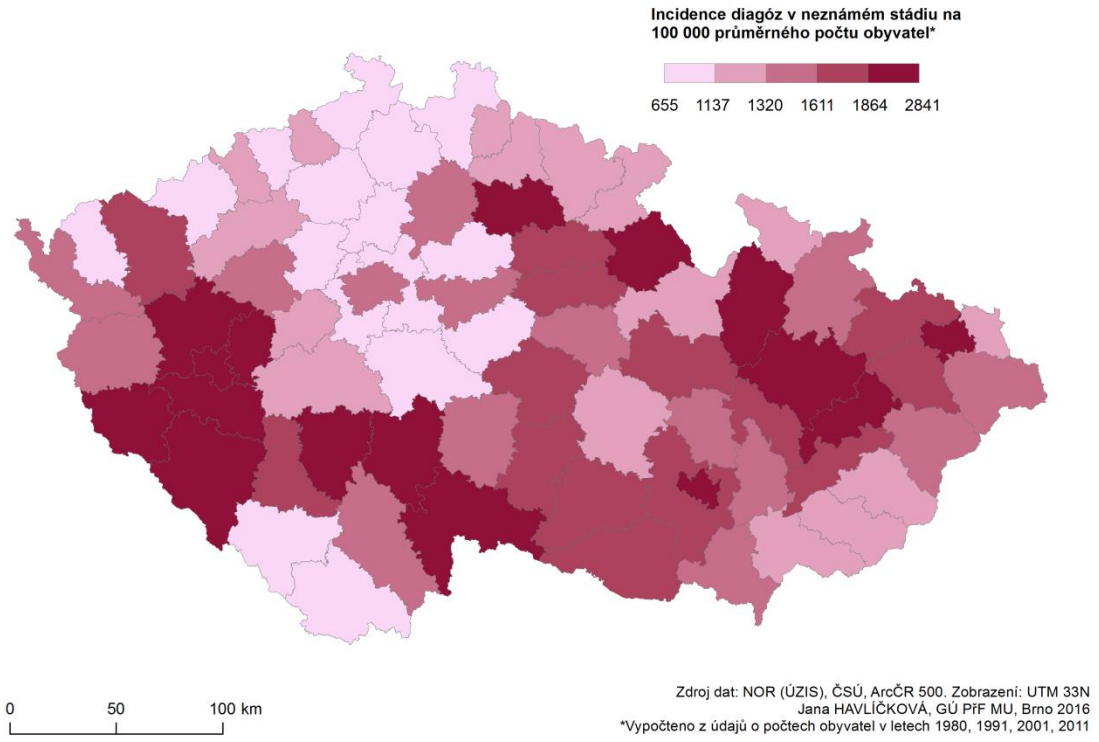
Obr. 23: Prostorová diferenciace podílů diagnóz v neznámém stádiu k celkovému počtu diagnóz v okresech České republiky v letech 1976-2010



Obr. 24: Porovnání prostorové diferenciace podílů diagnóz v neznámém stádiu k celkovému počtu diagnóz v okresech České republiky v obdobích let 1976-1994 (*mapové okno a*) a 1995-2010 (*mapové okno b*) metodou nepravého kartogramu

Na Obr. 25 je vyobrazena incidence počtu diagnóz v neznámém stádiu v okresech ČR v letech 1976-2010. Incidence je zde prezentována jako počet diagnóz v neznámém stádiu na 100 000 průměrného počtu obyvatel. Metodou prezentace byl zvolen jednoduchý homogenní pseudokartogram. Výsledné hodnoty jsou opět rozděleny do pěti

intervalů na základě výpočtu kvantilů. Touto metodou byly zaznamenány vysoké hodnoty incidence v okresech Plzeňského, Jihočeského a Olomouckého kraje. Oblastmi s nízkými hodnotami incidence byly identifikovány především okresy ve Středočeském a Libereckém kraji a dále dva okresy v kraji Jihočeském a po jednom okrese v kraji Ústeckém a Karlovarském.



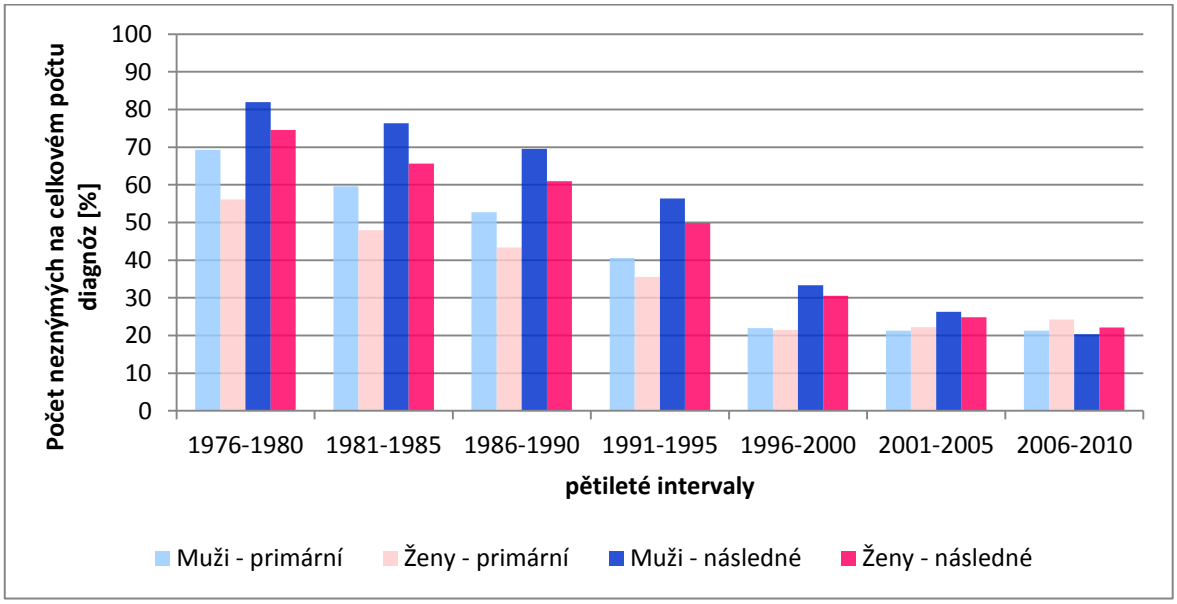
Obr. 25: Incidence diagnóz v neznámém stádiu v okresech České republiky v letech 1976-2010 v počtech případů na 100 000 průměrného počtu obyvatel

5.3 Analýza neznámých stádií na základě dalších kritérií - pohlaví, diagnóz, věku ad.

Předchozí podkapitoly měly poskytnout základní vhled do problematiky novotvarů s označením klinického stádia jako neznámé. Tato problematika bude v následujícím textu zkoumána a rozvíjena dále. Použité metody zjistily významný výskyt diagnóz v neznámém stádiu ve všech okresech a jsou tedy dostatečně reprezentativním vzorkem pro následné analýzy. Bylo zjištěno, že z celkového počtu 443 403 diagnóz VZN bylo 159 261 diagnóz v neznámém klinickém stádiu. To znamená, že se na celkovém počtu diagnóz podílejí 35,92 %. Zvolené metody však neukázaly bližší souvislosti. Cílem této podkapitoly je zjistit, zda a jak se počet diagnóz v neznámém klinickém stádiu liší mezi pohlavími, v jednotlivých letech, zda jsou u některých diagnóz neznámá stádia uváděna častěji atd.

Nejprve bylo zjišťováno, jak se lišily podíly neznámých stádií mezi muži a ženami v letech 1976-2010 a zda je rozdíl mezi primárními a následnými diagnózami. Výsledkem je graf na Obr. 26, který zobrazuje vývoj počtu primárních a následných diagnóz, jejichž stádium je označeno jako neznámé, u mužů a žen.

Obecně lze říci, že podíl diagnóz v neznámém stádiu na celkovém počtu diagnóz po většinu studovaného období převažoval u mužů a to v případech primárních i následných diagnóz. V posledním pětiletém období však již v podílu diagnóz v neznámém stádiu převládají ženy nad muži (jak u primárních, tak i následných případů). Zaznamenat lze také patrný klesající trend. Jak v případě primárních, tak i následných diagnóz v neznámém stádiu je patrný pokles podílu na celkovém počtu případů. Nejvýznamnější pokles nastal v letech 1996-2000. Tato skutečnost tak opět potvrzuje předchozí tvrzení, že po roce 1994 došlo ke změně v hlášení nových diagnóz a že došlo k významnému poklesu podílu neznámých klinických stádií diagnóz. Vyšší podíl neznámých stádií lze po většinu sledovaného období pozorovat v případě následných diagnóz u mužů i žen. To by mohlo poukazovat na skutečnost, že mohl být následný novotvar diagnostikován pozdě nebo byl nalezen až při pitvě. Bez zasvěceného názoru odborníka však tyto domněnky zůstávají pouze nepotvrzenou teorií. V posledních dvou pětiletých obdobích jsou již tyto podíly téměř vyrovnané. Pokles podílu neznámých u následných novotvarů v posledním sledovaném pětiletém období je však spíše než poklesu jejich počtu připisováno skutečnosti, že se další následné novotvary „nestihly“ u pacientů vyvinout. Jak uvádí GERYK a kol., (2008) a jak již bylo výše zmíněno, doba mezi vyvinutím následného novotvaru činí 6 let u mužů a 6,6 let u žen.



Obr. 26: Vývoj podílu primárních a následných diagnóz v neznámém stádiu u mužů a žen na celkovém počtu diagnóz pro pětileté intervaly v letech 1976-2010 (Zdroj dat: NOR (ÚZIS))

Počet neznámých stádií může být kromě pohlaví a roku diagnózy podmíněn také jejím typem. Je tedy pravděpodobné, že se podíl diagnóz v neznámém stádiu bude lišit mezi různými typy diagnóz. Tato teorie byla nejprve testována pro celou datovou sadu, tedy pro všechny typy diagnóz C00-D48. Nutno podotknout, že v databázi nejsou obsažena data pro diagnózy nezhoubných novotvarů D10-D36. Výsledky dokumentuje Tab. 5. Jednoznačně nevyšší podíl neznámých stádií vykazují diagnózy in situ (D00-D09), tedy tzv. neinvazivní novotvary v počátečním stádiu. Podíl neznámých stádií dosahuje téměř 100 %. Následují ZN nepřesně určených sekundárních a neurčených lokalizací a novotvary nejistého nebo neznámého chování, u kterých podíl neznámých přesahuje 90 %. Naopak nejnižší podíly neznámých stádií vykazují ZN prsu, ženských pohlavních orgánů, rtu, dutiny ústní a hltanu, dýchací soustavy, a trávicího ústrojí.

Dále byly testovány samostatně jednotlivé skupiny ZN (C00-C97). Podíly neznámých stádií v jednotlivých typech diagnóz shrnuje Příloha 1. První dvě pozice zaujímají diagnózy, u kterých bylo celých 100 % případů v neznámém klinickém stádiu. Jedná se o nepřesně určené ZN v dýchací soustavě a monocytickou leukémii. Také další typy leukémie se vyskytují na čelních pozicích tabulky. U 21 diagnóz přesahuje podíl neznámých stádií 90 % všech případů. Naopak diagnózou s nejnižším podílem případů v neznámém stádiu jsou ZN prsu. To může souviset s tím, že odhalení přesného stádia nádoru oka, mozku, mízní tkáně, kosti, chrupavky a dalších je pravděpodobně mnohem obtížnější než například odhalení klinického stádia nádoru prsu.

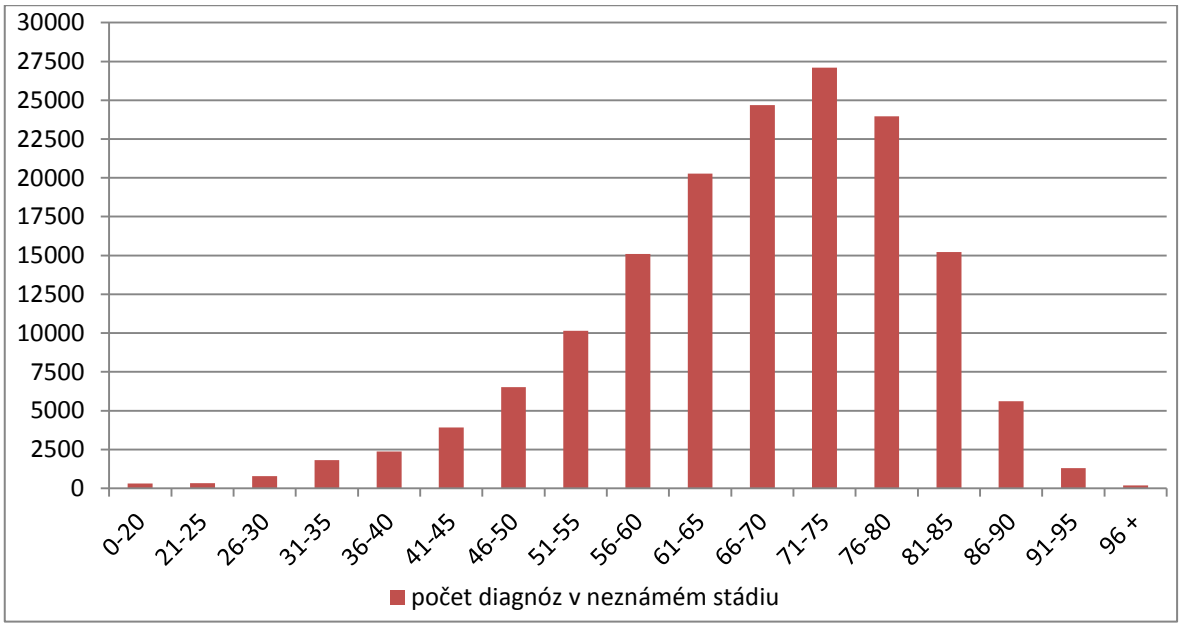
Tab. 5: Podíl diagnóz v neznámém stádiu na celkovém počtu diagnóz z hlediska jednotlivých diagnostických skupin dle MKN-10 v letech 1976-2010

Kód	Neznámé stádium	Celkem	Podíl [%]
D00-D09	15881	15920	99,76
C76-C80	4220	4379	96,37
D37-D48	6881	7148	96,26
C97	14	16	87,50
C69-C72	1880	2254	83,41
C81-C96	10623	14431	73,61
C40-C41	278	416	66,83
C45-C49	1137	2029	56,04
C73-C75	1310	2796	46,85
C60-C63	8309	20090	41,36
C64-C68	12376	30883	40,07
C43-C44	62178	197668	31,46
C15-C26	17887	60375	29,63
C30-C39	6543	26628	24,57
C00-C14	1640	6923	23,69
C51-C58	5086	22109	23,00
C50	3017	29337	10,28

