UNIVERZITA PARDUBICE
FAKULTA EKONOMICKO-SPRÁVNÍ
ÚSTAV MATEMATIKY A KVANTITATIVNÍCH METOD

METODY SNÍŽENÍ DIMENZE PRO MODELOVÁNÍ STAVU ZDRAVÍ

DISERTAČNÍ PRÁCE

Autor: Ing. Lucie Zapletalová

Školitel: prof. RNDr. Viera Pacáková, Ph.D.

Pardubice 2022

UNIVERSITY OF PARDUBICE
FACULTY OF ECONOMICS AND ADMINISTRATION
INSTITUTE OF MATHEMATICS AND QUANTITATIVE METHODS

DIMENSION REDUCTION METHODS FOR MODELLING HEALTH STATUS

DISERTAČNÍ PRÁCE

Author: Ing. Lucie Zapletalová

Supervisor: prof. RNDr. Viera Pacáková, Ph.D.

Pardubice 2022

Abstrakt

Stav zdraví patří mezi ukazatele měřící kvalitu života. Zdravotní stav populace je v rámci Evropské Unie dlouhodobě sledován a jsou přijímány strategie vedoucí k jeho zlepšení například na úrovni EU, jednotlivých států, popř. jejich regionů. Existuje celá řada determinantů stavu zdraví, které stav zdraví ovlivňují. Databáze Eurostatu, OECD a WHO poskytují rozsáhlé datové soubory obsahující jak ukazatele stavu zdraví, tak jeho determinantů v agregované formě na celostátní, popř. regionální úrovni. Problémem těchto databází jsou chybějící, popř. zastaralé hodnoty ukazatelů. Jedná se především o determinanty stavu zdraví, popř. ukazatele týkající se starších osob. Vzhledem k tomu, že ukazatele stavu zdraví a jeho determinantů jsou vícerozměrná data, je nutné snížit jejich rozměrnost pro další využití při modelování stavu zdraví a jeho determinantů. Hlavním cílem disertační práce je porovnání a vyhodnocení výsledků lineárních a nelineárních technik pro snížení rozměrnosti ukazatelů stavu zdraví a jejich determinantů v zemích EU-27 a využití takto předzpracovaných dat k posouzení nerovností ve stavu zdraví a identifikování skupin států s podobnou, resp. rozdílnou úrovní stavu zdraví. V této disertační práci je pomocí metod pro snížení rozměrnosti ukazatelů, metod shlukové analýzy a hybridního přístupu snížena rozměrnost použitých ukazatelů, jsou posouzeny nerovnosti ve stavu zdraví a jeho determinantech a jsou nalezeny státy EU-27 s podobnou úrovní stavu zdraví a jeho determinantů. Tyto státy jsou následně lineárně uspořádány. Výsledky jsou přehledně vizualizovány pomocí různých možností vizualizace včetně využití geografických dat. Nakonec jsou získané výsledky porovnány s již publikovanými.

Klíčová slova

Stav zdraví, determinanty stavu zdraví, snížení rozměrnosti ukazatelů, shlukování objektů, nerovnosti ve stavu zdraví, lineární uspořádání států

Abstract

Health status belong among indicators measuring quality of life. Health status of population is monitored within European Union for a long time and strategies leading to improving are accepted at the EU, individual countries or their regions levels. Many determinants exist which influence health status. Eurostat, OECD and WHO databases provide large datasets containing indicators of health status and its determinants in aggregated form at national or regional level. Missing or outdated values of indicators are problems of these databases. Determinants of health status or indicators related to older people pose this problem. Whereas, the health status indicators and their determinants are multidimensional categories and therefore the reducing their dimensionality is necessary for further use in the health status and its determinants modelling. The main goal of this Ph.D. thesis is comparison and evaluation of the results of linear and non-linear techniques for reducing dimensionality of health status indicators and their determinants in EU-27 countries and using such pre-processed data to assess inequalities in health status and identifying groups of states with similar or different levels of health status. In this Ph.D. thesis, the dimensionality of the used indicators is reduced, inequalities in the health status and its determinants are assessed and countries of EU-27 with similar levels of health status and its determinants are found by using methods for reducing dimensionality of indicators, methods of cluster analysis and a hybrid approach. After that these countries are arranged linearly. The results are clearly visualized by using different visualisation options including geographic data. Finally, the obtained results are compared with already published.

Keywords

Health status, determinants of health status, reducing dimenzionality of indicators, clustering of objects, health inequalities, linear arrangement of countries

Prohlašuji:

Práci s názvem Metody snížení dimenze pro modelování stavu zdraví jsem vypracovala samostatně. Veškeré literární prameny a informace, které jsem v práci využila, jsou uvedeny v seznamu použité literatury.

Byla jsem seznámena s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Beru na vědomí, že v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 7/2019 Pravidla pro odevzdávání, zveřejňování a formální úpravu závěrečných prací, ve znění pozdějších dodatků, bude práce zveřejněna prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne

Ing. Lucie Zapletalová.
(rozená Kopecká)

Poděkování:

Poděkování patří paní prof. RNDr. Viere Pacákové, Ph.D. za odborné vedení práce, rady a připomínky v průběhu studia. Dále bych ráda poděkovala svým nejbližším za podporu během celé doby studia.

OBSAH

Seznam obrázků

Seznam tabulek

Seznam použitých zkratek

AP	anti-vzor (objekt)
CAs	onkologická onemocnění
CADs	kardiovaskulární onemocnění
CDR	hrubá míra úmrtnosti
COVID-19	onemocnění koronavirus 2019
DALY	ztracená léta života v důsledku nemoci
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
d_i	bezrozměrná agregovaná míra
eps	velikost ε-ového okolí
EU	Evropská unie
EU-27	Evropská unie – 27 zemí (bez Spojeného království)
EU-28	Evropská unie – 28 zemí (se Spojeným královstvím)
EU-SILC	šetření EU o příjmech a životních podmínkách
Eurostat	statistický úřad EU
FA	faktorová analýza
FCM	metoda fuzzy k-průměrů
FSI	fuzzy koeficient siluet
HALE	očekávaná délka života ve zdraví
HIV	Human Immunodeficiency Virus
HLY_0	zdravé roky života při narození
k-NN	k-nejbližšího souseda (vzdálenost)
kPC	kernel hlavní komponenta
KPCA	kernel analýza hlavních komponent
LE_0	střední délka života při narození
LE_65	střední délka života ve věku 65 let
MDS	vícerozměrné škálování
minPts	minimální počet pozorování v oblasti určené velikostí ε-ového okolí
MKN	Mezinárodní klasifikace nemocí
MSA	dílčí míra adekvátnosti pro jednotlivé proměnné
NUTS 2	Nomenklatura územních statistických jednotek (region soudržnosti)
OECD	Organizace pro hospodářskou spolupráci a rozvoj
P	vzor (objekt)
PC	hlavní komponenta
PCA	analýza hlavních komponent
PC(k)	rozdělovací koeficient
PE	koeficient entropie rozkladu

PHI	index zdraví populace
MPC	modifikovaný rozdělovací koeficient
RBF	radiálně bázická funkce
RC	rotovaná komponenta
SDR	standardizovaná míra úmrtnosti
SPC	řídká hlavní komponenta
SPCA	řídká analýza hlavních komponent
SS_B	mezishlukový součet čtverců
SS_T	celkový součet čtverců
SVM	metoda podpůrných vektorů
WHO	Světová zdravotnická organizace

Seznam použitých zkratek států

AT	Rakousko
BE	Belgie
BG	Bulharsko
CY	Kypr
CZ	Česká republika
DE	Německo
DK	Dánsko
EE	Estonsko
EL	Řecko
ES	Španělsko
FI	Finsko
FR	Francie
HR	Chorvatsko
HU	Maďarsko
IE	Irsko
IT	Itálie
LT	Litva
LU	Lucembursko
LV	Lotyšsko
MT	Malta
NL	Nizozemsko
PL	Polsko
PT	Portugalsko
RO	Rumusnko
SE	Švédsko
SI	Slovinsko
SK	Slovenská republika

Úvod

Stav zdraví evropské populace je možné popsat mnoha ukazateli, což z něho činí vícerozměrnou kategorii. Vnímání stavu zdraví mapované od starověku po dnešek se zásadně změnilo. Dobré zdraví je to nejdůležitější, co každý člověk může vlastnit. Jedná se o největší bohatství každého z nás. Nicméně udržet si dobré zdraví, popřípadě řešit následky nemoci či invalidity, stojí nemalé finanční prostředky jak jednotlivce a domácnosti, tak i veřejné rozpočty. Německý filozof Arthur Schopenhauer kdysi řekl: *„Zdraví není vším, ale bez zdraví je všechno ničím.“* A je tomu skutečně tak! Pokud chtějí ekonomiky jednotlivých evropských států prosperovat, musí mít mimo jiné zdravé práceschopné obyvatelstvo. Není proto pouze v zájmu jednotlivců chránit své vlastní zdraví, ale jde o zájem celospolečenský. S přibývajícím věkem výskyt již existujících onemocnění narůstá, ale i nová onemocnění sužující evropskou populaci se objevují napříč generacemi. Mezi nejzávažnější nepřenosná onemocnění dnešní doby postihující evropskou populaci patří: kardiovaskulární onemocnění, rakoviny, diabetes, respirační onemocnění a duševní poruchy. Léčba těchto onemocnění je již dnes pro jednotlivce a veřejné rozpočty nákladná a evropská populace stále více stárne. V průběhu 20. a 21. století bylo vyvinuto mnoho indexů stavu zdraví, jejichž cílem je měřit stav zdraví konkrétních populací. Některé z nich jsou veřejně dostupné a jsou stále aktualizovány.

Vícerozměrnými kategoriemi jsou také determinanty, které stav zdraví populace ovlivňují. Každý z nás se denně setkává s mnoha faktory, které mají menší či větší vliv na naše zdraví, ať už jde o faktory související se životním stylem, dědičné, genderové, kulturní, politické, sociálně ekonomické nebo faktory, které souvisejí s životním prostředím a jeho znečišťováním. V některých případech je v silách jednotlivců redukovat faktory mající negativní dopad na jejich zdraví a posilovat faktory mající dopad pozitivní. K tomu, aby docházelo k individuálním snahám o posílení či zachování dobrého zdraví, je třeba dostatečná motivace, např. prostřednictvím příspěvků v rámci prevence výskytů závažných onemocnění.

Důvody pro věnování se možnostem snižování rozměrnosti ukazatelů stavu zdraví a jeho determinantů jsou především existence reálných datových souborů obsahujících velké množství ukazatelů. Bez použití vhodných analýz pro snížení rozměrnosti ukazatelů stavu zdraví a jeho determinantů dochází především k jejich nepřehlednosti a k obtížné manipulaci s nimi, což způsobuje problémy např. technikám shlukové analýzy, klasifikačním a regresním modelům apod. Existuje celá řada lineárních i nelineárních technik snižování rozměrnosti ukazatelů v datovém souboru. Tyto techniky nemají vést pouze k přehledné vizualizaci

zobrazovaných dat, ale také k úspoře času při trénování modelu, odstranění multikolinearity z modelu atd.

Kromě ukazatelů stavu zdraví, jeho determinantů a současným přístupům snižování rozměrnosti dat se tato disertační práce v prvních dvou kapitolách dále zabývá možnostmi využití datových souborů se zredukovanou dimenzí v rámci metod shlukové analýzy a hybridního přístupu kombinujícího vícerozměrné škálování s lineárním uspořádáním. Výsledky těchto analýz a již publikované výsledky stavu zdraví a jeho determinantů ukazují na vliv použití různých metod pro redukci rozměrnosti ukazatelů a poskytují tak jejich porovnání.

Disertační práce se zabývá také dostupností a získáním vhodných datových souborů obsahujících ukazatele stavu zdraví a jeho determinantů ohodnocujících 27 členských evropských zemí Evropské Unie (EU-27). Dále jsou použité datové soubory vizuálně prozkoumány pomocí metod pro snížení rozměrnosti ukazatelů. Následně jsou tyto státy rozděleny podle stavu zdraví a jeho determinantů pomocí metod shlukové analýzy. Výsledné shluky jsou doplněné o informaci týkající se celkové úrovně stavu zdraví a jeho determinantů v zemích EU-27 pomocí hybridního přístupu. V rámci hybridního přístupu jsou modelovány agregované míry stavu zdraví a jeho determinantů na základě vzdálenosti od uměle vytvořeného ideálního objektu. NUTS 2 regiony (Nomenclature of Units for Territorial Statistics) zemí EU-27 zde nejsou analyzovány z důvodu nízké kvality poskytovaných ukazatelů, především determinantů stavu zdraví (chybějící, neaktuální údaje).

Vzhledem k současnému stavu poznání v řešené problematice a vzhledem k dostupnosti dat v širokých veřejných databázích zní cíl disertační práce následovně: Porovnání a vyhodnocení výsledků lineárních a nelineárních technik pro snížení rozměrnosti ukazatelů stavu zdraví a jejich determinantů v zemích EU-27 a využití takto předzpracovaných dat k posouzení nerovností ve stavu zdraví a identifikování skupin států s podobnou, resp. rozdílnou úrovní stavu zdraví.

1 Stav zkoumání zdraví v zemích Evropy

Pojem zdraví má mnoho definic. Například Světová zdravotnická organizace (WHO – World Health Organization) definuje zdraví od roku 1948 následovně: *„Zdraví je stav úplné fyzické, duševní a sociální pohody, nejen nepřítomnost nemoci nebo vady.“* Jedná se o definici, která kromě fyzického a duševního zdraví, zahrnuje také sociální péči, jelikož zdraví je úzce spojeno se sociálním prostředím, životními a pracovními podmínkami (Svalastog a kol., 2017).

Aby bylo možné lépe pochopit postavení zdraví v současné společnosti, je nezbytné znát vývoj vnímání tohoto pojmu v minulosti, a to od chápání zdraví z hlediska rovnováhy těla a duše až po vnímání zdraví jako ekonomické kategorie (Svalastog a kol., 2017).

1.1 Historie vnímání stavu zdraví

Ve starověku zdraví spadalo především pod vliv náboženství a bylo ekvivalentní k získávání laskavosti u božstev prostřednictvím modliteb a obětí, které měly uklidnit jejich hněv a přinést vysněný lék (Badash a kol., 2017). Ve třetím století před naším letopočtem byly zřizovány chrámy, ve kterých pacienti přespávali a čekali na Boha, který jim měl předepsat lék (Porter, 2005).

Ve Starověkém Řecku, stejně jako ve starověké indické a čínské medicíně, bylo zdraví považováno za rovnováhu mezi člověkem a životním prostředím. Zdraví bylo vnímáno jako jednota těla a duše a nemoci měly přirozený původ. Čínští lékaři za dynastie Čou (1122-250 př. n. l.) považovali za prevenci nemocí cvičení, hluboké dýchání a střídmost. Spojovali fyzické zdraví s morální pohodou a duchovní vyrovnaností, což vedlo ke kosmické harmonii. Starověká egyptská medicína byla založena na víře, že nemoc vyplývá z nerovnováhy mezi světskou a duchovní existencí. Řecký básník Pindar definoval zdraví jako: *„harmonické fungování orgánů“* a řecký filozof Aristotelés zdůrazňoval potřebu regulace vztahů ve společnosti, aby bylo možné dosáhnout harmonického fungování a zachování zdraví jejich členů. Další z řeckých filozofů Démokritos spojoval zdraví s chováním lidí. Jeho vnímání zdraví vycházelo z nepochopení toho, proč se lidé obrací k Bohu, když mají zdraví pod vlastní kontrolou. V neposlední řadě nejslavnější starověký lékař Hippokratés uvedl souvislost zdraví se životním prostředím a životním stylem. I když není možné přesně určit, které nemoci ohrožovaly jednotlivé populace ve starověku, uvádí se, že lidé nejčastěji trpěli výskyty neštovic, spalniček, tyfu, záškrtu, cholery, kapavky, šarlatové horečky atd. (Porter, 2005; Svalastog a kol., 2017).

Vnímání zdraví silně ovlivňovala v období středověku církev, která byla jedinou nositelkou znalostí v této oblasti. Z tohoto důvodu došlo ve středověku k zanedbání širšího vnímání zdraví

a také znalostí v oblasti medicíny. K nápravě docházelo až v období renesance (Svalastog a kol., 2017). Od raného středověku začaly převládat jiné nemoci než uváděné v období antiky. Důvodem byla především chudoba, vyvolaná demografickým růstem mezi lety 800-1 300 n. l. a odklon pozornosti od hygienických podmínek, které byly v době antiky stěžejní. Populaci postihovaly hlavně křivice, kurděje, malomocenství, mor, tuberkulóza, chřipka, syfilis aj. (Petruševski, 2013; Porter, 2005).

V období průmyslové revoluce (18.-19. století) se zdraví stalo ekonomickou kategorií. Důvodem byla především potřeba zajistit pracovní schopnost lidí a redukovat dny pracovní neschopnosti. Hodnota zdraví byla především v tvorbě ekonomického zisku. V době, kdy především fyzická práce patřila mezi hlavní smysl života, bylo zdraví spojováno se silou a celkovými schopnostmi člověka. K dalším aspektům zdraví patřila schopnost jednotlivce adaptovat se na změny v životním prostředí (Svalastog a kol., 2017). V dnešní době je zdraví vnímáno rozsáhleji než pouhá absence choroby, jedná se o schopnost jednotlivce seberealizovat se a dojít k naplnění svým představ. Zdraví by se mělo sledovat nejen u jednotlivců, ale také u skupin a komunit kvůli interakci jednotlivců se sociálním prostředím (Svalastog a kol., 2017). Zdraví a nemoc jsou dynamické procesy a každý z nás se pohybuje na škále od optimálního fungování všech aspektů lidského života až po nemoc vrcholící smrtí. Uvádí se, že k získání dobrého zdraví je třeba hledat faktory, které ho podporují (Svalastog a kol., 2017).

Svalastog a kol. (2017) uvádějí, že pro zdravotníky i politiky jednotlivých zemí je důležité znát, jak laická veřejnost vnímá otázky týkající se zdraví, např. co považují za dobré zdraví a jaké faktory k němu přispívají. Na vnímání otázek týkajících se zdraví populace má stejně jako životní prostředí vliv pohlaví nebo věk (Millstein, Irwin, 1987).

Také Zahra a kol. (2015) zkoumali vnímání zdraví a faktorů ovlivňujících zdraví laickou veřejností. Lidé, kteří žijí v různých sociálních, demografických a ekonomických podmínkách, různě vnímají faktory ovlivňující zdraví. Aby bylo možné vyvinout programy vzdělávání v oblasti zdraví a prevence, je třeba identifikovat a pochopit faktory, které ovlivňují vnímání a chování jednotlivců s ohledem na zdraví. Nicméně nejvyšší procento populace se shodlo, že životní prostředí je rozhodujícím faktorem ovlivňujícím zdraví. Podle Svalastog a kol. (2017) je zdraví charakterizováno celistvostí, pragmatismem a individualismem. *Celistvost* znamená, že je třeba vzít v úvahu životní situaci jako celek, a ne pouze jako neexistenci nemoci. Zdraví je spojené se všemi ostatními aspekty života, mezi které patří každodenní pracovní, rodinný a společenský život. *Pragmatismus* představuje zdraví jako relativní jev. Zdraví je hodnoceno podle toho, co lidé očekávají vzhledem ke svému věku, zdravotnickým podmínkám

a sociální situaci. Poslední *individualismus* vyjadřuje zdraví jako vysoce osobní jev. Každý z nás je jedinečný, a proto by měly být strategie pro zlepšování zdraví co nejvíce individualizovány.

Souto a kol. (2018) se zaobírají vnímáním zdraví portugalskou populací ve věku 18 až 79 let. Na základě dotazníku týkajícího se vnímání zdraví, kterého se zúčastnilo 1 139 respondentů (především žen), byl skrze explorativní a konfirmační faktorovou analýzu nalezen model, při kterém vnímání stavu zdraví definují dva faktory – aktuální a předchozí zdraví. Dále uvádějí, že vlivem zvyšující se vzdělanosti je dnes platnost vnímání stavu zdraví vyšší než v minulosti.

V současné době je většina všech úmrtí způsobena nepřenosnými nemocemi, mezi něž patří především kardiovaskulární, onkologická, respirační onemocnění, diabetes a duševní poruchy. Především zvýšení střední délky života ve většině vyspělých zemí světa způsobilo, že za velkou část úmrtí mohou právě nepřenosná onemocnění v porovnání s infekčními chorobami. Nicméně v rozvinutých zemích, mohou za předčasnou úmrtnost a stagnaci střední délky života především alkohol, tabák a jiné návykové látky. Podstatný vliv na předčasnou úmrtnost mají i různé sociálně ekonomické podmínky (Rehm, Probst, 2018).

Jones a kol. (2008) analyzovali databázi 335 událostí nově vznikajících infekčních chorob mezi lety 1940-2004 a snažili se prokázat nenáhodné globální vzorce. Jejich výsledky potvrdily, že původ nově vznikajících infekčních chorob je významně korelován se sociálně ekonomickými, environmentálními a ekologickými faktory a poskytují základ pro identifikaci regionů, z nichž tato infekční onemocnění s největší pravděpodobností pocházejí. Jedná se například o tato onemocnění: tuberkulóza odolná vůči více léčivům, malárie odolná proti chlorochinu, HIV (Human Immunodeficiency Virus) a těžký akutní respirační syndrom (SARS) koronavirus. Koncem roku 2019 byl objeven nový kmen koronaviru (COVID-19 – Coronavirus Disease 2019), který doposud nebyl v lidské populaci detekován (WHO, 2020a).

1.2 Dostupnost dat

V současnosti je publikováno množství vědeckých studií a článků, které se zabývají problematikou zdraví z různých úhlů pohledu. Většina z nich využívá širokou škálu ukazatelů, týkajících se zdraví jednotlivců nebo zdraví různých skupin. Na zdraví jednotlivců a léčbu různých typů onemocnění jsou zaměřeny hlavně medicinské časopisy, využívající individuální data o pacientech. Možnosti získávání datových podkladů pro různé vědecké cíle se značně zvýšily zavedením elektronické databáze e-Health (European Commission, 2018a).

Jak je uvedeno ve sdělení Evropské komise z roku 2018, zavádění digitálních řešení pro oblast zdravotnictví se mezi členskými státy EU značně liší, a proto jsou přijímány opatření v těchto třech oblastech (European Commission, 2018a):

- bezpečný přístup občanů k údajům o zdraví a jejich sdílení přes hranice,
- lepší údaje pro pokrok ve výzkumu, prevenci nemocí, personalizovaném zdraví a péči,
- digitální nástroje pro posílení postavení občanů a péče zaměřené na člověka.

Agregovaná data, případně zprůměrovaná data pro různé kolektivy osob jsou využívána hlavně pro potřeby veřejného zdraví, resp. veřejného zdravotnictví.

Veřejné zdraví je *„zdravotní stav obyvatelstva a jeho skupin, přičemž tento zdravotní stav je určován souhrnem přírodních, životních a pracovních podmínek a způsobem života“* (definice podle zákona č. 258/2000 Sb., o ochraně veřejného zdraví). Podle NZIP (Národní zdravotnický informační portál, 2022) se veřejným zdravím rozumí zdraví společnosti jako celku, a lze jej měřit a hodnotit určitými kvantitativními i kvalitativními indikátory a analytickými procesy.

Cílem orgánů veřejného zdraví je posílit zdravotnické služby a snížit nerovnosti za účelem zlepšení a ochrany blahobytu jednotlivců, komunit a obyvatelstva. K dosažení tohoto cíle využívají odborníci v oblasti veřejného zdravotnictví vzdělávání a výzkum na podporu zdravého životního stylu. Provádějí také výzkum nemocí a vyvíjejí léčebné programy pro boj proti šíření infekčních onemocnění. Veřejné zdravotnictví také podporuje spravedlivé poskytování kvalitní zdravotní péče všem občanům.

Veřejné zdravotnictví je multidisciplinární medicínský obor, který využívá a integruje poznatky mnoha dalších vědních disciplín. Jeho základ tvoří celá řada medicínských a společenskovědních oborů, zejména sociální medicína, hygiena, ochrana a podpora veřejného zdraví, epidemiologie, organizace a řízení zdravotnictví, statistika, demografie, sociologie, psychologie, ekonomie, medicínské právo a další (Hamplová, 2019).

Aby bylo možné měřit úroveň stavu zdraví a nerovnosti ve stavu zdraví jak na úrovni jednotlivých zemí, tak na úrovni jejich regionů, je třeba mít k dispozici spolehlivé a srovnatelné ukazatele. Pro monitorování zdraví na úrovni EU byly definovány *Evropské základní zdravotní indikátory* (ECHI), které jsou výsledkem dlouhodobé spolupráce mezi zeměmi EU a Evropskou komisí. Jejich cílem je poskytnout srovnatelný systém zdravotních informací a znalostí.

Datový nástroj ECHI je grafický nástroj a interaktivní aplikace pro prezentaci relevantních a srovnatelných informací o zdraví na evropské úrovni. Nástroj představuje seznam 88 ukazatelů, které jsou seskupeny do pěti kapitol (European Commission, 2022):

- demografické a sociálně ekonomické faktory,
- zdravotní stav,
- determinanty stavu zdraví,
- zdravotní zásahy: zdravotní služby,
- podpora zdraví.

Většinu údajů pro všechny ukazatele poskytuje Eurostat, avšak mnohé ukazatele jsou čerpány z jiných zdrojů, jako jsou WHO, OECD, specifické programy a specializované databáze. Metadata poskytují podrobný přehled o zdrojích údajů a o tom, jak se určuje každý ukazatel.

Zdravotní databáze OECD nabízí nejkomplexnější zdroj srovnatelných statistik o zdraví a zdravotních systémech v zemích OECD. Je to nezbytný nástroj pro provádění srovnávacích analýz a čerpání ponaučení z mezinárodních srovnání různých zdravotních systémů. Publikované ukazatele jsou rozděleny do těchto základných kategorií:

- zdravotní výdaje a financování,
- zdravotní stav,
- nemedicínské determinanty zdraví,
- zdroje zdravotní péče,
- migrace pracovní síly ve zdravotnictví,
- využití zdravotní péče,
- indikátory kvality zdravotní péče,
- zdroje a využití dlouhodobé péče,
- sociální ochrana.

Rozsáhlý podrobný seznam ukazatelů poskytuje publikace OECD (2022). Databáze těchto ukazatelů se každoročně aktualizuje.

Důležitými publikacemi, které využívají zdravotní databáze OECD, jsou publikace *Stručný pohled na zdraví* (Health at a Glance), vycházející ve dvouletých intervalech od roku 2013. Pět dosavadních vydání poskytuje komplexní soubor ukazatelů zdraví populace a výkonnosti zdravotnických systémů v rámci členů OECD a klíčových rozvíjejících se ekonomik. Tyto publikace zahrnují zdravotní stav, rizikové faktory zdraví, přístup ke zdravotní péči, její kvalitu

a zdroje ve zdravotnictví. Poskytují pozoruhodné důkazy o velkých rozdílech mezi zeměmi v rámci ukazatelů zdravotního stavu a zdravotních rizik, jakož i ve vstupech a výstupech zdravotnických systémů.

Společná dvouletá publikace OECD a EU *Stručný pohled na zdraví obyvatel v zemích Evropy* (Health at a Glance: Europe), která vychází od roku 2012, představuje soubor klíčových ukazatelů zdravotního stavu, determinantů zdraví, zdrojů a aktivit zdravotní péče, kvality zdravotní péče, výdajů na zdravotnictví a jeho financování ve 35 evropských zemích. Výběr ukazatelů je založen převážně na užším seznamu zdravotních ukazatelů Evropského společenství (ECHI). Doplňují jej dodatečné ukazatele o výdajích na zdravotní péči a kvalitě péče, které vycházejí z odborných znalostí OECD v těchto oblastech. Každý ukazatel je prezentován v uživatelsky přívětivém formátu, který sestává z grafů znázorňujících odchylky mezi zeměmi v čase.

Costa a kol. (2019) se zabývají posouzením dostupnosti dat pro *index zdraví populace* (PHI – Population Health Index), který je vícerozměrnou mírou vyvinutou v rámci projektu EURO-HEALTHY financovaného EU, na úrovni 28 států EU a jejich 269 NUTS 2 regionů. Dále uvádějí, že PHI zvyšuje povědomí o nedostatku relevantních ukazatelů pro NUTS 2 regiony především v oblasti determinantů stavu zdraví, např. životního stylu, zdrojů zdravotní péče, zdravotních návyků aj. Podle Santana a kol. (2020b) je PHI vhodným nástrojem pro podporu monitorování veřejného zdraví. Další článek Santana a kol. (2020a) uvádí strukturu PHI, který se dále dělí na *index determinantů stavu zdraví* a *index výsledků stavu zdraví*. Do indexu výsledků stavu zdraví patří následující dimenze:

- délka života, mortalita – střední délka života při narození, kojenecká úmrtnost, zabranitelná úmrtnost,
- kvalita života, nemocnost – méně než dobré sebehodnocení stavu zdraví, věkově standardizovaná ztracená léta života v důsledku nemoci, nízká porodní váha.