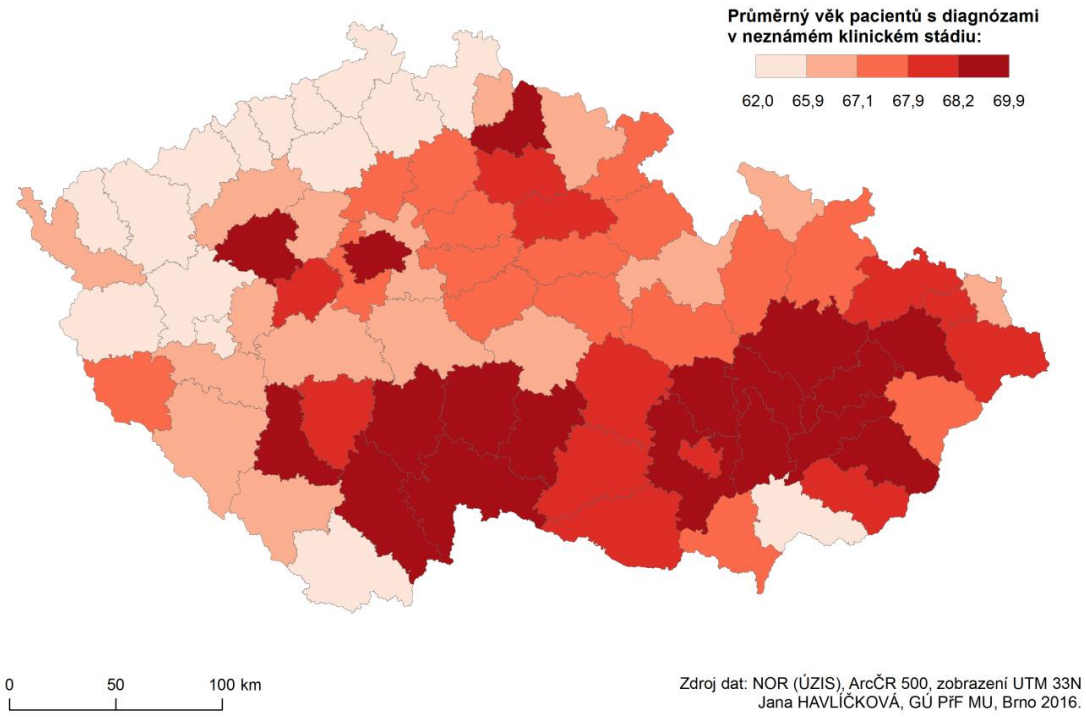
Pozn.: Vysvětlení kódů poskytuje Tab. 1 (Zdroj dat: NOR (ÚZIS))

V kapitole *5.1 Analýza neznámých stádií v čase* bylo prokázáno, že se podíl neznámých stádií rakoviny v jednotlivých věkových kategoriích pohybuje kolem 30 %. V nejnižších a nejvyšších věkových kategoriích se objevují jisté výkyvy, které jsou však spojeny s nízkým počtem diagnóz v dané kategorii. S ohledem na vyšší počty diagnotiskovaných novotvarů ve vyšších věkových kategoriích lze předpokládat, že i počty novotvarů v neznámém stádiu budou dosahovat vyšších hodnot u starších osob. Tuto myšlenku potvrzuje graf na Obr. 27 zobrazující věk osob s novotvary v neznámém stádiu. Ve věku do 30 let jsou počty novotvarů v neznámém stádiu poměrně nízké. Stejně tak je tomu u osob starších 96 let.Počty neznámých stádií plynule rostou až po jejich kulminaci u osob mezi 71 a 75 lety. Nejmladšími osobami s diagnózami v neznámém stádiu bylo osm dětí ve věku do jednoho roku (3 chlapci a 5 děvčat), nejstarší pak byla jedna žena ve věku 104 let. Průměrný věk osob s neznámými diagnózami je za celé sledované období 67,17 let.

V kontextu prostorové distribuce průměrného věku pacientů s diagnózou v neznámém stádiu v rámci okresů ČR, kterou zobrazuje mapa na Obr. 28, bylo zjištěno, že se pohybuje od 62,0 let (okres Sokolov) do 69,9 let (okres Tábor). Nejvyšší hodnoty průměrného věku se vyskytují v oblastech Jihočeského kraje, Kraje Vysočina a dále na pomezí krajů Jihomoravského, Zlínského, Olomouckého a Moravskoslezského. Nižší hodnoty jsou pak patrné v oblasti severozápadních Čech podél hranice s Německem. Toto prostorové rozložení přibližně kopíruje průměrný věk populace v okresech České republiky.



Obr. 27: Počet novotvarů v neznámém klinickém stádiu dle věkových intervalů v letech 1976-2010 v České republice. (Zdroj dat: NOR (ÚZIS))



Obr. 28: Průměrný věk pacientů s diagnózami VZN v neznámém klinickém stádiu v letech 1976-2010 v okresech České republiky (Zdroj dat: NOR (ÚZIS))

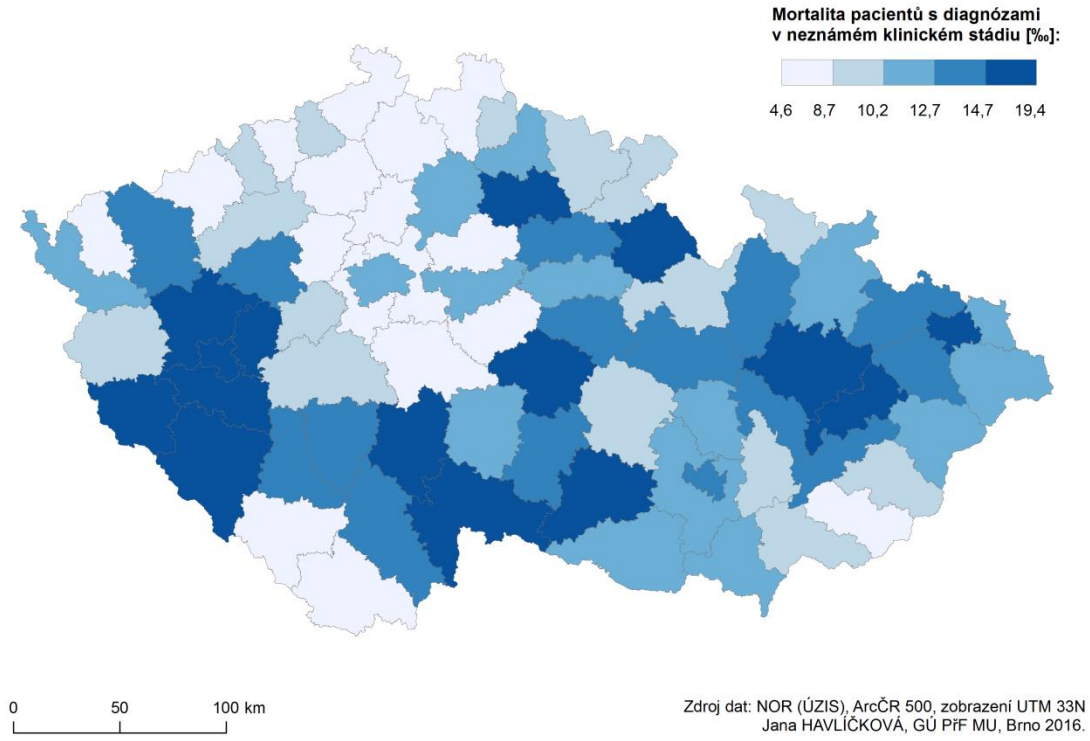
Onkologická onemocnění jsou po onemocněních oběhové soustavy druhou nejčastější příčinou úmrtí osob v České republice. Podíl úmrtí spojených právě se zhoubnými novotvary Od 70. let 20. století, kdy dosahoval podíl zemřelých na zhoubné novotvary (ZN) 18,5 %, tento podíl téměř soustavně roste. Dle publikace Zemřelí 2012 (ÚZIS [on-line], 2013) dosáhl podíl zemřelých na ZN celých 27 % všech zemřelých v daném roce. Patrný je opět i rozdíl mezi pohlavími, přičemž podíl zemřelých na ZN je vyšší u mužů.

U mužů převládají z hlediska mortality novotvary C33-C34 (ZN průdušnic, průdušek a plic) a C61 (ZN prostaty), u žen představují nejvýznamnější diagnózu taktéž novotvary C33-C34 a dále novotvary C50 (ZN prsu), (ÚZIS [on-line], 2013).

V kontextu mortality pacientů s VZN byla ve sledovaném období zaznamenána míra mortality v České republice téměř 27,0 ‰. Nejvyšší mortalitu pacientů s VZN byla zaznamenána v okrese Plzeň-město (47,5 ‰), dále v okresech Klatovy (45,6 ‰), Olomouc (44,5‰), Plzeň-jih (43,3 ‰), Plzeň-sever (41,9 ‰) a Rokycany (41,1 ‰).

Mortalita pacientů s diagnózou VZN v neznámém stádiu je za sledované období poněkud nižší a za celou Českou republiku dosahuje hodnoty 11,8 ‰. Mortalitu pacientů s diagnózami VZN v neznámém klinickém stádiu v okresech ČR demonstruje Obr. 29. Nejvyšších hodnot mortality zde dosahují pacienti v okrese Olomouc (19,3 ‰), dále pak v okresech Plzeň-město (18,8 ‰), Tábor (18,5 ‰), Klatovy (18,3 ‰). Nejnižší míru

mortality pak vykazuje okres Praha-západ (4,7 ‰). Z mapy je patrná koncentrace okresů s vyšší mortalitou v oblasti Plzeňského kraje, východní části Jihočeského kraje a v Olomouckém kraji. Lépe jsou na tom z hlediska úmrtnosti severní a střední Čechy. Podobně jako v případě průměrného věku pacientů s neznámou diagnózou můžeme i zde sledovat relativně nižší hodnoty mortality podél části hranice s Německem.



Obr. 29: Mortalita pacientů s diagnózami VZN v neznámém klinickém stádiu v letech 1976-2010 v okresech České republiky (Zdroj dat: NOR (ÚZIS))

6 APLIKACE VYBRANÝCH EXPLORAČNÍCH NÁSTROJŮ

Předchozí analýza poskytnutých dat prokázala, že se mezi celkovým počtem onkologických diagnóz nachází velmi významný podíl diagnóz v neznámém stádiu. Na základě tohoto zjištění byla tato charakteristika dále zkoumána pomocí různých nástrojů exploračních kartografie. Definice použitých metod je možno nalézt v kapitole *4 Metody pro analýzu agregovaných dat*. V následujících kapitolách je prezentována aplikace vybraných exploračních nástrojů ve studii onkologických diagnóz v neznámém klinickém stádiu v okresech České republiky v letech 1976-2010. Cílem této studie je poukázat na vhodnost prvotní explorace a analýzy dat pro získání nových objektivních informací, zjištění případné existence prostorových vzorů a závislostí a následnou tvorbu výsledných kartografických výstupů prezentujících výsledky hypotéz.

V následujícím textu bylo využito zejména explorativních a analytických nástrojů programu GeoDa. Kartografické výsledky pak byly vytvořeny v ArcMap 10.3.1.

6.1 GeoDa 1.6.7

Geoda je volně dostupný open-source multiplatformní software pro analýzu prostorových dat, geovizualizaci, prostorovou autokorelaci a prostorové modelování v programovacím jazyce C++. Verze 1.6.7 byla uvolněna 9. března 2015. Nástroj GeoDa je možno používat na Windows (XP - 8 resp. 8.1), Linux a Mac OS. Program byl původně vyvinut v Spatial Analysis Laboratory na University of Illinois. Hlavním vývojářem byl Luc Anselin. Vývoj se později přesunul na pracoviště v Arizona State University (GeoDa Center for Geospatial Analysis and Computation).

GeoDa nabízí širokou škálu nástrojů pro exploratorní prostorovou analýzu dat (Exploratory Spatial Data Analysis – ESDA). K dispozici je mnoho analytických funkcí od základních nástrojů popisné statistiky, jako je histogram, Box Plot, Scatter Plot, Conditional Plot atd., ale také nástroje pokročilé ESDA – univariantní a bivariantní Moranův Index, Lokální G statistiku, prostorovou regresi ad. V novějších verzích programu je navíc možná tvorba polygonového *.shp z gridu a textového souboru, dále anamorfní mapy (formou Dorlingovvých kartogramů), podmíněné kartogramy, podmíněné grafy, 3D Scatter Plot, nebo Parallel Coordinate Plot (PCPlot).

Program nabízí import různých formátů dat. Primárně *.shp, dále pak *.sqlite, *.dbf, *.csv, *.json, *.gml, *.kml, *.xml, *.tab, *.mif, *.mid. Zásadní výhodou tohoto programu je, že umožňuje plynulé provázání jednotlivých otevřených oken, ve kterých probíhají analýzy. Díky tzv. „*linkingu*“ se při výběru jednoho či několika prvků v jednom okně „vysvítí“ vybrané prvky i v ostatních oknech. Anselin L. et al., (2002) definuje linking jako dynamické propojení všech oken, které je obnovováno s vždy novým výběrem. Právě linking je součástí základní funkcionality interaktivní průzkumové analýzy dat, pro niž je GeoDa velmi silným nástrojem. Dynamickou formou linkingu je pak „*brushing*“. Ten ale nefunguje pro histogramy (Glosary of key terms [on-line], 2003). „*Paneling*“ pak otevírá zvláštní okno pro každou třídu nebo např. zvolený interval

(Andrienko et al., 2001). V mnoha ohledech je paneling lepší než brushing. Paneling může být použit taktéž pro histogramy, pro které není možné využiití brusingu. Tyto nástroje umožňují v okně analýzy snadno identifikovat odlehlé hodnoty (tzv. outliery), které jsou pak automaticky zvýrazněny v dalších oknech programu (v mapě nebo dalších nástrojích explorativní analýzy).

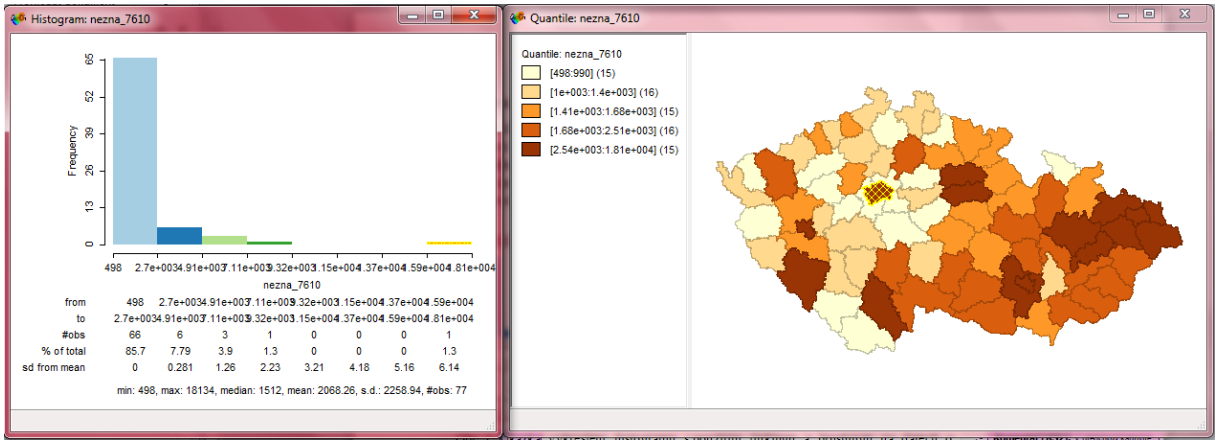
Na oficiálních stránkách programu Geoda jsou volně ke stažení podrobné tutoriály s cvičnými příklady a dokumentace k programu včetně zdrojového kódu.

6.2 Analýza prostorového vzoru výskytu diagnóz v neznámém stádiu

Metoda nepravého kartogramu, která byla použita v kapitole *5.2 Analýza neznámých stádií v prostoru* pro vizualizaci prostorové diferenciace výskytu diagnóz v neznámém stádiu nedokáže odhalit veškeré souvislosti a vztahy v datech. Proto bylo přikročeno k použití exploračních a analytických nástrojů, pro odhalení a pochopení hlubších souvislostí.

6.2.1 Identifikace odlehlých dat

K prvotnímu přiblížení datového souboru slouží výpočet základních statistických údajů (viz Tab. 6). Nejjednodušším způsobem, jak lze identifikovat v datovém souboru odlehlé hodnoty (outliery) je vykreslení histogramu. Jak je ukázáno na Obr. 30, nachází se v souboru dat výrazně odlehlá hodnota. Díky brushingu bylo zjištěno, že se jedná o okres Hlavní město Praha. Absolutně nejvíce diagnóz v neznámém stádiu se nachází právě v tomto okrese. Je to však dáno především vysokým počtem obyvatel a zároveň i celkovým vysokým počtem všech diagnóz.

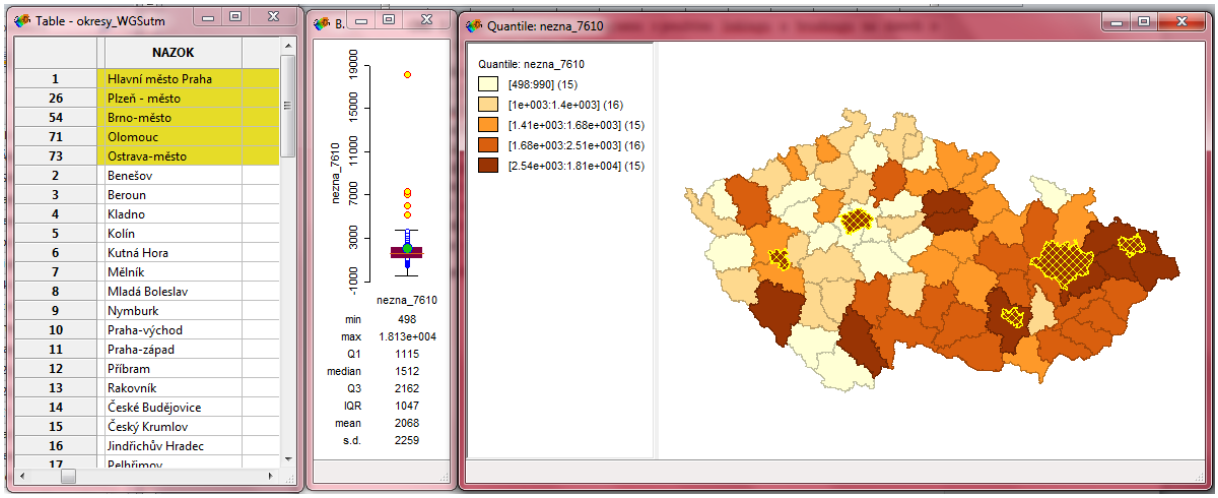


Obr. 30: Ukázka vykreslení histogramu s použitím brushingu na datech o neznámém stádiu rakoviny

Dalším způsobem detekce outlierů je krabicový diagram (viz Obr. 31). Z něho můžeme kromě detekce outlierů vyčíst také mnoho dalších měr variability. Dle DE SMITHA, (2015) ukazuje vrchní a spodní linie „krabice“ ukazují 25. a 75. percentil datového souboru. Vzdálenost mezi nimi je pak označována jako mezikvartilové rozpětí. Oranžová linie uprostřed krabice pak charakterizuje medián. Jelikož linie nedělí krabici na dva stejné díly, poukazuje to na šikmost dat. Vertikální linie pak ukazuje celý rozsah dat. Jako outliery jsou definovány takové hodnoty, které se nacházejí o 1,5 násobku mezikvartilového rozpětí dál od vrchní nebo spodní linie krabice. Díky krabicovému grafu jsme v datech identifikovali další odlehlé hodnoty, které odpovídají okresům Hlavní město Praha, Plzeň – město, Brno – město, Olomouc a Ostrava – Město, tedy opět těm nejlidnatějším okresům. Vypočtené statistiky jsou navíc zobrazeny pod grafem.

Tab. 6: Popisné statistiky výběrového souboru neznámých klinických stádií

Proměnná	Hodnota
Min	498
Max	18 134
Q_1	1 115
Medián	1 512
Q_3	2 162
Mezikvartilové rozpětí	1 047
Průměr	2 068
Směrodatná odchylka	2 259



Obr. 31: Ukázka použití krabicového grafu v programu Geoda na datech o neznámém stádiu rakoviny.

6.2.2 Korelace onkologických dat v neznámém stádiu

Korelační diagram neboli *scatter plot* je graf zobrazující v kartézských souřadnicích závislost dvou proměnných jako množinu bodů. S jeho pomocí lze zjistit

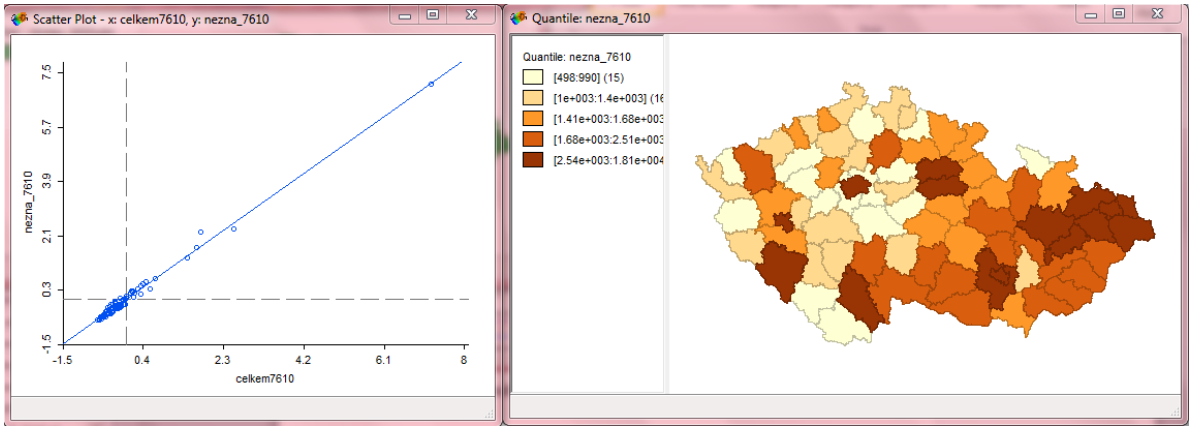
vzájemný vztah mezi dvěma proměnnými. V programu geoda byl vykreslen pro dvě proměnné – celkový počet diagnóz pro celé sledované období (nezávislá proměnná) a počet diagnóz v neznámém stádiu (závislá proměnná).

Původním předpokladem bylo, že mezi zvolenými proměnnými existuje závislost. Cílem této podkapitoly bylo zjistit, do jaké míry je výskyt diagnóz v neznámém stádiu podmíněn celkovým počtem diagnóz v jednotlivých okresech.

Bylo využito možnosti převést data na standardizovaná (data jsou převedena na jednotky směrodatné odchylky). Dle ANSELINA, (2005) lze v tomto případě označit všechny hodnoty vyšší než 2 jako outliery. Z toho vyplývá, že kromě detekování korelace s jinou proměnnou je možné díky scatter plotu také identifikovat odlehlé hodnoty. Na Obr. 32 je opět jasně identifikována pětice okresů s výrazně vyšším počtem neznámých stádií rakoviny. Jelikož zobrazujeme dvě neznámé, které jsou na sobě teoreticky přímo závislé, předpokládáme, že mezi nimi bude korelační vztah. Tento předpoklad potvrzuje rozložení bodů v grafu, jež je možné proložit přímkou. Linie proložená daty, která svírá s osou x úhel přibližně 45 ° je důkazem existence lineárního vztahu mezi studovanými daty. Body v grafu jsou k linii poměrně těsně přimknuty, což indikuje přímý korelační vztah.

Hlavním výsledkem korelačního diagramu je mimo grafického výstupu také výpočet *koeficientu determinace* R^2. Výpočet koeficientu determinace probíhá na základě vztahu [4.2]. V tomto případě se jeho hodnota se blíží 1 ($R^2 = 0,988$). Z toho vyplývá, že změna celkového počtu diagnóz silně podmiňuje změnu počtu diagnóz v neznámém stádiu. Vynásobíme-li tuto hodnotu 100, dostaneme informaci o tom, z kolika procent vysvětluje změna nezávislé proměnné závisle proměnnou. Tzn., že celkový počet diagnóz v okresech z 98,8 % podmiňuje výskyt diagnóz novotvaru v neznámém stádiu.

Na existence korelace však nelze usuzovat pouze pomocí vizuálního hodnocení korelačního diagramu a koeficientu determinace. Ačkoli oba zmíněné ukazatelé prokazují téměř jednoznačné výsledky, pro úplnost byl doplněn také výpočet *Pearsonova koeficientu korelace r*. Jeho výpočet je dán vzorcem [4.1]. *Korelační* koeficient v tomto případě nabývá hodnoty $r = 0,994$. Dle tabulky přibližné interpretace hodnot korelačního koeficientu, uvedené výše v kapitole *4.1 Korelační analýza a korelační diagram*, lze říci, že mezi studovanými daty je patrná velmi vysoká závislost.



Obr. 32: Ukázka použití korelačního diagramu v programu Geoda na datech o neznámém stádiu rakoviny.

6.3 Identifikace shluků v prostoru

Prostorový vzor výskytu nemoci, respektive v tomto případě výskytu diagnóz v neznámém stádiu lze identifikovat na základě vizuálního hodnocení běžných mapových výstupů (např. z Obr. 24). Metody prostorových statistických metod, které slouží k průzkumu prostorové autokorelace však pomáhají možné prostorové shluky správně popsat a ohodnotit jejich významnost v rámci studovaného území. Jedním z nejvýznamnějších a nejpoužívanějších nástrojů pro hodnocení prostorových vzorů a hledání prostorových shluků na globální i lokální úrovni je Moranovo I kritérium. Tento nástroj byl použit taktéž v této studii. Definice a vysvětlení identifikace prostorového shlukování a výpočet jednotlivých kritérií je obsahem kapitoly *4.1 Korelační analýza a* korelační diagram.

Analýze prostorového shlukování byla podrobena opět data z databáze NOR (ÚZIS) o počtu diagnóz novotvarů v neznámém stádiu v okresech České republiky v období let 1976-2010. Absolutní počty diagnóz v neznámém stádiu byly přepočteny na 100 000 obyvatel průměrné populace ve zkoumaném období. Postup výpočtu podrobněji popisuje kapitola *3.2 Charakteristika zpracovávaného datového souboru*. Výpočty jednotlivých kritérií byly prováděny na takto upravených datech.

Analýza byla provedena za obě pohlaví a celé studované období, dále zvlášť za muže a ženy v celém studovaném období a za obě pohlaví dohromady ve dvou časových obdobích - v letech 1976-1994 a 1995-2010. Cílem této analýzy bylo prokázat, že data mají ve studované oblasti tendenci k prostorovému shlukování, následně prokázat statistickou významnost tohoto shlukování a tyto shluky lokalizovat. Výstupy této analýzy byly následně zpracovány a prezentovány pomocí ArcMap 10.3.1.

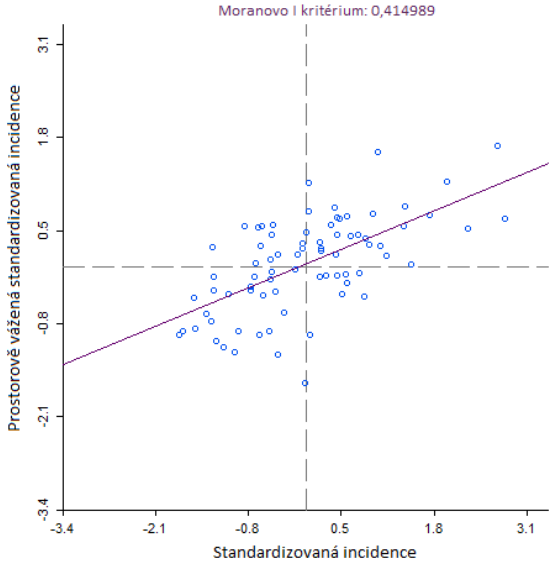
6.3.1 Hodnocení globálního prostorového vzoru

V prvé řadě byl popsán globální prostorový vzor, který popisuje převládající prostorový vzor ve zkoumaném území. Pomocí *globálního Moranova I kritéria* bylo zjištěno, jak silná je prostorová autokorelace a zda jsou prostorové jevy rozmístěné

náhodně, shlukovaně nebo pravidelně. V případě vyšetřování prostorové autokorelace plošných objektů je podobnost polohy hodnocena pomocí vztahů sousedství. Tyto vztahy jsou kvantifikovány tzv. maticí prostorových vah (DOBROVOLNÝ, 2015). Prostorové váhy jsou definovány počtem sousedů, typem sousedství, maximální vzdálenosti působení autokorelace nebo jejich vzájemnou kombinací. V případě hodnocení globálního prostorového vzoru incidence diagnóz v neznámém stádiu bylo zvoleno sousedství typu královna 1. řádu (queen's move, na základě pohybu figurky královny v šachách). Tento typ sousedství zahrnuje všechny regiony, které sdílejí alespoň část hranice. Na základě výpočtu sousedství má každý okres průměrně pět sousedů (min = 2, max = 8, medián = 5, průměr = 4,96, směrodatná odchylka = 1,66). Po výpočtu typu sousedství byl proveden výpočet globálního Moranova I kritéria. Výpočet Moranova I kritéria stejně jako interpretace možných výsledků je uveden v kapitole *4.1 Korelační analýza a korelační diagram.*

Výsledné Moranovo I kritérium bylo vypočteno na základě vzorce [4.3] je *I = 0,41*. Jelikož je hodnota kladná a liší se od nuly, znamená to, že data vykazují pozitivní prostorovou autokorelaci. Z toho vyplývá, že ve studovaném území *se vyskytuje prostorové shlukování*. Samotná velikost Moranova I kritéria však neindikuje statistickou významnost. Aby bylo možno zamítnout nulovou hypotézu o neexistenci prostorové autokorelace, musela být statistická významnost dále testována pomocí permutační procedury. Významnost tedy byla testována oproti očekávané hodnotě (tedy H_0: I = 0) v případě náhodného rozmístění pomocí 10 000 permutací. Výsledkem byla hodnota *p-value = 0,0001*. S ohledem na tvrzení, že čím nižší hodnoty p-value nabývá, tím menší je pravděpodobnost platnosti nulové hypotézy (viz kapitola *4.2.2 Hodnocení významnosti výsledků na základě p-value*), můžeme zamítnout stanovenou nulovou hypotézu a říci, že permutační procedura prokázala statistickou významnost existence prostorového shlukování na hladině významnosti $\alpha = 0,05$.

Grafické znázornění Moranova I kritéria je na Obr. 33. Na základě jeho zhodnocení bylo prokázáno, že ve studovaném území se objevují tendence ke shlukování. Hodnoty vyskytující se v I. resp. ve III. kvadrantu grafu poukazují na tendenci ke shlukování vysokých, resp. nízkých hodnot sledovaného jevu. Body ležící ve II. a IV. kvadrantu pak představují outliery (okresy s vyšší hodnotou sledovaného jevu ve svém okolí a naopak).



Obr. 33: Moranův diagram pro incidenci diagnóz v neznámém stádiu na 100 000 osob průměrného počtu obyvatel v okresech České republiky v letech 1976-2010

6.3.2 Hodnocení lokálního prostorového vzoru

Pro zjištění lokalizace konkrétních shluků slouží *lokální ekvivalent Moranova I kritéria*, které je v kontextu programu GeoDa implikováno jako *lokální indikátor prostorové autokorelace – LISA* – výpočet LISA je opět popsán v kapitole *4.1*. Indikátory LISA zohledňují příspěvek každého jednotlivého pozorování. Na základě analýzy LISA lze kategorizovat sledované jednotky (dle typu prostorové autokorelace) do čtyř skupin, které odpovídají čtyřem kvadrantům Moranova diagramu (MAŠKARINEC, 2013).