V rámci indexu determinantů stavu zdraví jsou zahrnuty následující dimenze:

- zaměstnanost – míra nezaměstnanosti, dlouhodobá míra nezaměstnanosti,
- příjmové a životní podmínky – disponibilní příjem domácností, lidé ohroženi chudobou nebo sociálním vyloučením, poměr disponibilního příjmu,
- sociální ochrana – výdaje na péči o nejstarší,
- bezpečnost – kriminalita,

- vzdělání – osoby s ukončeným středoškolským nebo vysokoškolským vzdělání, osoby s předčasně ukončeným vzděláním,
- stárnutí – osoby 65+ ohrožené chudobou, index stárnutí,
- životní styl – obezita u dospělých, každodenní kuřáci, čistá konzumace alkoholu, živě narození matkami mladšími 20 let,
- znečištění – roční průměr denních koncentrací jemných částic, skleníkový plyn, obyvatelstvo vystavené hluku z dopravy,
- extrémy počasí – obyvatelstvo postižené záplavami,
- bytový stav – průměrný počet pokojů na osobu, domácnosti bez vnitřního splachovacího WC, domácnosti bez centrálního topení,
- voda a hygiena – obyvatelstvo napojené na veřejný vodovod, obyvatelstvo napojené na čistírny odpadních vod,
- nakládání s odpady – míra recyklace komunálního odpadu,
- hustota obyvatel,
- bezpečnost na silnicích – oběti dopravních nehod, úmrtnost v důsledku dopravních nehod,
- zdroje ve zdravotnictví – lékaři, zdravotnický personál,
- výdaje na zdravotní péči – celkové výdaje do zdraví, vlastní výdaje do zdraví, veřejné výdaje do zdraví,
- výkon zdravotní péče – propuštění z nemocnice kvůli cukrovce, hypertenzi a astmatu, vyhnutelná úmrtnost prostřednictvím zdravotní péče.

Nejkomplexnější zdravotní informace poskytuje WHO, specializovaná agentura OSN odpovědná za mezinárodní veřejné zdraví. Databáze úmrtností WHO je kompilací údajů o úmrtnostech podle země a oblasti, roku, pohlaví, věku a příčiny smrti, jak je každoročně předávají národní orgány z jejich občanské registrace a systému vitální statistiky. Databáze obsahuje údaje od roku 1950 do dnešního dne. Od poloviny 80. let 20. století sdělují členské státy evropského regionu WHO základní statistiky týkající se zdraví do rodiny databází *Zdraví pro všechny* (HFA). Databáze HFA spojují ukazatele, které jsou součástí hlavních monitorovacích rámců relevantních pro region, jako jsou například *Zdraví 2020* a *Cíle udržitelného rozvoje*. Ukazatele pokrývají základní demografii, zdravotní stav, zdravotní determinanty a rizikové faktory, jakož i zdroje zdravotní péče, výdaje na zdraví a další.

1.3 Stav zdraví

Na regionální a národní úrovni, na úrovni členských zemí EU a OECD i na celosvětové úrovni jsou shromažďovány, pravidelně aktualizovány a on-line publikovány rozsáhlé datové soubory různých ukazatelů, týkajících se veřejného zdraví, jak bylo podrobněji specifikováno v předchozí podkapitole. Přes množství zdravotních ukazatelů neexistuje samostatný ukazatel, který by měřil přímo stav kolektivního zdraví obyvatel v určitém regionu v konkrétním čase. Často se k tomuto účelu na úrovni států využívá střední délka života při narození, resp. ve věku 65 let, jejíž hodnota je stavem zdraví ovlivněna, ale nejedná se přímo o ukazatel stavu zdraví.

K posouzení úrovně stavu zdraví v určité skupině obyvatel je nutné vhodně zvolit několik ukazatelů. Úroveň zdraví obyvatelstva je vyjadřována mírami *„pozitivního zdraví“*, tzn. délkou života nebo *„negativního zdraví“*, tzn. nemocností a úmrtností populace. V přehledu různých přístupů k měření stavu zdraví se dále zaměříme především na země Evropy, konkrétně EU-27, což koresponduje s dalším obsahem disertační práce. Důvodem výběru těchto zemí jsou dostupné aktuální ukazatele stavu zdraví v databázi Eurostatu. Lze očekávat, že pro země EU-27 budou tyto ukazatele aktualizovány, což by mělo umožnit porovnání stavu zdraví v době před a po onemocnění Covid-19.

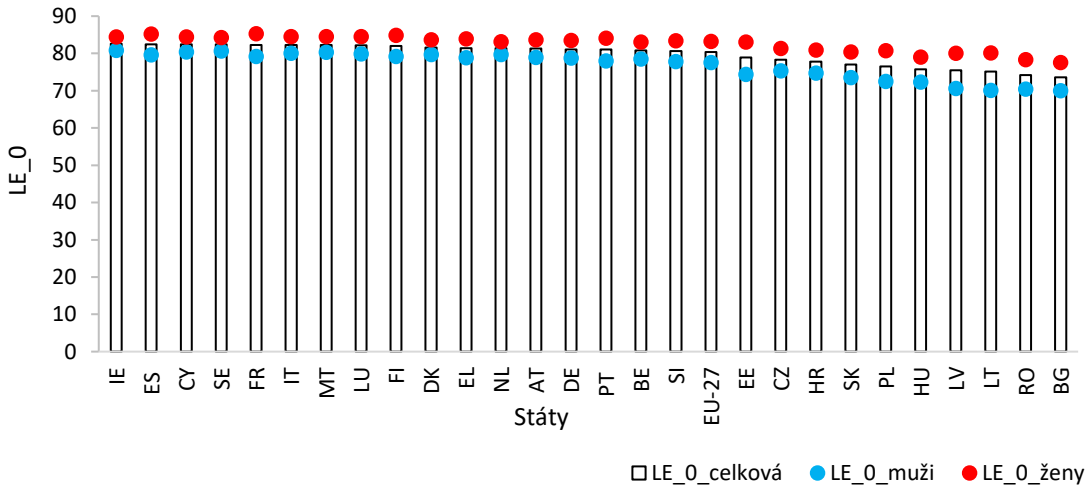
1.3.1 Střední délka života

V České republice (CZ) byla *střední délka života při narození* (LE_0) v roce 2020 o více než dva roky nižší než průměr EU. Vlivem pandemie COVID-19 se tento ukazatel stavu zdraví dočasně snížil v roce 2020 na úroveň roku 2013 o jeden rok na 78,3 let oproti roku 2019. V případě *střední délky života ve věku 65 let* (LE_65) došlo v CZ k poklesu tohoto ukazatele mezi roky 2019 a 2020 z 18,4 let na 17,3. Co se týká úmrtnosti na závažná onemocnění, mezi nejčastější příčiny úmrtnosti patří ischemická choroba srdeční, cévní mozková příhoda a rakovina plic. Diabetes představuje v CZ další významnou příčinu úmrtí (čtvrtá nejvyšší standardizovaná míra úmrtí v EU). Vysoká míra úmrtnosti je v CZ způsobena především nemocemi, kterým je možné předejít nebo jsou léčitelné. Růst LE_0 do roku 2019 lze přisuzovat úspěšnému snižování úmrtností na nejzávažnější nemoci. Slovenská republika (SK) je na tom ve stavu zdraví podle LE_0 hůře než CZ. LE_0 (77) zaostává o více než rok za CZ. Stejně jako v CZ došlo i zde ke snížení LE_0 v roce 2020 oproti roku 2019 vlivem pandemie COVID-19. V SK patří LE_0 k nejnižším v celé EU. Významný má SK také rozdíl v LE_0 mezi muži a ženami, kde ženy (80,4) žijí téměř o 7 let déle než muži (73,5). V případě LE_65 došlo v SK mezi lety 2019 a 2020 k poklesu hodnot tohoto ukazatele u mužů z 15,7 let na 14,8 a u žen

z 19,7 let na 18,9 (Eurostat, 2022a; OECD/European Observatory on Health Systems and Policies, 2021).

LE_0 je definována jako průměrný počet let žití, které může novorozenec očekávat, pokud bude po celý život vystaven současným podmínkám úmrtnosti (věkově specifické pravděpodobnosti úmrtnosti). LE_65 je zase definována jako průměrný počet let žití, které může osoba ve věku 65 let očekávat, pokud bude vystavena současným podmínkám úmrtnosti. Pro detaily viz. Eurostat (2022a). Na obrázku 1 je vyobrazena LE_0 podle pohlaví pro země EU-27 pro rok 2020. Státy Evropy jsou seřazeny podle celkové LE_0.

Obrázek 1: Střední délka života při narození pro země EU-27 (2020)



Zdroj: vlastní zpracování na základě Eurostat (2022a)

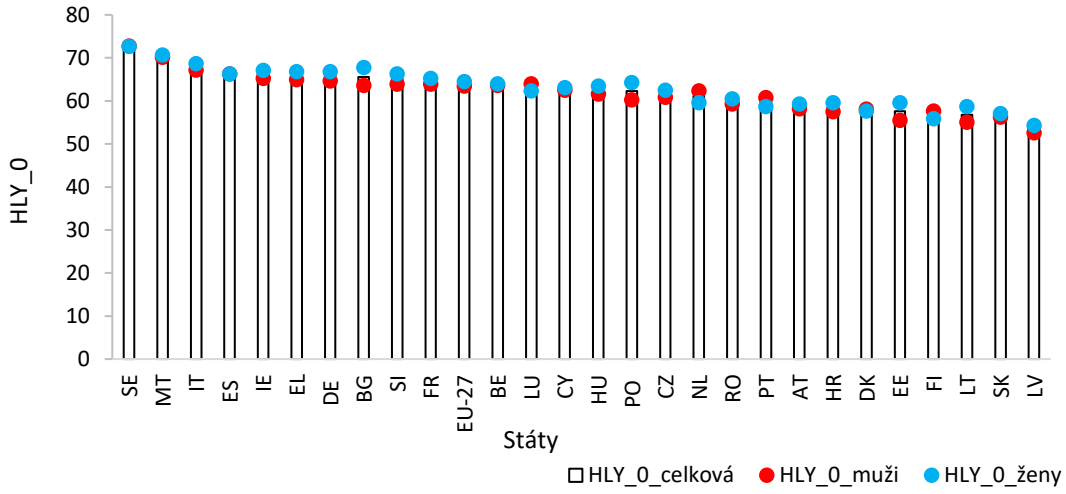
1.3.2 Zdravé roky života a očekávaná délka života ve zdraví

Některé indexy stavu zdraví jsou založeny na LE, jejíž odhady nejsou citlivé na zdravotní stav populace (Wolfson, 1996). K měření zdravotního stavu populace nestačí zaměřit se pouze na délku života, ale je třeba zahrnout i jeho kvalitu (Robine a kol., 2013). Informace o tom, v jakém stavu zdraví jsou léta navíc prožitá, je důležitá např. pro řízení výdajů na veřejné zdravotnictví nebo rozhodování o nastavení důchodového věku (Hlaváček, Lakotová, 2019) aj. *Zdravé roky života* (HLY – Healthy Life Years) jsou ukazatelem Eurostatu, který monitoruje zdraví jako produktivní, popřípadě ekonomický faktor. Podle Eurostat (2022a) HLY měří počet zbývajících let, u kterých se očekává, že je osoba v určité věku prožije bez závažných nebo středně závažných zdravotních problémů. Tento ukazatel bývá někdy nazýván jako střední délka života bez zdravotního postižení a kombinuje údaje o úmrtnosti s údaji o zdravotním stavu na základě *dlouhodobých omezení aktivit* (GALI - Global Activity Limitation Indicator),

které jsou zjišťovány v rámci každoročního průzkumu EU-SILC (šetření EU o příjmech a životních podmínkách). Respondentům je v této souvislosti položena následující otázka: *„Do jaké míry jste byli posledních šest měsíců omezováni kvůli zdravotnímu problému v činnostech, které lidé obvykle dělají?“* (di Lego, 2021).

Konkrétně HLY_0 měří počet let, po které se očekává, že osoba při narození ještě prožije ve zdravém stavu. Ukazatel HLY_65 zase měří počet let, po které se očekává, že osoba ve věku 65 let ještě prožije ve zdravém stavu (viz. Eurostat, 2022a). Obrázek 2 vizualizuje HLY_0 (zdravé roky života při narození) podle pohlaví pro země EU-27 v roce 2020, jak uvádí Eurostat (2022a). Státy Evropy jsou seřazeny podle hodnot HLY_0 pro obě pohlaví společně.

Obrázek 2: Zdravé roky života při narození pro země EU-27 (2020)



Zdroj: vlastní zpracování na základě Eurostat (2022a)

Široce používanými ukazateli očekávané délky života ve zdraví se zabývá di Lego (2021). Mezi tyto ukazatele řadí HLY a *očekávanou délku života ve zdraví* (HALE – Health-Adjusted Life Expectancy). Ukazatel HALE poskytuje WHO (2022d) a je definován jako průměrný počet let, po které může člověk očekávat, že bude žít při *„plném zdraví“*, pokud jsou brány v úvahu roky neprožité v *„plném zdraví“* kvůli nemoci nebo zranění.

Oba ukazatele stavu zdraví HLY a HALE jsou subjektivními ukazateli z důvodu, že kombinují demografická data *(„tvrdá data“)* s daty získanými skrze dotazníkové šetření, kde může dojít k zahrnutí např. pesimismu respondentů. Vzhledem k tomu je důležité, zabývat se především vývojem těchto ukazatelů v čase namísto absolutního počtu roků prožitých ve zdraví (Hlaváček, Lakotová, 2019). Podle di Lego (2021) mají oba uvedené ukazatele limity a oba se od sebe liší. Na rozdíl od HLY ukazatel HALE více koreluje s úmrtnostmi, a proto je indikátor HALE

užitečnější při nastavování politik týkajících se léčby a spojených se smrtelnějšími stavy. Na druhou stranu ukazatel HLY slouží spíše pro nastavování politik z hlediska prevence, jelikož odráží úroveň dobrého zdraví bez ohledu na úmrtnost. Ukazatel HLY nabývá nižších hodnot než HALE a zároveň jeho hodnoty v čase kolísají, což bylo v CZ jedním z důvodů rozhodnutí nezvyšovat důchodový věk nad 65 let, i když LE neustále do roku 2019 rostla (Hlaváček, Lakotová, 2019). Eurostat poskytuje ukazatel HLY pro osoby ve věku 65 let stejně jako u LE a WHO poskytuje ukazatel HALE při narození (HALE_0) a pro osoby ve věku 60 let.

1.3.3 Úmrtnosti na nejzávažnější onemocnění

Podle Výkladového slovníku termínů v epidemiologii (Göpfertová, Šmerhovský, 2015) je *úmrtnost* (mortality) definována jako počet úmrtí na dané onemocnění ve vztahu k počtu osob daného populačního celku a času. Mezi nejzávažnější onemocnění ohrožující evropskou populaci, která se objevují na předních příčkách v úmrtnostech, patří *onemocnění kardiovaskulární* (CADs – Cardiovascular Diseases), následovaná *onemocněními onkologickými* (CAs – Cancers). Mezi další závažná nepřenosná onemocnění patří *onemocnění respirační, duševní poruchy* a *diabetes*. Vybraná závažná onemocnění sloužící k hodnocení stavu zdraví ve státech a regionech Evropy jsou klasifikována podle *Mezinárodní klasifikace nemocí 10* (MKN-10), kde lze nalézt popis jednotlivých skupin těchto onemocnění (viz. ÚZIS, 2020).

Podle OECD (2021a) jsou CADs nejčastější příčinou úmrtí v zemích OECD. V roce 2019 CADs způsobily každé třetí úmrtí a CAs způsobily jedno ze čtyř úmrtí v těchto zemích. Narůstající počty úmrtí následkem CADs je možné vysvětlit mimo jiné i stárnutím populace. Další již zmíněná respirační onemocnění představovaly 10 % všech úmrtí, a to vlivem rizikových faktorů kouření, znečištění ovzduší, vystavování se prachu, výparům a chemikáliím. Za 9 % všech úmrtí mohly Alzheimerova choroba a další demence. V případě diabetu se jednalo o 3 % ze všech úmrtí. Obecně je možné uvést, že se různé sociálně ekonomické skupiny obyvatel liší v hlavních příčinách úmrtí, a to především u onemocnění, kterým lze předejít. OECD (2021a) analyzuje hlavní příčiny úmrtí pro poslední dostupný rok 2019, čili poslední rok, který nebyl poznamenán pandemií COVID-19. Toto nové onemocnění způsobuje více úmrtí, než tomu bylo v předchozích letech a lze očekávat, že bude mít vliv na ukazatele stavu zdraví v letech následujících.

Podle Wilkins a kol. (2017) CADs způsobily každý rok 3,9 milionů úmrtí v Evropě (45 % všech úmrtí) a přes 1,8 milionů úmrtí v EU (37 % všech úmrtí). V roce 2015 žilo více než 85 milionů lidí s CADs v Evropě a z toho 49 milionů lidí v EU. Odhaduje se, že CADs stojí ročně ekonomiku EU 210 miliard EUR, z toho 53 % jde na zdravotní péči, 26 % na ztrátu produktivity a 21 % na neformální péči o osoby s CADs vykonávanou rodinnými příslušníky. Na základě Fitzmaurice a kol. (2017) jsou CAs druhou nejčastější příčinou úmrtí téměř po celém světě. V roce 2015 bylo zaznamenáno 17,5 milionů případů výskytu CAs na světě a 8,7 milionů úmrtí na toto onemocnění. V rámci deseti let mezi roky 2005 až 2015 došlo k nárůstu počtu případů CAs o 33 %. Dále uvádějí, že CAs jsou hlavní příčinou předčasné úmrtnosti ve 28 z 53 zemí v Evropě.

Porovnáním úmrtností způsobených kritickými nemocemi v zemích EU se zabývá článek Kopecká, Jindrová (2017). Zde je porovnáváno 28 členských zemí EU na základě ukazatele týkajícího se LE_0 a ukazatelů *standardizovaných měr úmrtností* (SDR – Standardized Death Rates) na 100 000 obyvatel pro věkovou skupinu 0-64 let. Pro vysvětlení viz. podkapitola 4.2.1. Data byla převzata z databáze WHO pro rok 2015. Podle většiny zvolených ukazatelů byl nejhorší stav zdraví v Bulharsku, Litvě, Lotyšsku, Maďarsku a Rumunsku s velkým odstupem od ostatních států. Pomocí aplikace vícerozměrných statistických metod byly zjištěny další poznatky o závažnosti situace v úmrtnostech na závažná onemocnění, různé příčinné souvislosti a regionální rozdíly v zemích EU.

Článek Kopecká (2018b) se věnuje porovnání úmrtností způsobenými závažnými onemocněními v regionech CZ. Pro porovnání stavu zdraví jsou použity úmrtnosti týkající se CAs a CADs. Porovnáním úmrtností způsobenými závažnými onemocněními na regionální úrovni se věnuje také článek Pacáková, Kopecká (2019b). Mimo úmrtnosti způsobené CADs a CAs, jsou dále zahrnuté úmrtnosti způsobené duševními chorobami, které mají rostoucí trend vlivem stárnutí populace. Tento článek je zaměřen na NUTS 2 regiony zemí *Visegrádské skupiny* (V4 – Visegrad Group). Také v článku Pacáková, Kopecká (2018c) jsou použity ukazatele SDR, které popisují stav zdraví ve vybraných evropských zemích. Mezi zahrnuté příčiny úmrtí patří opět CADs a vybraná CAs. Úmrtnostmi na tři nejčastější závažná onemocnění v zemích EU-28 jako jsou CADs, CAs a respirační onemocnění se věnuje článek Pacáková a kol. (2020), ve kterém jsou použity SDR na 100 000 obyvatel. Zdravotní rizika dlouhověkosti řeší článek Pacáková, Kopecká (2018b). V tomto článku je posuzován stav zdraví a jeho nerovnosti osob starších 65 let skrze ukazatele týkající se LE_65, HLY_65, SDR pro CADs a CAs.

Důležitým ukazatelem stavu zdraví je *vyhnutelná úmrtnost* (avoidable mortality), která se dále dělí na *léčitelnou (odstranitelnou) úmrtnost* (treatable mortality) a *zabranitelnou (předvídatelnou) úmrtnost* (preventable mortality). Vyhnutelná úmrtnost stojí na myšlence, že některým úmrtím je možné se *„vyhnout“* u lidí mladších 75 let (Eurostat, 2021 a OECD, 2021a). Eurostat (2021) poskytuje ukazatele vyhnutelné úmrtnosti vyjádřené na 100 000 obyvatel a definuje léčitelnou a zabranitelnou úmrtnost následovně:

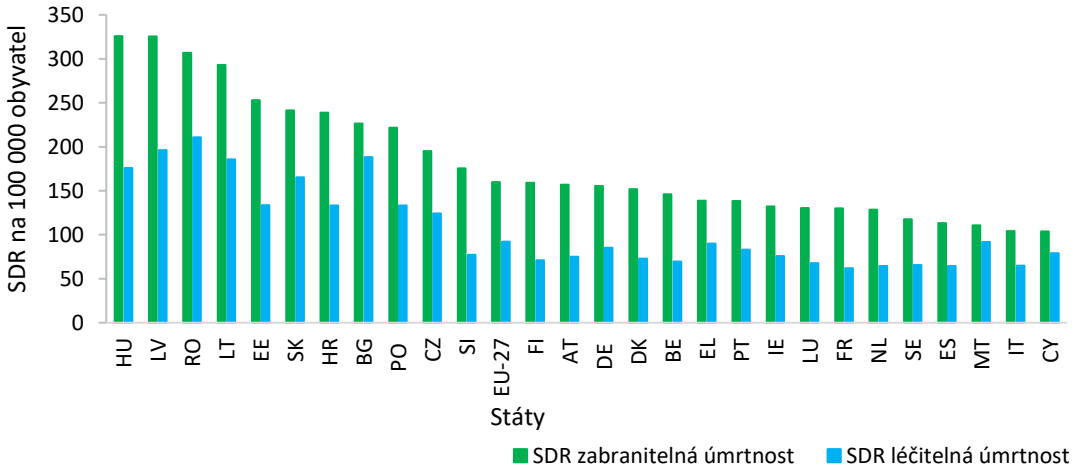
- léčitelná úmrtnost je úmrtnost, které by bylo možné předejít optimální kvalitou zdravotní péče,
- zabranitelná úmrtnost je úmrtnost, které by bylo možné předejít skrze zásahy veřejného zdraví, zaměřenými na širší okruh determinantů veřejného zdraví, jako jsou faktory chování a životního stylu, faktory sociálně ekonomické a životního prostředí.

Ukazatele vyhnutelné úmrtnosti představují *„výchozí bod“* pro hodnocení účinnosti systémů zdravotní péče a systémů veřejného zdraví (OECD, 2021a). OECD (2021a) také uvádí, že v roce 2019 bylo možné předejít kolem tří milionů předčasných úmrtí lidí mladších 75 let, což představuje jednu čtvrtinu všech úmrtí. Zhruba 1,9 milionů z těchto úmrtí je možné označit za zabranitelné úmrtnosti a 1 milion úmrtí představuje léčitelnou úmrtnost. V případě obou vyhnutelných úmrtností jsou muži více ohroženou skupinou než ženy.

Mezi nejčastější příčiny zabranitelné úmrtnosti patřily v roce 2019 v následujícím pořadí rakovina plic, dopravní nehody, sebevraždy, srdeční infarkty, mozkové mrtvice, respirační onemocnění (chřipka, chronická obstrukční plicní nemoc) a úmrtí související s drogami. V případě léčitelné úmrtnosti k nejčastějším patřily nemoci oběhové soustavy (srdeční infarkt, mozková mrtvice), kolorektální rakovina, rakovina prsu, onemocnění dýchacích cest (zápal plic, astma), cukrovka a další onemocnění endokrinního systému. (OECD, 2021a)

Na obrázku 3 je možné porovnat země EU-27 z hlediska léčitelných a zabranitelných SDR pro rok 2018. Pouze data pro Francii a EU-27 jsou dostupná pro rok 2017. Eurostat (2022a) poskytuje roční údaje o těchto dvou vyhnutelných úmrtnostech buď v absolutních číslech, nebo jako SDR. Graf na obrázku 3 poskytuje SDR stanovené na 100 000 obyvatel mladších 75 let. SDR se stanovují na základě *evropské standardní populace* (ESP – European Standard Population).

Obrázek 3: SDR pro léčitelnou a zabranitelnou úmrtnost pro země EU-27 (2018)



Zdroj: vlastní zpracování na základě Eurostat (2022a)

Nejvyšší léčitelná a zabranitelná úmrtnost je podle obrázku 3 ve východní části Evropy. Úmrtí, kterým je možné zabránit včasnou prevencí, tvoří větší část vyhnutelných úmrtností oproti léčitelným úmrtnostem ve všech monitorovaných zemích Evropy. Na celkové vyhnutelné úmrtnosti má téměř stejný podíl léčitelná a zabranitelná úmrtnost na Maltě, Kypru a v Bulharsku. Vzájemná lineární závislost mezi léčitelnou a zabranitelnou úmrtností měřená jednoduchým Pearsonovým korelačním koeficientem je 0,93 (silná pozitivní lineární závislost).

Kossarova a kol. (2013) uvádějí některá omezení vyhnutelné úmrtnosti. Úmrtí z jakékoliv příčiny je výsledkem mnoha událostí, i když se jedná o vyhnutelnou úmrtnost. Například sem vstupuje faktor chování osoby vyhledávající zdravotní péči. Dále ukazatele vyhnutelné úmrtnosti neodrážejí informace o výskytech závažných onemocnění, tzn., systémy zdravotní péče a veřejného zdraví se musejí těmto zvýšeným výskytům přizpůsobit. Ukazatel úmrtnosti nemusí být vždy adekvátním ukazatelem úspěšnosti zdravotního systému, a to v případech, kdy se nezaměřujeme přímo na odvracení smrti, ale zmírnění bolesti, popř. zlepšení kvality zdravotní péče. Kruk a kol. (2018) zase uvádějí, že v roce 2016 v zemích s nízkými a středními příjmy se vyskytlo 15,6 milionů vyhnutelných úmrtí. Z těchto úmrtí 8,6 milionů bylo způsobeno buď nekvalitní zdravotní péčí, nebo jejím nevyužíváním. V rámci článku Pacáková a kol. (2021) byly zjišťovány mimo jiné nerovnosti v léčitelné a zabranitelné úmrtnosti v zemích EU-27 pro rok 2018.

1.3.4 Kojenecká úmrtnost

Další úmrtností patřící do PHI, konkrétně do indexu výstupů stavu zdraví, je kojenecká úmrtnost (infant mortality). Podle Eurostat (2022a) je tato úmrtnost definována jako podíl počtu

zemřelých dětí do jednoho roku k počtu živě narozených v tomto roce. Hodnoty tohoto ukazatele, který poskytuje Eurostat (2022a), a kde jsou data o kojenecké úmrtnosti dostupná pro rok 2019 pro většinu zemí Evropy, jsou vyjádřeny na 1 000 živě narozených dětí. V zemích EU byla v roce 2018 v průměru méně než 3,5 úmrtí na 1 000 živě narozených dětí. Pouze Malta, Rumunsko, Bulharsko a SK vykazují kojeneckou úmrtnost nad 5 úmrtí na 1 000 živě narozených. V rámci zemí EU došlo v posledních několika desetiletích k rapidnímu snížení kojenecké úmrtnosti, kde k nejvýraznějšímu poklesu došlo v Bulharsku a Rumunsku. Za úmrtnost během prvního měsíce života mohou především vrozené anomálie a nedonošenost. Po prvním měsíci života je příčinou zejména syndrom náhlého úmrtí (OECD/European Union, 2020).

1.3.5 Ztracená léta života v důsledku nemoci

Dalším ukazatelem stavu zdraví, který je také obsažen v PHI, jsou *ztracená léta života v důsledku nemoci* (DALY – Disability Adjusted Life Year). Podle WHO (2022a) jsou DALYs definovány jako součet let života ztracených v důsledku předčasné úmrtnosti a let prožitých se zdravotním postižením. WHO poskytuje tento ukazatel věkově standardizovaný, tzn. očištěný o rozdíly ve věkovém rozdělení populace. Díky kombinaci roků života ztracených v důsledku předčasné úmrtnosti a roků života ztracených v důsledku stavů zdraví, které jsou horší než plné zdraví, nebo v důsledku zdravotního postižení, vytváří ukazatel DALY lepší obraz odrážející zátěž nemocemi než úmrtnost samotná (WHO, 2022a).

Například Kříž (2016) se zabývá ukazatelem DALY v CZ a srovnáním s vybranými státy Evropy, kterými jsou Bulharsko, Dánsko, Rakousko, Rusko, Slovensko a Švýcarsko pro rok 2012. V tomto srovnání se CZ nacházela přesně uprostřed, tedy ukazatel DALY dosahoval hodnoty mediánu. Zheng a kol. (2020) konstatují, že ve světle dožívání se vyššího věku v Číně došlo prostřednictvím oddálení věku odchodu do důchodu k mnoha zdravotním problémům, zejména onemocněním smyslových orgánů a bolestí krku a kříže, což by mohlo odbourat pokrok ve snižování DALY, který je v současné době nedostatečný.

1.3.6 Incidence a prevalence závažných onemocnění

Dalšími důležitými ukazateli stavu zdraví jsou ukazatele *nemocnosti* (morbidity), a to *incidence* a *prevalence*. Nemocnost je stav vykazování příznaků nebo nezdraví kvůli nemoci (Hernandez, Kim, 2022). Göpfertová, Šmerhovský (2015) definují incidenci jako počet nových případů, vydělený celkovým osobočasem (suma času, kterým osoba přispěla k celkové době sledování) za dané období. Prevalence je míra frekvence onemocnění v populaci specifikovaná místně

a časově. Čitatel je tvořen počtem všech existujících onemocnění v určitém období (intervalová prevalence) nebo k určitému datu (okamžiková prevalence). Jmenovatelem je počet osob v riziku (Göpfertová, Šmerhovský, 2015).

V posledních třiceti letech byl zaznamenán malý pokles ve věkově standardizované incidenci CADs v 57 členských státech *Evropské kardiologické společnosti* (ESC – European Society of Cardiology), ale také nárůst v sedmi z těchto členských zemí se středními příjmy. Odhady věkově standardizovaného výskytu hlavních složek CADs ischemické choroby srdeční a cévní mozkové příhody byly dvakrát vyšší v zemích se středními příjmy v porovnání se zeměmi s vysokými příjmy. Dle pohlaví byl u CADs zaznamenán dvakrát vyšší výskyt v případě mužů než žen. U cévní mozkové příhody byla míra výskytu podobná u obou pohlaví. V roce 2019 tato dvě nejzávažnější CADs představovala v členských zemích ESC 70 milionů ztracených let života v důsledku nemoci (Timmis a kol., 2022).