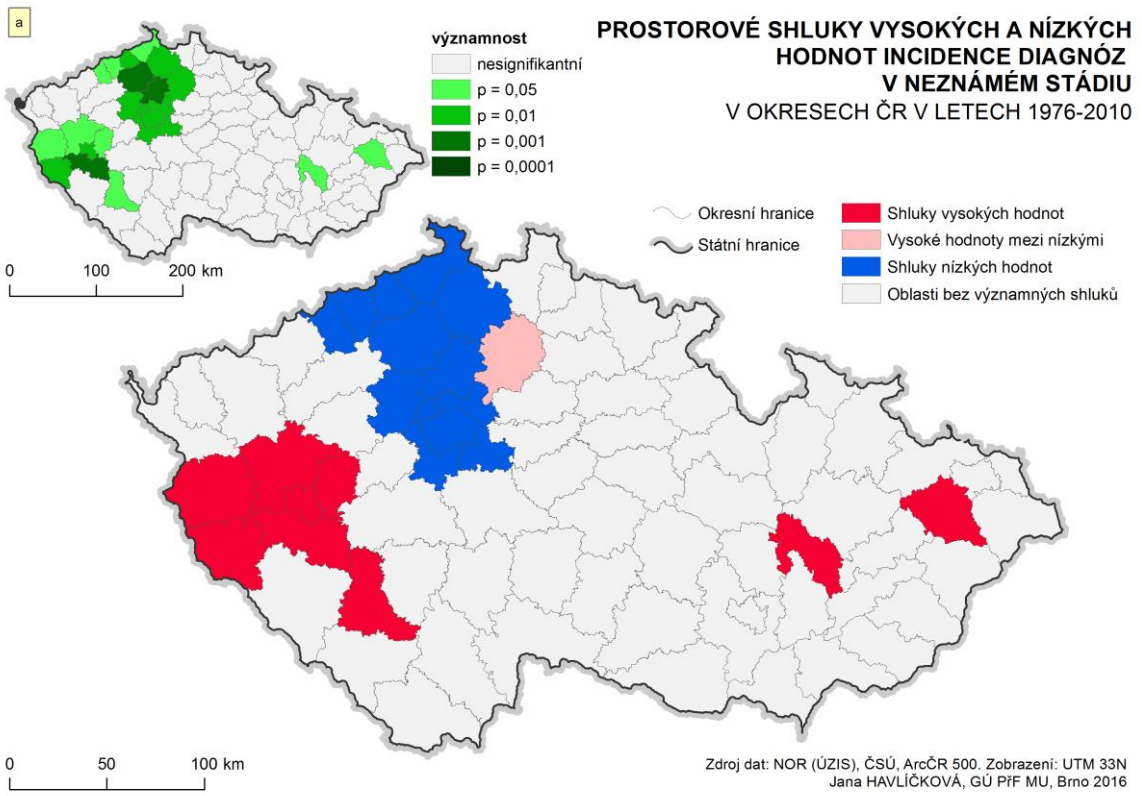
Stejně jako v případě globálního hodnocení prostorové autokorelace i v případě LISA indikátoru je v prvé řadě třeba definovat typ sousedství a s ním související prostorové váhy. I zde bylo zvoleno sousedství typu královna. V prostředí programu GeoDa je možno zvolit si tři typy výstupů – Moranův diagram, mapa shluků a odlehlých hodnot a dále mapa hodnotící statistickou významnost výsledných identifikovaných shluků.

Obr. 34 představuje ukázku prvního z konečných výstupů lokálního identifikátoru prostorové autokorelace. Červená barva představuje shluky vysokých hodnot (high-high) incidence diagnóz v neznámém stádiu. Jsou to tedy okresy, ve kterých je hodnota incidence vyšší než v jejich okolí. Naopak modré regiony označují shluky nízkých hodnot (low-low), tedy okresy s nižší hodnotou incidence než v okolí. Světle červená barva označuje oblast negativní autokorelace (high-low), které jsou dle ANSELINA, (2005) označovány jako odlehlé hodnoty (spatial outliers). Jedná se tedy o vysoké hodnoty obklopené shluky nízkých hodnot. Ve studovaném území se nenachází shluk typu (low-high). Vedlejší mapové okno *a* představuje mapu signifikance, tedy mapu statistické významnosti výsledných shluků, která vznikla po 10 000 permutací. Nutno podotknout, že shluky s významností p = 0,05 jsou poněkud nespolehlivé, protože pravděpodobně

nereflektují problémy spojené s vícenásobným porovnáváním, a nelze je tedy označit za statisticky významné (MAREK, 2015).

Výsledky hodnocení lokálního Moranova I kritéria odhalili v prostoru významné shluky. Významné shluky okresů s nízkými hodnotami incidence se nachází na severu a severozápadě České republiky. Za statisticky významné shluky nízkých hodnot byly označeny okresy Hlavní město Praha, Kladno, Mělník, Praha-východ, Praha-západ, Česká Lípa, Litoměřice a Ústí nad Labem. Statisticky významné shluky vysokých hodnot byly identifikovány na východu České republiky, jedná se o okresy Domažlice, Plzeň-město a Plzeň jih. Okres Mladá Boleslav byl identifikován jako okres s vyššími hodnotami incidence než okresy v jeho okolí. Ostatní okresy identifikované jako shluky a zobrazené v hlavním mapovém okně na Obr. 34 byly pomocí mapy signifikance označeny jako statisticky nesignifikantní.

V porovnání s metodou jednoduchého homogenního pseudokartogramu (viz Obr. 25) přináší výsledky hodnocení lokálního prostorového shlukování mnohem podrobnější a srozumitelnější informace o existujícím prostorovém vzoru v datech. Zatímco v případě hodnocení incidence bylo identifikováno několik oblastí s její vysokou hodnotou Plzeňský, Olomoucký a Jihočeský kraj, lokální Moranovo I kritérium detekovalo pouze jeden statisticky významný shluk vysokých hodnot incidence. Velmi podobně tomu bylo i v případě shluků nízkých hodnot.



Obr. 34: Prostorové shluky vysokých a nízkých hodnot a mapa významnosti generované pomocí lokálního Moranova I kritéria (LISA)

Stejná analýza byla dále provedena pro incidenci diagnóz v neznámém stádiu pro obě pohlaví v období let 1976-1994 a 1995-2010 a zvlášť za celé sledované období pro muže a ženy. Ve všech případech byla prokázána existence statisticky významného prostorového shlukování na hladině významnosti $\alpha = 0{,}05$. Jednotlivé výsledné prostorové shluky vysokých a nízkých hodnot včetně map významnosti jsou obsahem Příloha 2 - Příloha 6.

Cílem tvorby série těchto analýz bylo poukázat na významnou změnu existence a lokalizace prostorových vzorů ve zmíněných časových obdobích a rozdíl mezi pohlavími.

S ohledem na předchozí tvrzení, že po roce 1994 došlo ke změně pravidel hlášení nových onkologických případů, byla analýza provedena právě za období před a po roce 1994. Před rokem 1994 byl identifikován významný prostorový shluk vysokých hodnot incidence diagnóz v neznámém stádiu na východě ČR. Jedná se především o okresy Moravskoslezského a Olomouckého kraje (Prostějov, Olomouc, Přerov, Bruntál, Opava, Nový Jičín a Karviná). Shluky nízkých hodnot se vyskytují podobně jako v předchozím popsaném případě na severu a severozápadě státu. Po roce 1994 byla zaznamenána významná změna v lokalizaci shluku vysokých hodnot incidence. Nevyskytuje se již na východě ČR, ale naopak na západě Většinu okresů Plzeňského kraje zde bylo identifikováno jako statisticky významný shluk vysokých hodnot incidence. Umístění shluků nízkých hodnot zůstává přibližně stejné, ačkoli při zhodnocení jejich statistických signifikancí lze říci, že ve druhém sledovaném období je statistická významnost většiny identifikovaných okresů nižší.

V případě hodnocení prostorové autokorelace u mužů byl zjištěn významný shluk vysokých hodnot v okresech Plzeň-město, Plzeň-jih a Domažlice. Statisticky významný shluk nízkých hodnot je pak zaznamenám na pomezí krajů Středočeského, Libereckého a Ústeckého kraje. Velmi podobné shluky jsou identifikovány i v případě žen. Navíc se zde vyskytují dva okresy (Hlavní město Praha, Mladá Boleslav), které jsou označeny jako outliery vysokých hodnot mezi nízkými a vyskytují se mezi okresy, které jsou označeny jako shluk nízkých hodnot.

6.3.3 Mapová koncepce shlukových map

Cílem této práce byl kromě analýzy vybraných dat o VZN návrh vhodné kartografické prezentace zjištěných výsledků, která bude respektovat veškerá pravidla správné mapové koncepce. Hodnocení lokálního prostorové autokorelace v programu GeoDa přineslo výsledky v podobě shluků vysokých a nízkých hodnot zvolené charakteristiky – v tomto případě incidence diagnóz v neznámém klinickém stádiu na 100 000 průměrného počtu obyvatel. Vedle informace o typu shluku program generoval také informaci o jejich významnosti. Úkolem bylo tedy najít vhodný typ prezentace pro shluky, ale také pro doplňující informaci o jejich signifikanci.

V první části bylo třeba vyřešit, jak zobrazit jednotlivé shluky. Pro jejich prezentaci tedy byla použita shluková mapa. Barvy byly zvoleny tzv. asociativním způsobem (ČERBA, 2007), tedy tak aby měly logickou návaznost na prezentované údaje. Shluky vysokých hodnot jsou tedy prezentovány sytě červenou barvou, shluky nízkých hodnot naopak sytě modrou barvou. Vysoké outliery mezi nízkými hodnotami jsou pak zobrazeny světle červenou barvou a nízké outliery mezi vysokými hodnotami světle modrou barvou. Oblasti bez shluků byly ponechány světle šedou barvou.

V druhé části bylo nutné zamyslet se nad tím, jak zobrazit statistickou významnost jednotlivých shluků, jelikož bez této informace by byla mapa nekompletní, což by mohlo vést k mylným interpretacím. Především bylo nutné, aby bylo patrné, které shluky mají významnost vyšší než p = 0,05. Jak bylo výše uvedeno, shluky s významností p = 0,05 mohou být poněkud nespolehlivé.

Pro co nejjednodušší interpretaci, které by byl schopný i laik bez hlubších kartografických znalostí byla nakonec zvolena „metoda vedlejšího zobrazení“ (*adjacent display method*), kterou ve své publikaci diskutuje MACEACHREN, A et al., [on-line], (1998). MacEachren zde uvádí, že toto zobrazení je složeno z mapové dvojice. Hlavní mapa zobrazuje dané téma, druhá, menší mapa zobrazuje binární spolehlivost (spolehlivé/nespolehlivé). V našem případě však nezobrazujeme pouze binární informaci o spolehlivosti, ale dělíme okresy do jednotlivých intervalů spolehlivosti, jak byla vypočtena v programu GeoDa.

Hranice intervalů byly určeny na základě běžných hladin testování. Barevné schéma bylo zvoleno zelené – od nejtmavší barvy, která reprezentuje ty nejvíce spolehlivé intervaly, po nejsvětlejší pro méně významné intervaly. Nejsvětlejší oblasti zde tedy zastupují okresy, se statistickou významností p = 0,05 a mohou být tedy interpretovány jako nespolehlivé. Oblasti bez shluků byly opět ponechány světle šedou barvou.

Díky přítomnosti mapy statistické významnosti v jednom mapovém okně společně se shlukovou mapou, může být čtenář mapy schopen rychle a přesně odlišit významné shluky od těch nevýznamných. Tento přístup je vhodný právě z důvodu jednoduché, ale účelné interpretace. Čtenář je okamžitě schopen identifikovat okresy, ve kterých oblastech jsou stádia diagnóz nejčastěji označována jako neznámá a kde je naopak situace lepší. Na základě tohoto hodnocení je pak možné přistoupit ke vhodným opatřením, která by vedla ke zlepšení situace v oblasti klasifikace klinického stádia novotvarů.

6.4 Analýza hlavních komponent (PCA)

Kapitola *5.3 Analýza neznámých stádií na základě dalších kritérií - pohlaví, diagnóz, věku ad.* prokázala jistou návaznost výskytu onkologických diagnóz v neznámém klinickém stádiu na další charakteristiky onkologických onemocnění případně charakteristiky demografické. Cílem této kapitoly je zjistit možné asociace mezi

neznámými diagnózami a dalšími vybranými vlastnostmi, případně zjistit sílu těchto vztahů.

Přestože je možné závislosti zkoumat pomocí metod korelace či prostorové autokorelace, mají tyto metody také jistá omezení. Klíčovým omezením je skutečnost, že studují pouze lineární závislosti mezi dvojící proměnných. Korelační koeficient, který byl již použit výše, není konkrétní kvantifikací zkoumaného vztahu, nýbrž pouze určuje sílu a směr lineárního vztahu mezi proměnnými.

Tyto nedostatky jsou eliminovány použitím analýzy hlavních komponent. Tuto analýzu již dříve aplikovala ve své práci BÁČOVÁ, (2012) pro posouzení vlivu vybraných faktorů v případové studii bilaterálního karcinomu prsu. Autorka prováděla analýzu v softwaru SpaceStat a výsledkem této analýzy byla mapa zobrazující prostorovou kumulaci ženské populace se zvýšeným rizikem úmrtí na bilaterální karcinom prsu v krajích České republiky v letech 1979-2005.

V tomto případě byla pro analýzu zvolena data z databáze NOR (ÚZIS), charakterizující počty diagnostikovaných novotvarů v neznámém klinickém stádiu v okresech České republiky v období let 1976-2010. Jak bylo již dříve vzpomenuto, neznámá stádia novotvarů dosud nebyla nikde blíže charakterizována ani analyzována.

BÁČOVÁ, (2012) uvádí, že PCA je robustní statistická metoda, jejíž výsledky jsou málo citlivé k hrubým chybám ve vstupních datech. Použitím této metody pro analýzu vlivu vybraných faktorů můžeme dosáhnout alespoň částečné eliminace vlivu případných chyb vstupních souborů. BAČOVÁ a kol., (2013) také zmiňují, že je PCA obvyklý algoritmus, který se běžně používá k redukci datového souboru. Z těchto důvodů byla tato analýza použita i v případě dat charakterizující počty VZN v neznámém stádiu. Cílem analýzy je zajistit redukci zvolených proměnných bez ztráty podstatných informací a zjistit, které z proměnných vysvětlují variabilitu zvoleného datového souboru.

6.4.1 Program a programovací jazyk R

Jak již napovídá název podkapitoly, mluvíme-li o R, můžeme mít na mysli programovací jazyk, ale i software pro provádění statistických výpočtů a tvorbu grafických výstupů. Nespornou výhodou je to, že jde o open-source program. V případě programovacího jazyka R se v podstatě jedná o volně dostupnou implementaci programovacího jazyka S, který je používán profesionálními statistickými programy.

Do základního systému je možné jednoduše doinstalovat velké množství balíčků funkcí, které rozšiřují základní funkcionalitu programu (Matematický software R [on-line], 2009). Základní sada balíčků je implementována v samotné instalaci programu. Další balíčky je možné si stáhnout ve formě zip archivů z oficiálních stránek R. Pro účely této diplomové práce bylo nutné nainstalovat balíčky *openxlsx, xlsx* a *stats*.

6.4.2 Analýza hlavních komponent (PCA) diagnóz v neznámém klinickém stádiu v okresech České republiky v letech 1976-2010

Z dat poskytnutých z databáze NOR bylo vybráno 5 charakteristik, které vhodně reprezentují původní datovou sadu. Tyto pak byly doplněny daty o průměrných počtech mužů a žen vypočtených dle údajů ze sčítání lidu, domů a bytů z let 1980, 1991, 2001 a 2011 Českého statistického úřadu. Tyto charakteristiky se staly vstupními parametry pro PCA. Konkrétně byly pro další využití vybrány tyto charakteristiky:

- Celkový počet diagnóz v neznámém stádiu
- Počty primárních novotvarů
- Počty následných novotvarů
- Mortalita pacientů s diagnostikovanými novotvary v neznámém klinickém stádiu
- Průměrný věk pacientů s onkologickým onemocněním
- Průměrný počet žen v letech 1980-2011
- Průměrný počet mužů v letech 1980-2011

Vzhledem k tomu, že žádný z volně dostupných GIS softwarů nenabízí výpočet analýzy hlavních komponent (s výjimkou ArcGIS, jehož implementace PCA nevyhovuje požadavkům této studie, jelikož je zde výpočet PCA implementován pouze pro rastrová data), byl výpočet jednotlivých komponent proveden ve volně dostupném programu pro statistickou analýzu dat R.

V tomto případě se nejedná o provedení analýzy pomocí předem připravených funkcí, nýbrž o výpočet funkce v jednotlivých krocích pomocí programovacího jazyka R. Nejprve bylo nutné vytvořit si vstupní soubor dat, který bude podroben analýze. Byl tedy vytvořen soubor *.xlsx s výše uvedenými charakteristikami a ID okresu. Tento soubor byl do programu naimportován pomocí funkce *read.xlsx*. Dále byla pomocí funkce *cor* vypočtena korelační matice všech charakteristik (s výjimkou pole ID).

Pro samotný výpočet hlavních komponent byla využita funkce *prcomp* z balíčku funkcí *stats*. Výpočet PCA je v tom případě implementován jako singulární rozklad hodnot (SVD - singular value decomposition), který dává lepší numerickou přesnost než jiné nabízené funkce. V souladu s doporučeními pro průběh PCA byla data sedmi vybraných proměnných nejprve normalizována na rozsah ⟨-1,1⟩. Toho bylo dosaženo využitím vztahu: $\frac{x-\overline{x}}{\max(|x-\overline{x}|)}$ (MAREK, 2015).

V podmínce funkce bylo použito vycentrování a škálování datového souboru, které je vyžadováno pro správný výsledek funkce. Vycentrování a škálování souboru může být provedeno taktéž samostatně před samotnou analýzou, ale vzhledem k možnosti provést centrování a škálování v rámci jednoho příkazu spolu s PCA analýzou, bylo přikročeno k této možnosti.

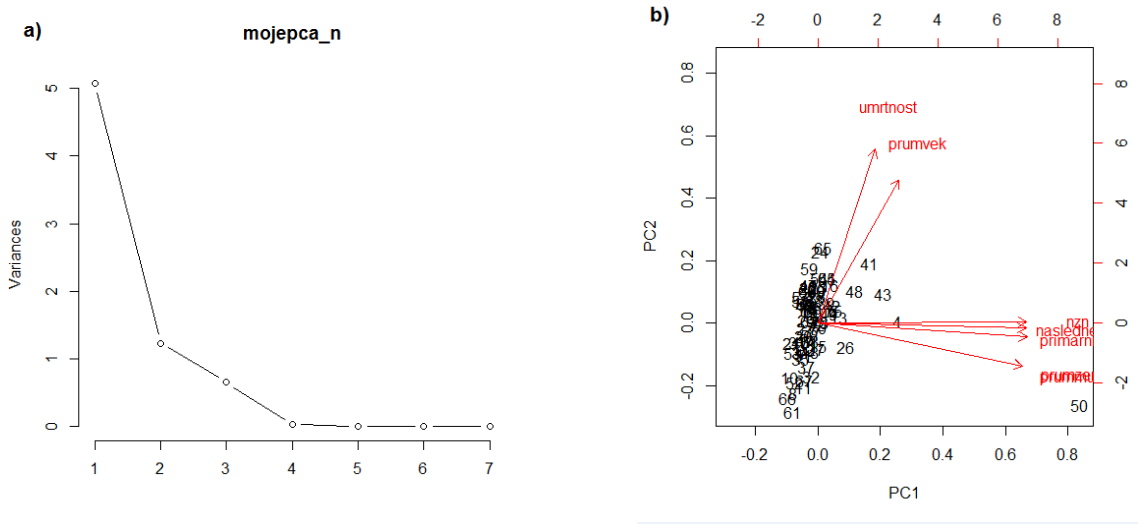
Výsledkem funkce *prcomp* je matice vah - zátěží (*eigenvectors*) a hodnoty komponentního skóre každé z vypočtených komponent (*eigenvalues*) pro každý okres. SCHROEDER, (2013) uvádí, že komponentní skóre popisuje pozici pozorované jednotky

v atributovém prostoru ve smyslu její vzdálenosti podél osy hlavních komponent, která je determinována lineární transformací vstupních hodnot použitím komponentních vah jako koeficientů. Výsledné vektory každé z komponent jsou shrnuty v Tab. 7.

Výpočet vlastních vektorů hlavních komponent vysvětluje kapitola *4.3 Analýza hlavních komponent*. Tabulka podává informace o zátěži, jakou představují jednotlivé proměnné pro vypočtené komponenty. Zvýrazněny jsou zde nejvýznamnější zátěže v každé komponentě. V první komponentě (*PC1*) je zřejmý dominantní vliv počtu neznámých diagnóz a počtu primárních a následných diagnóz, ale také průměrného počtu mužů a žen. Tuto komponentu bychom mohli označit jako *úroveň rozvoje nádorových onemocnění*. V druhé a třetí komponentě (*PC2 a PC3*) pak vykazuje dominanci mortalita pacientů s neznámým stádiem diagnózy a průměrný věk pacientů s onkologickým onemocněním a můžeme ji označit jako *mortalita a průměrný věk onkologických pacientů*. Tato komponenta tedy poukazuje na úzkou souvislost výskytu neznámých stádií rakoviny s úmrtností a průměrným věkem onkologických pacientů.

Analýza tedy potvrzuje významný vliv počtu primárných a následných diagnóz, a průměrného počtu mužů a žen na počty diagnóz v neznámém stádiu, zatímco vliv mortality a průměrného věku pacientů lze svým významem zařadit až na druhé místo. Tyto výsledky se pochopitelně mohou lišit v prostoru. Pro hodnocení místních odlišností je vhodné vytvořit srozumitelnou kartografickou prezentaci dosažených výsledků.

Zjištěné závěry byly taktéž podloženy pomocí grafického znázornění výsledků analýzy na Obr. 35. Grafy byly opět vygenerovány v programu R pomocí funkcí *screeplot (data, type="lines"*) a biplot *(princomp (data))*. Obrázek *a* znázorňuje tzv. scree plot neboli indexový graf úpatí vlastních čísel. Na tomto grafu je možno jednoznačně identifikovat zlomové místo (neboli tzv. scree), která představuje zlom mezi významnými a nevýznamnými komponentami. Dle grafu jsou nejvýznamnějšími první tři komponenty, které byly následně taktéž vybrány pro výslednou prezentaci. Obrázek *b* pak zobrazuje tzv. biplot neboli dvojný graf. Tento graf kombinuje zobrazení komponentních vah a komponentního skóre. Délka jednotlivých průvodičů je přímo úměrná komponentní váze, tedy příspěvku původního znaku do hlavní komponenty. Dle grafu tedy můžeme soudit, že nejvýznamnější příspěvky do hlavních komponent přináší celkový počet neznámých diagnóz, průměrný počet mužů a žen a počty primárních a následných diagnóz. Úhel mezi jednotlivými průvodiči je pak nepřímo úměrný velikosti korelace mezi těmito proměnnými. Zmíněné proměnné, které mají nejvyšší komponentní váhy, vykazují také vysokou korelaci.



Obr. 35: Grafické znázornění výsledků analýzy hlavních komponent: a) scree plot, b) biplot

Tab. 7: Shrnutí matice vektorů sedmi hlavních komponent vzešlých z analýzy hlavních komponent (tučně jsou označeny nejvýznamnější zátěže v komponentě)

Proměnná	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Neznámé	**0,44108**	0,00393	-0,11717	0,19886	**0,85627**	-0,13419	-0,03091
PPN	**0,44176**	-0,05999	-0,06236	0,24812	-0,40649	**-0,74706**	0,11161
PNN	**0,44003**	-0,02136	-0,08958	**0,58674**	-0,28120	0,61014	-0,04773
Mortalita	0,11995	**0,74348**	**-0,60793**	-0,23287	-0,08913	0,03310	-0,00023
Průměrný věk	0,17013	**0,61257**	**0,77138**	0,02537	0,00786	-0,00874	0,00058
Muži*	**0,43193**	-0,18471	0,07041	-0,50728	-0,03542	0,21742	**0,68447**
Ženy**	**0,43217**	-0,18385	0,06923	-0,49236	-0,11505	0,05632	**-0,71820**

(*Pozn: PPN = počty primárních novotvarů, PNN = počty následných novotvarů, * průměrný počet mužů, ** průměrný počet žen.*)

Z původního souboru sedmi proměnných vzniklo výpočtem sedm hlavních komponent. První tři výše zmíněné komponenty byly zvoleny jako vhodná náhrada původní sady proměnných a to především s ohledem na to, že dohromady vysvětlují celkem 99,4 % celkové variability vstupního souboru. Zastoupení jednotlivých hlavních komponent na celkovém rozptylu je shrnuto v Tab. 8. Tyto tři komponenty tedy mohou být zvoleny jako vhodná redukce původního datového souboru, neboť bylo studiem literatury zjištěno, že k objektivnímu vysvětlení zdrojové matice dat a její redukci bez výrazné ztráty informace stačí dosáhnout celkové míry variability 80 % (BÁČOVÁ, 2012).

Tab. 8: Podíl jednotlivých hlavních komponent na celkovém rozptylu, směrodatná odchylka a hodnoty komponentního skóre (eigenvalues) každé vypočtené komponenty

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Komponentní skóre (eigenvalues)	5,0686	1,2362	0,6544	0,0341	0,0048	0,0017	0,0002
Směrodatná ochylka	2,2514	1,1118	0,8090	0,1846	0,0692	0,0418	0,0141
Podíl celkového rozptylu (%)	72,410	17,660	9,349	0,487	0,068	0,025	0,003
Kumulativní součet rozptylu (%)	**72,410**	**90,070**	**99,417**	99,904	99,972	99,997	100

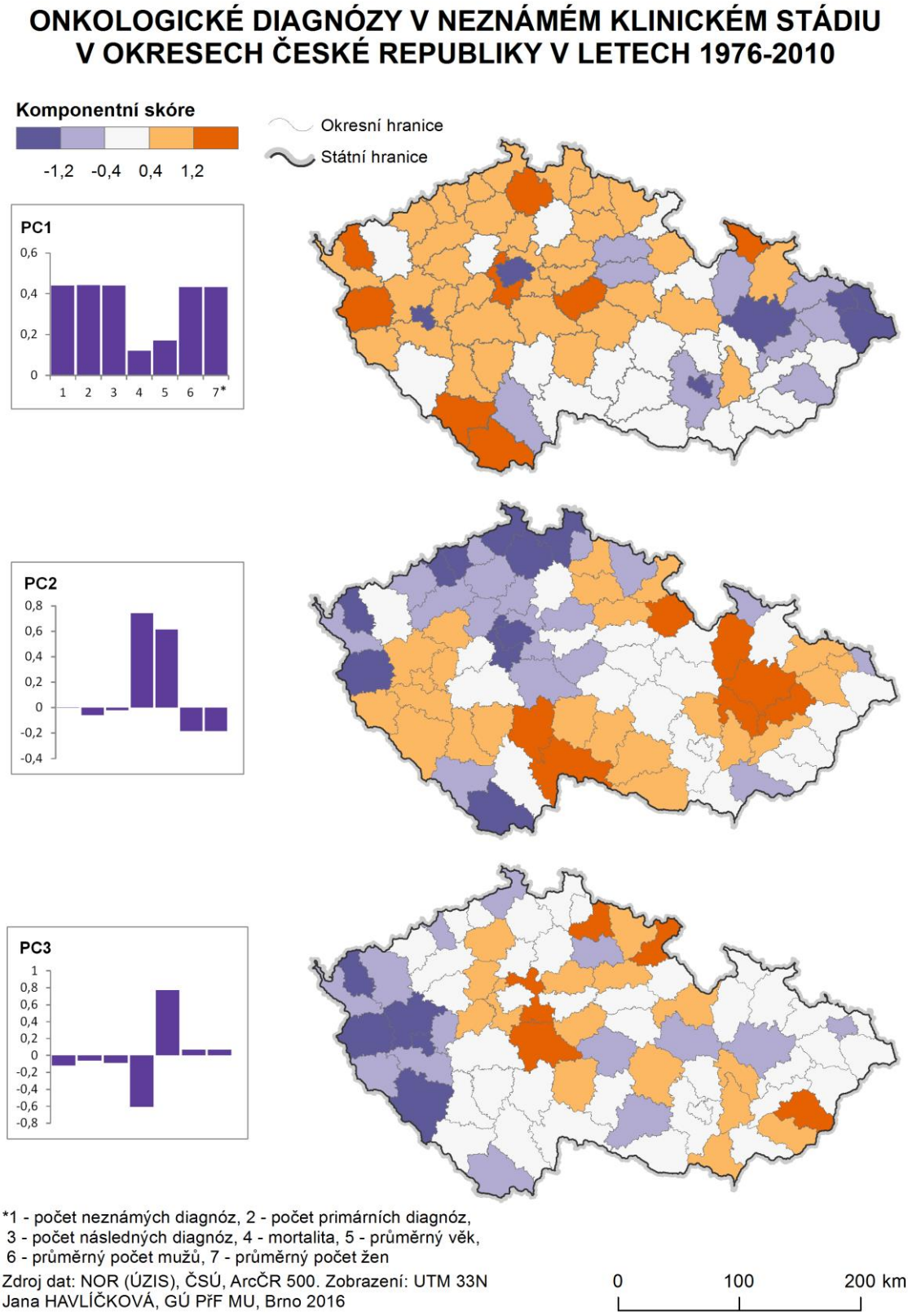
Tyto tři komponenty byly taktéž vybrány pro následnou kartografickou prezentaci. Právě použití analýzy hlavních komponent přispělo významnou měrou ke snížení rozměrnosti původního souboru dat a zároveň výrazně usnadnilo úkol kartografického vyjádření vícerozměrného datového souboru. Volbou vhodné kartografické prezentace, v níž byly vzájemně zkombinovány první tři komponenty, bylo možné identifikovat oblasti, kde jsou koncentrováni potenciální pacienti, u kterých je stádium onkologického onemocnění označeno jako neznámé.

Díky možnosti vygenerovat konkrétní hodnoty komponentního skóre pro jednotlivé okresy v programu R, bylo možné vytvořit prezentaci, která je na Obr. 36. Ze série map zobrazujících komponentní skóre prvních tří hlavních komponent lze identifikovat potenciálně rizikové okresy. Hodnoty zátěží všech původních datových souborů jsou zde zobrazeny pomocí sloupcových grafů.