Timmis a kol. (2022) uvádějí zdroje dat týkající se CADs. Například WHO poskytuje data o úmrtnostech na CADs a rizikových faktorech. Dále studie *Globální zátěže nemocemi* (GBD – Global Burden of Diseases) přinesla odhady prevalence CADs. Institute for Health Metrics and Evaluation (2022) uvádí, že GBD poskytuje nástroj pro kvantifikaci ztrát zdraví v důsledku nemocí, zranění a rizikových faktorů. European Commision (2018) uvádí, že se výskyty závažných onemocnění jako jsou CADs, CAs, respirační onemocnění, diabetes, astma a duševní poruchy zvyšují, nicméně na druhou stranu u většiny z nich dochází ke snižování úmrtností až na výjimky, mezi které patří např. duševní poruchy nebo rakovina slinivky břišní.

Podle OECD (2021a) i přes zvyšující se výskyt závažných onemocnění došlo díky rozsáhlým screeningovým programům např. u kolorektálního karcinomu k poklesu výskytu tohoto onemocnění u starších osob. Na druhou stranu je pozorován nárůst incidence tohoto onemocnění u mladších pacientů, což je zapříčiněno nezdravým životním stylem. Incidence rakoviny prsu, představující druhou nejčastější příčinu úmrtí u žen, v posledních deseti letech rostla, nicméně úmrtnost na toto onemocnění díky včasným diagnózám klesá. Mezi další závažná onemocnění patří bezesporu diabetes. Diabetes způsobuje další závažné zdravotní stavy např. kardiovaskulární onemocnění, slepotu, selhání ledvin nebo amputaci dolních končetin. S diabetem žilo v roce 2019 6,7 % dospělé populace, avšak odhaduje se, že dalších 39 milionů dospělých má diabetes nediagnostikovaný (OECD, 2021a).

Podle OECD (2021a) je pro zdravý a ekonomicky produktivní život zásadní podmínkou udržovat dobré duševní zdraví. V roce 2020 oproti roku 2019 byly zaznamenány vyšší úrovně

úzkosti a deprese v zemích OECD, zřejmě vlivem narušení dosavadního způsobu života kvůli pandemii COVID-19. Prevalence úzkosti a deprese dramaticky vzrostla zejména u mladých lidí na přelomu roků 2020 a 2021. Vyšší míra úzkosti a deprese byla zaznamenána u nezaměstnaných osob a lidí ve finanční tísni. Tento trend předchází pandemii COVID-19, nicméně výrazně zrychlil.

V zemích OECD byly vlivem pandemie COVID-19 narušeny zejména programy pro včasnou diagnostiku rakovin. Předpokládá se, že kvůli zpoždění diagnostiky u některých druhů rakovin dojde ke snížení míry přežití. V souvislosti s pandemií COVID-19 nejčastější závažná onemocnění CADs, CAs, chronická obstrukční plicní nemoc a cukrovka zvyšují riziko závažnějšího průběhu COVID-19, který může vést ke smrti (OECD, 2021a).

Problematikou odhadů pravděpodobností výskytů a úmrtností na rakovinu plic a jejich vývojem mezi lety 1985 a 2014 se zabývá článek Kopecká (2018a). V případě CZ byl evidentní mírný rostoucí trend incidencí, ale klesající trend úmrtností. Data byla získána z databáze WHO pro incidence a databáze OECD pro úmrtnosti. Nicméně článek Kopecká, Pacáková (2017) věnující se odhadům pravděpodobností výskytů rakoviny plic a rakoviny tlustého střeva a jejich vývojem od roku 2000 do roku 2014, ukazuje na mírný, avšak klesající trend incidencí. Další článek Kopecká (2019) používá pro posouzení stavu zdraví v zemích EU-28, ukazatele LE, HLY a SDR na dvě nejzávažnější onemocnění na 100 000 obyvatel a jejich výskyty. Zde použitá data jsou získána pro poslední možný dostupný rok a mezi zdroje patří databáze Eurostatu, OECD, WHO a publikace Wilkins a kol. (2017). Výskytům chronických závažných onemocnění v rámci zemí Evropy a krajů CZ se dále věnují články Kopecká (2018c), Kopecká a kol. (2020).

1.3.7 Sebehodnocení stavu zdraví

OECD (2021a) se zabývá také *sebehodnocením stavu zdraví* (self-perceived health). Sebehodnocení stavu zdraví odráží celkové vnímání vlastního zdravotního stavu jednotlivce. Respondentům průzkumu byla položena otázka, která obvykle zněla: *„Jaké je vaše zdraví obecně?“ („How is your health in general?“)*. Jedná se o otázku doporučenou WHO. Srovnání sebehodnocení stavu zdraví mezi zeměmi OECD má však mnohá úskalí. Mezi hlavní tři úskalí patří:

- subjektivita způsobená sociálně ekonomickými rozdíly,
- stárnutí populace – vyšší podíl starších lidí v některých zemích,
- odlišnosti v otázkách a kategoriích odpovědí v průzkumu.

V roce 2019 80 % dospělých s nejvyššími příjmy hodnotilo své zdraví jako dobré nebo velmi dobré, na rozdíl od dospělých s nejnižšími příjmy, kde ani ne 60 % dospělých hodnotilo své zdraví jako dobré nebo velmi dobré v zemích OECD. Největší sociálně ekonomické rozdíly v sebehodnocení stavu zdraví byly zjištěny v Lotyšsku, Estonsku, CZ a Litvě.

Na základě Eurostat (2020) ukazatel sebehodnocení stavu zdraví vyjadřuje subjektivní hodnocení svého zdravotního stavu respondentem. Ukazatel sebehodnocení stavu zdraví je možné využít pro hodnocení celkového stavu zdraví, nerovností ve stavu zdraví nebo pro potřeby zdravotní péče. Odpovědi na otázku hodnotící obecně vnímané zdraví zní pro EU následovně: velmi dobré (very good)/dobré (good)/průměrné (fair)/špatné (bad)/velmi špatné (very bad). V rámci EU 68 % osob starších 16 let vnímalo své zdraví jako velmi dobré nebo dobré v roce 2019. Nejhorší vnímání stavu zdraví bylo zjištěno v Litvě a Lotyšsku, nejlepší vnímání stavu zdraví bylo ve Španělsku, Švédsku, na Kypru, v Řecku a Irsku. Z hlediska příslušnosti k pohlaví muži měli větší tendenci hodnotit své zdraví lépe než ženy. Negativní vnímání stavu zdraví rostlo s věkem, převažovalo u lidí s nižšími příjmy a s nižším stupněm vzdělání. Touto problematikou se zabývá článek Pacáková a kol. (2021). Ukazatelé týkající se sebehodnocení stavu zdraví jsou také použity v článku Pacáková, Kopecká (2018a) k posouzení odchylek mezi vlastním hodnocením zdraví obyvatel a objektivní situací stavu zdraví ve vybraných zemích Evropy. V článku Jindrová, Labudová (2020) je analyzován vztah mezi sociálními a demografickými charakteristikami a sebehodnocením zdraví populace v zemích EU-28 v roce 2018. Pro posouzení dopadu vybraných sociálně ekonomických a demografických charakteristik na vlastní hodnocení stavu zdraví obyvateli SK je použit model logistické regrese.

1.3.8 Indexy stavu zdraví

Existuje mnoho indexů stavu zdraví. Jejich přehled je k dispozici např. v článcích Kaltenthaler a kol. (2004) a Ashraf a kol. (2019). V publikaci Kaltenthaler jsou zahrnuty indexy stavu zdraví zahrnující alespoň dva zdravotní ukazatele. Přehled se skládá z 18 studií pocházejících z evropských zemí a Severní Ameriky. Ashraf a kol. (2019) publikují přehled 26 studií (27 indexů), kde 21 indexů měří zdraví celkové populace a zbylých šest měří zdraví pouze její části. Mezi indexy, které obě tyto publikace uvádějí, patří:

- *H index* – průměrná doba trvání zdraví v rozmezí 0 až 1 (čím zdravější je populace, tím vyšší je hodnota H),

- *Q index* – pomocí vypočtených hodnot Q bylo porovnáno 17 tříd onemocnění podle MKN v cílové populaci ve srovnání s populací referenční (vyšší hodnoty Q představují vyšší priority),
- *Očekávaná délka života bez zdravotního postižení* – počet let, které populace očekává, že bude žít bez zdravotního postižení,
- *K index* – jedná se o index kvality zdravotní péče (nejnižší hodnota indexu představuje systém s nejkvalitnější zdravotní péčí),
- *G index* – slouží k měření zdraví znevýhodněných skupin v populaci. Hodnota indexu G se pohybuje od nuly představující paritu zdravotního stavu mezi cílovou a referenční populací, až po kladné číslo představující rozsah a závažnost rozdílu.
- *Očekávaná délka života bez odvratitelné úmrtnosti* – jedná se o očekávaný počet let bez odvratitelné úmrtnosti.

Ashraf a kol. (2019) uvádějí dále v přehledu indexů zdraví na rozdíl od Kaltenthaler a kol. (2004) například: *obecný index zdraví* (General Index of Health), *index dětské úmrtnosti* (Index of Child Mortality), *index vícenásobné deprivace* (Index of Multiple Deprivation), *index nerovnosti ve zdraví* (Inequity-in-Health Index) aj. Mezi ukazatele stavu zdraví, které poskytují databáze Eurostatu nebo WHO a jsou v těchto publikacích zařazeny, patří HLY a DALY.

Podle Ashraf a kol. (2019) obecný index zdraví byl vyvinut k identifikaci prioritní geografické oblasti pro distribuci zdrojů ve zdravotnictví ve městě Vancouver. Tento index byl konstruován na základě dat týkajících se vnějších příčin úmrtnosti, úmrtností mezi 15 až 64 lety, výskytu nízké porodní váhy. Skóre indexu nabývá hodnot v rozmění 0 až 30 bodů a porovnává 12 geografických oblastí ve městě Vancouver. Konstrukce tohoto indexu probíhá následovně: každému objektu (oblasti) je přiděleno skóre mezi 0 až 10 body. Oblasti s nejnižším výskytem je přiděleno 10 bodů a ostatní jsou následně seřazeny podle decilů. Ve finále se skóre každé oblasti sečtou, čímž je získán již zmiňovaný index (viz. Rumel, Contanzo, 1992). Index dětské úmrtnosti byl podle Ashraf a kol. (2019) zkonstruován pro dlouhodobé hodnocení zdravotního stavu dětí v Indii pomocí faktorové analýzy na základě dat týkajících se porodnosti, perinatální úmrtnosti, novorozenecké úmrtnosti, kojenecké úmrtnosti a úmrtnosti pod pět let. Zjišťovaný index je následně součtem faktorových skóre a procent rozptylu vysvětleného každým faktorem (viz. Satyanarayana a kol., 1995). Další index vícenásobné deprivace byl použit pro srovnání volebních okrsků ve Velké Británii vzhledem k deprivaci a zdraví. K měření tohoto indexu byla použita data: srovnávací úmrtnostní poměry pro muže a ženy ve věku do 65 let, podíl osob pobírajících příspěvek na živobytí pro osoby se zdravotním postižením z celkové populace,

podíl osob v produktivním věku pobírajících dávky v pracovní neschopnosti nebo příspěvek pro těžké zdravotní postižení a podíl nízké porodní váhy. K výpočtu indexu je použita faktorová analýza generující váhy pro kombinování použitých ukazatelů (viz. Kaltenthaler a kol., 2004). Poslední zde zmíněný index nerovnosti ve zdraví představuje dvourozměrný složený index získaný prostřednictvím analýzy hlavních komponent, který umožňuje kvantitativně odhadnout a graficky znázornit nerovnosti ve zdraví na úrovni zemí a regionů (Ashraf a kol., 2019). Mezi ukazatele, které byly pro konstrukci tohoto indexu použity, patří: děti s podváhou, dětská úmrtnost, úmrtí způsobené malárií u dětí ve věku 0 až 4 roky, úmrtí na malárii ve všech věkových kategoriích, porody za účasti kvalifikovaného zdravotnického personálu, očkování proti spalničkám. Vysoké hodnoty tohoto indexu představují větší nerovnosti ve zdravotních výsledcích (viz. Ashraf a kol. 2019). Pro detaily viz. Eslava-Schmalbach a kol. (2008), kteří tento index zavedli.

1.4 Determinanty stavu zdraví

Determinanty stavu zdraví jsou příčiny a podmínky, které komplexně působí na zdraví člověka. Zdraví člověka je tedy složitým způsobem podmíněno (determinováno) kladným i záporným spolupůsobením souboru vnitřních a vnějších vlivů. Jde o širokou škálu osobních, sociálních a ekonomických faktorů i charakteristik životního prostředí. Stejně jako v případě stavu zdraví je možné jejich determinanty popsat mnoha ukazateli, což z nich činí také vícerozměrné kategorie. Některé z těchto determinantů, především místo, kde žijeme, stav životního prostředí, genetika, příjem domácností nebo úroveň vzdělání mají významný dopad na stav zdraví.

Podle WHO (2017) determinanty stavu zdraví zahrnují:

- sociálně ekonomické prostředí,
- fyzické prostředí,
- individuální vlastnosti a chování člověka.

Vyjmenované determinanty, které ovlivňují zdraví pozitivně či negativně, zahrnují následující faktory stavu zdraví (WHO, 2017):

- příjem a sociální postavení,
- vzdělání,
- fyzické prostředí (nezávadná voda, čistý vzduch, zaměstnání, pracovní podmínky aj.)
- podpora přátel, rodiny a komunit, kultura,
- genetika,

- zdravotnické služby,
- pohlaví.

V literatuře se nejednoznačně používají pojmy determinanty stavu zdraví a faktory stavu zdraví. Někdy se tyto pojmy chápou jako synonyma, jiné publikace jejich význam striktně rozlišují. Pojem *faktory stavu zdraví*, přesněji rizikové faktory zdraví, se používají v souvislosti se zdravím jednotlivců ovlivnitelným jednotlivci, např. kouření, spotřeba alkoholu, užívání drog, stravování, nedostatek pohybu. Jako *determinanty zdraví* jsou chápány faktory veřejného zdraví v konkrétním územním celku a čase. Podle publikace *Health at a Glance 2021* (OECD, 2021a) se jedná o financování zdravotnictví, sociálně ekonomickou situaci, kvalitu životního prostředí a úroveň zdravotní péče. Tyto determinanty stavu zdraví může jednotlivec ve snaze o dosažení dobrého zdraví jen těžko ovlivnit. V této práci je upřednostněn druhý uvedený přístup a determinanty stavu zdraví jsou chápány ve vztahu k veřejnému zdraví. Souvisí to také s využitím průřezových agregovaných dat pro země EU-27.

Již v článku Dahlgren, Whitehead (1991) je uveden model hlavních determinantů stavu zdraví, kde většinu z nich lze ovlivnit politickými zásahy. Tyto determinanty jsou rozděleny do pěti kategorií, protože tyto kategorie vyžadují odlišné zásahy v rámci jednotlivých politik. Dahlgren, Whitehead (1991) v rámci determinantů stavu zdraví na prvním místě uvádějí obecné sociálně ekonomické, kulturní a environmentální podmínky. Další úroveň tvoří životní a pracovní podmínky, tzn. pracovní prostředí, vzdělání, zemědělství a výroba potravin, nezaměstnanost, voda a hygiena, služby ve zdravotnictví a bydlení. Tyto úrovně determinantů stavu zdraví vyžadují dlouhodobé strukturální změny skrze politické zásahy. Úroveň nazvaná sociální a komunitní sítě zahrnuje vzájemnou podporu rodiny, přátel a místní komunity. Politiky se na této úrovni snaží podporovat zdraví jednotlivce posilováním sociální a komunitní podpory. Další úroveň zahrnuje individuální faktory životního stylu, které mohou být podpořeny skrze politiky ovlivňováním životního stylu a postojů prostřednictvím zdravotnické osvěty, nebo podpory skupin s nezdravým životním stylem. Poslední úroveň determinantů stavu zdraví se zaměřuje na věk, pohlaví a genetickou výbavu jednotlivce. Tuto oblast není možné ovlivnit žádnou politikou. Jakýkoli politický zásah může být navržen na kterékoli úrovni determinantů stavu zdraví, kde je možné tyto zásahy provádět.

Po třiceti letech od zavedení Dahlgren-Whitehead modelu (tzv. duhového modelu) je v článku Dahlgren, Whitehead (2021) zmapován vývoj tohoto modelu a dále je uvedena jeho důležitost. Duhový model je hlavně užitečný pro odborníky a tvůrce politik mimo sektor zdravotnictví. Umožňuje jim přemýšlet nad možnostmi přijímání zásahů majících vliv na stav zdraví

populace. Dále se tento model osvědčil při navazování spolupráce v různých sektorech. Dříve bylo právě úlohou zdravotnického sektoru navazovat spolupráci s ostatními a poskytovat jim podporu. Výhodou je také celostní pohled na determinanty stavu zdraví spojený s jeho relativní jednoduchostí. Na rozdíl od medicínských modelů, které se zaměřují především na příčiny konkrétní nemoci, duhový model se zabývá pouze determinanty celkového stavu zdraví. Díky zaměření se na jeden determinant stavu zdraví je možné vyvinout komplexní strategii. U medicínských modelů existuje riziko fragmentace preventivních opatření, jestliže je pro různá onemocnění nalezen jeden rizikový faktor. Na druhou stranu podle Dahlgren, Whitehead, 2021 se nejedná o model determinantů nerovností ve stavu zdraví, ale pouze konceptualizuje hlavní determinanty stavu zdraví pro celou populaci. Tuto skutečnost vysvětlují na příkladu týkajícího se nebezpečných pracovních podmínek. V dnešní době nejsou nebezpečné pracovní podmínky hlavním faktorem ovlivňujícím zdravotní stav populace jako celku, ale jedná se o faktor nerovností ve zdraví mezi osobami v kvalifikovaných pozicích a osobami v nekvalifikovaných pozicích.

Existuje řada významných studií, které prokázaly jasnou souvislost mezi sociálně ekonomickým zázemím a zdravím, např. Marmot a kol. (1984) a obecnější diskuse v Marmot (2002). Deatonův článek (2003) zkoumá teoretický základ a empirické důkazy pro souvislost mezi příjmovou nerovností a zdravím mezi chudými i bohatými zeměmi a diskutuje řadu dalších teoretických souvislostí mezi ekonomickou nerovností a zdravím. Kromě příjmových nerovností patřících mezi sociálně ekonomické faktory stavu zdraví používá Carrilero a kol. (2021) ve svém článku úroveň vzdělání, stav bydlení, zaměstnanost, profesní skupinu, rodinný stav, sociální třídu, úroveň deprivace oblasti, národnost atd. jako faktory ovlivňující stav zdraví. Beckfield, Krieger (2009) zhodnotili 45 studií zabývajících se vztahem mezi determinanty sociální politiky a nerovnostmi ve zdraví.

Golembilewski a kol. (2019) se zabývají kombinací neklinických determinantů stavu zdraví s klinickými daty. Neklinické determinanty stavu zdraví mají pro poskytování zdravotní péče a zdravotní politiku stále větší význam díky rostoucímu zájmu o lepší řešení nemedicínských problémů pacientů. Jako hlavní a často prezentované neklinické determinanty stavu zdraví se uvádí:

- *sociálně ekonomické a materiální podmínky* – příjem, chudoba, přístup k jídlu, zaměstnání, životní podmínky, rasa a etnikum, pohlaví a stav pojištění,
- *způsob chování* – konzumace tabáku a alkoholu, strava, užívání návykových látek, dodržování medikace a fyzická aktivita,

- *vybudované prostředí* – infrastruktura,
- *životní prostředí* – kvalita vzduchu, znečištění a klima,
- *veřejná politika* – zdravotní a sociální politika, právo a regulace,
- *zdravotnické služby a podmínky* – přístup a využití zdravotní péče, zdravotní gramotnost a prevalence nemocí,
- *sociální poměry* – rodina, sociální podpora, pečovatelé, rodinný stav, občanská participace a komunitní stigma.

Konkrétní determinanty stavu zdraví použité v této disertační práci jsou uvedeny v podkapitole 4.2.2 Použité determinanty stavu zdraví.

1.5 Nerovnosti ve zdravotním stavu evropské populace

Determinanty stavu zdraví mají zásadní význam při zkoumání nerovností ve stavu zdraví ve zvolených územních celcích a v identifikaci příčin těchto nerovností. I přes pokrok ve zdravotním stavu populace, který je v rámci zemí Evropy zaznamenán, např. snižování úmrtnosti, prodlužování LE_0, existují nerovnosti ve stavu zdraví nejen mezi jednotlivými evropskými zeměmi, ale i v rámci regionů těchto zemí (European Commission, 2018b). Mezi faktory, které způsobují nerovnosti ve stavu zdraví, patří zejména expozice k rizikovým faktorům způsobujícím závažná onemocnění, přístup ke zdravotní péči, zdroje ve zdravotnictví, ale také sociálně ekonomické faktory atd.

Problematikou nerovností ve stavu zdraví se zabývalo a do současné doby i zabývá množství autorů, např. Rodgers (1979); Marmot a kol. (1984), Deaton (2003), Preston (2007), Beckfield, Krieger (2009), d'Hombres a kol. (2013), Jayasinghe (2015), Lundberg a kol. (2016), Pacáková, Kopecká (2018a), Pacáková, Kopecká (2018c), Pacáková a kol. (2019), Bambra a kol. (2020), Gavurova a kol. (2020), Lebano a kol. (2020), Carrilero a kol. (2021), Wilkinson (1992) a Wilkinson, Pickett (2006).

Bylo zjištěno, že rozdělení příjmů silně souvisí s úmrtností (Rodgers, 1979). Wilkinson (1992) našel napříč státy vztah mezi různými mírami příjmové nerovnosti a věkově upravenými příčinami úmrtí. Wilkinson, Pickett (2006) identifikovali 168 analýz ve 155 dokumentech, které uvádějí výzkumná zjištění o souvislosti mezi distribucí příjmů a zdravím populace a klasifikovali je podle toho, do jaké míry jejich zjištění podporují hypotézu, že větší rozdíly v příjmech jsou spojeny s nižšími standardy zdraví populace. d'Hombres a kol. (2013) poskytuje vícerozměrnou analýzu vlivu nerovností v příjmech na zdravotní stav, sociální

kapitál a štěstí. Analýza dat podle této studie však nemůže potvrdit hypotézu silného a významného vlivu nerovností v příjmech na stav zdraví.

Kromě vlivu příjmů na nerovnosti ve stavu zdraví bylo dále zdokumentováno mnoho případů pozitivní korelace mezi zdravím a sociálně ekonomickými ukazateli (Marmot a kol., 1984; Marmot, 2002). Článek Deaton (2003) zkoumá vztah nerovností ve stavu zdraví mezi chudými a bohatými zeměmi. Vztahem národního důchodu ke stavu zdraví se zabývá Preston (2007), který uvádí, že právě národní důchod je nejlepším samostatným ukazatelem životní úrovně v zemích, jelikož zahrnuje hodnotu všech konečných produktů (statků a služeb) a je středem zájmu růstových modelů, od nichž se odvíjejí ekonomická opatření. Beckfield, Krieger (2009) zase zhodnotili 45 studií, které se zabývaly vztahem mezi determinanty sociální politiky a nerovnostmi ve stavu zdraví. Další analýza, kterou uvedl Lundberg a kol. (2016), naznačuje, že nižší míra chudoby je spojena s nižší úmrtností jak u malých dětí, tak u dospělých. Aplikací principů systémového přístupu ke konceptualizaci sociálních determinantů nerovností v oblasti zdraví se zabývá Jayasinghe (2015). Z výše zmíněného vyplývá, že odstranění nerovností v oblasti sociálně ekonomické je významnou podmínkou k odstranění nerovností ve stavu zdraví.

V článku Pacáková, Kopecká (2018c) na základě hodnot vybraných 15 proměnných ve 22 monitorovaných evropských zemích v letech 2000 a 2015 byly aplikací faktorové analýzy získány tři společné faktory: obecný faktor sociální a zdravotní situace, faktory úmrtnosti na závažná onemocnění a faktor nezaměstnanosti. Společně vysvětlili 85,1 % variability původních proměnných v roce 2000 a 89 % variability v roce 2015. Hlavním cílem tohoto článku bylo posoudit a kvantifikovat nerovnosti ve zdravotním stavu obyvatel v závislosti na sociálně ekonomické situaci v evropských zemích v letech 2000 a 2015. Data pro analýzy byla čerpána z databází Eurostatu, OECD a WHO. Jednalo se o agregovaná data týkající se především stavu zdraví, a to LE_0, LE_65, SDR CADs a vybraných CAs na 100 000 obyvatel doplněné o sociálně ekonomické ukazatele, mezi které patří míra nezaměstnanosti, míra dlouhodobé nezaměstnanosti, medián příjmu a výdaje na zdravotní péči na osobu. Tři společné faktory byly následně využity pro posouzení nerovností ve stavu zdraví metodami shlukové a vícerozměrné porovnávací analýzy. V obou letech 2000 a 2015 byl zřejmý podle extrahovaných faktorů značný rozdíl mezi post-socialistickými zeměmi a zbytkem Evropy, ale i částečná redukce těchto rozdílů mezi lety 2000 a 2015, což je příznivý výsledek analýzy.

V článku Pacáková a kol. (2019) zvolených 19 ukazatelů zdravotního stavu obyvatel, výdajů na zdravotní péči, zdrojů zdravotní péče a sociální situace v zemích Evropy umožnilo pomocí

faktorové analýzy extrahovat tři společné faktory, jmenovitě *F1* – faktor zdravotního stavu a výdajů na zdravotní péči (vysvětluje 54,28 % variability původních dat), *F2* – faktor sociálních determinantů zdraví (vysvětluje 16,46 % variability původních dat) a *F3* – faktor personálních a technických zdrojů zdravotní péče (vysvětluje 8,61 % variability původních dat). Tyto tři faktory společně vysvětlují 79,35 % variability původních dat. Grafické zobrazení 25 monitorovaných zemí Evropy ve dvourozměrném souřadnicovém systému s osami dvojic faktorů umožňuje rychle vyhodnotit pozorovanou situaci v každé zemi a také porovnat situaci v různých zemích. V dvourozměrném souřadnicovém systému faktorů *F1* a *F2* je zřejmá přímá závislost faktorů *F1* a *F2*, což znovu potvrzuje významný vliv sociálně ekonomické situace na stav zdraví v monitorovaných zemích Evropy a současně značné nerovnosti v úrovni podle obou faktorů v těchto zemích. Lze pozorovat tři skupiny zemí, jednu s vysokými hodnotami obou faktorů, obsahující všechny *„staré členské země“* EU, druhou s nízkými hodnotami obou faktorů, obsahující pět *„nových členských států“* EU a třetí se střední úrovní faktoru *F1* a nízkou až střední úrovní faktoru *F2*. Extrahované společné faktory byly využity k posouzení nerovností v monitorovaných zemích a dále byly země analyzovány pomocí shlukové analýzy Pro detaily viz. Pacáková a kol. (2019).

Hodnocení vlivu sociálně ekonomické situace na zdravotní stav obyvatel v zemích EU se věnuje rovněž článek Pacáková, Kopecká (2019a). Zdravotní stav je zde charakterizován ukazateli incidence závažných onemocnění a zdravých let života. Pro posouzení sociálně ekonomické situace byly zvoleny ukazatele GDP, průměru a mediánu disponibilního příjmu a míry chudoby a materiální deprivace. Aplikací metod faktorové analýzy byl opět potvrzen jejich významný vzájemný vztah a skrze Wardovu metodu byly vizualizovány shluky těchto zemí s podobnou situací na základě použitých ukazatelů.

Důležitým determinantem stavu zdraví jsou zdravotnické služby. Článek Jindrová, Kopecká (2017a) posuzuje rizikové faktory ve vztahu k úmrtnostem na závažná onemocnění. Jedná se o faktory výdajů na zdravotní péči, životního stylu (konzumace alkoholu) a sociální faktory (např. míra chudoby). V článku je potvrzen významný vliv výdajů do zdravotnictví na zdravotní stav populace.

Zhodnocením vlivu rizikových faktorů chování a faktorů vzdělání na výkonnost systémů zdravotní péče v členských nebo kandidátských zemí EU pomocí metod shlukové analýzy se zabývají Balçık a kol. (2021). Na základě tohoto článku bylo prokázáno, že faktory chování a vzdělání mají vliv na výkonnost systému zdravotní péče. Z výsledků je zřejmé, že především

post-socialistické státy se nachází vždy společně v jednom shluku, popř. tvoří shluky spolu se státy jižní Evropy.

Dále Gavurova a kol. (2020) se ve svém článku věnuje hodnocení efektů vybraných indikátorů stavu zdraví na konkurenceschopnost rozvinutých zemí OECD. Na základě poznatků v tomto článku je konstatováno, že by se rozvinuté země měly zaměřit na snižování genderových nerovností ve stavu zdraví, protože snížení těchto nerovností vede ke zvýšení jejich konkurenceschopnosti. Lebano a kol. (2020) se zabývají poskytováním zdravotní péče migrantům a jejich přístupem k ní ve vybraných evropských zemích. Existují nerovnosti v přístupu ke zdravotní péči mezi migranty a nemigrujícími osobami. Zejména se jedná o komunikační bariéry, nadužívání pohotovostních služeb, nevyužívání primární zdravotní péče a diskriminace. Carrilero a kol. (2021) přezkoumali zprávy vlád EU-15, které se týkají sociálně ekonomických nerovností ve zdravotním stavu obyvatelstva. Mezi nejdůležitější sociálně ekonomické ukazatele mající vliv na zdraví uvádějí úroveň vzdělání, sociální třídu a národnost. Pandemii COVID-19 a její dopady na nerovnosti v oblasti zdraví řeší článek Bambra a kol. (2020). V minulosti způsobovaly pandemie vyšší míru infekce a úmrtnosti mezi nejvíce znevýhodněnými komunitami, což se odráží v pandemii COVID-19 i dnes. COVID-19 má syndemickou povahu, tzn., interaguje s již existujícími sociálními nerovnostmi u chronických onemocnění a sociálními determinanty stavu zdraví. COVID-19 snižuje LE, nicméně u některých znevýhodněných skupin například v UK (Spojené království) se LE snižovala už před pandemií.

V článku Rosicova a kol. (2015) autoři provedli *Poissonovu regresní analýzu,* ve které použili úmrtnost jako vysvětlovanou proměnnou a počet obyvatel podle věkové a genderové kategorie jako prediktory. Do modelu pak přidali následující charakteristiky: úroveň vzdělání, nezaměstnanost, příjem a podíl Romů. Pomocí Poissonovy regrese bylo potvrzeno, že rizika úmrtnosti byla nejvyšší ve vybraných lokalitách SK, ve kterých byl vysoký podíl populace s nízkým vzděláním a podíl Romů. Výsledky poukazují na nutnost řešit zdravotní potřeby deprivovaných městských oblastí na příkladu Bratislavy a Košic. Zlepšování každodenních životních podmínek je především úkolem místní samosprávy a občanské společnosti, podporované národní vládou, s cílem vytvořit mechanismy místní participativní správy při vytváření zdravějších a bezpečnějších měst.

Snahy o snížení sociálních nerovností ve zdraví se často zaměřují na geografické rozdíly, protože politika se řídí nejsnáze ve správních jednotkách, jako jsou místní samosprávy.

Nerovnosti ve zdraví mezi sociálně ekonomickými a etnickými skupiny patří mezi hlavní výzvy pro veřejné zdraví po celém světě a jsou v poslední době předmětem mnoha studií.

Mnoho publikací se věnuje vztahům mezi nerovnostmi v příjmech a stavem zdraví jak na agregované, tak na individuální úrovni. Například článek Fiscella, Franks (2000) zkoumá vztahy mezi nerovnostmi v příjmech, vlastním hodnocením zdraví a úmrtností v USA. Uvádějí, že při úpravě dat týkajících se individuálních příjmů podle věku a pohlaví má příjmová nerovnost mírný vliv na úroveň deprese a vlastní hodnocení zdraví, ale žádný vliv na nemocnost nebo úmrtnost. Na druhou stranu má individuální příjem větší vliv na úroveň deprese, vlastní hodnocení zdraví, nemocnost a úmrtnost. Podle Leon-Gonzalez, Tseng (2011) většina studií používajících agregované ukazatele je kritizována za opomenutí nelineárních vazeb mezi zdravím (úmrtnostmi) a příjmem na individuální úrovni. Gerdtham, Johannesson (2004) se zabývají také vztahy mezi příjmy a úmrtnostmi. Mimo jiné uvádějí, že vztah mezi příjmem a úmrtností je nelineární s klesajícím efektem příjmu u vyšších příjmových skupin. Analýza provedená v článku Beaujot, Niu (2005) ukazuje na větší důležitost ukazatelů příjmu, vzdělání a pracovního postavení při hodnocení stavu zdraví na individuální úrovni než na agregované. I přes důležitost ukazatelů na individuální úrovni jsou ukazatele na agregované úrovni považované za relevantní, pokud se jedná o menší oblasti, u kterých k agregaci dochází. Například článek Marra a kol. (2011) uvádějí, že existuje souvislost mezi nižším sociálně ekonomickým statusem a horším stavem zdraví měřeným pomocí příjmu a vzdělání jak na individuální, tak na agregované úrovni.

Poznání a následná řešení nerovností v oblasti zdraví může učinit společnosti inkluzivnějšími. K tomu mohou systémy veřejného zdraví přijmout širokou škálu politických řešení, od přechodu k poskytování primární péče zaměřené na pacienta, přes rozšíření pokrytí zdravotní péče, zlepšení zdravotní gramotnosti, prevenci a intervence v oblasti veřejného zdraví. Kromě zdravotnictví mohou k řešení nerovností ve zdraví přispět také politiky týkající se trhu práce, vzdělávání, životního prostředí, bydlení a sociální politiky. Lze očekávat, že politiky zaměřené na snižování rozdílů v různých oblastech a hodnocení klíčových determinantů zdraví budou mít pozitivní vliv na zdraví obyvatelstva.

2 Současné přístupy ke snižování rozměrnosti dat

Většina reálných datových souborů týkající se stavu zdraví, popř. jeho determinantů, které poskytují databáze Eurostatu, OECD, WHO a další, obsahují mnoho proměnných (ukazatelů). Existuje celá řada možností snižování rozměrnosti na úrovni proměnných v datovém souboru, které představují předzpracování datového souboru před samotným trénováním modelu. Pramoditha (2021) uvádí 11 technik snižování rozměrnosti, které jsou dnes používané. Mezi tyto techniky řadí metody, které zachovávají pouze nejdůležitější proměnné, např. *zpětná eliminace* (Backward Elimination) a *dopředná selekce* (Forward Selection). Další techniky jsou založeny na aplikaci vhodné transformace na množině původních proměnných. Tyto techniky je možné rozdělit na lineární, mezi které patří *analýza hlavních komponent* (PCA – Principal Component Analysis), *faktorová analýza* (FA – Factor Analysis), *řídká analýza hlavních komponent* (SPCA – Sparse Principal Component Analysis), *lineární diskriminační analýza* (LDA – Linear Discriminant Analysis), *zkrácený rozklad singulární hodnoty* (Truncated SVD - Singular Value Decomposition) a nelineární (Manifold Learning), kam se řadí například *jádrová analýza hlavních komponent* (KPCA – Kernel Principal Component Analysis), *vícerozměrné škálování* (MDS – Multidimensional Scaling) a *izometrické mapování* (Isomap – Isomentric mapping). V článku Van Der Maaten a kol. (2009) uvádějí, že nelineární techniky extrakce proměnných fungují lépe na uměle vytvořených úlohách než na úlohách každodenního života.

Pramoditha (2021) uvádí, že i přes ztrátu informací v podobě variability původních dat, přináší snižování rozměrnosti původních proměnných několik důležitých výhod pro metody strojového učení. Různé techniky pro snižování rozměrnosti mohou generovat 2D (popř. 3D) projekce a umožnit tak vizuální průzkum klastrových struktur vícerozměrných datových souborů (Xia a kol., 2021). Například ve studii Han, Ge (2020) je zjišťován vliv snižování rozměrnosti dat na výběr akcií prostřednictvím shlukové analýzy v různých tržních situacích.

Anowar a kol. (2021) považují metody snižování rozměrnosti dat za užitečné pro algoritmy strojového učení, protože při existenci mnoha ukazatelů (měření) v datovém souboru je pro rozhodování užitečná jen jejich část. U modelu, který je trénovaný na mnoha ukazatelích, jeho přesnost a výkon klesá kvůli zapojení nevýznamných nebo silně korelovaných ukazatelů. Dále se zabývají porovnáním vybraných algoritmů extrakce proměnných. Nejprve jsou algoritmy koncepčně porovnány a následně empiricky vyhodnoceny a porovnány výsledky algoritmů extrakce proměnných na různých souborech dat v binárním a vícetřídním nastavení na základě korelačních metrik a vizualizace dat. V článku Cao a kol. (2003) se autoři zabývají

aplikací a porovnáním PCA, *analýzy nezávislých komponent* (ICA – Independent Component Analysis) a KPCA v rámci *metody podpůrných vektorů* (SVM – Support Vector Machine). Aplikace metod pro snížení rozměrnosti ukazatelů zlepšují výsledky v rámci SVM a metody ICA a KPCA dosahují lepších výsledků než PCA. Autoři Xie a kol. (2016) uvádějí, že přesnost klasifikace je možné vylepšit prostřednictvím kombinací metod pro snížení rozměrnosti ukazatelů. Porovnání výsledků metod pro snížení rozměrnosti ukazatelů je také řešeno ve článku Niskanen, Silvén (2003), kde jsou tyto výsledky kvantitativně hodnoceny prostřednictvím speciálních klasifikačních případů.

Porovnáním technik snížení rozměrnosti ukazatelů pro využití v metodách shlukové analýzy se zabývají Araújo a kol. (2011). Toto porovnání provádějí skrze *upravený Randův index,* který zahrnuje znalost tříd každého objektu v datovém souboru. Článek Xiang a kol. (2021) porovnává deset metod redukujících rozměrnost vysoko-rozměrných jednobuněčných RNA-seq dat. Pro porovnání metod pro snížení rozměrnosti ukazatelů je dále aplikována metoda shlukové analýzy, metoda *k*-průměrů. Následně jsou použity známé buněčné populace k výpočtům metrik upraveného Randova indexu, *normalizované vzájemné informace* a *Silhouetova koeficientu*. Problematice porovnání algoritmů snižování rozměrnosti ukazatelů pro shlukování arabského textu se věnuje Mohamed (2020). Využívá dvou hodnotících kritérií: *přesnost algoritmu shlukové analýzy* (AC – Accurancy of Clustering) a normalizovanou vzájemnou informaci (MI – Mutual Information). Postup Mohamed (2020) uvádí následující:

- snížení rozměrnosti ukazatelů prostřednictvím zvolených algoritmů,
- aplikace metody *k*-průměrů na původní datové matice a matice získané pomocí algoritmů pro snížení rozměrnosti,
- získání správných označení shluků pomocí tzv. *Maďarského algoritmu* (Hungarian Algorithm),
- výpočet AC a MI.

Gracia a kol. (2014) se zabývají porovnáváním algoritmů pro snížení rozměrnosti ukazatelů ve smyslu ztráty kvality. Navrhují způsob porovnání a analyzování různých algoritmů pro snížení rozměrnosti ukazatelů z hlediska ztráty kvality, kterou tato redukce způsobí.

2.1 Snižování rozměrnosti ukazatelů

Tato podkapitola má za cíl vysvětlit význam snižování rozměrnosti ukazatelů a představit některé z využívaných metod a nové přístupy na základě rozsáhlé odborné literatury. Jak bude dále uvedeno, výstupy metod sloužících pro snížení rozměrnosti ukazatelů jsou vhodné

k vizualizaci dat v nízko-rozměrném prostoru, ale také jako vstupní proměnné pro metody shlukové analýzy nebo pro další modelování sledovaných jevů. Podrobnému popisu metod pro snižování rozměrnosti ukazatelů využitých v této práci je věnována podkapitola 5.1 v kapitole 5 Použité metody.

Snižování rozměrnosti dat patří mezi důležité kroky jejich přípravy pro další analýzy. Podle Dash a kol. (1997) je snižování rozměrnosti důležité pro účinnou manipulaci *s velkými soubory dat* (big data), což vyžadují např. *nástroje hloubkové analýzy dat* (data miningu). Na základě Pramoditha (2021) a Terek a kol. (2010) snižování rozměrnosti dat má vést k těmto výhodám:

- úspoře času při trénování modelu za pomoci algoritmů strojového učení (ve vysoko-rozměrném prostoru je většina datových bodů daleko od sebe, a proto algoritmy strojového učení nemohou efektivně a účinně trénovat na těchto datech),
- odstranění problému *„přeučení“* (overfitting) modelu (ve vysoko-rozměrném prostoru se modely stávají složitějšími a mají tendenci se *„přeučit“*),
- odstranění šumu v datech (zlepší se přesnost modelu),
- přehledné vizualizaci dat,
- řešení problému multikolinearity,
- transformaci nelineárních dat na lineárně oddělitelné.

K redukci rozměrnosti dat může docházet jak na úrovni případů (objektů), tak na úrovni proměnných v datovém souboru (Terek a kol., 2010).

Existují dva způsoby snižování rozměrnosti dat. Je to buď *výběr proměnných* (Feature Selection), nebo *extrakce proměnných* (Feature Extraction) (Dash a kol., 1997). Skrze výběr proměnných je stanovena podmnožina proměnných z původní množiny. Tato podmnožina proměnných je v ideálním případě nezbytná a dostatečná pro popis cílového konceptu. Cílový koncept může být dán příslušností k nějaké třídě. Data, v nichž existuje informace o příslušnosti prvků k určitým skupinám, jsou označována jako *data s učitelem* (Supervised Data). V případě, že jsou k dispozici *data bez učitele* (Unsupervised Data), tzn., neexistují informace o příslušnosti prvků k určitým skupinám, používají se metody extrakce proměnných. Tyto metody vytvářejí nové proměnné, které jsou nekorelované a zachovávají co nejvíce rozptylu původních dat.

2.1.1 Složené ukazatele

Složené ukazatele (CI – Composite Indicators) se stávají stále populárnější, protože poskytují komplexní pohled na jev, který nelze zachytit jedním ukazatelem. Poskytují srovnání např.

území (států, regionů, měst aj.), která lze použít k ilustraci ekonomických, sociálních, environmentálních, společenských, technologických a jiných problémů. CIs však mohou vést k přijímání špatných rozhodnutí, pokud jsou špatně konstruovány nebo interpretovány. Při konstrukci CIs je třeba zvážit výběr proměnných, metod, vah proměnných a řešení chybějících hodnot.

Problematikou CIs na úrovni NUTS 2 regionů se zabývají např. Staníčková, Melecký (2018). Existuje celá řada metod pro vytváření CIs a množství jejich aplikací v četných publikacích polských statistiků, např. Hellwig (1968); Grabiński (1992); Grabiński a kol. (1983); Zeliaś, Malina (1997), Młodak (2006); Pawełek (2008); Kuc, (2012). CIs týkající se stavu zdraví nebo jeho determinantů uvádějí také např. články Kopecká (2019); Pacáková, Kopecká (2018a); Pacáková, Kopecká (2018b) nebo Pacáková a kol (2020).

Grabiński (1984) uvádí, že *vícerozměrná porovnávací analýza* se zabývá metodami a technikami porovnávání vícerozměrných objektů. Jedním z konkrétních problémů je zde stanovení lineární hierarchie (lineárního uspořádání) množiny objektů v multidimenzionálním prostoru proměnných z hlediska určitých komplexních charakteristik, které nelze měřit přímým způsobem (stav zdraví, sociálně ekonomická úroveň, ekonomický vývoj, životní úroveň atd.). Vstupem pro tuto metodu je matice *n* objektů (řádků) a *p* proměnných (sloupců). Podle Stankovičová, Vojtková (2007) je na rozhodnutí analytika, kterou metodu vícerozměrné porovnávací analýzy použije, popř. je možné použít více těchto metod a porovnat je. Jednou z možností je použít *bodovací metodu*, u které je třeba každou hodnotu proměnné obodovat. Pro každou proměnnou je nalezen objekt, který je podle dané proměnné nejlepší a tomu je přiřazeno 100 bodů. Pokud jsou žádoucí vysoké hodnoty proměnné, jedná se o proměnnou, která se nazývá v polské literatuře *stimulant* a např. v publikaci Stankovičová, Vojtková (2007) *„proměnná typu +“*. Pokud jsou žádoucí nízké hodnoty proměnné, jedná se o proměnnou, která se nazývá v polských aplikacích *destimulant,* resp. *„proměnná typu –„* (Stankovičová, Vojtková, 2007). Zbývajícím objektům přiřadíme tolik bodů, kolik procent tvoří hodnota dané proměnné z nejlepší hodnoty. Výsledkem této metody je tzv. *syntetická proměnná*, nejčastěji vyjádřena jako aritmetický průměr počtu bodů pro každý objekt přes všechny proměnné.

Bodovací metoda je jednou z metod standardizace proměnných, charakterizujících vícerozměrné objekty. Hodnoty těchto proměnných mají často odlišnou úroveň a jsou měřeny v různých měrných jednotkách, což neumožňuje jejich prostou agregaci. Standardizací proměnných dosáhneme jejich porovnatelnosti a možnosti agregace, resp. vytvoření syntetické proměnné (viz. např. Pawełek, 2008). Specifický způsob standardizace, tzv. *unitarizace* mění

rozsah každé proměnné na konstantní, jednotkový interval. Hodnota proměnné nebo její vzdálenost od jednoho z limitů původního rozsahu se dělí rozpětím rozsahu.

Úspěšnou aplikaci kvantifikace syntetických ukazatelů stavu zdraví a jeho různých determinantů, jako jsou sociálně ekonomická situace, výdaje na zdravotní péči, personální a technické zdroje zdravotní péče, determinanty předčasných úmrtí v evropských zemích, prezentují publikace Pacáková a kol. (2016); Pacáková, Žáková (2019); Pacáková, Kopecká (2018a), Pacáková, Kopecká (2018c), Kopecká (2019d), Pacáková a kol. (2020); Pacáková a kol. (2021).

2.1.2 Analýza hlavních komponent a její možnosti

Často používanou metodou pro snižování rozměrnosti dat na úrovni proměnných je analýza hlavních komponent (PCA), vytvářející nové *hlavní komponenty* (PCs – Principal Components), které jsou lineárními kombinacemi původních proměnných při co nejmenší ztrátě informace (rozptylu). Hlavním cílem PCA je spíše než snížení rozměrnosti úlohy zjištění jejího skutečného rozměru. PCA nepotřebuje označení skupin pro extrahování hlavních komponent, ale pouze hodnoty původních proměnných (Dash a kol., 1997; Stankovičová, Vojtková, 2007).

PCA má za cíl transformovat původní proměnné do menšího počtu nových proměnných (Holčík a kol., 2015). Coste a kol. (2005) se zabývají metodologickými otázkami při určování rozměrnosti (dimenze) zdravotních CIs pomocí PCA. Uvádějí, že stanovení počtu PCs silně ovlivňuje faktorový model. Jejich cílem bylo ilustrovat proměnlivost faktorových modelů získaných použitím různých publikovaných pravidel pro stanovení počtu PCs. Hudrliková (2013) se zabývá CIs výkonnosti členských států EU, které mohou sloužit pro mezinárodní porovnávání. Jednou z uvedených metod je PCA, která je vhodná v případě silných korelací v původních datech.

V článku Latifoğlu a kol. (2008) je PCA součástí návrhu lékařského diagnostického systému pro diagnostiku ateroskleróz (kornatění cév). Pomocí PCA je snížen rozměr úlohy z 61 proměnných týkajících se ateroskleróz na čtyři PCs. Dále jsou tyto čtyři PCs váženy a použity pro klasifikaci v rámci klasifikátoru *umělého imunitního rozpoznávacího systému* (AIRS – Artificial Immune Recognition System), který klasifikuje pacienty na zdravé nebo s aterosklerózou. Rodrigues a kol. (2014) řeší problematiku stárnutí populace. Díky silným korelacím mezi sledovanými ukazateli, jako jsou fyzická soběstačnost, fyzická aktivita, zdravotní potíže, vnímání zdravotního stavu aj., byla pomocí PCA umožněna identifikace

skupin lidí se společnými vlastnostmi. PCA je využitá také pro odvození stravovacích návyků u starších lidí a jejich porovnání společně se shlukovou analýzou (Thorpe a kol., 2016). Článek zkoumá souvislosti mezi těmito návyky a sociálně demografickým a zdravotním chováním. Úmrtnostmi mezi různými sociálně ekonomickými vrstvami v thajské společnosti se zabývají Aungkulanon a kol. (2017). PCA použili ke konstrukci indexu sociálně ekonomické deprivace. Následně byla použita shluková analýza pro seskupení sociálně ekonomického stavu a úmrtností z jednotlivých příčin.

Rotované komponenty

Jak již bylo zmíněno, pro výpočet PCs jsou potřebné hodnoty původních proměnných. Každá PC je lineární kombinací původních proměnných, to znamená, že komponentní zátěže jsou obvykle nenulové, což komplikuje jejich interpretaci. Jednou z možností, jak dosáhnout snadnější interpretace PCs, je využít rotací (viz. Zou a kol., 2006).

Podle Hebák a kol. (2015) nadstavbou PCA je faktorová analýza (FA), která v nejpoužívanějších variantách přímo z PCA vychází. FA je označením skupiny metod pro zpracování a analýzu dat, které jsou často vnímané jako metody explorativní (popisné). Stejně jako PCA slouží FA ke snížení počtu proměnných, nicméně má širší metodický aparát. Hlavním úkolem FA je vysvětlit závislosti mezi původními proměnnými (Hebák a kol., 2015; Stankovičová, Vojtková, 2007). U FA dochází k rotaci faktorů, aby tyto faktory co nejsnadněji popisovaly původní proměnné, čehož je dosaženo v případě, kdy jsou faktory co nejblíže skupině silně korelovaných proměnných. V této situaci se však může stát, že jsou faktory do určité míry navzájem korelovány (na rozdíl od PCA). Nejznámější rotací faktorů je *ortogonální varimax rotace* (viz. Holčík a kol., 2015).

Často se mylně uvádí, že PCA s rotací je zvláštním případem FA, což je vidět i u některých počítačových softwarů (STATISTICA, SPSS, STATGRAPHICS), kde se využívá PCA jako jedna z možností pro odhad parametrů faktorového modelu. Nicméně FA reprodukuje faktory a PCA hlavní komponenty. Je pravdou, že rotací komponent dochází k jejich snadnější interpretaci (stejně jako v případě rotace faktorů), ale jak je uvedeno v balíku (package) *psych* v rámci funkce *principal* v programu R, rotované komponenty už nejsou *komponentami* (PC), ale *rotovanými komponentami* (RCs – Rotated Components). U PCA bez rotace se původní souřadnicový systém natáčí ve směru největší variability. Nicméně v případě PCA s rotací sice zůstává celkový rozptyl v rámci nového podprostoru nezměněný, tzn. stejný jako u PCA

bez rotace, ale je rozložen mezi RCs rovnoměrněji. To znamená, že informace o povaze skutečně dominantních komponent může být ztracena (Jolliffe, 2002; Revelle, 2020).

Základem pro použití jak metody PCA, tak metody FA je existence silných korelací mezi původními proměnnými, protože pak dávají výsledky těchto metod smysl (Tabachnick, Fidell, 2007).

Například Bountziouka, Panagiotakos (2012) uvádějí, že PCA má široké využití ve výživové epidemiologii pro extrahování vzorů ve stravování. V této studii je použita jak PCA bez rotace, tak PCA s vybranými rotacemi na data dvou dotazníkových šetření týkající se stravování. PCA bez rotace je úspěšnější než PCA s rotacemi. V případě PCA bez rotace dochází k dobré shodě mezi PCs odvozenými pro obě dotazníková šetření. To znamená, že v obou případech jsou extrahovány PCs obsahující stejné složení potravin, které pokaždé vzhledem k jejich interpretaci vysvětlují podobná procenta variability. Na druhé straně RCs jsou extrahovány opět se stejným složením potravin, ale takto interpretované RCs vykazují pokaždé odlišná procenta variability. Dalšímu využití PCA se věnují Livesley a kol. (1998). Konkrétně identifikují struktury poruch osobností zkoumáním fenotypových a genetických struktur opět pomocí PCA s rotací.

Řídká analýza hlavních komponent

Výše uvedené metody patří k základním a nejpoužívanějším metodám pro snížení počtu proměnných (nalezení správné rozměrnosti dat). Stankovičová, Vojtková (2007) uvádějí, že PCA byla navrhnuta K. Pearsonem už v roce 1901. Nicméně podle Erichson a kol. (2018b) existují modernější verze PCA, např. SPCA. Tato verze vytváří lépe interpretovatelné *řídké hlavní komponenty* (SPC – Sparse Principal Component) skrze *řídké váhové vektory* (zátěže), které mají pouze několik *„aktivních“* nenulových hodnot. SPCA vytváří SPCs jako lineární kombinace několika původních proměnných. Podle Luss, d'Aspremont (2010) je možné SPCA použít pro shlukování a výběr proměnných. Jestliže se SPCA použije jako nástroj ke shlukování, SPCs umožní identifikovat shluky pouze podle několika původních proměnných.

Chang a kol. (2015) se zabývají *„řídkým“* modelováním prostorových proměnných spojených s výskytem astmatu. SPCA je zde využita jako první krok pro redukci dimenze 1 117 proměnných týkajících se životního prostředí, kterými bylo ohodnoceno 199 220 pacientů před testováním na astma. Metoda SPCA zlepšuje interpretaci PCs, ale také minimalizuje ztrátu

vysvětleného celkového rozptylu. SPCs jako lineární kombinace několika původních proměnných jsou zbaveny proměnných, které nepřispívají k dobré interpretaci PCs u PCA.

2.1.3 Vícerozměrné škálování a jeho možnosti

Další skupinou metod, které redukují vícerozměrný prostor, je vícerozměrné škálování (MDS). Cílem těchto metod je snížit počet dimenzí v datovém souboru a zobrazit objekty (popř. proměnné) v novém souřadnicovém systému, ve kterém je možné blíže prozkoumat jejich vzájemné vztahy. MDS může sloužit k identifikaci přirozených shluků objektů, tzn., plní podobné cíle jako např. PCA. Rozdíl mezi PCA a MDS je v charakteru vstupních dat, kde MDS pracuje s mírami podobnosti (nepodobnosti), které nemusejí vycházet ze vztahů popsaných pomocí korelační nebo kovarianční matice (Hebák a kol., 2015).

Přehledem možností MDS se zabývají např. Hout a kol. (2013). Uvádějí, že MDS je možné členit podle toho, zda implementuje metrické nebo nemetrické algoritmy. Rozdíl mezi metrickým (např. klasickým) a nemetrickým MDS je ve způsobu zacházení s nepodobnostmi. V případě metrického MDS je pracováno s původními hodnotami proměnných, zatímco v případě nemetrického MDS s pořadím hodnot (Hebák a kol., 2015). Pokud jsou k dispozici kvantitativní data (intervalová, poměrová) je možné použít metrické MDS, nicméně pokud jsou k dispozici data kvalitativní (ordinální data) používá se nemetrické MDS, jak uvádí např. Hout a kol. (2013). Výběr počtu dimenzí v případě klasického metrického MDS je ekvivalentní s výběrem hlavních komponent v PCA. Další metody MDS představují *modely individuálních preferencí* např. (INDSCAL – Individual Difference Scaling), u kterých je uvažováno více vstupních matic, tzn. každá matice pro jeden subjekt, jeden rok apod. (viz. Hebák a kol., 2015).

Existuje celá řada metod MDS, jejichž postupy se liší. Pomocí těchto metod je možné získat více než jen zajištění vizualizace dat v prostoru nižší dimenze oproti dimenzi původních dat. PCA je matematicky identická s metrickým MDS založeným na euklidovské vzdálenosti. Na rozdíl od PCA, která se zaměřuje na samotné dimenze, MDS se více zaměřuje na vztahy mezi objekty (Hout a kol., 2013).

Aplikací MDS se zabývají např. Spruyt a kol. (2006). Řeší komplexní vztahy mezi poruchami spánku u dětí a každodenním chováním pomocí MDS. Tímto způsobem byly extrahovány tři dimenze, kde první představovala dimenzi chování, druhá dimenzi zdraví dítěte a třetí situační aspekty předchozí i současné. Využití MDS společně se shlukováním je užitečné například u spojení nových mutací HIV se selháním léčby a jejich seskupení do nových mutačních komplexů, čemuž se věnují Sing a kol. (2005). Tyto metody však neodpovídají na otázku, zda

nové mutace přímo přispívají ke zvýšené rezistenci, a proto dále využívají klasifikační modely, a to rozhodovací stromy a SVM. Dickes, Valentová (2013) se zabývají vícerozměrným měřením sociální soudržnosti v 47 zemích Evropy za využití individuálních dat. Pomocí MDS bylo stanoveno, že sociální soudržnost se skládá z ukazatelů důvěry v instituce, solidarity, sociální a kulturní účasti a účasti politické. INDSCAL naznačuje, že uvedené ukazatele sociální soudržnosti jsou ve všech 47 zemích Evropy rovnocenné. Dále bylo MDS využito ke stanovení pozic a shluků 47 zemí v 2D prostoru.

2.1.4 Kernel analýza hlavních komponent a její možnosti

Dalším rozšířením PCA je kernel analýza hlavních komponent (KPCA). Metody využívající jádro patří mezi metody strojového učení, které je možné použít k řešení nelineárního problému. Hlavní myšlenkou těchto metod je předzpracování dat a jejich promítnutí do prostoru vyšší dimenze, kde je pravděpodobnější, že lineární metody budou lépe fungovat (Pilario a kol., 2020).

PCA využívá lineární projekce pro výpočet reprezentace dat v nižších dimenzích. Problém nastává, pokud jsou vztahy mezi původními proměnnými nelineární. U PCs je rozptyl maximálně zachován, což není vhodné v případě nelineárních vztahů. Proto byla vyvinuta metoda KPCA (Gutmann, 2017).

Existuje mnoho publikací, věnujících se konkrétním KPCA a jejich využití, např. Ezukwoke, Zareian (2019); Hoffman (2007); Rathi a kol. (2006); Reverter a kol. (2014); Romdhani a kol. (1999); Schölkopf a kol. (1998); Sheng a kol. (2016); Wang a kol. (2017). Základní myšlenkou KPCA je zobrazit původní data do prostoru vyšší dimenze skrze konkrétní funkci a následně aplikovat lineární PCA. Lineární PCA ve vysoko-rozměrném prostoru odpovídá nelineární PCA v původním vstupním prostoru. Pokud je k dispozici vstupní datová matice o velikosti n x p, pak PCA může najít až p nenulových vlastních čísel na rozdíl od KPCA, která může extrahovat až n nenulových vlastních čísel. Nicméně často v obou případech stačí vybrat několik PCs (popř. kPCs – *kernel hlavní komponenty*) (Kallas a kol., 2012).

Otázku výběru *kernel funkcí* řeší Pilario a kol. (2020), kteří se dále zabývají způsoby výběru parametrů těchto funkcí a problémy, týkajícími se použití kernel metod. Mimo jiné v tomto článku je studováno 230 publikací, týkajících se používání kernel metod a jsou identifikovány problémy související s jejich používáním. Gutmann (2017) uvádí dvě často používané funkce jádra, kterými jsou *polynomiální* a *gaussovské*. Mezi další funkce jádra patří např. *lineární, sigmoidní funkce jádra* (Linear and Sigmoid Kernel Function), *Besselova funkce jádra prvního*

druhu (Bessel Function of the First Kind Kernel) aj. (viz. Karatzoglou a kol., 2004; Pilario a kol., 2020; Schölkopf a kol., 1998 nebo Shawe-Taylor, Sun, 2014). Mezi nejpoužívanější funkce jádra patří *Gaussova RBF* (Radial Basis Function). Oblíbenost této metody spočívá především v její hladkosti a flexibilitě (Pilario a kol., 2020).