Jak již bylo uvedeno, nejvýznamnější příspěvek v první hlavní komponentě tvoří celkový počet neznámých novotvarů, počty primárních a následných novotvarů a průměrné počty mužů a žen. V mapě první hlavní komponenty, jsou oranžovou barvou zobrazeny okresy s vysokými hodnotami výše zmíněných proměnných, jež indikují vysoké riziko, že novotvar bude diagnostikován v neznámém stádiu. Fialovou barvou jsou naopak zobrazeny okresy, ve kterých je počet neznámých stádií nízký navzdory vysokým hodnotám proměnných, které tvoří nejvýznamnější příspěvek komponenty. Z této mapy je tedy patrné, že riziková populace je identifikována ve více než polovině území České republiky. Nejvýznamnější riziko přitom můžeme nalézt v okresech Sokolov, Tachov, Prachatice, Český Krumlov, Praha-západ, Česká Lípa, Kutná Hora a Jeseník. Příspěvek první komponenty vysvětluje celých 72,41 % variability vstupního souboru.

V dalších dvou komponentách převládá vliv průměrného věku onkologických pacientů a mortalita pacientů s neznámými stádii. Komponentní skóre v druhé a třetí komponentě zobrazují následující dvě mapy. V kontextu druhé komponenty je riziková populace lokalizována v části Olomouckého kraje (okresy Šumperk, Olomouc, Prostějov, Přerov), Jihočeského kraje (okresy Jindřichův Hradec a Tábor) a dále okres Rychnov nad Kněžnou. Třetí komponenta pak lokalizuje rizikové oblasti v okresu Praha-východ, Benešov, Semily, Náchod a Zlín. V těchto oblastech má tedy na výskyt neznámých

diagnóz nejvýznamnější vliv průměrný věk a mortalita pacientů. Se zřetelem k tomu, že tyto dvě komponenty vysvětlují souhrnem 27,009 % celkové variability, lze tento příspěvek považovat taktéž považovat za významný.



Obr. 36: Výsledky použití analýzy hlavních komponent na data charakterizující počty diagnóz v neznámém stádiu v okresech ČR v letech 1976-2010

6.4.3 Mapová koncepce analýzy hlavních komponent

I v případě hodnocení výsledků analýzy hlavních komponent bylo třeba navrhnout nejvhodnější způsob, jak prezentovat, jak se hodnocené prostorové jednotky liší ve smyslu komponentního skóre hlavních komponent. Tento úkol byl náročnější než prezentace shlukových map, vzhledem k relativně obtížné interpretaci výsledků PCA.

Cílem bylo opět vytvořit co nejjednodušší prezentaci, ze které budou zřejmé všechny důležité souvislosti. Jedním z nejčastěji užívaných přístupů je tvorba série map, z nichž každá ilustruje komponentní skóre pro jednu komponentu (SCHROEDER, 2010). Komponentní skóre zde definuje pozici pozorované jednotky v atributovém prostoru ve smyslu vzdálenosti od osy hlavních komponent. Tento přístup byl zvolen i pro vysvětlení variability dat o diagnózách v neznámém stádiu. Na základě jednotlivých hodnot komponentního skóre (eigenvalues), generovaných z programu R, byly vytvořeny tři mapy pro první tři hlavní komponenty. Jelikož je zde hodnocena právě pozice v atributovém prostoru, je zřejmé, že hodnoty nabývají kladných i záporných hodnot. Pro vzájemnou srovnatelnost musely být intervaly komponentního skóre pro všechny komponenty stejné.

Barevné stupnice pro mapy byly generovány pomocí ColorBrewer 2 (ColorBrewer 2.0 [on-line]). Barvy byly zvoleny tak, aby byly vhodné i pro barvoslepé čtenáře mapy. Byla zvolena symetrická divergentní barevná stupnice v kombinaci barev fialové (záporné hodnoty) a oranžové (kladné hodnoty). Tím bylo zajištěno, aby byly na první pohled patrné kladné a záporné hodnoty komponentního skóre pro každý okres a tím byla usnadněna interpretace.

Jak již bylo uvedeno výše, výsledné komponenty vznikají jako lineární kombinace původních hodnot. Z toho důvodu byla série map pro jednodušší interpretaci doplněna grafy pro jednotlivé komponenty, jež vyjadřují vektory (zátěže), jimiž jednotlivé proměnné přispívají do jednotlivých hlavních komponent. Tyto grafy byly vytvořeny v Microsoft Excel a následně byly implementovány do ArcMap 10.3.1, kde vhodně doplnili celkovou koncepci mapy. Aby grafy nerušily barevný koncept, byl pro ně zvolen stejný odstín fialové barvy, jako nejtmavší fialový odstín intervalu komponentního skóre.

Význam jednotlivých sloupců grafů je z důvodu úspory prostoru shrnut hromadně v levém dolním rohu nad tiráží

7 ZÁVĚR

V dnešní době, kdy má věda a výzkum v oblasti medicíny a konkrétně onkologie nezastupitelný význam a díky neúnavné práci výzkumníků je činěn významný pokrok, je paradoxní, že incidence rakovinných onemocnění každým rokem roste. Nezpochybnitelný vliv má na to pochopitelně prodlužující se délka lidského života, se kterou roste pravděpodobnost rozvoje nádorového onemocnění. Speciálním případem onkologických onemocnění jsou vícečetné zhoubné novotvary, které mohou postihnout téhož pacienta po ukončení léčby prvního novotvaru, a které jsou také předmětem této diplomové práce. S tím je také spjata nutnost o rizicích informovat nejen širokou veřejnost, ale také politické aktéry, kteří jsou činní v oblasti veřejného zdraví. Nejen včasná a dostatečná informovanost populace, ale i výzkum šíření chorob v závislosti na různých faktorech prostředí, může mít vliv zavedení vhodných preventivních opatření. Pro tyto účely může být velmi vhodná kartografie, která umožňuje nejen prezentaci „hrubých“, nezpracovaných dat, ale také jejich podrobnou exploraci a odhalování hlubších souvislostí. Právě díky kartografii a jejím exploračním metodám je možné odhalit v datech prostorové vzory a další místní odlišnosti, díky čemuž je možné zvolit regionálně diferencované plánování v oblasti prevence onkologických onemocnění. Možnostmi využití metod explorační kartografie v oblasti onkologických dat se zabývá i tato diplomová práce.

V kontextu výzkumu šíření onkologických onemocnění v závislosti na prostoru a dalších faktorech nutno podotknout vznik nových významných vědních oborů jako je prostorová epidemiologie či nádorová epidemiologie.

Cílem této práce bylo analyzovat a zhodnotit využívané metody analýzy zdravotních dat o vícečetných zhoubných novotvarech. Práce zjistila jen velmi malé množství prací, které se věnují přímo problematice statistické analýzy dat o vícečetných zhoubných novotvarech ani jejich kartografické exploraci. Významný zdroj představuje monografie *New malignancies among cancer survivors* (CURTIS, FREEDMAN et al., 2006), která popisuje a kvantifikuje riziko vzniku nových malignit mezi více než dvěma miliony přeživších rakoviny v období let 1973-2000 ve Spojených státech amerických.

Dále práce přinesla přehled čtyř různých aplikací, které různým způsobem umožňují analýzu či zobrazení prostorových zdravotních dat.

V další části práce podává vysvětlení pojmu „vícečetné zhoubné novotvary“ a hodnotí a kvantifikuje jejich výskyt v krajích České republiky a vývoj od roku 1976 do roku 2010. Od počátku sledovaného období byl zjištěn rapidní nárůst počtu VZN. Zahrnuto je také porovnání situace mezi muži a ženami. Dále práce hodnotí zastoupení klinických stádií primárních a následných diagnóz. Jelikož jsou zpracovávaná onkologická data rozčleněna na základě mezinárodní klasifikace nemocí, přináší práce také stručné vysvětlení této i dalších klasifikací, které jsou v onkologii používány.

Důraz byl v práci kladen taktéž na vysvětlení metod statistické analýzy onkologických dat, které byly následně využity pro analýzu poskytnutých dat, a které

odhalily v datech zajímavý trend – významný podíl vícečetných novotvarů bylo diagnostikováno v neznámém klinickém stádiu. Tento fenomén byl dále zobrazen pomocí běžných kartografických přístupů a později také analyzován a kvantifikován prostřednictvím teoreticky popsaných analytických metod.

Problematice neznámých klinických stádií diagnóz v České republice a jejich souvislostem s dalšími parametry se věnuje kapitola pátá. Tato kapitola taktéž hodnotí zastoupení neznámých stádií u jednotlivých diagnostických typů (dle mezinárodní klasifikace nemocí) i u všech diagnóz jednotlivě. Výsledek je pak shrnut v Příloze 1.

Praktickou část práce tvoří aplikace vybraných statistických metod pro hlubší analýzu dat o neznámých stádiích. Práce analyzuje data jak metodami základními (korelace), tak i metodami sofistikovanějšími (prostorová autokorelace, analýza hlavních komponent). Výsledkem prostorové autokorelace jsou mapy prostorových shluků, které byly identifikovány za pomoci analytických nástrojů programu GeoDa a jsou publikovány v rámci Příloh 2 – 6. Program GeoDa se ukázal jako silný nástroj pro posouzení existence a statistické významnosti prostorového shlukování.

Zvláštní důraz byl v práci kladen na posouzení vlivu vybraných proměnných na variabilitu výskytu neznámých stádií diagnóz a na redukci dimenzionality vstupního datového souboru. Toho bylo docíleno aplikací analýzy hlavních komponent. Zjištěné závěry byly prezentovány pomocí série map zobrazujících komponentní skóre prvních tří hlavních komponent, které nejlépe vysvětlují celkovou variabilitu souboru. Pro jednoduchou interpretaci byla zvolena co nejjednodušší forma prezentace.

Vzhledem k náročné interpretaci a nutnosti pochopení celého konceptu analýzy hlavních komponent není tato metoda příliš vhodná pro širokou laickou veřejnost. Přesto v sobě tato analýza skrývá obrovský potenciál ve smyslu studování vícerozměrných datových souborů s cílem jeho redukce a nalezení bližších souvislostí. Tyto postupy by v budoucnu mohly být častěji využívány v například epidemiologických analýzách.

Nejen pouhé mapování zdravotního stavu může být významným nástrojem pro šíření informací. Do popředí se v poslední době dostávají metody, které kladou důraz na zjišťování hlubších, na první pohled nepatrných souvislostí. Jisté úskalí lze spatřit v tom, že mnoho využitelných analýz není vhodné použít pro širokou veřejnost. To ale jistě překoná fakt, že tyto metody mohou pomoci odborníkům z řad onkologů, epidemiologů, preventistů a dalších nalézat například na první pohled nezřetelné vztahy mezi různými diagnózami navzájem nebo mezi diagnózami a vybranými faktory životního prostředí. V tomto směru lze také spatřovat další možný vývoj problematiky.

Z hlediska výzkumu diagnóz v neznámém klinickém stádiu pomocí dostupných analytických funkcí by jistě bylo zajímavé právě zapojení názoru specialistů. Díky tomu by bylo možné nahlédnout na zjištěné závěry nejen z pohledu kartografického, ale také z epidemiologického či medicínského.

SEZNAM POUŽITÝCH ZDROJŮ

Tištěná literatura

BÁČOVÁ, R. (2012): Možnosti využití metod prostorové analýzy pro zpracování zdravotních dat. Diplomová práce. Masarykova univerzita v Brně. 77 s. (diplomová práce).

BÁČOVÁ, R., KUBÍČEK, P., KONEČNÝ, M. (2013): Příklady využití kartografické vizualizace nádorových onemocnění v Česku. Informace ČGS, 32, č. 2, s. 1-12.

BARCELLOS, C. (2001): The specificities of spatial health data analysis. Cadernos de Saúde Pública, 17, n. 5, p. 1079-180.

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

DOS SANTOS SILVA. I. (1999): Cancer epidemiology principles and methods. 2nd ed. Lyon: IARC. ISBN 978-92-832-0405-3, 441 p.

ELLIOT, P., WARTENBERG, D. (2004): Spatial Epidemiology: Current Approaches and Future Challenges. Environmental health perspectives, 112, n. 9, p. 998-1006.

GRIFFITH, d., ARBIA, G (2010) cit. podle MAREK, L. (2015): Prostorové a vícerozměrné statistické analýzy epidemiologických dat. Dizertační práce. Univerzita Palackého v Olomouci. ISBN 978-80-244-4820-6, 169 s.

CHRÁSKA, M. (2000): Základy výzkumu v pedagogice. Olomouc: Univerzita Palackého.

LUKÁŠOVÁ, K., (2012): Metoda hlavních komponent v klasifikaci shluků patogenů lýkožrouta smrkového (Ips typographus; Coleoptera: Curculionidae). Závěrečná práce. Univerzita Pardubice. 25 s.

MAREK, L. (2015): Prostorové a vícerozměrné statistické analýzy epidemiologických dat. Dizertační práce. Univerzita Palackého v Olomouci. ISBN 978-80-244-4820-6, 169 s.

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

MEZINÁRODNÍ STATISTICKÁ KLASIFIKACE NEMOCÍ A PŘIDRUŽENÝCH ZDRAVOTNÍCH PROBLÉMŮ: MKN-10: DESÁTÁ REVIZE: AKTUALIZOVANÁ DRUHÁ VERZE K 1.1.2009 (2008): 2., aktualiz. vyd. Praha: Bomton Agency, ISBN 978-80-904259-0-3.

RUMSEY, D. J. (2011): Statistics For Dummies. 2nd ed. Hoboken, N.J.: Wiley, ISBN 978-0-470-91108-2, 384 p.

SPURNÁ, P. (2008): Prostorová autokorelace – všudypřítomný jev při analýze prostorových dat? Sociologický časopis, 44, č. 4, s. 22.

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

TOBLER, W. R. (1970): A computer movie simulating urban growth in the Detroit region. Economic Geography,46, č. 332, s. 234-240.

Elektronické zdroje

ABOUT SEER. 2016 Surveillance, Epidemiology, and End Results Program. [on-line] Dostupné z: http://seer.cancer.gov/registries/.

ANDRIENKO, G., ANDRIENKO, N., SAVINOV, A., (2001): Choropleth maps: Classification revisited. Dostupné z: http://geoanalytics.net/and/papers/ica01.pdf.

ANSELIN, L. (1995): Local Indicator of Spatial Association-LISA. Geographical Analysis 27, No. 2, p. 93-115. Dostupné z: http://isites.harvard.edu/fs/docs/icb.topic868440.files/Anselin1995%20LISA.pdf.

ANSELIN, L. (2005): Exploring Spatial Data with Geoda [on-line]. Dostupné z: http://www.csiss.org/clearinghouse/GeoDa/geodaworkbook.pdf.

ANSELIN, L., SYABRI, I., SMIRNOV, O., (2002): Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows. Dostupné z: https://geodacenter.asu.edu/pdf/multi_lisa.pdf.

BEDÁŇOVÁ, I., VEČEREK, V., (2007): Základy statistiky pro studující veterinární medicíny a farmacie: skripta [on-line]. Brno: Veterinární a farmaceutická univerzita. Dostupné z: http://cit.vfu.cz/statpotr/POTR/Skripta.pdf.

BEDNÁŘ, J. (2006): Testování statistických hypotéz. [on-line]. Dostupné z: file:///C:/Users/PC/Downloads/M4Testhypotez.pdf.

BRANDUSESCU, A., R. SIEBER, N. SCHUURMAN (2011): The use of geovisualization to public health, in the context of open source applications and digital earths: an effective representation? [on-line]. Dostupné z: http://rose.geog.mcgill.ca/ski/system/files/fm/2011/ABrandusescu.pdf.

COLORBREWER 2.0 [on-line]. Color Brewer: Color Advice for Maps. Dostupné z: http://www.colorbrewer2.org.