Důležitou otázkou, kterou je třeba řešit v rámci KPCA, je tedy výběr funkce jádra a jeho parametrů. Výběr funkce jádra je důležitá záležitost, protože ovlivňuje dosažené výsledky. Nejpoužívanějšími funkcemi jádra jsou již zmiňované funkce v tomto pořadí: RBF, polynomická a sigmoidní. V rámci jader samotných dochází k vytváření nových alternativ, např. smíšených jader (Pilario a kol., 2020).

Podle Pilario a kol. (2020) neexistuje pro určení parametrů jader teoretický základ, nicméně jejich hodnoty musí být specifikovány již před samotnou aplikací metody. Dále uvádí, že nejčastěji dochází k výběru parametrů jednotlivých jader *empiricky* nebo neuvedením postupu určení hodnot těchto parametrů. Méně často se parametry stanovují na základě *křížové validace* nebo *optimalizace*. Empirický výběr parametrů znamená, že například pro jádro RBF se jeho parametr σ může určit např. na základě rozptylu dat nebo jejich rozměrnosti, pro detail viz. Lee a kol. (2004); Lee a kol. (2008). Například Godoy a kol. (2014) navrhují stanovit tento parametr jako dvakrát počet proměnných ($2*p$).

Další možností pro výběr parametrů jader je křížová validace, popř. *k-násobná křížová validace* (*k*-fold cross-validation) (Pilario a kol., 2019; Pilario a kol., 2020). Při *k*-násobné křížové validaci je datový soubor rozdělen do *k* skupin, z nichž *k*-1 skupin je určeno pro trénování, zatímco zbývající skupina je určena pro testování. Tento postup se opakuje tak, že každá část je použita pro testování právě jednou (Holčík a kol., 2015). KPCA může být použita např. pro regresi a klasifikaci, to znamená, vhodnost parametrů je možné hodnotit např. na základě chyb klasifikace (Alam, Fukumizu, 2014). V případě metod *učení s učitelem* (Supervised Learning), kdy je ke vstupní množině dat určen výstup (např. u regrese nebo klasifikace), se křížová validace běžně používá pro výběr *hyperparametrů* algoritmů jádra. Cílem článku Alam, Fukumizu (2014) je navržení metody výběru hyperparametrů v KPCA pro jádro a počet komponent, založené na křížové validaci pro srovnatelné chyby rekonstrukce předobrazů v původním prostoru. Výběr počtu komponent se odvíjí právě od stanovení hyperparametrů. Tito autoři dále uvádějí, že výběr vhodného jádra je sice nezbytný pro příznivé výsledky, avšak v případě *učení bez učitele* (Unsupervidsed Learning), kdy není ke vstupní množině dat určen výstup, nebyly zavedeny žádné dobře podložené metody. Jedná se například o případ KPCA,

která není použita pro regresi nebo klasifikaci, to znamená, že vhodnost parametrů nelze hodnotit na základě chyb klasifikace. Problematikou důležitosti hyperparametrů algoritmů strojového učení se zabývá Probst a kol. (2019), kteří uvádějí, že se jedná o parametry, k jejichž nastavení dochází před spuštěním algoritmu, to znamená, že na rozdíl od parametrů modelu je nelze získat při trénování modelu. Elgeldawi a kol. (2021) vysvětlují rozdíl mezi parametry modelu a hyperparametry. Pokud se při automatickém učení algoritmu upravují vnitřní parametry na základě dat, jedná se o *parametry modelu*. Jestliže však nedochází u parametrů k jejich nastavování během procesu učení, jedná se o hyperparametry. Hyperparametry musí být nakonfigurovány již před procesem učení.

Křížová validace přináší lepší odhady parametrů jádra než přístup empirický (Fu a kol., 2017). Podle Pilario a kol. (2020) je další možností stanovení parametrů jádrové funkce optimalizace. Například Lázaro a kol. (2015) používají pro odhad parametru σ jádrové RBF funkce optimalizační algoritmus typu Quasi-Newton. Dále Williams (2002) představuje možnost výběru parametrů jádra skrze maximalizaci podílu vysvětleného rozptylu pomocí prvních r vlastních čísel. V případě, že hodnota parametru σ (šířka pásma jádra RBF) je příliš vysoká, model ztratí schopnost objevovat nelineární vzory. V opačném případě, pokud je hodnota parametru σ příliš nízká, je model příliš citlivý na šum v množině trénovacích dat. Mezi další autory, kteří se zabývají volbou parametrů jádrové funkce, patří např. Han a kol. (2011) nebo Kallas a kol. (2014).

Jednu z hlavních nevýhod nelineární KPCA oproti lineární PCA uvádějí Shawn (2006); Shiokawa, Kikuchi (2018). Jedná se o ztížený návrat k původnímu vstupnímu prostoru při použití jádra, to znamená, že nelineární kPCs je obtížné interpretovat, protože neodpovídají vektorům ve vstupním prostoru.

Porovnáním mezi lineární PCA a nelineární KPCA se zabývá mimo jiné článek Ezukwoke, Zareian (2019). Tito autoři přinášejí porovnání těchto metod při využití různých jader u KPCA na různé datové soubory např. kruhy (circles), měsíce (moons) nebo kosatec (iris) aj. Avšak ne každé jádro dává uspokojivé výsledky, pokud pracuje s různými typy dat. Raschka a kol. (2014) uvádějí příklad týkající se tvarů půlměsíce. Vstupem do analýz je nelineární 2D datový soubor, ve kterém jsou viditelné dvě skupiny vyznačené červeně a modře. Nicméně tyto dvě skupiny nejsou lineárně oddělitelné. Aplikace lineární PCA na těchto datech selhala (nenabídla dobrou reprezentaci dat). Podle první PC dochází částečně k překrytí červené a modré části půlměsíců stejně jako v původním nelineárním 2D datovém souboru, to znamená, že nedošlo

k redukci dimenze. Na druhou stranu RBF v rámci KPCA je schopná při vhodném nastavení parametru σ tyto dvě skupiny půlměsíců oddělit podle první PC.

V současné době se těší velké oblibě *umělé neuronové sítě* (ANN – Artificial Neural Networks), např. v oblasti klasifikace, redukce rozměrnosti atd. Qiu a kol. (2012) se zabývají různými implementacemi a algoritmy neuronových sítí pro PCA a její rozšíření (např. KPCA). Pilario a kol. (2020) uvádí metody, které kombinují kernel metody s neuronovými sítěmi. Dále pak Wilson a kol. (2016) dospěli k závěru, že vztah mezi kernel metodami a ANN má být spíše doplňkový, nikoliv konkurenční.

Ve článku Wang a kol. (2017) je KPCA spolu se SVM použita pro identifikaci a následnou klasifikaci lidské bytosti skrze zeď pomocí UWB (ultra-širokopásmového) radaru. Nelineární extrakce proměnných zde probíhá skrze KPCA, jejíž výsledky jsou následně využity pro klasifikační algoritmus SVM. Problematikou použití SVM klasifikace pro diagnostiku abnormalit srdečního rytmu (klasifikace dvou různých abnormalit a normálního rytmu) po provedení extrakce proměnných pomocí KPCA na EKG signálech se zabývají Kallas a kol. (2012). Zde je ukázáno, že SVM v kombinaci s KPCA funguje lépe než bez této extrakce proměnných. Du a kol. (2016) uvádějí klasifikaci pacientů s ADHD (hyperkinetická porucha), kde opět jako v předchozích publikacích využívají kombinaci KPCA a SVM.

2.2 Možnosti shlukování objektů

Mezi významné statistické metody, které navazují na problematiku předchozí podkapitoly 2.1, patří metody *shlukové analýzy* (cluster analysis). Shlukovou analýzu poprvé použil v roce 1939 R. C. Tryon (viz. Stankovičová, Vojtková, 2007). Shluková analýza klasifikuje objekty do stejnorodých shluků. Hlavním cílem této analýzy je klasifikace objektů do skupin, kde jsou si objekty ve stejné skupině co nejvíce podobné a objekty mezi skupinami co nejvíce odlišné (Hebák a kol., 2007). Ve většině případů je shlukování dat spojeno se shlukováním objektů (Řezanková a kol., 2009). Nicméně může být také spojeno se shlukováním proměnných (snížení počtu otázek v dotazníku aj.) nebo skupin podobných kategorií nominální proměnné pomocí dvourozměrné tabulky četností (Hebák a kol., 2015).

V případě shlukování objektů jsou metody shlukové analýzy často rozlišovány podle toho, zda zařazují objekty do určitého počtu shluků (*nehierarchické metody shlukování*), nebo zda vytváří hierarchie shluků (*hierarchické metody shlukování*). V případě hierarchických metod shlukové analýzy je možné výsledky zobrazit pomocí *dendrogramu*, což funguje především v případě menšího počtu objektů. Na základě tohoto grafu je dále možné určit

optimální počet shluků, nejčastěji heuristickým přístupem. Na druhou stranu v případě metod nehierarchické shlukové analýzy se předpokládá apriorní znalost počtu shluků. Metody nehierarchické shlukové analýzy jsou vhodné, je-li k dispozici velký počet objektů. V tomto případě je nutné uchýlit se k metodám určujícím optimální počet shluků (Řezanková a kol., 2009). K určení optimálního počtu shluků a dále k určení do jaké míry bylo dosaženo cílů shlukové analýzy, slouží matice vnitroshlukové a mezishlukové variability a další charakteristiky (Everitt, Dunn, 2001; Löster, 2016; Stankovičová, Vojtková, 2007).

2.2.1 Wardova metoda, metoda k-průměrů a jejich možnosti

Wardova metoda (Ward‘s Method) patří mezi nejpoužívanější hierarchické metody shlukové analýzy. Shluky se tvoří maximalizací vnitroshlukové homogenity (Hair a kol., 1992; Stankovičová, Vojtková, 2007). Například ve článku Hands, Everitt (1987) je porovnána Wardova metoda s dalšími hierarchickými metodami shlukové analýzy, v rámci kterých si Wardova metoda vedla nejlépe při vytváření původní shlukovací struktury. Wardova metoda je složitější, ale přesnější metodou než jiné známé hierarchické metody shlukové analýzy (např. metoda nejbližšího souseda, metoda nejvzdálenějšího souseda, metoda průměrné vzdálenosti aj.) (viz. Eszergár-Kiss, Caesar, 2017).

Mezi nehierarchické metody shlukové analýzy patří např. metoda *k-průměrů* (k-means). Jedná se o optimalizační metodu. Metoda *k*-průměrů je založená na výchozím počátečním rozdělení objektů do *k* shluků (shlukových centroidů). Tato metoda je vhodná pro velké datové soubory, protože zde není nutné pracovat s maticí vzdáleností jako v případě hierarchických metod shlukové analýzy (Řezanková a kol., 2009; Stankovičová, Vojtková, 2007).

2.2.2 Algoritmus fuzzy k-průměrů a jeho možnosti

Obě již zmíněné metody shlukové analýzy přiřazují každý objekt právě k jednomu shluku, to znamená, že objekt, který patří do shluku A nepatří do shluku B a naopak. Tyto metody způsobují, že i odlehlé objekty, které jsou od ostatních objektů velmi vzdáleny, mohou být zařazeny právě k jednomu ze shluků, což může být následně ve výsledcích těchto analýz matoucí. Problém v zařazování odlehlých objektů do shluků částečně řeší další nehierarchická metoda shlukování, metoda *fuzzy k-průměrů* (FCM – Fuzzy C-Means). Výhodou FCM algoritmu je umožnění přiřazení jednoho objektu k více než jednomu shluku pomocí stupňů příslušnosti (Dunn, 1973; Bezdek, 1981). García-Escudero a kol. (2016) uvádějí, že ve fuzzy shlukové analýze mohou jednotlivé objekty patřit do více shluků skrze stupně příslušnosti indikující sílu asociace mezi objektem a konkrétním shlukem. U fuzzy shlukování se tedy může

stát, že objekt, který je odlehlý, získá nízké stupně příslušnosti u všech shluků, což znamená, že tento objekt nelze jednoznačně přiřadit ani do jednoho ze shluků. Dále uvádějí, že algoritmus FCM není robustní, což může být problémem, pokud data obsahují právě tyto odlehlé objekty.

Chuang a kol. (2006) se věnují zahrnutí prostorové informace do funkce příslušnosti, jelikož původní FCM algoritmus nevyužívá tuto informaci plně. Hathaway, Bezdek (2001) zase uvádějí modifikované verze algoritmu FCM pro shlukování souboru s chybějícími daty. Další modifikace FCM algoritmu řeší García-Escudero a kol. (2018), kteří kombinují fuzzy shlukovou analýzu s robustními statistickými odhady. Dále mimo jiné diskutují o volbě „fuzzifikačního" koeficientu (ovlivňujícího míru „fuzzifikace") a počtu shluků. Alptekin (2014) skrze FCM algoritmus zkoumá postavení Turecka ve srovnání se zeměmi EU pomocí ukazatelů zdravotní péče. Ve článku Hu a kol. (2019) byla pomocí fuzzy shlukové analýzy studována rizika 105 chemických přísad používaných při hydraulickém štěpení na životní prostředí a lidské zdraví. Tímto způsobem byly detekovány shluky s vysokým, středním a nízkých rizikem chemických přísad.

2.2.3 DBSCAN algoritmus a jeho možnosti

Hebák a kol. (2015) se v publikaci Statistické myšlení a nástroje analýzy dat zabývají speciálními metodami shlukové analýzy, mezi které patří algoritmy založené na hustotě, např. *algoritmus DBSCAN* (Density-Based Spatial Clustering of Applications with Noise). DBSCAN algoritmus patří mezi algoritmy strojového učení, konkrétně učení bez učitele. V tomto případě je shluk považován za množinu objektů spojených na základě hustoty, tzn. četností a vzdáleností objektů v sousedství. Tento algoritmus odstraňuje některé nedostatky tradičních algoritmů shlukové analýzy, nicméně jeho nevýhodou je citlivost na vstupní parametry, které jsou zadávané uživatelem. Jedná se o *poloměr sousedství* a *minimální počet objektů v sousedství* (viz. Hebák a kol., 2015).

Určením minimálního počtu objektů v sousedství se zabývají např. Sander a kol. (1998), kteří navrhují v případě datového souboru s více než dvěma proměnnými nastavit tento vstupní parametr na hodnotu rovnou dvakrát počet dimenzí (2**p*). Po určení minimálního počtu objektů následuje stanovení poloměru sousedství. Technika stanovení tohoto parametru je popsaná také v článku Rahmah, Sitanggang (2016). Jedná se o výpočet průměrné vzdálenosti mezi každým pozorováním a jeho *k-nejbližšími sousedy* (*k*-NN – *k*-Nearest Neighbour). Optimální hodnota poloměru sousedství je následně stanovena v bodě maximálního zakřivení grafu, který znázorňuje průměrné *k*-vzdálenosti ve vzestupném pořadí.

Kromě již zmíněných se DBSCAN algoritmu věnují také Chakraborty, Nagwani (2014) a Ester a kol. (1996). DBSCAN algoritmus je podle nich navržen pro objevování shluků různých tvarů a velikostí, to znamená, že je vhodný pro práci s rozsáhlými soubory dat. Podrobnější popis DBSCAN algoritmu je uveden např. ve článku Berkhin (2006). Aplikaci DBSCAN algoritmu pro hloubkovou analýzu v prostorových datech popisují Sharma a kol. (2016) a zavádějí vylepšený algoritmus pro shlukování. Dudik a kol. (2015) se zabývají porovnávací analýzou DBSCAN algoritmu a metody *k*-průměrů. Z jejich výsledků vyplývá, že algoritmus DBSCAN díky vyšší citlivosti lépe segmentuje než metoda *k*-průměrů. Khanteymoori, Kumar (2021) porovnávají hierarchickou Wardovu metodu, nehierarchickou metodu *k*-průměrů a DBSCAN algoritmus na datech týkajících se kruhů a měsíců. Z jejich výsledků je zřejmé, že nejlépe jsou odděleny shluky pomocí DBSCAN algoritmu. Hahsler a kol. (2019) popisují implementaci a použití balíku *dbscan* v programu R.

Konkrétním využitím DBSCAN algoritmu v oblasti zdraví se zabývá článek Pasin a Ankarali (2015). Tito autoři navrhují využít větší počet algoritmů v oblasti zdraví, které mohou přinést pozitivní vývoj poznání jak v oblasti zdraví, tak v oblasti medicíny. Shlukování založené na hustotě je důležité např. pro seskupování pacientů s podobnými příznaky do smysluplných shluků (Al-Shammari a kol., 2019).

2.3 Kombinace vícerozměrného škálování s lineárním uspořádáním

Metody shlukové analýzy uvedené v předchozí podkapitole poskytují pouze jednotlivé shluky, uvnitř kterých jsou si objekty z hlediska sledovaného jevu podobné. Kombinace vícerozměrného škálování (MDS) s lineárním uspořádáním, se kterou přichází prof. Walesiak a kol. z *Wroclaw University of Economics and Business* pod názvem *hybridní přístup* (hybrid approach), přináší kromě vizuální prezentace skupin objektů také nalezení objektů se stejnou nebo podobnou celkovou úrovní sledovaného jevu (ale s odlišnou konfigurací hodnot vstupních proměnných) pomocí vzdálenosti od ideálního vzorového objektu. Takto určenou celkovou úrovní sledovaného jevu pro jednotlivé objekty mohou být doplněny výsledky shlukové analýzy, což může být dále užitečné v případě existence odlehlých objektů, které vytvářejí samostatné shluky. Nakonec hybridní přístup přináší lineární uspořádání objektů podle agregovaného ukazatele, vytvořeného na základě výsledků MDS a vzdálenosti od ideálního vzorového objektu.

Mezi pojmy, které s se v případě hybridního přístupu používají, patří *stimulanty*, *destimulanty*, *nominanty*, *objekt vzor* (P – Pattern) a *objekt anti-vzor* (AP – Anti-Pattern). Tyto pojmy

vycházejí z publikace Hellwig (1979). Podobně jako v případě složených ukazatelů (viz. 2.1.1) lze stimulanty označit jako proměnné, jejichž vysoké hodnoty jsou žádoucí, a jejichž nízké hodnoty jsou nežádoucí. U destimulantů je tomu naopak. V případě nominantů jejich vysoké i nízké hodnoty nejsou žádoucí. Nominanty se v rámci hybridního přístupu převádějí na stimulanty, což je nutností pro stanovení objektu P a AP. Co se týká objektů P a AP, jedná se o uměle vytvořené objekty, kde objekt P je ohodnocen maximy z hodnot stimulantů a minimy z hodnot destimulantů a objekt AP pak nabývá minim z hodnot stimulantů a maxim z hodnot destimulantů (Walesiak, 2016).

V posledních letech vzniklo více publikací, týkajících se kombinace MDS s lineárním uspořádáním, např. Dehnel, Walesiak (2019); Dehnel a kol. (2019); Walesiak, Dehnel (2019); Walesiak, Dehnel (2020). Zjišťováním podobné, popř. stejné celkové úrovně stavu zdraví regionů a jejich lineárním uspořádáním se v posledních letech věnují např. články Kopecká a kol. (2020) nebo Pacáková, Kopecká (2019b).

Vizualizací lineárního uspořádání dat pomocí MDS se zabývá Walesiak (2016). V prvním kroku jsou za využití MDS vizualizovány objekty (29 okresů Dolnoslezského vojvodství) ve 2D prostoru na základě proměnných týkajících se turistické atraktivnosti. V dalším kroku jsou takto vizualizované objekty lineárně uspořádány pomocí euklidovské vzdálenosti od ideálního objektu P. V dalších článcích Dehnel a kol. (2019); Walesiak, Dehnel (2018); Walesiak, Dehnel (2019) je hybridní přístup použit k hodnocení ekonomické efektivnosti malých podniků ve vojvodství Velkopolsko a k porovnání polských provincií z hlediska sociální soudržnosti.

3 Cíle disertační práce

V souladu se současným stavem monitorování zdraví a jeho determinantů, v návaznosti na dosavadní vlastní publikační činnost je disertační práce v dalších kapitolách zaměřena na prohloubení aplikace vícerozměrných statistických metod, jejich doplnění o další pokročilé metody statistiky a informatiky, které v této oblasti zkoumání byly dosud využity málo, resp. vůbec. Motivací k tomu je poskytnout z množství sledovaných ukazatelů veřejného zdraví kvalitní a méně roztříštěné informace pro kompetentní řídicí složky. **Hlavním cílem disertační práce je porovnání a vyhodnocení výsledků lineárních a nelineárních technik pro snížení rozměrnosti ukazatelů stavu zdraví a jejich determinantů v zemích EU-27 a využití takto předzpracovaných dat k posouzení nerovností ve stavu zdraví a identifikování skupin států s podobnou, resp. rozdílnou úrovní stavu zdraví.** Aby bylo možné dosáhnout takto specifikovaného cíle, je nutné splnit následující dílčí cíle:

C1: Vhodnými metodami snížit rozměrnost ukazatelů stavu zdraví pro země EU-27.

C2: Vybranými metodami posoudit nerovnosti ve stavu zdraví v 27 evropských státech a identifikovat skupiny států s podobnou situací ve stavu zdraví.

C3: Nalézt státy s podobnou celkovou úrovní stavu zdraví, ale s odlišným uspořádáním (konfigurací) hodnot původních proměnných a následně provést jejich uspořádání pomocí hybridního přístupu.

C4: Rozšířit aplikaci zvolených metod na identifikaci hlavních determinantů stavu zdraví pro zkvalitnění politik veřejného zdraví.

C5: Vizualizovat výsledky analýz v rámci států pomocí různých možností vizualizace včetně využití geografických dat.

C6: Propojit získané výsledky stavu zdraví a jeho determinantů a následně je porovnat s již publikovanými.

4 Metodologie a použitá data

Disertační práce se zabývá především snížením rozměrnosti ukazatelů stavu zdraví a následně i jeho determinantů. Ukazatele stavu zdraví a ukazatele jeho determinantů jsou v této disertační práci odděleně studovány.

Nástroje data miningu vyžadují snížení rozměrnosti dat nejen pro jejich přehlednější vizualizaci, ale všeobecně pro snadnější a účinnou manipulaci s rozsáhlými datovými soubory. Snížení rozměrnosti dat např. na straně proměnných je užitečné mimo jiné také pro techniky shlukové analýzy. Výsledky shlukové analýzy se tak stávají lépe interpretovatelnými a mohou sloužit k posouzení a identifikaci sledovaných jevů. Výsledky, získané skrze shlukovou analýzu, jsou ovlivněny nejen zvolenou technikou shlukové analýzy, ale také volbou techniky pro snížení rozměrnosti dat, tzn. vstupní datovou maticí.

4.1 Použitá metodologie

Nejprve je pozornost věnovaná získání vhodných datových souborů obsahujících ukazatele stavu zdraví a jeho determinantů ohodnocujících země EU-27. Dále je u vybraných ukazatelů použitých v této práci snížena jejich rozměrnost pomocí metod pro snížení rozměrnosti ukazatelů. Následně jsou země EU-27 rozděleny podle stavu zdraví a jeho determinantů pomocí metod shlukové analýzy. Výsledné shluky jsou doplněné o informaci týkající se celkové úrovně stavu zdraví a jeho determinantů v zemích EU-27 pomocí hybridního přístupu. V rámci hybridního přístupu jsou modelovány agregované míry stavu zdraví a jeho determinantů na základě vzdálenosti od uměle vytvořeného ideálního objektu.

Aby bylo řešení studované problematiky jednodušší a úspěšné, je vhodné použít některou z metodologií data miningu, mezi které patří například metodologie SEMMA (viz. Terek a kol., 2010). Jedná se o metodologii softwarového produktu firmy SAS – Enterprise Miner. Tato metodologie zahrnuje následujících pět kroků, jejichž počáteční písmena tvoří název metodologie (Terek a kol., 2010; Azevedo, Santos, 2008; Woodside, 2016):

- *Sample* – tvorba vzorku,
- *Explore* – vizuální prozkoumání dat a jejich redukce,
- *Modify* – seskupování dat,
- *Model* – tvorba modelů,
- *Assess* – porovnání modelů.

Nejnáročnější součástí každé analýzy je získání dat. Existuje celá řada databází obsahující data týkající se zdravotního stavu populací, popř. jeho determinantů. Databáze Eurostatu, OECD nebo WHO obsahují agregované datové soubory pro jednotlivé evropské země za jednotlivé roky. Hlavním problémem těchto databází jsou však chybějící údaje. Čím menší jsou studované územní celky (např. NUTS 2 regiony), tím větší nastává problém s chybějícími údaji, popř. neaktuálními údaji, což se odráží na kvalitě získaných dat.

Rozsáhlost dat týkajících se stavu zdraví a jeho determinantů způsobuje jejich nepřehlednost a tím pádem míra poskytnutých informací bez dalších úprav je minimální. Proto je nutné snížit rozměrnost těchto dat při co nejmenší ztrátě informace obsáhlé v původních datech. Tímto způsobem je možné docílit přehlednější vizualizace objektů (např. států). Techniky vizuálního prozkoumání dat spočívají především ve snížení původní rozměrnosti dat (počtu proměnných) na 3D, v lepším případě 2D data. Většina článků týkající se tohoto problému uvádí pro snížení rozměrnosti dat PCA jako metodu průzkumnou, která je dále využitelná pro metody rozdělení dat a jejich modelování (viz. podkapitola 2.1.2).

V současné době jsou k dispozici metody, které z původní PCA vycházejí a umožňují odstranit některé její nedostatky. Toho je možné dosáhnout zavedením různých rotačních algoritmů, jejichž pomocí jsou původní komponentní zátěže transformovány na zátěže rotované, což pomáhá snadnější interpretaci. Další možností, jak zlepšit interpretaci v rámci PCA, je zavedení SPCA a její obměn, které přinášejí do původní PCA řídkost na úrovni komponentních zátěží a tím snadnější interpretaci. Pro více detailů viz. část 2.1.2 a část 5.1.2 a 5.1.3 v kapitole 5 Použité metody.

Tyto uvedené metody však předpokládají existenci lineárních závislostí mezi původními proměnnými, což není vždy splněno. Proto byly vyvinuty nelineární metody pro snížení rozměrnosti dat, např. KPCA, která patří mezi metody varietního učení (viz. část 2.1.4 a podkapitola 5.1.4 Kernel analýza hlavních komponent).

K rozdělení objektů podle stavu zdraví evropské populace a jeho determinantů je možné použít metody shlukové analýzy, ale již při snižování rozměrnosti úlohy na straně ukazatelů jsou některé skupiny objektů zjevné. Pomocí metod shlukové analýzy aplikované na nově zavedené PCs, RCs, SPCs a kPCs je možné tyto skupiny přesně identifikovat a na základě nových latentních vstupních proměnných posoudit nerovnosti stavu zdraví ve státech Evropy. Důležitou součástí shlukové analýzy je také detekování odlehlých objektů.

Kromě např. Wardovy metody nebo metody *k*-průměrů, které každý objekt přiřadí do jednoho ze shluků, existuje fuzzy shlukování přiřazující objektům stupně příslušnosti ke každému shluku (viz. podkapitoly 2.2.1 a 2.2.2 a podkapitoly 5.2.1 a 5.2.2 v kapitole 5 Použité metody). Mezi modernější metody shlukové analýzy patří tzv. DBSCAN algoritmus, který na rozdíl od předchozích metod vytváří shluky definované jako množiny objektů spojené na základě hustoty (viz. podkapitola 2.2.3 a podkapitola 5.2.1 a 5.2.3 v kapitole 5 Použité metody).