CURTIS, R. E., FREEDMAN D. M., R. E, RIES L. A. G., HACKER, D. G., EDWARDS, B. K., TUCKER, M. A., FRAUMENI, J. F. Jr. (eds), (2006). New Malignancies Among Cancer Survivors: SEER Cancer Registries, 1973-2000. National Cancer Institute, NIH Publ. No. 05-5302. Bethesda, MD. Dostupné z: http://seer.cancer.gov/archive/publications/mpmono/MPMonograph_complete.pdf.

ČERBA, O. (2007): Kvalitativní areály. [online prezentace]. Plzeň: Západočeská univerzita. Dostupné z WWW: http://old.gis.zcu.cz/studium/tka/Slides/kvalitativni_arealy.pdf.

DOBROVOLNÝ, P. (2015): Statistická analýza plošných jevů. Prezentace k přednáškám z geostatistiky, [on-line prezentace]. Dostupné z: https://is.muni.cz/auth/el/1431/jaro2016/Z6101/um/39007348/Geostatistika_2015_08_sa_II.pdf.

EPI INFO. (2014): Epi Info ™ - Community Edition [on-line]. Dostupné z: https://epiinfo.codeplex.com/.

EPI INFO. (2015): Introducing Epi Info™ 7 [on-line]. Dostupné z: https://wwwn.cdc.gov/epiinfo/index.htm.

FELLER, L. a LEMMER, J. (2012): New 'second primary' cancers [on-line]. Dostupné z: http://www.ncbi.nlm.nih.gov/pubmed/23198353.

FREISL, M. (2014): Pravděpodobnost a statistika hypertextově. [on-line]. Dostupné z: http://home.zcu.cz/~friesl/hpsb/phodn.html.

GEOSPATIAL ANALYSIS – 5TH EDITION. de Smith, M. J., Goodchild, Longley. 2015 [on-line]. Dostupné z: http://www.spatialanalysisonline.com/HTML/index.html?eda__esda_and_estda.htm.

GERYK, E., P. KUBÍČEK, R. ŠTAMPACH a kol. (2008): Vícečetné zhoubné novotvary: Ukazatel zdraví a nákladů péče v onkologii. Zdravotnictví v České republice. Praha: Asociace pro rozvoj sociálního lékařství a řízení péče o zdraví, 2, č. 11, s. 50-55. ISSN 1213-6050. Dostupné z: http://www.zdravcr.cz/archiv/zcr-2-2008.pdf.

GERYK, E., P., DÍTĚ, M., PEŠEK, J., KOZEL (2009): Následné primární novotvary u 125 262 onkologicky nemocných v České republice 1976–2005. Onkologie, 3, č. 3, s. 181-189. Dostupné z: http://www.solen.cz/savepdfs/xon/2009/03/10.pdf.

GERYK, E., P., DÍTĚ, J., KOZEL, J., TRNA, M., KONEČNÝ (2010): Klinická stadia u nemocných s vícečetnými novotvary. Onkologie, 4, č. 6, s. 357-361. Dostupné z: http://www.onkologiecs.cz/pdfs/xon/2010/06/09.pdf.

GLOSSARY OF KEY TERMS [on-line]. Dostupné z: https://geodacenter.asu.edu/node/390.

GORE, A. (1998): The Digital Earth: Understanding our planet in the 21st Century [on-line]. Dostupné z: http://www.hunagi.hu/G/pub/Globalis/DE-AlGore.pdf.

GOOVAERTS, P. (2010): Three-dimensional Visualization, Interactive Analysis and Contextual Mapping of Space-time Cancer Data [on-line]. Dostupné z: http://www.agile-online.org/Conference_Paper/CDs/agile_2010/ShortPapers_PDF/89_DOC.pdf.

HEALTHMAP (2015): About [on-line]. Dostupné z: http://www.healthmap.org/site/about.

HOLČÍK, J., KOMENDA, M. a kol., (2015): Matematická biologie: e-learningová učebnice [on-line]. 1. vydání. Brno: Masarykova univerzita. ISBN 978-80-210-8095-9. Dostupné z: http://portal.matematickabiologie.cz/.

HOŠKOVÁ, P., (2006): Matematická statistika II. Přednášky z matematické statistiky [on-line]. Dostupné z: http://pef-info.wz.cz/download/MSIIa_prednasky.pdf.

INTERNATIONAL SOCIETY FOR DIGITAL EARTH, (2014): Statute 2 – Definition and Vision [on-line]. Dostupné z: http://www.digitalearth-isde.org/statues/110.

KONEČNÝ, M. (2015): Digital Earth concepts [on-line prezentace]. Dostupné z: https://is.muni.cz/auth/el/1431/podzim2015/Z8121/um/Digital_Earth_Concepts.ppt?studium=694928.

MACEACHREN, A. M., BREWER, C. A., PICKLE, L. W. (1998): Visualizing georeferenced data: reprezenting reliability of health statistics [on-line]. Environment and Planning A vol. 30, p. 1547-1561. Dostupné z: http://www.geovista.psu.edu/publications/MacEachren/MacEachren_Visualizing_98.pdf.

MALÝ, M., (2015): Ilustrační příklad odhadu LRM v SW Gretl. Studijní podklady [on-line]. Praha: Česká zemědělská fakulta. Dostupné z: http://pef.czu.cz/~maly/Odhad%20LRM.pdf.

MATEMATICKÝ SOFTWARE R. (2009): LinuxEXPRES [on-line]. Dostupné z: http://www.linuxexpres.cz/software/matematicky-software-r-s-nim-je-kazda-statistika-hezci.

MELOUN, M., MILITKÝ, J. (2002): Kompendium statistického zpracování dat: metody a řešené úlohy včetně CD. Vyd. 1. Praha: Academia. ISBN 80-200-1008-4.

MELOUN, M., (2011): Počítačová analýza vícerozměrných dat v oborech přírodních, technických a společenských věd. Skripta [on-line]. Pardubice: Univerzita Pardubice. Dostupné z: http://www.crr.vutbr.cz/system/files/brozura_05_1106.pdf.

SUBSEQUENT NEOPLASMS. (2016) National Cancer Institute [online]. Dostupné z: http://www.cancer.gov/cancertopics/pdq/treatment/lateeffects/HealthProfessional/page2.

PICKLE, L. W., (2003): Usability Testing of Map Designs [on-line]. Dostupné z: http://www.galaxy.gmu.edu/interface/I03/I2003Proceedings/PickleLinda/PickleLinda.paper.pdf.

SCHROEDER, J. P., (2010): Bicomponent Trend Maps: A Multivariate Approach to Visualizing Geographic Time Series. Cartography and Geographic Information Science, 37(3), p. 169–187. Dostupné z: http://www.tandfonline.com/doi/abs/10.1559/152304010792194930.

ŠTĚPÁNOVÁ, R. (2011): DISEASE MAPPING: Renální karcinom v ČR zkoumaný metodami disease mappingu. Diplomová práce. Masarykova univerzita v Brně. Dostupné z: https://is.muni.cz/auth/th/184531/prif_m/Stepanova_diplomova_prace.pdf.

TNM KLASIFIKACE ZHOUBNÝCH NOVOTVARŮ (2013), 7. Vydání (originál 2011) [on-line]. Dostupné z: http://www.uzis.cz/book/export/html/259.

TNM SYSTÉM/TNM KLASIFIKACE. (2015) Linkos [on-line]. Dostupné z: http://www.linkos.cz/slovnicek/tnm-system-tnm-klasifikace/.

ÚZIS (2013): Zemřelí 2012 [on-line] Praha: Ústav zdravotnických informací a statistiky České republiky. ISSN 1210-9967. Dostupné z: www.uzis.cz/system/files/demozem2012.pdf.

VELKÝ LÉKAŘSKÝ SLOVNÍK (2008): Velký lékařský slovník [on-line]. Dostupné z: http://lekarske.slovniky.cz/pojem/morbidita.

VELKÝ LÉKAŘSKÝ SLOVNÍK (2008): Velký lékařský slovník [on-line]. Dostupné z: http://lekarske.slovniky.cz/pojem/mortalita.

WALLER, L. A., C. A. GOTWAY (2004): Applied spatial statistics for public health data. 1st ed. Hoboken: Wiley-Interscience. ISBN 0-471-38771-1, 494 p.

SEZNAM POUŽITÝCH ZKRATEK

AJAX – Asynchronous JavaScript and XML

API – Application Programming Interface

CDC – Centers for Disease Control and Prevention

CIESIN – Center for International Earth Science Information Network

CSV – Comma-separated Values

DOD – Department of Defence

ECDC – European Centre for Disease Prevention and Control)

ESDA – Exploratory Spatial Data Analysis

FAO – Food and Agriculture organization

HHS – Health and Human Services

KML – Keyhole Markup Language

KMZ – Keyhole Markup Zip Format

MySQL – databázový systém vytvořený švédskou firmou MySQL AB

NCI – National Cancer Institute

OIE – World Organisation for Animal Health

PCA – Principal Component Analysis, analýza hlavních komponent

PC1 – 1. hlavní komponenta

PC2 – 2. hlavní komponenta

PC3 – 3. hlavní komponenta

PCPlot – Paralel Coordinate Plot

PHP – Hypertext Preprocessor

PNG – Portable Network Graphics

PDF – Portable Document Format

VZN – Vícečetné zhoubné novotvary

WHO – World Health Organization

ZN – Zhoubné novotvary

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM PŘÍLOH

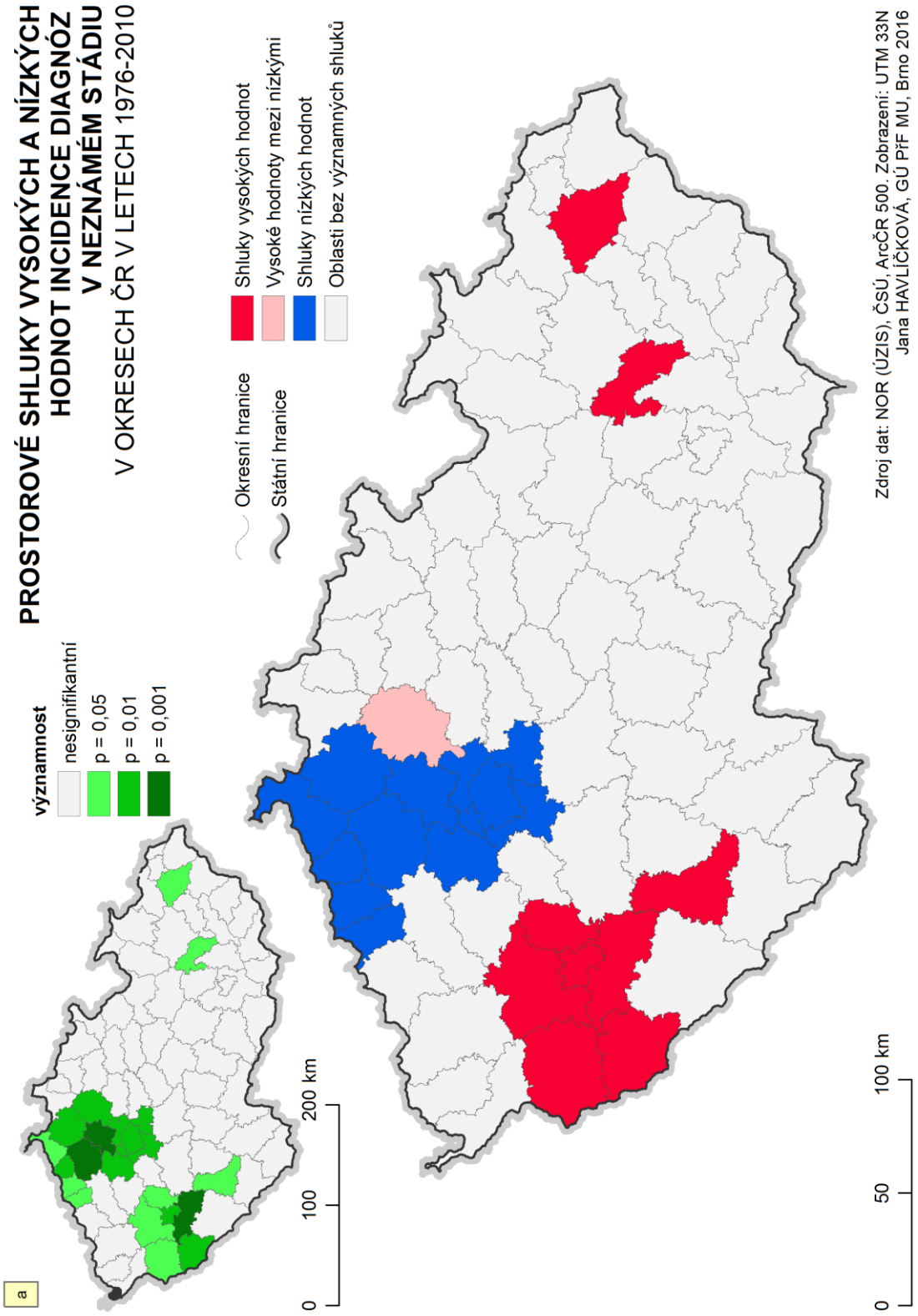
Příloha 1: Procentuální zastoupení neznámých klinických stádií u jednotlivých diagnóz (C00-C97) v letech 1976-2010 (Zdroj: NOR (ÚZIS))

Kód	Neznámé	Celkem	Podíl [%]	Novotvar
C39	65	65	100,00	ZN jiných a nepřesně určených lokalizací v dýchací soustavě a nitrohrudních orgánech
C93	63	63	100,00	Monocytická leukemie
C95	205	206	99,51	Leukémie neurčeného buněčného typu
C78	1145	1153	99,31	Sekundární ZN dýchací a trávicí soustavy
C92	1473	1492	98,73	Myeloidní leukemie
C94	74	75	98,67	Jiné leukemie určených buněčných typů
C96	93	95	97,89	Jiné zhoubné novotvary mízní, krvetvorné a příbuzné tkáně
C77	600	614	97,72	Sekundární a neurčený ZN mízních uzlin
C72	83	85	97,65	ZN míchy, mozkových nervů a jiných částí centrální nervové soustavy
C79	735	753	97,61	Sekundární ZN jiných a neurčených lokalizací
C80	1295	1361	95,15	ZN bez určení lokalizace
C88	114	120	95,00	Maligní imunoproliferativní nemoci
C90	1488	1601	92,94	Mnohočetný myelom a plazmocytární novotvary
C91	3751	4039	92,87	Lymfoidní leukemie
C74	153	165	92,73	ZN nadledviny
C70	89	96	92,71	ZN mozkomíšních plen
C75	47	51	92,16	ZN jiných žláz s vnitřní sekrecí a příbuzných struktur
C26	410	446	91,93	ZN novotvar jiných a nepřesně určených trávicích orgánů
C58	11	12	91,67	ZN placenty
C71	1324	1460	90,68	ZN mozku
C76	445	498	89,36	ZN jiných a nepřesně určených lokalizací
C97	14	16	87,50	ZN mnohočetných samostatných lokalizací
C37	93	109	85,32	ZN brzlíku
C46	84	100	84,00	Kaposiho sarkom
C38	298	367	81,20	ZN srdce, mezihrudí a pohrudnice
C48	307	401	76,56	ZN retroperitonea a peritonea
C55	142	186	76,34	ZN dělohy
C68	257	337	76,26	ZN jiných a neurčených močových orgánů
C63	114	153	74,51	ZN jiných a neurčených mužských pohlavních orgánů
C45	115	155	74,19	Mezoteliom
C41	171	246	69,51	ZN kosti a kloubní chrupavky a jiných neurčených lokalizací
C30	139	204	68,14	ZN nosní dutiny a středního ucha
C22	1529	2416	63,29	ZN jater a intrahepatálních žlučových cest
C40	107	170	62,94	ZN kloubní chrupavky končetin
C69	384	613	62,64	ZN oka a očních adnex
C14	50	80	62,50	ZN jiných a nepřesně určených lokalizací rtu, ústní dutiny a hltanu