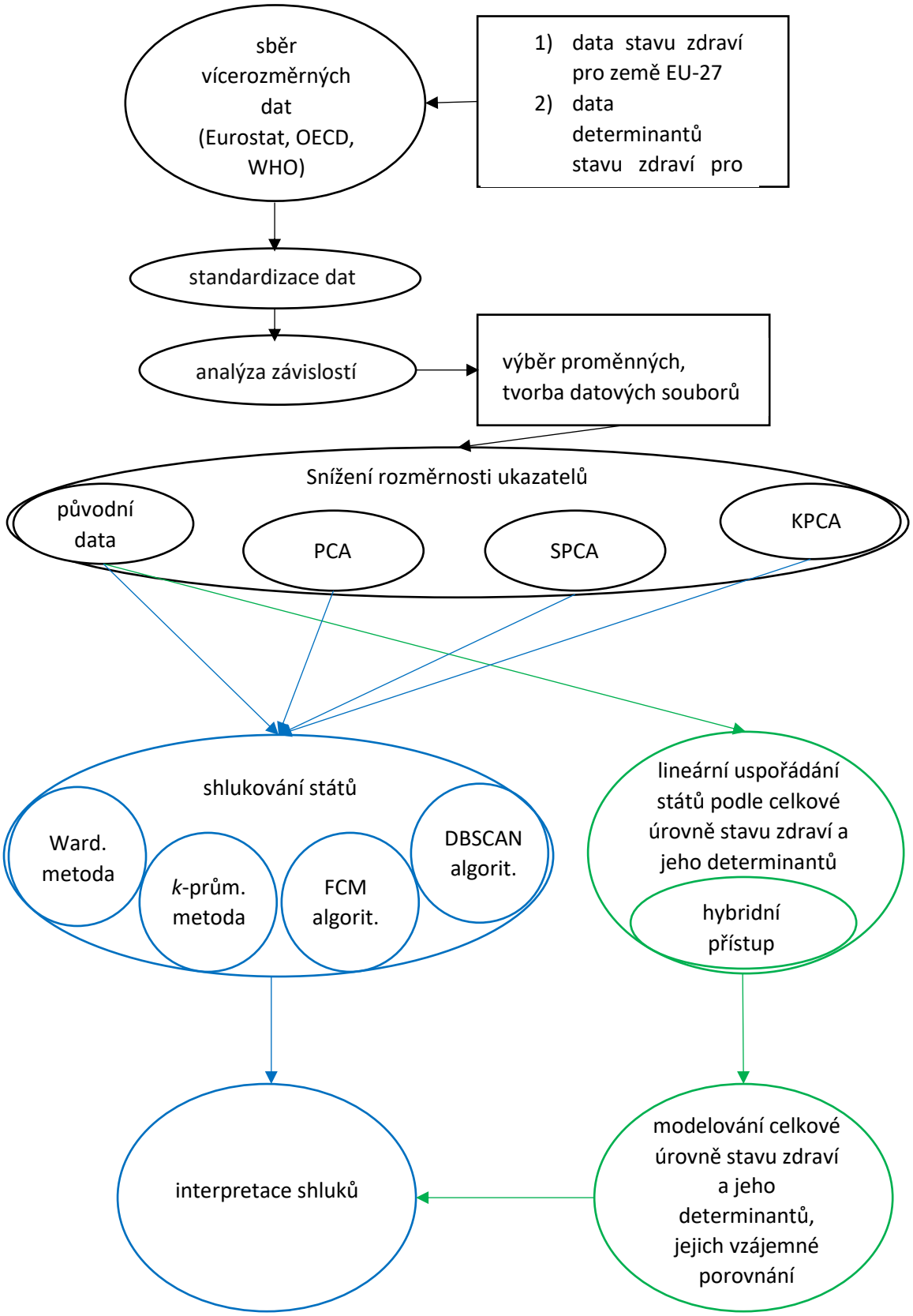
Existují závažná onemocnění, která některé evropské státy postihují více než jiné. Metody shlukové analýzy detekují především shluky, ve kterých jsou zařazeny státy s podobnou konfigurací hodnot ukazatelů týkající se stavu zdraví. Walesiak (2016) popisuje tzv. hybridní přístup kombinující vícerozměrné škálování (MDS) s lineárním uspořádáním. Tento přístup mimo jiné umožňuje zjistit, jestli jsou objekty na podobné celkové úrovni stavu zdraví vzhledem ke vzorovému ideálnímu objektu, i když mají odlišné uspořádání hodnot těchto ukazatelů (tzn. některé populace trpí méně na kardiovaskulární onemocnění, ale více na onemocnění dýchacích cest než jiné), což je výhodou oproti algoritmům shlukové analýzy. Pomocí tohoto přístupu je vytvořen jeden agregovaný ukazatel na základě vzdálenosti od ideálního objektu popisující celkovou úroveň sledovaného jevu (viz. podkapitola 2.3 a podkapitola 5.3 Hybridní přístup). Hybridní přístup nepatří mezi algoritmy shlukové analýzy, nicméně jeho pomocí lze snížit rozměrnost ukazatelů, vizuálně detekovat skupiny objektů a vytvořit pořadí těchto objektů podle celkové úrovně stavu zdraví.

Obrázek 4 představuje schéma zahrnující použitá data (viz. podkapitola 4.2 Použitá data), aplikované metody (viz. kapitola 5 Použité metody) a vyhodnocení výsledků. V uvedeném schématu jsou použity tyto dvě datové matice:

- hodnot ukazatelů stavu zdraví pro země EU-27,
- hodnot determinantů stavu zdraví pro země EU-27.

Součástí metod pro snižování rozměrnosti ukazatelů a metod shlukové analýzy je i nastavování jejich hyperparametrů, pokud to použité metody vyžadují.

Obrázek 4: Schéma použitých dat, metod a vyhodnocení výsledků



Zdroj: vlastní zpracování na základě dostupných dat a použité literatury

4.2 Použitá data

V této podkapitole jsou popsány proměnné obsažené v použitých datových maticích v kapitole 6 Aplikace metod pro snížení rozměrnosti ukazatelů stavu zdraví a jeho determinantů. Tato data slouží pro splnění hlavního cíle disertační práce. Jedná se o proměnné týkající se stavu zdraví, které charakterizují jednotlivé země EU-27, a které vypovídají o stavu zdraví této populace (viz. podkapitola 1.3). Dále jsou použita agregovaná data popisující determinanty stavu zdraví ovlivňující evropskou populaci na úrovni států (viz. podkapitola 1.4). Vzhledem k nízké kvalitě determinantů stavu zdraví pro NUTS 2 regiony zemí EU-27 (chybějící ukazatele, chybějící hodnoty ukazatelů, neaktuální data aj.) (viz. Costa a kol., 2019), nejsou tyto ukazatele na regionální úrovni použity. Determinanty stavu zdraví použité v této práci mohou být do jisté míry ovlivněny politickými zásahy (viz. Dahlgren, Whitehead, 1991). Z tohoto důvodu nevstupují mezi determinanty stavu zdraví pohlaví a genetické dispozice, které těmito zásahy samozřejmě nelze ovlivnit.

Data o zdravotním stavu populace a jeho determinantech v zemích Evropy jsou v agregované formě k dispozici především v databázích Eurostatu, OECD a WHO. Tato data jsou následně v přehledných grafech či tabulkách uváděna v publikacích, které vycházejí v rámci těchto databází (viz. kapitola 1).

Podle Hebák a kol. (2015) jakákoli provedená analýza nemůže poskytnout hodnotné výsledky, pokud jsou použita nevhodná data. Aby získaná data odpovídala skutečnosti, musí být dodržena jejich (Hebák a kol., 2015; Molnár a kol., 2012):

- *objektivita* – spočívá v nezávislosti dat na osobě pořizovatele, uživatele a zkoumané jednotce (především u osob),
- *validita* – je nutné zjistit, zda bylo skutečně měřeno, co se mělo měřit např. porovnáním s informacemi získanými jiným způsobem,
- *reliabilita* – spolehlivost metody udává, do jaké míry je schopna dosáhnout za stejných podmínek stejného výsledku.

V případě ukazatelů týkajících se zdravotního stavu populace a jeho determinantů, které poskytují databáze Eurostatu, OECD nebo WHO, se mohou hodnoty těchto ukazatelů, popř. celé datové matice lišit především kvůli chybějícím pozorováním (objektům), ale i v odlišnostech uváděných dat na straně proměnných. Ne všechna data v různých databázích je možné pro konkrétní roky získat. Dalším problémem mohou být u jednotlivých ukazatelů (např. délky života prožité ve zdraví) jejich různé definice v rámci jmenovaných

databází. Pro příklad databáze Eurostatu poskytuje HLY, ale databáze WHO uvádí obdobný ukazatel HALE (viz. kapitola 1). Je tedy nutné dodržet při doplňování chybějících údajů ke konkrétním objektům z jiné databáze, aby získané údaje odpovídaly zjišťované skutečnosti.

4.2.1 Použité ukazatele stavu zdraví

Definice použitých ukazatelů stavu zdraví jsou převzaty z databází a publikací Dyba a kol. (2021), Eurostat (2022a), OECD/European Union (2020), Timmis a kol. (2022) a WHO (2020b), WHO (2022c), WHO (2022d). Na základě PHI (Population Health Index) jsou zde popsány následující skupiny ukazatelů stavu zdraví: délka života, úmrtnost, kvalita života a nemocnost (viz. podkapitola 1.2).

Databáze Eurostatu poskytuje ukazatele úmrtnosti vyjádřené pomocí počtů úmrtí v jednotlivých letech (absolutní ukazatele) nebo hrubými a standardizovanými mírami úmrtností (relativní ukazatele). *Hrubá míra úmrtnosti* (CDR – Crude Death Rate) popisuje úmrtnost ve vztahu k celkové populaci. Vyjádřena v počtu úmrtí na 100 000 obyvatel je počítána jako počet úmrtí zaznamenaných v populaci za dané období vzhledem k počtu obyvatel ve stejném období násobené 100 000. CDR pro všechny věkové kategorie je váženým průměrem měr úmrtnosti podle věku. Váhovým faktorem je věkové rozdělení populace, jejíž úmrtnost je pozorována. Věková struktura populace tedy silně ovlivňuje tento ukazatel pro široké věkové kategorie. V relativně *„staré“* populaci bude více úmrtí než v populaci *„mladé“*. CDR je počítána pro 5leté věkové kategorie, ale i pro osoby mladší nebo starší 65 let. Aby bylo srovnání států nebo regionů z hlediska CDR (stavu zdraví) smysluplné, je třeba použít užší věkové kategorie. Vliv věku je možné zahrnout pomocí standardní populace. Jedná se o standardizovanou míru úmrtnosti (SDR), která je váženým průměrem věkově specifické míry úmrtnosti. Váhovým faktorem je věkové rozdělení *standardní referenční populace* (ESP). Vzhledem k tomu, že se většina příčin úmrtí výrazně liší v závislosti na věku, je pro srovnání zemí EU-27 vhodnější použít SDR, která je použita i v této práci pro nejzávažnější onemocnění postihující evropskou populaci (viz. podkapitola 1.3.3). Pro detaily viz. Eurostat (2022a).

Stav zdraví je pro země EU-27 analyzován pomocí následujících ukazatelů popsaných v podkapitolách 1.3.1–1.3.7:

- *HLY_0 zdravé roky života při narození, HLY_65 zdravé roky života ve věku 65 let* – viz. podkapitola 1.3.2 (Eurostat, 2022a),
- *LE_0 střední délka života při narození, LE_65 střední délka života ve věku 65 let* – viz. podkapitola 1.3.1 (Eurostat, 2022a),

- *HALE_0 očekávaná délka života ve zdraví při narození*, *HALE_60 očekávaná délka života ve zdraví ve věku 60 let* – viz. podkapitola 1.3.2 (WHO, 2022d),
- *SDR – zhoubné novotvary* podle MKN-10 (C00 – C97) pro všechny věkové kategorie a věkovou kategorii 65+ (Eurostat, 2022a),
- *SDR – diabetes mellitus* pro všechny věkové kategorie a věkovou kategorii 65+ (Eurostat, 2022a),
- *SDR – duševní poruchy a poruchy chování* podle MKN-10 (F00 – F99) pro všechny věkové kategorie a věkovou kategorii 65+ (Eurostat, 2022a),
- *SDR – nemoci nervového systému a smyslových orgánů* podle MKN-10 (G00 – H95) pro všechny věkové kategorie a věkovou kategorii 65+ (Eurostat, 2022a),
- *SDR – nemoci oběhového systému* podle MKN-10 (I00 – I99) pro všechny věkové kategorie a věkovou kategorii 65+ (Eurostat, 2022a),
- *SDR – nemoci dýchacího systému* podle MKN-10 (J00 – J99) pro všechny věkové kategorie a věkovou kategorii 65+ (Eurostat, 2022a),
- *zabranitelná úmrtnost* – viz. podkapitola 1.3.3 (Eurostat, 2022a),
- *léčitelná úmrtnost* – viz. podkapitola 1.3.3 (Eurostat, 2022a),
- *kojenecká úmrtnost* – viz. podkapitola 1.3.4 (Eurostat, 2022a),
- *DALY pro všechny příčiny nemoci a všechny věkové kategorie* – Ukazatel je ovlivněn velikostí populace v jednotlivých zemích, a proto je pro následující analýzy přepočítán na osobu (viz. podkapitola 1.3.5) (WHO, 2022c),
- *prevalence kardiovaskulárních onemocnění* – věkově standardizovaná prevalence kardiovaskulárních onemocnění na 100 000 obyvatel (Timmis a kol., 2022),
- *incidence rakovin* – incidence rakovin na 100 000 obyvatel (Dyba a kol., 2021; WHO, 2022b),
- *podíl dospělých s diabetem* – věkově standardizovaná prevalence diabetu prvního nebo druhého typu u dospělých ve věku 20–79 let (v %) (viz. Timmis a kol., 2022),
- *prevalence symptomů psychické tísně* – Prevalence symptomů psychické tísně je založena na průzkumu EU-SILC, vypočítává se z odpovědí na pět otázek např.: *„Byl jste během posledních čtyř týdnů velmi nervózní?“* Odpovědi se vybírají z pěti bodové stupnice (0–4) od *„v žádném okamžiku“* až po *„vždy“*. Skóre může dosáhnout maximální hodnoty 20, která se následně vynásobí pěti, aby bylo dosaženo maximálního skóre 100. Osoby se symptomy psychické tísně dosahují skóre větší než 50. Položky se týkají pocitu nervozity, sklíčení, pocitu klidu, pocitu skleslosti nebo deprese a pocitu štěstí. Prevalence je vážena velikostí populace. Ukazatel prevalence symptomů

psychické tísně je uveden v % z populace ve věku 16+. Pro detaily viz. OECD/European Union (2020).

- *osoby hlásící astma* – podíl osob trpících astmatem v % (viz. Eurostat, 2022a),
- *osoby hlásící chronické onemocnění dolních cest dýchacích* – Podíl osob hlásících chronické onemocnění dolních cest dýchacích (vyjma astma) v %. Hlavní příčinou těchto onemocnění je především kouření (viz. Eurostat, 2022a).
- *osoby vnímající svůj zdravotní stav jako „dobrý" nebo „velmi dobrý"* – Představují podíl populace ve věku 16+, která vnímá svůj zdravotní stav jako *„dobrý"* nebo *„velmi dobrý"* (% z populace ve věku 16+). Údaje pocházejí ze statistik EU-SILC (viz. Eurostat, 2022a).

4.2.2 Použité determinanty stavu zdraví

Definice použitých determinantů stavu zdraví jsou převzaty z databází a publikací Eurostatu, 2022a; OECD/European Union (2020). Mezi použité determinanty stavu zdraví jsou zařazeny ty, které je možné alespoň z části ovlivnit prostřednictvím politických zásahů, popř. zásahů orgánů veřejné správy (viz. podkapitola 1.4). Determinanty stavu zdraví jsou pro země EU-27 analyzovány pomocí následujících ukazatelů:

Emise jemných částic ($PM_{2,5}$) na osobu

Jemné částice ($PM_{2,5}$) o průměru menším než 2,5 mikronu, rozptýlené ve vzduchu, způsobují závažnější zdravotní problémy než emise primárních pevných částic vznikajících např. spalováním paliv nebo vytápěním domácností. Zdravotními problémy, které emise jemných částic způsobují, se rozumí především kardiovaskulární, respirační onemocnění, rakovina plic aj. (OECD/European Union, 2020).

Míra dospělých kouřících denně v %

Procento osob ve věku 15+, kteří hlásí, že kouří každý den (OECD/European Union, 2020).

Nadměrná konzumace alkoholu dospělými v litrech na osobu

Jedná se o roční prodej čistého alkoholu v litrech na osobu ve věku 15+ (OECD/European Union, 2020).

Hlášení míry obezity mezi dospělými v %

Obezita je definována jako nadměrná hmotnost představující zdravotní rizika z důvodu vysokého podílu tělesného tuku. Obezita je měřena pomocí *indexu tělesné hmotnosti* (BMI –

Body Mass Index). Podle WHO jsou osoby s BMI vyšším nebo rovným 30 definováni jako obézní. Míru obezity lze hodnotit pomocí odhadů BMI, které osoby hlásí ve zdravotních průzkumech nebo jsou naměřeny ze zdravotních vyšetření (OECD/European Union, 2020).

Výdaje do zdraví (health expenditure) na osobu v Eurech při zohlednění parity kupní síly (PPP – Purchasing Power Parity)

Výdaje do zdraví jsou měřeny konečnou spotřebou zdravotnického zboží a služeb. Jedná se o běžné výdaje na zdravotnické zboží a služby, veřejné zdraví, preventivní programy a celkovou správu poskytování a financování zdravotní péče. Existují dva způsoby financování zdravotní péče, mezi které patří vládní/povinné a dobrovolné/vlastní. V průměru představují výdaje vládní nebo plynoucí z povinného veřejného či soukromého zdravotního pojištění tři čtvrtiny celkových výdajů do zdraví napříč zeměmi EU (OECD/European Union, 2020).

Výdaje do zdraví jako podíl na HDP (vládní/povinné) – viz. definice výdaje do zdraví (OECD/European Union, 2020).

Výdaje do zdraví jako podíl na HDP (dobrovolné/vlastní) – viz. definice výdaje do zdraví (OECD/European Union, 2020).

Výdaje „z vlastní kapsy" (out-of-pocket spending) do zdraví vyjádřené jako podíl na konečné spotřebě domácností

Výdaje *„z vlastní kapsy"* jsou výdaje hrazené přímo pacientem v případě, že pojištění nepokrývá plnou cenu zdravotnického zboží nebo služeb. Jedná se o sdílení nákladů a dalších výdajů hrazených přímo domácnostmi (OECD/European Union, 2020).

Výdaje do zdraví na dlouhodobou péči na osobu v Eurech

Jedná se o výdaje do zdraví na dlouhodobou zdravotní péči. Dlouhodobá zdravotní péče zahrnuje řadu služeb lékařské a osobní péče, které jsou spotřebovávány s primárním cílem zmírnit bolest a utrpení, a tak zvládnout zhoršování zdravotního stavu u pacientů s určitým stupněm dlouhodobé závislosti (Eurostat, 2022a).

Výdaje do zdraví na dlouhodobou péči % z HDP – viz. výdaje do zdraví na dlouhodobou péči (Eurostat, 2022a).

Lékaři na 1 000 obyvatel

Do údajů týkajících se počtu lékařů na 1 000 obyvatel spadají lékaři, kteří poskytují zdravotní péči. Dále sem mohou být zahrnuti i lékaři, kteří vykonávají činnost administrativní, manažerskou a činnost akademických a výzkumných pracovníků (OECD/European Union, 2022).

Zdravotní sestry na 1 000 obyvatel

Do údajů týkajících se počtu sester na 1 000 obyvatel jsou zahrnuty sestry pečující o pacienty. Dále sem mohou být zahrnuty i sestry, které vykonávají činnost manažerskou, vychovatelskou nebo výzkumnou. Pod tento ukazatel spadají jak *„profesionální sestry“*, tak *„pomocné sestry“*, které mají nižší úroveň vzdělání než sestry profesionální, ale jsou registrované jako zdravotní sestry (OECD/European Union, 2020).

Nemocniční lůžka na 1 000 obyvatel

Lůžka v nemocnicích na 1 000 obyvatel představují všechna lůžka, která jsou pravidelně udržovaná a obsazená a jsou okamžitě k dispozici. Jedná se o lůžka ve všeobecných nemocnicích, v nemocnicích pro duševní zdraví, nemocnicích pro uživatele návykových látek a v dalších specializovaných nemocnicích. Do tohoto ukazatele nejsou zahrnuta lůžka v pečovatelských a pobytových zařízeních. Údaje týkající se lůžek nepokrývají v některých zemích všechny uvedené nemocnice (OECD/European Union, 2020).

Průměrná délka pobytu v nemocnici (dny)

Průměrná délka pobytu v nemocnici představuje průměrný počet dní, které pacienti v nemocnici stráví. Tento ukazatel je pro jednotlivé státy měřen tak, že se celkový počet dní, které všichni hospitalizovaní pacienti v nemocnici během roku stráví, vydělí počtem přijetí nebo propuštění. Jednodenní hospitalizace nejsou do tohoto ukazatele zahrnuty (OECD/European Union, 2020).

Hlášení neuspokojených potřeb lékařské prohlídky osob 16+ (příliš drahé, příliš daleko, dlouhé čekací doby) v %

Vlastní hlášení osob o tom, zda potřebovaly vyšetření nebo léčbu kvůli zdravotnímu problému, ale nedostaly ji nebo ji nevyhledávaly. Otázky týkající se neuspokojených potřeb lékařského vyšetření jsou zahrnuty v průzkumu EU-SILC (Eurostat, 2022a).

Hlášení neuspokojených potřeb lékařské prohlídky osob 65+ (příliš drahé, příliš daleko, dlouhé čekací doby) v % - viz. definice hlášení neuspokojených potřeb lékařské prohlídky (Eurostat, 2022a).

Hlášení používání služeb domácí péče při těžkých úrovních obtíží v %

Podíl osob, kteří v posledních dvanácti měsících využily služeb domácí péče při těžkých úrovních obtíží (Eurostat, 2022a).

Celková míra nezaměstnanosti % z pracovní síly

Míra prezentuje počet nezaměstnaných jako procento z pracovní síly (Eurostat, 2022a).

Medián ekvivalizovaného disponibilního příjmu (PPS)

Celkový disponibilní příjem domácnosti se vypočítá z osobního příjmu všech členů domácnosti plus příjmu obdrženého na úrovni domácnosti. Disponibilní příjem se skládá z položek, mezi které patří: mzda, plat, výdělek ze samostatně výdělečné činnosti, soukromé příjmy z investic a majetku, transfery mezi domácnostmi a všechny sociální transfery. Tento příjem je *„ekvivalizován“* pomocí upravené stupnice OECD, aby se zohlednily velikost a složení domácnosti. Údaje jsou uvedeny v PPS (Purchasing Power Standards), které zohledňují rozdíly v cenové hladině mezi zeměmi (Eurostat, 2022a).

Giniho koeficient ekvivalizovaného disponibilního příjmu %

Giniho koeficient slouží k měření příjmové nerovnosti v jednotlivých zemích. Nízká hodnota tohoto koeficientu odráží rovnoměrnější rozdělení příjmů (Eurostat, 2022a).

Osoby ohrožené chudobou nebo sociálním vyloučením %

Jedná se o hlavní ukazatel sledování chudoby a sociálního vyloučení v zemích EU. Tento ukazatel odpovídá součtu osob, které jsou ohroženy chudobou nebo jsou těžce materiálně a sociálně deprivovaní (mají neuspokojené materiální a sociální potřeby) nebo žijí v domácnosti s nízkou pracovní intenzitou. Osoby jsou do tohoto ukazatele zahrnuty pouze jednou, i když se nacházejí ve více než jedné vyjmenované situaci. Tento ukazatel se následně vyjádří jako podíl na celkové populaci, která je ohrožena chudobou nebo sociálním vyloučením (to znamená, že nyní jsou osoby zahrnuty do všech vyjmenovaných situací, ve kterých se skutečně nacházejí) (Eurostat, 2022a).

Míra hmotné a sociální deprivace %

Jedná se o míru, která ukazuje na vynucený nedostatek nezbytných položek k vedení přiměřeného života. Konkrétně je definovaná jako podíl osob, které trpí vynuceným nedostatkem alespoň sedmi ze třinácti deprivačních položek, kde šest se týká jednotlivců a sedm domácností. Seznam položek zahrnuje následující: schopnost čelit neočekávaným výdajům, schopnost zaplatit si týdenní dovolenou mimo domov jednou za rok, schopnost splácet dluhy a závazky, schopnost dovolit si každý druhý den jídlo s masem, schopnost adekvátně pečovat o domácnost, mít přístup k osobnímu automobilu, schopnost vyměnit opotřebený nábytek, mít přístup k internetu, možnost vyměnit obnošené oblečení za nové, mít dva páry dobře padnoucích bot, schopnost utrácet každý týden za sebe malou částku peněz, možnost mít pravidelné volnočasové aktivity a možnost scházet se s přáteli v restauraci alespoň jednou za měsíc. Tato míra je ukazatelem EU-SILC (Eurostat, 2022a).

Průměrná velikost domácnosti

Jedná se o průměrný počet osob v domácnosti.

Hrubá míra uzavření sňatků na 1000 obyvatel

Jedná se o podíl počtu sňatků během roku k průměrnému počtu obyvatel v daném roce na 1000 obyvatel (Eurostat, 2022a),

Hrubá míra rozvodovosti na 1000 obyvatel

Jedná se o podíl počtu rozvodů během roku k průměrnému počtu obyvatel v daném roce na 1000 obyvatel (Eurostat, 2022a),

Terciální vzdělání osob ve věku 15–64 let v %

Zahrnuje krátkodobé terciální vzdělání, bakalářský nebo ekvivalentní stupeň, magisterský nebo ekvivalentní stupeň, doktorský nebo ekvivalentní stupeň. Ukazatel je vyjádřen v procentech (Eurostat, 2022a).

Méně než základní, základní a nižší sekundární vzdělání osob ve věku 15–64 let %

Zahrnuje vyjmenované úrovně vzdělání. Ukazatel je vyjádřen v procentech (Eurostat, 2022a).

5 Použité metody

Data mining je disciplínou ležící na rozhraní statistiky, databázových technologií, rozpoznávání vzorů, strojového učení a dalších oblastí (Hand, 1998). Tato disciplína se zabývá sekundární analýzou velkých datových souborů, která má za cíl nalézt skryté vztahy zajímavé pro vlastníky databází (např. veřejnou správu). Statistika jako taková je důležitou součástí data miningu, jelikož se zabývá právě primární analýzou datových souborů, při které dochází ke shromažďování dat za účelem odpovědi na konkrétní otázky. (Hand, 1998)

Nástroje data miningu vyžadují snížení rozměrnosti dat pro jejich přehlednější vizualizaci, snadnější a účinnou manipulaci s rozsáhlými datovými soubory atd. Jedná se o jednu z prvních fází metodologie data minigu SEMMA. Díky digitalizaci vzniká obrovské množství dat (např. rozsahem a rychlostí generování) ve zdravotnictví, organizacích veřejné správy aj. Algoritmy data miningu, např. strojového učení, je možné použít k předpovědím užitečným pro přijímání rozhodování na manažerské úrovni. Pro proces trénování modelů však většinou nejsou vhodné všechny atributy (ukazatele, proměnné) v těchto datových sadách. Důležitost snížení rozměrnosti získaných ukazatelů spočívá ve snížení zátěže pro algoritmy strojového učení. (Reddy a kol., 2020)

Další fáze metodologie SEMMA jsou v této disertační práci zastoupeny metodami shlukové analýzy a hybridním přístupem kombinujícím vícerozměrné škálování (MDS) s lineárním uspořádáním. U těchto analýz je možné pozorovat odlišnosti v jejich výsledcích při použití různých vstupních datových souborů získaných skrze metody pro snížení rozměrnosti dat.

5.1 Metody pro snížení rozměrnosti ukazatelů

Snížení rozměrnosti dat vzhledem k proměnným je důležitou vícerozměrnou technikou jak z hlediska prosté vizualizace, tak z důvodu efektivnějšího využití dalších metod. Mezi zde již uvedené metody snížení rozměrnosti ukazatelů stavu zdraví patří lineární metody PCA, SPCA, ale také nelineární varianty jako je KPCA. Pro získání nového datového souboru nižší dimenze vzhledem k proměnným při co nejmenší ztrátě informace v původním datovém souboru jsou v případě PCA a SPCA důležitým předpokladem silné lineární závislosti mezi původními proměnnými.

Datové soubory získané pomocí metod pro snížení rozměrnosti ukazatelů stavu zdraví mohou dále sloužit jako vstup pro metody shlukové analýzy. Za prvé je možné velký počet proměnných pomocí těchto metod snížit a dále je možné vysvětlit vzájemné vztahy mezi těmito proměnnými. Například jednou z výhod PCA je i to, že nově vytvořené PCs jsou nekorelované,

což není možné často u původních proměnných zajistit bez úplného vyřazení některé z původních proměnných (Hebák a kol., 2015; Stankovičová, Vojtková, 2007).

5.1.1 Závislosti mezi proměnnými

Vzájemný vztah neboli souvztažnost dvou proměnných lze měřit pomocí *jednoduchých (párových) korelačních koeficientů*, mezi které patří např. *Pearsonův a Spearmanův korelační koeficient*. Korelace však není důkazem příčin různých událostí (Hebák a kol., 2015).

Pearsonův korelační koeficient $r_{X,Y}$ dvou náhodných veličin X a Y je definován jako vztah:

$$r_{X,Y} = \frac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}, \tag{1}$$

kde $C(X,Y)$ představuje kovariance náhodných veličin X a Y a jmenovatel je vyjádřen jako součin směrodatných odchylek X a Y.

Obecně platí, že tento korelační koeficient nabývá hodnoty z intervalu $\langle -1,1 \rangle$. Čím blíže je korelační koeficient k hraničním hodnotám (+1 značí úplnou přímou lineární závislost, pozitivní korelace; -1 značí úplnou nepřímou lineární závislost, negativní korelace), tím je korelace mezi X a Y silnější. Pokud je však korelace rovná 0, jedná se o vzájemně lineárně nezávislé veličiny X a Y (Hebák a kol., 2015).

Pearsonův korelační koeficient měří pouze míru lineární závislosti mezi veličinami a měl by být aplikován na datové soubory větších rozsahů neobsahující odlehlá pozorování. V případě, že jsou pochybnosti o splnění těchto předpokladů nebo mezi zkoumanými veličinami mohou existovat nelineární vazby, je vhodné pro určení míry závislosti místo Pearsonova korelačního koeficientu využít tzv. Spearmanův korelační koeficient. Tento neparametrický korelační koeficient je možné použít také v případě ordinálních náhodných veličin. Spearmanův korelační koeficient je založen na stanovení pořadí hodnot náhodných veličin. Pokud se nevyskytují průměrná pořadí hodnot náhodných veličin lze Spearmanův korelační koeficient vyjádřit vztahem (Hebák a kol., 2015):

$$r_s = 1 - \frac{6\sum_r^n d_r^2}{n(n^2-1)}, \tag{2}$$

kde d_r jsou diference pořadí jednotek vzhledem k veličinám X a Y. Při výskytu průměrných pořadí je nutné vztah (2) modifikovat na vztah (3), kde $C = \frac{1}{2}\left[\sum_k \left(h_{x,k}^3 - h_{x,k}\right) + \sum_{k'} \left(h_{y,k'}^3 - h_{y,k'}\right)\right]$ a $h_{x,k}$ (resp. $h_{y,k'}$) představuje četnosti k-té (resp. k'-té) skupiny stejných

hodnot proměnné *X* a (resp. *Y*). Tato modifikace je běžnou součástí drtivé většiny statistických softwarů včetně R (Pacáková a kol., 2015).

$$r_s = 1 - \frac{6\sum_r^n d_r^2}{n(n^2-1) - C} \tag{3}$$

Tento korelační koeficient opět nabývá hodnot z intervalu $\langle -1,1 \rangle$, nicméně je vhodný i v případě, že míry sledovaného znaku jsou pořadová čísla (ordinální náhodná veličina). Krajních hodnot intervalu dosahuje tento koeficient v případě dokonalé shody pořadí ve čtvercové kontingenční tabulce (Hebák a kol., 2015). Obecně lze říci, že Spearmanův korelační koeficient odráží, jak dobře vztah sledovaných veličin odpovídá monotónní funkci, která kromě lineární může být také nelineární (Holčík a kol, 2015).