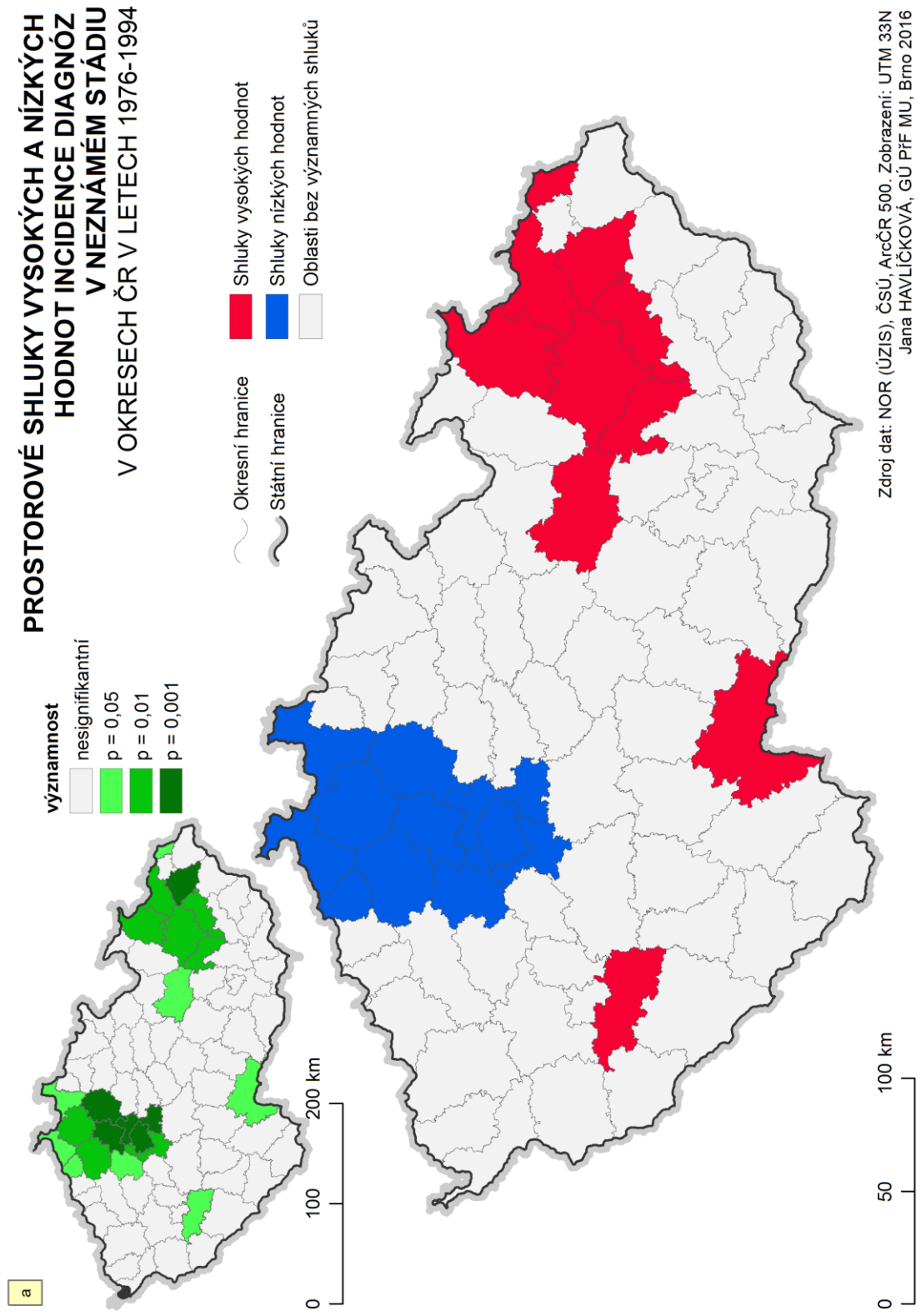
Kód	Neznámé	Celkem	Podíl [%]	Novotvar
C24	761	1247	61,03	ZN jiných a neurčených částí žlučových cest
C85	829	1411	58,75	Non-Hodgkinův lymfom, jiných a neurčených typů
C57	201	346	58,09	ZN jiných a neurčených ženských pohlavních orgánů
C47	32	57	56,14	ZN periferních nervů a autonomní nervové soustavy
C62	600	1078	55,66	ZN varlete
C23	1166	2113	55,18	ZN žlučníku
C17	398	725	54,90	ZN tenkého střeva
C82	561	1035	54,20	Folikulární lymfom
C25	2973	5487	54,18	ZN slinivky břišní
C84	210	388	54,12	Lymfom ze zralých T/NK buněk
C31	95	185	51,35	ZN vedlejších dutin
C07	358	751	47,67	ZN příušní žlázy
C83	1175	2566	45,79	Non-folikulární lymfom
C08	101	221	45,70	ZN jiných a neurčených slinných žláz
C49	599	1316	45,52	ZN jiné pojivové a měkké tkáně
C33	87	193	45,08	ZN průdušnice
C81	587	1340	43,81	Hodgkinův lymfom
C60	183	423	43,26	ZN pyje
C73	1110	2580	43,02	ZN štítné žlázy
C64	5297	12949	40,91	ZN ledviny mimo pánvičku
C21	242	592	40,88	ZN řiti a řitního kanálu
C61	7412	18436	40,20	ZN předstojné žlázy - prostaty
C65	507	1274	39,80	ZN ledvinné pánvičky
C67	6132	15785	38,85	ZN močového měchýře
C66	183	538	34,01	ZN močovodu
C16	2754	8161	33,75	ZN žaludku
C15	486	1488	32,66	ZN jícnu
C44	60727	187636	32,36	Jiný zhoubný novotvar kůže
C52	120	372	32,26	ZN pochvy
C11	72	272	26,47	ZN nosohltanu
C06	68	266	25,56	ZN jiných a neurčených částí úst
C00	423	1713	24,69	ZN rtu
C34	5381	21943	24,52	ZN průdušky
C54	2367	9692	24,42	ZN těla děložního
C20	2186	9733	22,46	ZN konečníku
C51	238	1092	21,79	ZN vulvy
C10	61	283	21,55	ZN ústní části hltanu
C56	1157	5715	20,24	ZN vaječníku
C13	63	320	19,69	ZN hypofaryngu
C02	118	612	19,28	ZN jiných a neurčených částí jazyka
C53	850	4694	18,11	ZN hrdla děložního
C19	939	5225	17,97	ZN rektosigmoideálního spojení

Kód	Neznámé	Celkem	Podíl [%]	Novotvar
C18	4043	22742	17,78	ZN tlustého střeva
C12	18	106	16,98	ZN pyriformního sinu
C03	33	216	15,28	ZN dásně
C43	1451	10032	14,46	Zhoubný melanom kůže
C01	55	387	14,21	ZN kořene jazyka
C09	134	944	14,19	ZN mandle
C05	35	276	12,68	ZN patra
C32	385	3562	10,81	ZN hrtanu
C04	51	476	10,71	ZN ústní spodiny
C50	3017	29337	10,28	ZN prsu

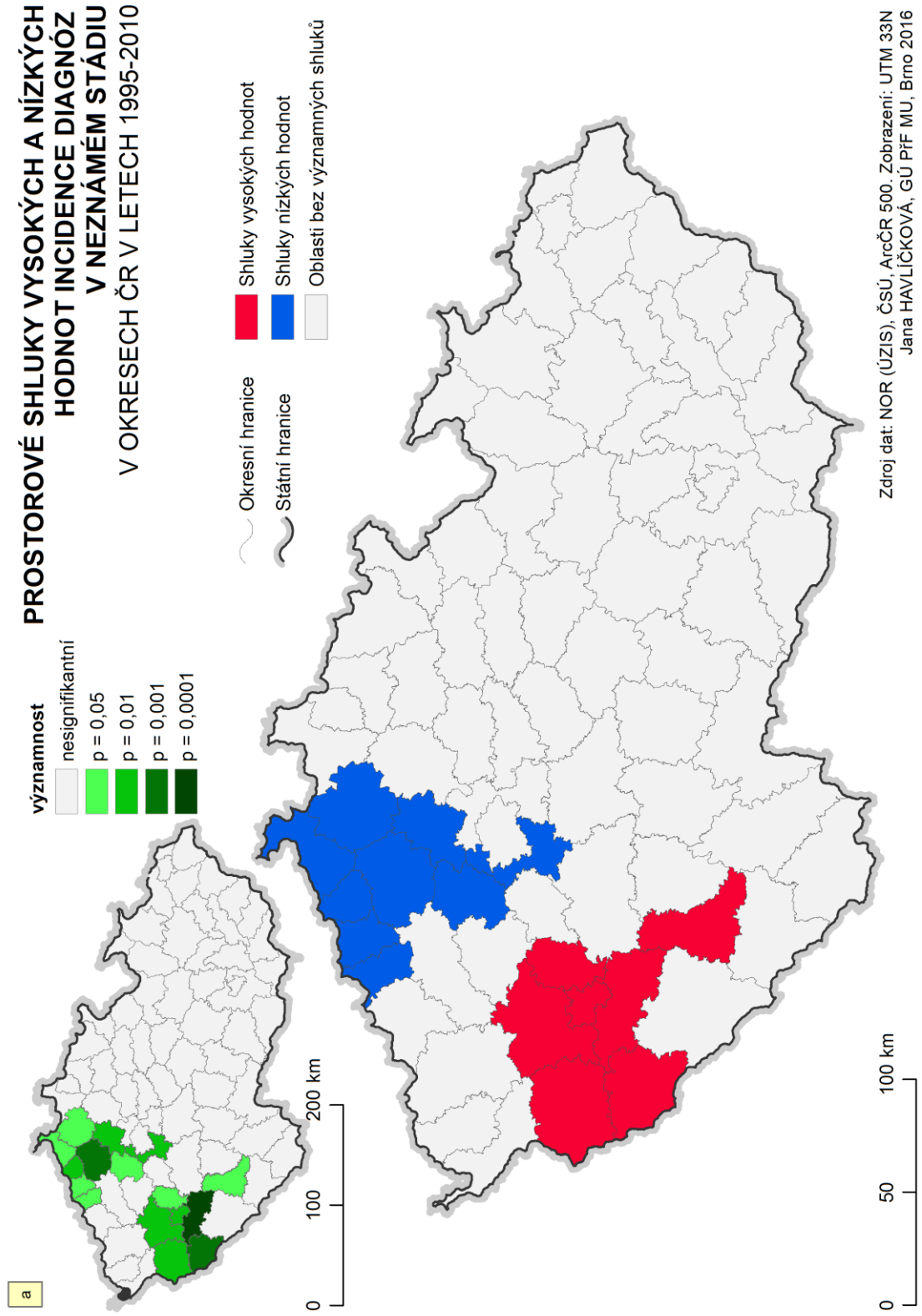
Příloha 2: Prostorové shluky vysokých a nízkých hodnot incidence diagnóz v neznámém stádiu v okresech ČR v letech 1976-2010



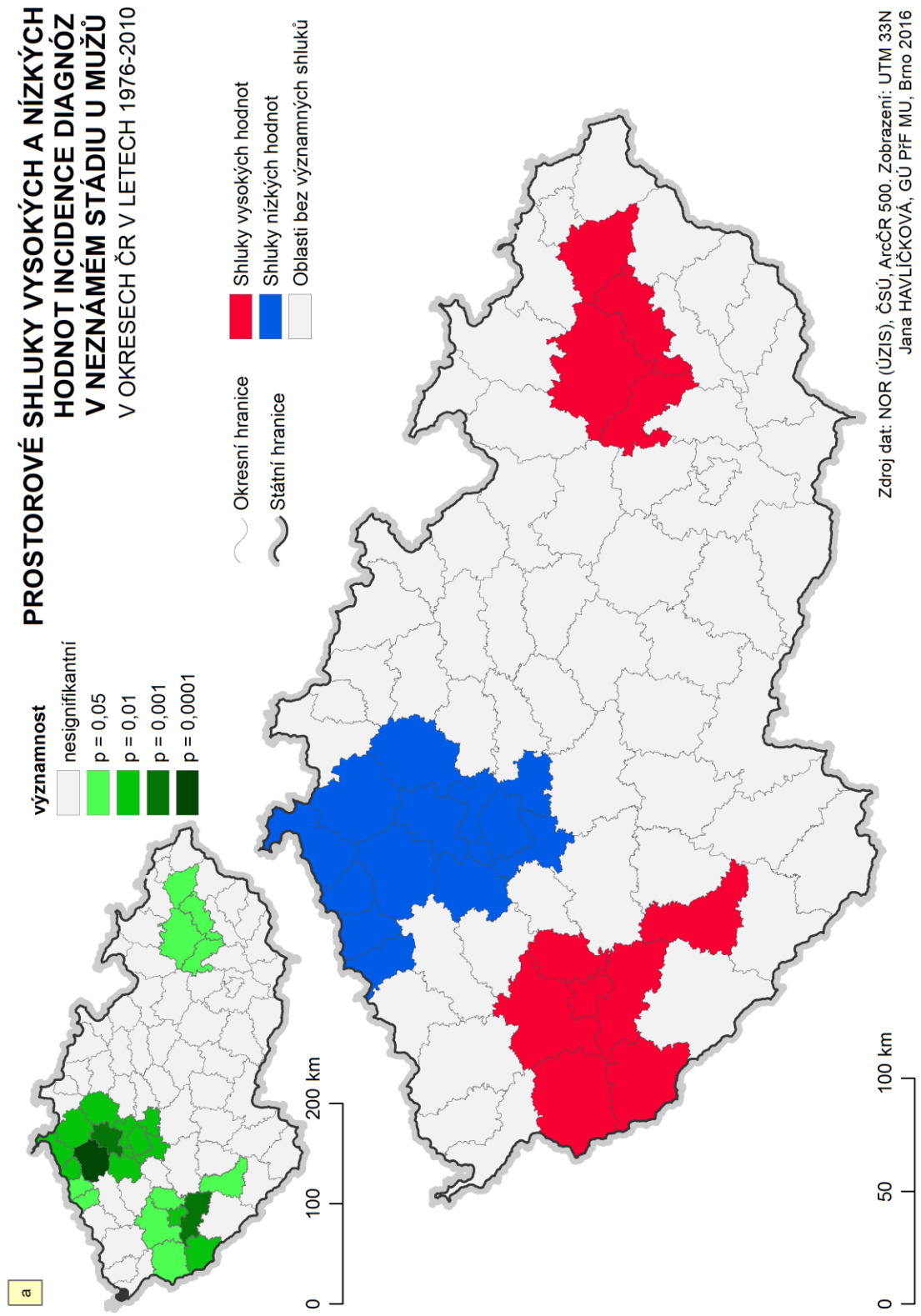
Příloha 3: Prostorové shluky vysokých a nízkých hodnot incidence diagnóz v neznámém stádiu v okresech ČR v letech 1976-1994



Příloha 4: Prostorové shluky vysokých a nízkých hodnot incidence diagnóz v neznámém stádiu v okresech ČR v letech 1995-2010



Příloha 5: Prostorové shluky vysokých a nízkých hodnot incidence diagnóz v neznámém stádiu u mužů v okresech ČR v letech 1976-2010



Příloha 6: Prostorové shluky vysokých a nízkých hodnot incidence diagnóz v neznámém stádiu u žen v okresech ČR v letech 1976-2010