V publikaci de Vaus (2002) je uvedena klasifikace hodnot korelačních koeficientů pro společenské vědy. Například hodnota korelačního koeficientu v intervalu $\langle 0{,}5;\, 0{,}69 \rangle$ je označena jako podstatná, v intervalu $\langle 0{,}7;\, 0{,}89 \rangle$ jako velmi silná a v intervalu $\langle 0{,}9;\, 0{,}99 \rangle$ jako téměř perfektní.

Testováním významnosti korelačních koeficientů se také zabývá Holčík a kol. (2015). Test významnosti korelačního koeficientu je důležité provést v případě, pokud realizujeme náhodný výběr a dosažené výsledky na základě tohoto výběru jsou zobecňovány na základní soubor.

Podle Hebák a kol. (2007) a Hebák a kol. (2015) je možné měřit vzájemné závislosti skupiny proměnných pomocí tzv. *Kaiser–Meyer–Olkin (KMO)* indexu. Tento index porovnává jednoduché Pearsonovy korelační koeficienty s parciálními korelačními koeficienty, což je vyjádřeno vztahem (Stankovičová, Vojtková, 2007):

$$KMO = \frac{\sum_{i\neq j}^p \sum_{i\neq j}^p r_{ij}^2}{\sum_{i\neq j}^p \sum_{i\neq j}^p r_{ij}^2 + \sum_{i\neq j}^p \sum_{i\neq j}^p r_{parc.,ij}^2}, \tag{4}$$

kde r_{ij} představuje jednoduchý korelační koeficient a $r_{parc.,ij}$ *parciální (dílčí) korelační koeficient*.

Hebák a kol. (2015) uvádí zjednodušenou interpretaci $r_{parc.,ij}$ dílčích korelačních koeficientů jako jednoduchý korelační koeficient mezi veličinami Y a X_1 při vyloučení vlivu všech ostatních veličin $X_2, \ldots, X_p$.

Čím větší jsou rozdíly mezi odpovídajícími jednoduchými a dílčími korelačními koeficienty, tím silnější závislosti existují ve skupině proměnných (Stankovičová, Vojtková, 2007). KMO index nabývá hodnot z intervalu $\langle 0;1 \rangle$. Pokud se hodnoty tohoto indexu blíží k hodnotě 1, jedná

se o silnou vzájemnou závislost mezi proměnnými. Naopak nízké hodnoty tohoto indexu ukazují na slabou vzájemnou závislost. Analogickou mírou KMO, kterou je možné použít pro každou proměnnou samostatně, je tzv. *dílčí míra adekvátnosti pro jednotlivé proměnné* (MSA – Kaiser's Measure of Sampling Adequacy), která opět nabývá hodnot z intervalu $\langle 0;1 \rangle$ a může pomoci při výběru vhodných proměnných pro PCA (Hebák a kol., 2015; Stankovičová, Vojtková, 2007).

Hodnota KMO indexu je využívána pro určení vhodnosti dat pro PCA, popř. FA. Pro hodnoty KMO indexu je zde uvedena tabulka 1, která vypovídá o adekvátnosti dat podle hodnot KMO indexu a kterou pro účely interpretace výsledků vyvinul H. F. Kaiser (Stankovičová, Vojtková, 2007). Tento index je možné získat prostřednictví softwarů SPSS, Statistica, R, SAS aj. Pro podrobnosti viz. Hebák a kol. (2007); Hebák a kol. (2015) a Stankovičová, Vojtková (2007).

Tabulka 1: Hodnoty KMO míry adekvátnosti dat

Hodnoty KMO indexu	Adekvátnost dat
$\langle 0; 0{,}5)$	nedostatečná
$\langle 0{,}5; 0{,}6)$	slabá
$\langle 0{,}6; 0{,}7)$	průměrná
$\langle 0{,}7; 0{,}8)$	středně užitečná
$\langle 0{,}8; 0{,}9)$	chvályhodná
$\langle 0{,}9; 1\rangle$	vynikající

Zdroj: Stankovičová, Vojtková (2007)

5.1.2 Metoda hlavních komponent a rotované komponenty

PCA patří mezi nejstarší a nejčastěji využívané techniky redukce dimenze, jejíž teoretickými aspekty se zaobírá mnoho autorů, např. Hebák a kol. (2015); Stankovičová, Vojtková (2007). Podle nich PCA vytváří nové PCs, které jsou lineárními kombinacemi původních dat při co nejmenší ztrátě informace (rozptylu). PCA byla zavedena na začátku 20. století K. Pearsonem a dalším rozvojem této metody se zabýval např. H. Hotelling (Bro, Smilde, 2014).

Stankovičová, Vojtková (2007) se věnují nejen výběru počtu PCs, které vhodně reprezentují původní proměnné, ale také vlastním číslům, vlastním vektorům, koeficientům korelací proměnných s komponentami (komponentními zátěžemi), komponentním skóre atd.

Podle Dash, Dubey (2012) může být PCA chápána jako analýza, která, transformuje původní proměnné na nové latentní proměnné, které jsou spolu nekorelované a redukuje data tak, že převede n x p datovou matici na matici typu n x r, kde $r \leq p$.

Matematickou stránkou PCA se zabývají např. Dash, Dubey (2012); Hebák a kol. (2015) a Stankovičová, Vojtková (2007). Původně p proměnných je pomocí PCA transformováno na p PCs podle vztahu (5).

$$\begin{aligned} PC_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ PC_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\ldots \\ PC_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p, \end{aligned} \tag{5}$$

kde X_j označuje p původních proměnných a a_{ij} představují váhy (saturace), které jsou odhadovány tak, aby se vzájemná poloha pozorování v p-rozměrném prostoru nezměnila a aby každá PC byla lineární kombinací původních p proměnných. PCs jsou spolu nekorelované. První PC vysvětluje největší část celkového rozptylu původních dat a každá následující PC vysvětluje co největší část zbylého rozptylu. Nakonec celkový rozptyl všech p PCs zůstává stejný jako v případě původních proměnných (Stankovičová, Vojtková, 2007).

Řešení PCA vyplývá z Jordanovy dekompoziční věty. Z vektoru $\boldsymbol{X} = \left(X_1, X_2, \ldots, X_p\right)^{\mathrm{T}}$ je možné získat vektor středních hodnot $\boldsymbol{\mu} = \left(\mu_1, \ldots, \mu_p\right)^{\mathrm{T}}$ a kovarianční matici $\sum$ typu p x p. Následně je možné kovarianční matici vyjádřit vztahem (6) (Stankovičová, Vojtková, 2007):

$$\sum = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^{\mathbf{T}}, \tag{6}$$

kde $\boldsymbol{\Lambda}$ představuje diagonální matici s vlastními čísly matice $\sum$ označenými jako λ_i a i-tý sloupec matice $\boldsymbol{A}$ je i-tým vlastním vektorem matice $\sum$ značeným $\boldsymbol{v_i}$. Poté matice $\boldsymbol{PC}$ vycházející z vektoru $\boldsymbol{X}$ představuje ortogonální transformaci uvedenou vztahem (Stankovičová, Vojtková, 2007):

$$\boldsymbol{PC} = \boldsymbol{A}^{\mathbf{T}}(\boldsymbol{X} - \boldsymbol{\mu}). \tag{7}$$

Před samotnou analýzou je třeba rozhodnout, zda se bude vycházet z datové matice v původních nebo standardizovaných jednotkách např. podle vztahu (Hebák a kol., 2015; Stankovičová, Vojtková, 2007):

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}, i = 1, 2, \ldots, n \text{ a } j = 1, 2, \ldots, p, \tag{8}$$

kde x_{ij} představuje původní hodnotu pro i-tý objekt a j-tou proměnnou, $\overline{x_j}$ je aritmetický průměr pro j-tou proměnnou a s_j představuje směrodatnou odchylku pro j-tou proměnnou. Jedná se o standardizaci pomocí průměru a směrodatné odchylky, dále značené jako *„z-skóre“*. Vlivem této standardizace mají výsledné proměnné nulovou střední hodnotu a jednotkový

rozptyl. Standardizace pomocí vztahu (8) patří mezi nejpoužívanější, avšak může způsobit ztrátu informací v datech. Tato standardizace není vhodná v případě, kdy se v datovém souboru vyskytují odlehlá pozorování a proměnné nepocházejí z normálního rozdělení pravděpodobnosti. (Holčík a kol., 2015)

Dále je pro účely této práce uvedena standardizace dat rozpětím vyjádřená vztahem (9), dále značená jako *„min-max“*, popř. unitarizace. (Holčík a kol., 2015)

$$u_{ij} = \frac{x_{ij} - x_{min,j}}{x_{max,j} - x_{min,j}}, i = 1, 2, \dots, n \text{ a } j = 1, 2, \dots, p, \tag{9}$$

U standardizace rozpětím představují $x_{min,j}$ minimální hodnotu pro *j*-tou proměnnou a $x_{max,j}$ maximální hodnotu pro *j*-tou proměnnou. Tato standardizace mění rozsah každé proměnné na konstantní, jednotkový interval. Doporučuje se použít v případě, když proměnné nepochází z normálního rozdělení pravděpodobnosti a obsahují odlehlá pozorování. (Holčík a kol., 2015)

PCA může vycházet jak z kovarianční matice, tak z korelační matice. Určení PCs z kovarianční matice lze využít, pokud jsou náhodné veličiny v porovnatelných měřících jednotkách a z hlediska rozptylu se příliš neliší. V opačné případě má smysl provádět PCA na základě korelační matice (Hebák a kol., 2015).

Na základě vztahu (10) lze *i*-tou hlavní komponentu vyjádřit jako:

$$\boldsymbol{PC_i} = \boldsymbol{v_i^T}(\boldsymbol{X} - \boldsymbol{\mu}). \tag{10}$$

Výběrem počtu *„smysluplných“* hlavních komponent (PCs) se mimo již zmíněné autory zabývají také Bro, Smilde (2014). Podle nich je důležité pochopit vztah mezi vlastními čísly a PCs. Mírou významu *i*-té hlavní komponenty z hlediska vysvětleného celkového rozptylu původních proměnných je podíl:

$$\lambda_i / st(\sum) = \lambda_i / st(\boldsymbol{\Lambda}), \tag{11}$$

kde symbol st značí stopu matice. Pokud vlastní vektory $\boldsymbol{v_i}$ vynásobíme odmocninou z odpovídajícího vlastního čísla λ_i získáme vektory komponentních zátěží (Hebák a kol., 2015):

$$\boldsymbol{a_i} = \boldsymbol{v_i}\sqrt{\lambda_i}, \tag{12}$$

Prvky a_{ij} vektoru komponentních zátěží, tzv. *komponentní zátěže*, jsou kovariancí *j*-té původní proměnné a *i*-té hlavní komponenty. Tudíž lze na základě nich hodnotit vztah původních proměnných a získaných hlavních komponent.

Stankovičová, Vojtková (2007) uvádějí pro určení počtu PCs následující pravidla:

- *Kaiserovo pravidlo:* použít počet PCs, jejichž vlastní čísla jsou větší než 1, to znamená, že jsou větší než průměr všech vlastních čísel,
- počet PCs se vybere podle toho, zda spolu vysvětlí dostatečný celkový rozptyl původních dat,
- podle zlomu v *sutinovém grafu* (Scree plot) (osa x představuje pořadí PCs, osa y hodnoty vlastních čísel), použity jsou PCs po zlomový bod.

K tomu, aby byly PCs dobře interpretovatelné, je důležité použít vhodnou rotaci. Hebák a kol. (2015) uvádějí, že jednoduchá struktura modelu komponentní analýzy s r ($r < p$) zvolenými PCs by měla vypadat následovně:

- každý řádek matice zátěží by měl obsahovat alespoň jednu nulu,
- každý sloupec matice zátěží by měl obsahovat alespoň r nul,
- pro každou dvojici sloupců těchto zátěží by měly převládat proměnné, které s jednou komponentou mají nulovou zátěž a s ostatními zátěže vysoké,
- pro více než čtyři PCs by mělo být v každé dvojici sloupců matice zátěží co nejvíce proměnných s nulovými zátěžemi v obou sloupcích,
- pro každou dvojici sloupců matice zátěží by mělo být málo proměnných s vysokými zátěžemi v obou sloupcích.

Dále uvádí, že rotací neboli transformací PCs se rozumí výpočetní operace, pomocí které se z matice původních (nerotovaných) zátěží získá nová matice. Před samotným zavedením rotace je důležité se rozhodnout, zda zvolit rotační algoritmus ortogonální (pravoúhlý), který vede k řešení s nekorelovanými PCs nebo algoritmus šikmý, při němž dochází k výskytu korelací mezi PCs. Problematikou nejčastěji používané ortogonální Varimax rotace se zabývají např. Byung (2018); Richman (1986) a Weide, Beauducel (2019).

Jestliže je matice odhadnutých zátěží před rotací označena $\boldsymbol{B}_{pxr}$, kde p je počet proměnných a r počet zvolených rotovaných PCs, pak nově odhadnuté rotované zátěže, lze určit pomocí vztahu (Hebák a kol., 2015; Stankovičová, Vojtková, 2007):

$$\widetilde{\boldsymbol{B}} = \boldsymbol{BT}, \tag{13}$$

kde $\boldsymbol{T}$ je ortogonální matice typu r x r, tzv. transformační matice.

Normalizovanou varimax rotaci v roce 1958 navrhl H. F. Kaiser. Prvky transformační matice $\boldsymbol{T}$ jsou stanoveny maximalizací vztahu (Hebák a kol., 2015):

$$\sum_{r=1}^{R}\left[\frac{p\sum_{j=1}^{p}\left(b_{jr}^{2}\right)^{2}-\left(\sum_{j=1}^{p}b_{jr}^{2}\right)^{2}}{p^{2}}\right]. \tag{14}$$

Důležitými ukazateli kvality výsledků PCA jsou *komunality* (communality), *jedinečnost* (uniqueness) a *komplexita* (complexity). Komunalita (*h2*) je podle Revelle (2020) součet čtverců komponentních zátěží, který se odhaduje pro každou proměnnou. Maximální hodnota komunality dosahuje jedné. Komunalita uvádí na kolik procent je původní proměnná vysvětlena pomocí RCs. Pro detaily viz. Košťál (2013).

Pomocí $1-h2$ (pokud byly původní proměnné standardizované) dostáváme tzv. jedinečnost (*u2*). Jedinečnost na základě jejího výpočtu lze chápat jako procento variability konkrétní proměnné, která nemá vztah ke komponentám, podle Škaloudová (2010).

Nakonec komplexita (*com*) je podle Revelle (2020) uvedena jako *Hoffmanův index komplexity* (Hoffman's index), který je vypočten pomocí rotovaných komponentních zátěží a_{rj} pro každou původní proměnnou vztahem:

$$\frac{\left(\sum_{R}a_{rj}^{2}\right)^{2}}{\sum_{R}a_{rj}^{4}}. \tag{15}$$

Pettersson, Turkheimer (2010) uvádějí, že index komplexity představuje původní počet latentních proměnných (RCs) potřebných k zohlednění původních proměnných. Výsledek tohoto indexu roven jedné patří původní proměnné, která má vysokou komponentní zátěž pouze s jednou komponentou a s ostatními téměř mizivou. Pokud však index přesáhne hodnotu jedna, daná proměnná má významné zatížení s více než jednou komponentou.

5.1.3 Řídká analýza hlavních komponent

SPCA je modernější variantou PCA a stejně jako rotace PCs slouží ke snadnější interpretaci nově získaných latentních proměnných. Zou a kol. (2006) vyvinuli SPCA z PCA zavedením omezení pro regresní koeficienty (váhy) k získání tzv. *řídkých komponentních zátěží*. Hlavní nevýhodou PCA je to, že každá PC je lineární kombinací všech původních proměnných, což vede k obtížné interpretaci výsledků. Na druhou stranu SPCA se pokouší najít řídké komponentní zátěže neboli řídké váhové vektory, což znamená, že komponentní zátěže jsou aktivní (nenulové) pouze s několika původními proměnnými v rámci jedné komponenty. Stejně jako PCs jsou SPCs (řídké hlavní komponenty) vytvořeny jako lineární kombinace, ale tentokrát pouze několika původních proměnných. Výhodou však není pouze snadnější interpretace SPCs, ale také schopnost SPCA vyhnout se *„přeparametrizování modelu"*, které

může nastat v případě, že počet proměnných *p* je vyšší než počet pozorování *n*, podle Erichson a kol. (2018b); Zou a kol. (2006).

K původní datové matici $\boldsymbol{X}_{nxp}$ se SPCA snaží minimalizovat následující účelovou funkci vyjádřenou vztahem (Erichson a kol., 2018b):

$$f(\boldsymbol{A},\boldsymbol{B}) = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^{\mathrm{T}}\|_{\mathrm{F}}^{2} + \psi(\boldsymbol{B}) \quad \text{za podmínky } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{I}, \tag{16}$$

kde $\boldsymbol{B}_{nxp}$ je matice řídkých vah se značením prvků *b*, tzn. zátěží a $\boldsymbol{A}_{nxp}$ je matice ortonormální. Symbol ψ představuje funkci (regulaci), která způsobuje řídkost, a to buď *LASSO* (Least Absolute Shrinkage and Selection Operator) nebo *elastickou síť*. LASSO je jednou z technik výběru proměnných (Tibshirani, 1996). Jejím zobecněním je právě zmiňovaná elastická síť (Zou, Hastie, 2005; Zou a kol., 2006). SPCA je mimo jiné schopna snížit počet explicitně použitých proměnných.

Nechť $\boldsymbol{Y} = (y_1, y_2, \dots, y_n)^T$ je vektorem odezvy (v našem případě PCs) a $\boldsymbol{X}_{nxp}$ je matice prediktorů, pak odhad parametrů *b* pomocí elastické sítě je možné popsat vztahem:

$$\widehat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda_1\|\boldsymbol{b}\|_1 + \lambda_2\|\boldsymbol{b}\|^2, \tag{17}$$

kde $\|\boldsymbol{b}\|_1 = \sum_{j=1}^{p}|b_j|$, λ jsou nezáporné. Výraz $\lambda_1\|\boldsymbol{b}\|_1$ vytváří řídký model, část $\lambda_2\|\boldsymbol{b}\|^2$ odstraňuje omezení, která vzniknou, pokud $p > n$ a dále provede seskupený výběr proměnných na rozdíl od LASSO. V případě, že $\lambda_2 = 0$ stává se LASSO speciálním případem elastické sítě (Zou a kol., 2006).

Část $\arg\min_{\boldsymbol{b}}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|^2$ (viz. vztah (17)) odhadu parametrů *b* představuje regresní metodu pocházející z aproximace *obyčejnými nejmenšími čtverci* (OLS – Ordinary Least Squares), kde vektor $\boldsymbol{Y}$ je aproximován prediktory v $\boldsymbol{X}$. Koeficienty pro každou proměnnou v matici $\boldsymbol{X}$ jsou obsaženy v $\boldsymbol{b}$. Pokud k tomuto výrazu přidáme $\lambda_2\|\boldsymbol{b}\|^2$ dostaneme tzv. *hřebenovou regresi* (ridge regression). Jakékoliv pozitivní λ_2 způsobí snížení koeficientů $\boldsymbol{b}$. Pokud k tomuto výrazu přidáme výraz $\lambda_1\|\boldsymbol{b}\|_1$ jedná se o již zmiňované LASSO. Pro více detailů viz. Sjöstrand a kol. (2006).

Pokud označíme $\alpha = \frac{\lambda_2}{\lambda_2+\lambda_1}$, pak je možné regulaci pomocí elastické sítě parametrů $\boldsymbol{b}$ uvést vztahem (Zou, Hastie, 2005):

$$J(\boldsymbol{b}) = \alpha\|\boldsymbol{b}\|^2 + (1-\alpha)\|\boldsymbol{b}\|_1. \quad (18)$$

Je zřejmé, že u SPCA se jedná o kompromis mezi řídkostí a variabilitou zachycenou SPCs. Hlavní výzvou je proto výběr parametrů podporující řídkost nebo volba počtu nenulových zátěží (Gajjar a kol., 2016). Podle Zou a kol. (2006) je vhodné nastavit penalizační parametry tak, aby SPCs vysvětlovaly téměř stejnou variabilitu jako PCA. Gajjar a kol. (2016) se přiklánějí k volbě počtu nenulových zátěží tak, aby každá SPC v SPCA zachycovala přibližně stejnou variabilitu jako každá PC v PCA.

V programu R je pro SPCA dostupná funkce *spca* v rámci balíku *sparsepca*. Zde je použita právě elastická síť (Erichson a kol., 2018a). SPCs jsou získány ze vztahu (Erichson a kol., 2018b).

$$\boldsymbol{SPC} = \boldsymbol{XB}. \quad (19)$$

5.1.4 Kernel analýza hlavních komponent

PCA a SPCA využívá lineární projekce pro snížení rozměrnosti ukazatelů. Jestliže však existují nelineární vztahy mezi původními proměnnými, je vhodné použít nelineární techniky snížení rozměrnosti dat, např. KPCA (Gutmann, 2017). Schölkopf a kol. (1998) popisují tuto metodu jako nelineární formu PCA. Základem KPCA je zobrazení původních dat do prostoru vyšší dimenze skrze konkrétní funkci, kterou může být např. funkce gaussovská, polynomiální nebo sigmoidní (viz. Kallas a kol., 2012; Pilario a kol., 2020).

Du, Swamy (2013) a Schölkopf a kol. (1998) uvádějí, že KPCA zavádí kernel (jádrové) funkce do PCA. Je dána vstupní množina pozorování $x_i,\ i = 1, 2, \ldots, N$ a $x_i \in \mathbb{R}^P$. Φ: $\mathbb{R}^P \rightarrow \mathbb{R}^Q$ je nelineární zobrazení původního vstupního prostoru do prostoru vyšší dimenze $\mathbb{R}^Q$. Kovariančí matice $\boldsymbol{C}$ v $\mathbb{R}^Q$ je pak vyjádřena vztahem:

$$\boldsymbol{C} = \frac{1}{N}\sum_{i=1}^{N} \Phi(x_i)\Phi(x_i)^{\mathrm{T}}. \quad (20)$$

Hlavní komponenty se vypočítají pomocí řešení problému vlastních čísel λ_i, vztah (Du, Swamy, 2013):

$$\lambda\boldsymbol{v} = \boldsymbol{C}\boldsymbol{v} = \frac{1}{\mathrm{N}}\sum_{j=1}^{N} \left(\Phi\left(x_j\right)^{\mathrm{T}}\boldsymbol{v}\right)\Phi\left(x_j\right), \quad (21)$$

kde všechna řešení $\boldsymbol{v}$ s $\lambda \neq 0$ leží v rozpětí zobrazovaných dat $\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_N)$ a $\boldsymbol{v}$ je možné vyjádřit vztahem (Schölkopf a kol., 1998):

$$\boldsymbol{v} = \sum_{i=1}^{N} \alpha_i \Phi(x_i), \tag{22}$$

kde α_i jsou složky vlastních vektorů matice $\boldsymbol{K}$.

Následně kombinací vztahů (21) a (22), tzn. pronásobením obou stran vztahu (22) pomocí $\Phi(x_j)$ dostáváme vztah (Du, Swamy, 2013):

$$\boldsymbol{K\alpha} = \lambda\boldsymbol{\alpha}, \tag{23}$$

kde λ a $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)^{\mathrm{T}}$, v tomto pořadí, jsou vlastní čísla a k nim odpovídající vlastní vektory matice $\boldsymbol{K}$, která je jádrovou maticí o rozměrech NxN s prvky podle vztahu (Du, Swamy, 2013):

$$k_{ij} = k(x_i, x_j) = \left(\Phi(x_i)^{\mathrm{T}} . \Phi(x_j)\right). \tag{24}$$

Způsob výpočtu skalárního součinu pro některé funkce Φ vztahem (24) je dán teorií *reprodukčního kernel Hilbertova prostoru* (RKHS). Pro některé funkce není třeba znát transformované datové body $\Phi(x_i)$, aby bylo možné vypočítat skalární součin mezi nimi, ale stačí znát k_{ij}. Toto se nazývá *trikem jádra* (kernel trick), který lze použít k výpočtu *Gramove matice* (Gram matrix) k_{ij}, což sníží náročnost výpočtu (Gutmann, 2017).

Jednou z možných jádrových funkcí je Gaussova RBF uvedená vztahem (Du, Swamy, 2013; Gutmann, 2017 a Schölkopf a kol., 1998):

$$k_{ij} = \exp\left(-\sigma\|x_i - x_j\|^2\right), \tag{25}$$

kde $\sigma > 0$ je parametrem tohoto rozdělení.

Funkce jádra představují transformační funkci splňující nutnou a postačující podmínku tzv. *Mercerův teorém* (Mercer's theorem). To znamená, že funkce jádra musí být pozitivně definitní (viz. Genton, 2001; Schölkopf a kol., 1998).

Gaussova radiálně bazická funkce jádra je Mercerovo jádro. Tato funkce je spojitá symetrická a pozitivně definitní funkce, což znamená, že matice jádra musí být semidefinitní (obsahuje pouze pozitivní vlastní čísla) (Du, Swamy, 2013).

Další často používanou jádrovou funkcí je funkce polynomiální uvedená vztahem (Gutmannm, 2017; Pilario a kol., 2020):

$$k_{ij} = \left(\langle x_{i,} x_j \rangle + 1\right)^d, \tag{26}$$

kde *d* je parametrem tohoto rozdělení (stupněm polynomu) a $\langle x_{i,} x_j \rangle$ skalární součin.

Poslední zde uvedenou funkcí jádra je funkce hyperbolický tangens, která je vyjádřená vztahem (Lin, Lin, 2003; Pilario a kol., 2020):

$$k_{ij} = \tanh\left(s\langle x_{i,} x_j \rangle + c\right), \tag{27}$$

kde *s* představuje škálový parametr (scale) a *c* je parametr posunu.

Nové kPCs původních dat jsou extrahovány promítnutím zobrazovaného vzoru $\Phi(x)$ do $\boldsymbol{v}_k$ podle vztahu:

$$\left(\boldsymbol{v}_{\mathrm{k}}^{\mathrm{T}} . \Phi(x)\right) = \sum_{j=1}^{N} \alpha_{k,j} k\left(x_j, x\right),\ k = 1,2, \ldots, Q \tag{28}$$

kde $\alpha_{k,j}$ je j-tý prvek $\boldsymbol{\alpha}_k$.

Program R umožňuje aplikaci KPCA pomocí balíku *kernlab*, který obsahuje funkci *kpca*. Jednou z funkcí jader, kterou funkce *kpca* poskytuje, je nejznámější a nejpoužívanější radiální bazická (Gaussovská) funkce s jedním parametrem σ. Dále jsou v rámci funkce *kpca* k dispozici často ve článcích zmiňované výše uvedené funkce jádra, např. polynomiální a hyberbolický tangens.

5.2 Metody shlukování objektů

Dle metodologie SEMMA samotnému modelování předchází fáze spočívající v seskupování případů. Existuje celá řada metod shlukových analýz sloužících k seskupování objektů nebo proměnných. Tato práce se zaměřuje pouze na shlukování objektů (států). Shlukovou analýzu poprvé použil v roce 1939 R. C. Tryon (Stankovičová, Vojtková, 2007). Shluková analýza klasifikuje objekty do stejnorodých shluků. Hlavním jejím cílem je klasifikovat objekty do skupin tak, aby si objekty v určité skupině byly co nejvíce podobné (Hebák a kol., 2007).

Každá metoda shlukové analýzy má své silné a slabé stránky. Některé z nedostatků jedné konkrétní metody je možné odstranit použitím metody jiné. Avšak u sofistikovanějších metod shlukové analýzy je vyžadováno nastavení hyperparametrů jako apriorní informace před

samotným spuštěním algoritmu. Nastavení těchto hyperparametrů se může odvíjet od cílů, ke kterým má shluková analýza sloužit.

5.2.1 Hierarchické a nehierarchické metody shlukové analýzy

Jak již bylo zmíněno, cílem shlukové analýzy je zařazování objektů do skupin, kde jsou si tyto objekty co nejvíce podobné. Součástí shlukové analýzy je stanovení způsobu určení podobnosti nebo nepodobnosti mezi objekty. V případě kvantitativních dat se používají míry vzdálenosti. Mezi nejznámější vzdálenosti patří *euklidovská vzdálenost* (euclidean distance), kterou je možné vyjádřit vztahem (Řezanková a kol., 2009).

$$\mathrm{D_E}(x_i, x_j) = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}, \tag{29}$$

kde x_{ik} je hodnota *k*-té proměnné pro *i*-tý objekt a x_{jk} je hodnota *k*-té proměnné pro *j*-tý objekt.

Jak bylo zmíněno v části 2.2.1 mezi nejpoužívanější hierarchické metody shlukování patří Wardova metoda. Poprvé byla tato metoda popsána v článku Ward (1963). Shluky se tvoří maximalizací vnitroshlukové homogenity (Hair a kol., 1992; Stankovičová, Vojtková, 2007). Výhodou této metody je především to, že odstraňuje malé shluky obsahující pouze jedno pozorování a vytváří shluky podobné velikosti (Holčík a kol., 2015). Zřejmě i z tohoto důvodu se jedná o jednu z nejpoužívanějších metod hierarchického shlukování, která je proto i v této práci upřednostněna před ostatními metodami shlukové analýzy.

Problematikou algoritmů, které implementují Wardovo kritérium, se zabývají Murtagh, Legendre (2014). Uvádějí funkci danou následujícím vztahem, kterou je třeba minimalizovat.

$$\delta^2(c_1, c_2) = \frac{|c_1||c_2|}{|c_1| + |c_2|} \|c_1 - c_2\|^2 \tag{30}$$

Tímto dochází k měření změny celkového součtu čtverců vycházejícího ze sloučení tříd c_1 a c_2, kde c_1 a c_2 mohou být vektory původních dat nebo průměrů.

Pro shlukování Wardovou metodou jsou známé dva algoritmy shlukové analýzy v programu R. Jedná se o algoritmy *„ward.D“* a *„ward.D2“*. Tyto dva algoritmy se liší použitím vstupní matice vzdáleností, která má v případě *„ward.D“* podobu čtvercové euklidovské vzdálenosti a v případě *„ward.D2“* podobu euklidovské vzdálenosti (Murtagh, Legendre, 2014).

Murtagh, Legendre (2014) popisují vztah pro algoritmus *„ward.D2“* následovně:

$$\delta_{(ij)k} = \left(\frac{n_1 + n_3}{n_1 + n_2 + n_3} \delta^2(C_i, C_k) + \frac{n_2 + n_3}{n_1 + n_2 + n_3} \delta^2(C_j, C_k) - \frac{n_3}{n_1 + n_2 + n_3} \delta^2(C_i, C_j) \right)^{1/2} \tag{31}$$

kde n_1, n_2 a n_3 představují velikosti shluků C_i, C_j a C_k.

Székely, Rizzo (2005) uvádějí hierarchický aglomerativní algoritmus shlukové analýzy zvaný *Lance-Williams algoritmus* rekurentním vzorcem, vztah (32). Vztah (31) je odvozený na základě Lance-Williamsova algoritmu.

$$\delta_{(ij)k} = \alpha_i \delta_{ik} + \alpha_j \delta_{jk} + \beta \delta_{ij} + \gamma \left| \delta_{ik} - \delta_{jk} \right| \tag{32}$$

Předpokládá se, že δ_{ij}, δ_{ik} a δ_{jk} představují párové vzdálenosti mezi shluky C_i, C_j a C_k. Poté $\delta_{(ij)k}$ značí vzdálenost mezi novým shlukem $C_i \cup C_j$ a C_k (Székely, Rizzo, 2005). Ve vztahu (31) závisí parametry α_i, α_j, β a γ na velikosti shluku.

Protože existuje celá řada možností měření vzdáleností mezi objekty a mnoho metod hierarchické aglomerativní shlukové analýzy, je možné na základě dendrogramu získat mnoho odlišných výsledných shluků. Porovnáním dendrogramu pomocí kofenetického koeficientu korelace (cophenetic correlation coefficient) se zabývají např. Sokal, Rohlf (1962).

Carr a kol. (1999) uvádějí, že kofenetická korelace představuje korelaci mezi skutečnými vzdálenostmi získanými z původního datového souboru (matice vzdáleností) a vzdálenostmi, které je možné vyčíst z dendrogramu (kofenetická matice). Kofenetická korelace představuje míru toho, jak dobře dendrogram modeluje skutečnost. Kofenetický koeficient korelace lze vyjádřit vztahem (Holgerson, 1997):

$$\rho = \frac{\sum_{i<j} (d_{ij} - \bar{d})(d_{ij}^* - \overline{d^*})}{\sqrt{\sum_{i<j} (d_{ij} - \bar{d})^2 \sum_{i<j} (d_{ij}^* - \overline{d^*})^2}}, \tag{33}$$

kde d_{ij} představuje původní vzdálenost mezi *i*-tým a *j*-tým objektem a d_{ij}^* korespondující minimální vzdálenost mezi *i*-tým a *j*-tým objektem v dendrogramu. $\bar{d}$ a $\overline{d^*}$ jsou odpovídající průměrné vzdálenosti.

Mezi nehierarchické metody shlukování patří např. metoda *k*-průměrů (*k*-means). Jedná se o optimalizační metodu. Tato metoda je založená na počátečním rozdělení objektů

do *k* shluků (shlukových centroidů). Metoda *k*-průměrů je vhodná pro velké datové soubory, protože není nutné pracovat s maticí vzdáleností (Řezanková a kol., 2009 a Stankovičová, Vojtková, 2007).

V rámci výchozího nastavení programu R je u metody *k*-průměrů použit algoritmus popsaný v publikaci Hartigan, Wong (1979). Kromě těchto zmíněných autorů se algoritmem Hartigan-Wong zabývají také např. Morissette, Chartier (2013).

5.2.2 Metoda Fuzzy *k*-průměrů (Fuzzy *C*-means)

Tento iterativní algoritmus pro shlukování objektů vyvinul Dunn (1973) a následně byl vylepšen v článku Bezdek (1981). Na rozdíl od běžně používaných hierarchických a nehierarchických metod shlukování FCM algoritmus umožňuje přiřazení jednoho objektu k více než jednomu shluku pomocí stupně příslušnosti. FCM algoritmus minimalizuje následující účelovou funkci:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2, 1 \leq m < \infty, \tag{34}$$

kde x_i je *i*-tý prvek v *p*-rozměrném datovém prostoru, m představuje fuzzifikační koeficient, μ_{ij} je stupeň příslušnosti prvku x_i ve shluku *j*, dále c_j je *p*-rozměrný střed shluku a $\left\| x_i - c_j \right\|$ je norma, která vyjadřuje podobnost mezi naměřenými daty a středem.

Algoritmus je založen na optimalizaci účelové funkce (34) při aktualizaci stupňů příslušnosti a center shluků pomocí následujících vztahů:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}, \tag{35}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}}. \tag{36}$$

Iterace se zastaví, pokud bude splněno, že $\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, kde ε je kritérium ukončení ležící v intervalu $\langle 0; 1 \rangle$ a *k* jsou iterační kroky.

Program R nabízí v rámci balíku *ppclust* funkci *fcm* sloužící pro rozdělení objektů pomocí FCM algoritmu.

5.2.3 DBSCAN algoritmus

DBSCAN algoritmus patří mezi algoritmy založené na hustotě, kde hustota je v tomto případě chápána ve smyslu četností a vzdáleností objektů v jejich okolí. Podle Hebák a kol. (2015) je shluk definován jako množina objektů, které jsou spojeny na základě této hustoty.

Nutností je zadání dvou vstupních parametrů, konkrétně se jedná o poloměr okolí a minimální počet pozorování patřící do tohoto okolí. Výsledný počet shluků následně závisí právě na nastavení těchto parametrů (Hebák a kol., 2015). Velikost poloměru okolí se nazývá ε-ové okolí bodu (pozorování). Je k dispozici datová matice n x p. Velikost ε-ového okolí pozorování označeného *eps* je vyjádřena vztahem (Ester a kol., 1996):

$$eps = \{s \in N \mid dist(r, s \leq \varepsilon)\}, \tag{37}$$

kde r, s jsou pozorování a $r, s \in N$ a $\varepsilon \in \mathbb{R}^+$.

Jak již bylo zmíněno, je nutné také nastavení minimálního počtu pozorování patřícího do ε-ového okolí značeno jako *minPts*. Pozorování náležející jednotlivým shlukům lze rozdělit na *jádrová* a *hraniční* (viz. obrázek 5 - vlevo). Pozorování, která nepatří ani do jednoho shluku jsou označována jako *šumová* pozorování. Obecně platí, že ε-ové okolí hraničního pozorování obsahuje méně pozorování než ε-ové okolí jádrového pozorování. Je proto vyžadováno, aby pro každé pozorování $r \in C$ existovalo pozorování $s \in C$ takové, že pozorování r je uvnitř ε-ového okolí s a eps obsahuje alespoň minimální počet pozorování (viz. Ester a kol., 1996; Hahsler a kol., 2019).

Podle Hahsler a kol. (2019) je důležitým pojmem v rámci algoritmu DBSCAN *přímá dosažitelnost hustotou* (directly density-reachable). Pozorování $s \in eps$ je *přímo dosažitelné hustotou* z pozorování $r \in eps$ vzhledem k *minPts*, jestliže:

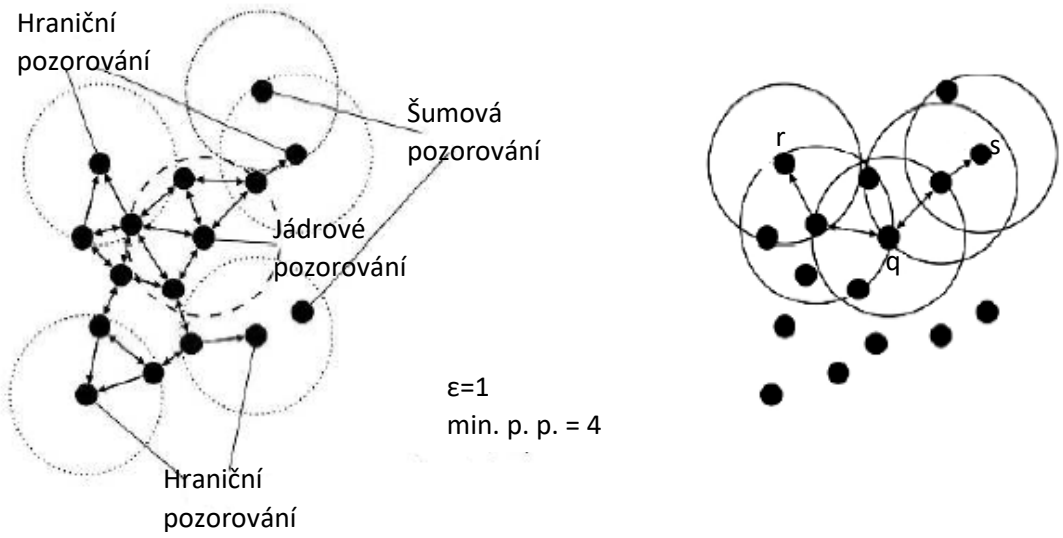
- $|eps(r)| \geq minPts$,
- $s \in eps(r)$.

Pozorování r je *hustotou dosažitelné* (density-reachable) z pozorování s, jestliže existuje uspořádaná posloupnost pozorování $(r_1, r_2, \dots, r_n) \in eps$, kde $s = r_1$ a $r = r_n$, u které platí, že $r_i + 1$ je přímo hustotou dosažitelné z r_i pro všechna $i \in \{1, 2, \dots, n-1\}$. Pozorování $r \in eps$ je *hustotou spojené* (density-connected) s pozorováním $s \in eps$, pokud existuje pozorování $q \in eps$ takové, že pozorování r a s jsou hustotou dosažitelné z pozorování q. Pro detaily viz. Hahsler a kol. (2019).

Výsledný shluk C je tedy neprázdná podmnožina *eps*, která splňuje podmínku *maximality* a *konektivity*. Maximalitou se rozumí, jestliže pozorování $r \in C$ a s jsou hustotou dosažitelná z pozorování r, pak pozorování $s \in C$. Konektivita znamená, že pro všechna pozorování $r, s \in C$ platí, že pozorování r je hustotou spojeno s pozorováním s (Hahsler a kol., 2019).

Obrázek 5 – vpravo zobrazuje koncept dosažitelnosti hustotou a spojení hustotou. Pozorování r a s jsou dosažitelná od pozorování q. Proto také pozorování r a s jsou hustotou spojená. Velikost ε-ového okolí je v tomto případě stanovena na hodnotu 1 a minimální počet pozorování patřící do ε-ového okolí jsou čtyři.

Obrázek 5: Koncepty používané DBSCAN algoritmem



Zdroj: zpracováno podle Hahsler a kol. (2019)

5.2.4 Metody pro stanovení optimálního počtu shluků

Mezi nejobtížnější úkoly shlukové analýzy patří stanovení optimálního počtu shluků. Existuje celá řada možností, jak stanovit optimální počet shluků. Nejjednodušším způsobem je heuristický přístup, kdy je možné např. u aglomerativní hierarchické metody shlukové analýzy z dendrogramu vyčíst výrazné shluky. Další možností je zjištění různých indexů a statistik pro různé počty shluků. Optimální počet shluků je pak určen maximální nebo minimální hodnotou těchto statistik (Hebák a kol., 2015). Stanovením optimálního počtu shluků ve shlukové analýze se mimo jiné zabývá Löster (2016). Podle něho je jednou z možností pro hodnocení počtu shluků princip analýzy rozptylu, kdy je maximalizován poměr mezishlukové a celkové variability:

$$\max\left(\frac{SS_B}{SS_T}\right), \tag{38}$$

kde SS_B představuje mezishlukový součet čtverců a SS_T celkový součet čtverců. Maximalizace tohoto poměru pro stanovení optimálního počtu shluků se využívá u metody *k*-průměrů a FCM algoritmu. Následující uvedené koeficienty pro stanovení optimálního počtu shluků se používají u FCM algoritmu. Jedná se o vztahy (39) až (44).

Rozdělovací koeficient (PC(*k*) – Partition Coefficient) měří, jak blízko je fuzzy řešení k tzv. tvrdému řešení. Tento koeficient je dán vztahem (Říhová, Říha, 2019):

$$\mathrm{PC}(k) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\mu_{ij}^2, \tag{39}$$

kde n je počet objektů, k počet shluků a μ_{ij} je stupeň příslušnosti *i*-tého objektu pro *j*-tý shluk. Rozdělovací koeficient nabývá hodnot z intervalu $\langle\frac{1}{k}; 1\rangle$ Koeficient se rovná hodnotě $\frac{1}{k}$, pokud všechny stupně příslušnosti nabývají hodnoty $\frac{1}{k}$. Druhého extrému, hodnoty 1, nabude koeficient v případě, jestliže stupeň příslušnosti každého pozorování nabývá hodnoty 1 pouze u jednoho ze shluků a u zbytku hodnoty 0. Pro detaily viz. např. Aria (2014); Siddique a kol. (2018). Podle Dunn (1973) je možné najít optimální počet shluků řešením následujícího vztahu:

$$max_{2\leq k\leq n-1}\mathrm{PC}(k) \tag{40}$$

Modifikovaný rozdělovací koeficient (MPC – Modified Partition Coefficient) je dalším koeficientem validity. Jedná se o předchozí modifikovaný koeficient vyjádřený vztahem (Říhová, Říha, 2019):

$$\mathrm{MPC} = 1 - \frac{k}{k-1}\big(1 - \mathrm{PC}(k)\big). \tag{41}$$

Tento koeficient na rozdíl od předchozího nabývá hodnot z intervalu $\langle 0; 1\rangle$.

Koeficient entropie rozkladu (PE – Partition Entropy) opět bere v potaz pouze stupně příslušnosti jako předchozí koeficienty. Je možné ho vyjádřit následujícím vztahem (Gan a kol., 2007):

$$\mathrm{PE} = -\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\mu_{ij}\log_a(\mu_{ij}), \tag{42}$$

Podle Gan a kol. (2007) nabývá tento koeficient hodnot z intervalu $\langle 0, \log_a k \rangle$. FCM algoritmus dosáhne pro předem stanovený počet shluků dobrého výsledku, pokud je blízko hodnoty 0, v opačném případě není přítomna shlukovací struktura nebo ji algoritmus FCM nedokázal najít.

Fuzzy koeficient siluet (FSI – Fuzzy Silhouette index) je mírou, která vybírá dva shluky, ve kterých pozorování x_i nabývá nejvyšších stupňů příslušnosti. Index FSI je popsán následujícími vztahy:

$$\mathrm{FSI} = \frac{\sum_{i=1}^{n}(\mu_{1i} - \mu_{2i})S(x_i)}{\sum_{i=1}^{n}(\mu_{1i} - \mu_{2i})}, \tag{43}$$

$$S(x_i) = \frac{\beta(x_i, \mathrm{gt}_j) - \delta(x_i, \mathrm{gt}_j)}{max\{\beta(x_i, \mathrm{gt}_j) - \delta(x_i, \mathrm{gt}_j)\}}. \tag{44}$$

Pokud je pozorování x_i součástí shluku označeného gt_j, $\mathrm{gt}_j \in (\mathrm{gt}_1, \mathrm{gt}_2, \ldots, \mathrm{gt}_k)$, je $\delta(x_i, \mathrm{gt}_j)$ průměrná vzdálenosti mezi x_i a všemi ostatními pozorováními, které patří do stejného shluku. Vzdálenost mezi x_i a shlukem, ke kterému je gt_j nejblíže je označena $\beta(x_i, \mathrm{gt}_j)$. K získání optimálního počtu shluků je potřebné tento koeficient maximalizovat. Problematikou zde popsaného indexu FSI se zabývají Mallik, Zhao (2019); Rawashdeh, Ralescu (2012); Subbalakshmi a kol. (2015).

V případě DBSCAN algoritmu je možné optimální počet shluků získat pomocí vzdáleností *k* nejbližších sousedů (*k*-NN distance). V rámci tohoto algoritmu závisí výsledný počet shluků na nastavení parametrů týkajících se poloměru okolí a minimálního počtu pozorování patřících do tohoto okolí. Prostřednictvím *k*-NN vzdáleností je možné vybrat vhodný parametr poloměru okolí na základě *k*-NN grafu, který vizualizuje závislost *k*-NN vzdáleností na pozorování seřazených podle *k*-NN vzdáleností. Za vhodnou se považuje velikost poloměru okolí v místě zlomu tzv. *„knee“*.

5.3 Hybridní přístup

Hybridním přístupem je v této práci nazvána kombinace vícerozměrného škálování (MDS) s lineárním uspořádáním (viz. podkapitola 2.3). MDS slouží k redukci vícerozměrného prostoru pozorování a k analýze vztahů mezi nimi (Hebák a kol., 2007). Tato metoda má za cíl nalézt nový souřadnicový systém, ve kterém je možné co nejlépe vizualizovat podobnosti nebo odlišnosti mezi vícerozměrnými objekty. MDS může vycházet z matice podobností, vzdáleností, korelací nebo kovariancí, čímž se liší např. od PCA. V této práci je vstupní datová matice pro MDS vždy matice euklidovských vzdáleností.

Důležitým úkolem v rámci MDS je ověření kvality nového zobrazení např. pomocí *Shepardova diagramu*, který zobrazuje vztah mezi původními a v rámci MDS reprodukovanými vzdálenostmi a idealizovaný vztah mezi původními a v MDS reprodukovanými vzdálenostmi (*D-hat funkce*). Dále je možné k ověření kvality nového zobrazení použít hodnotu, tzv. *stresu* (stress), která je stanovena jako součet čtverců odchylek v MDS reprodukovaných a *D-hat* idealizovaných vzdáleností (Hair a kol., 1992; Hebák a kol., 2007; Hebák a kol., 2015). Hebák a kol. (2015) uvádí tabulku Kruskalem doporučených hodnot, podle nichž je možné posoudit kvalitu modelu MDS. Hodnota stresu by neměla být vyšší než 0,2; což by vypovídalo o slabém modelu, jak je uvedeno v tabulce 2.

Tabulka 2: Kvalita modlu MDS podle Kruskala

STRESS	Kvalita modelu
$> 0{,}2$	slabá
$(0{,}1; 0{,}2\rangle$	uspokojivá
$(0{,}05; 0{,}1\rangle$	dobrá
$(0{,}025; 0{,}05\rangle$	vynikající
$\langle 0; 0{,}025\rangle$	perfektní

Zdroj: zpracováno podle Hebák a kol. (2015)

Základem hybridního přístupu je datová matice typu n x p, kde n je počet objektů a p je počet proměnných. Pro každou proměnnou je stanoveno, zda se jedná o stimulant, destimulant nebo nominant. Na základě již existujících proměnných jsou dále vytvořeny dva umělé objekty, a to objekt vzor (P) a objekt anti-vzor (AP) (viz. podkapitola 2.3). V programu R v rámci balíku *clusterSim* je možné pro stanovení objektů P a AP použít funkci *pattern.GDM1* (Walesiak, Dudek, 2022).

Z původní datové matice rozšířené o objekty P a AP je dále vytvořena matice euklidovských vzdáleností a následně je aplikováno MDS, čímž jsou získány nové souřadnice pro každý objekt a pro zvolený počet dimenzí. Pro vizualizaci jsou vhodné dvě dimenze. V rámci 2D jsou objekty P a AP spojeny přímkou, která se nazývá *osa souboru* (axis of the set). Dále jsou na základě 2D zobrazení v rámci MDS určeny *izokvanty rozvoje* (isoquants of development) na základě vzdálenosti od objektu P rozdělením osy souboru na několik stejných částí (teoreticky jich může být až nekonečno). Pomocí agregované míry d_i jsou následně objekty lineárně uspořádány podle vzdálenosti od objektu P (Walesiak, 2016; Walesiak, Dehnel, 2018):

$$d_i = 1 - \frac{\sqrt{\sum_{j=1}^{R}\left(v_{ij} - v_{+j}\right)^2}}{\sqrt{\sum_{j=1}^{R}\left(v_{+j} - v_{-j}\right)^2}}, \tag{45}$$

kde v_{ij} je *j*-tá souřadnice pro *i*-tý objekt v *R*-rozměrném prostoru vytvořeném pomocí MDS, v_{+j} a v_{-j} je *j*-tá souřadnice pro objekt P a AP. Čitatel zlomku představuje euklidovskou vzdálenost mezi *i*-tým objektem a objektem P a jmenovatel euklidovskou vzdálenost mezi objektem P a AP.

5.4 Vizualizace geografických dat v programu R

Existuje celá řada možností vizualizace výsledků získaných pomocí analýz popsaných v předchozích podkapitolách. Často se získané výsledky vizualizují prostřednictvím 2D, popř. 3D souřadnicových systémů. Pokud je možné analyzovaná pozorování charakterizovat pomocí geografické polohy, je vhodné využít geografických dat pro vizualizace výsledků. Tango (2010) uvádí, že mapování nemocí je nejzákladnějším způsobem vizualizace prostorového rozložení těchto nemocí ve vymezené oblasti. Disertační práce není prioritně zaměřena na aplikaci GIS, ale prostřednictvím geografických dat jsou zde vizualizovány zjištěné výsledky použitých analýz v programu R.

Pojem *geocomputation* je podle Lovelace a kol. (2019) mladým pojmem, který se začal používat v roce 1996. I přes to je tento termínu ovlivněn starými myšlenkami. Abrahart a kol. (2000) uvádějí, že geocomputation je o používání různých typů geografických dat a o vývoji geografických nástrojů v celkovém kontextu vědeckého přístupu. Jedná se o použití dostupných nástrojů statistiky, matematiky, umělé a výpočetní inteligence, které mají za cíl odvést práci přínosnou nebo užitečnou v praxi.

Geografický informační systém Komise (GISCO – Geographic Information System of the COmmission) je v rámci Eurostatu zodpovědný za plnění geografických informačních potřeb Evropské komise na úrovni EU, členských zemí a jejich regionů. Úkolem GISCO je tvorba statických a jiných tematických map, spravování databáze geografických informací a poskytování souvisejících služeb pro Evropskou komisi. Spravovaná databáze obsahuje geodata pokrývající celou Evropu a tematické geoprostorové informace. Za nejdůležitější lokalizační rámec na úrovni EU je považována klasifikace NUTS. Soubory geografických dat jsou poskytovány ve dvou formátech, tzv. *shapefile* a *osobní geodatabáze* Eurostat (2022b).

Program R nabízí v rámci balíku *eurostat*, který je nástrojem sloužícím ke stahování dat z databáze Eurostatu a nástrojem pro vyhledávání a manipulaci s daty, funkci *get_eurostat_geospatial* stahující geodata z GISCO (viz Lathi a kol., 2022).

Důvodem pro použití programu R v této oblasti jsou jeho pokročilé schopnosti analýzy, modelování a vizualizace skrze širokou škálu balíků. Program R je výkonným jazykem pro geocomputation, ale ne jediným. Mezi další je možné zařadit C++, Java, Python aj. Existuje mnoho způsobů, jak zacházet s geografickými daty v programu R prostřednictvím desítek balíků. Vzhledem k faktu, že program R je otevřeným softwarem (open source), je pravděpodobný rychlý vývoj v rámci jednotlivých balíků, jehož výhodou je možnost navázání na již existující balíky a tím pádem vyhnutí se znovuobjevování (Lovelace a kol., 2019). I přes všechny zde uvedené výhody programu R může být obtížné setkání s tímto programovacím jazykem pro nové uživatele. Práce v programu R vyžaduje zkušenosti s psaním skriptů a programováním (Bivand a kol., 2008).

V prvních prostorových balících vyvinutých v 90. letech 20. století se vyvíjely prostorové schopnosti R v *jazyce S* prostřednictvím četných S-skriptů a balíků pro prostorovou statistiku. Velký posun v prostorových schopnostech R přinesly balíky *rgdal* a *sp*. Skrze balík *rgdal* došlo ke zlepšení schopností importování dat z dříve nedostupných datových formátů. První verze z roku 2003 podporovala pouze import rastrových formátů souborů, avšak následná vylepšení umožnila import také vektorovým formátům. V roce 2005 druhý zmíněný balík *sp* umožnil rozlišení prostorových a neprostorových objektů, to znamená, že zeměpisné souřadnice už nebyly považovány za jakékoliv jiné číslo. V roce 2010 byl odstraněn problém s neschopností programu R provádět geometrické operace prostřednictvím balíku *rgeos*. Na začátku vývoje prostorových schopností programu R se nekladl velký důraz na vizualizaci. Možnosti vizualizace se změnily uvedením balíku *ggmap*, pomocí kterého lze usnadnit vizualizace pomocí balíku *ggplot2* se zaměřením na vektorová data (viz. Lovelace a kol., 2019). Kahle, Wickham (2013) zavádějí balík *ggmap* jako nový nástroj pro prostorovou vizualizaci s *ggplot2*. Balík *ggmap* kombinuje prostorové informace z Google Maps, OpenStreetMap, Stamen Maps nebo CloudMade Maps s grafickou implementací *ggplot2*. Vizualizace rastrových dat je podporována balíkem *raster*. Od roku 2018 získal *ggplot2* nové prostorové schopnosti díky balíku *ggspatial*, mezi které patří přidávání prvků prostorové vizualizace nebo prostorové animace aj. (Lovelace a kol., 2019)

Článek Pebesma (2018) představuje balík *sf* (sf – simple features), který slouží ke čtení, zápisu a manipulaci s jednoduchými funkcemi v programu R. Jedná se o moderní alternativu pro části rodiny balíků *sp*. Poskytuje nové základní třídy pro manipulaci s vektorovými daty v programu R. Při implementaci tohoto balíku byl zachován koncept oddělení geometrií a atributů, byly

vytvořeny nová propojení balíků např. *dplyr, ggplot2* aj. Kromě zachování osvědčených konceptů byly implementovány nové prostorové indexy.

Možnosti balíku *ggplot2* jsou popsány v dokumentu Wickham a kol. (2021). Tento balík poskytuje následující funkce:

- *aes* – jedná se o funkci estetického mapování (aesthetic mappings) popisující, jak jsou proměnné v datovém souboru vizualizovány skrze geometrii,
- *coord_map* – tato funkce promítá část země, která je přibližně sférická, na plochou 2D rovinu,
- *geom_path* – spojuje pozorování v pořadí, ve kterém se objevují v datovém souboru,
- *geom_point* – používá se k vytváření bodových grafů, které jsou užitečné k vyjádření vztahu mezi dvěma spojitými proměnnými, pokud přidáme do tohoto grafu třetí proměnnou ovlivňující velikost bodů, jedná se o bublinový graf,
- *geom_polygon* – tato funkce je podobná funkci *geom_path* s tím rozdílem, že počáteční bod je spojen s bodem koncovým do tvaru polygonu (mnohoúhelníku) a jeho vnitřek je obarven výplní, funkce *aes* mimo jiné určuje, které případy jsou spojeny do polygonu,
- *ggplot* – tuto funkci je možné použít ke grafickému zobrazení vstupního datového souboru a ke specifikaci estetického mapování, které má být společné ve všech vrstvách, pokud nedojde k jeho přepsání,
- *labs* – funkce je důležitou součástí grafu z hlediska jeho zpřístupnění širšímu publiku, jedná se zejména o název grafu, názvy os, legendy aj.,
- *lims* – jedná se o funkci sloužící ke změně měřítka os u grafu,
- *scale_colour_gradient* – vytváří vícebarevné stupnice pro spojité proměnné,
- *scale_manual* – lze použít k vytvoření vlastní diskrétní stupnice,
- *theme* – funkce slouží k úpravě nedatových komponent grafu, tzn. název grafu, označení os, legendy atd.

Jedním z balíků, který je propojen s balíkem *ggplot2*, je balík *dplyr*. Balík *dplyr* je popsán v dokumentu Wickham a kol. (2022) a představuje gramatiku a manipulaci s daty. V této disertační práci byly použity následující funkce:

- *filter* – funkce se používá pro podmnožinu datového souboru a uchovává všechny řádky, které splňují zadané podmínky, pro zachování řádků je nutné splnění všech podmínek vytvářejících hodnotu TRUE,

- *mutate-join* – jedná se o funkci přidávající sloupce jednoho datového souboru k jinému datovému souboru na základě klíče.

Možnostmi vizualizace v programovacím jazyku R se zabývá také Lakićević (2021). Existuje celá řada balíků, které činí proces vizualizace snazší a uživatelsky přívětivější. Lakićević (2021) představuje nejběžněji používané balíky *leaflet, ggplot2* a *ggmap* a porovnává výsledné vizualizace získané prostřednictví těchto balíků. Dále uvádí, že výběr balíku závisí nejen na uživatelských dovednostech programování v R, ale také na získané vizuální kvalitě. Z hlediska uživatelských dovedností je balík *leaflet* určen především pro začátečníky, *ggplot2* pro středně pokročilé a *ggmap* pro nejpokročilejší uživatele. Další článek Lemenkova (2021) řeší problematiku používání balíků *tmap, raster* a *ggmap* pro kartografickou vizualizaci, kde mimo jiné uvádí regionální vizualizaci Itálie pomocí balíků *ggmap* a *ggplot2*.

Analyzováním stavu zdraví nebo jeho determinantů se zabývají např. Loop a kol. (2017) nebo Marí-Dell'Olmo a kol. (2021). Například tvorbě heat map závažných onemocnění (hypertenze a diabetes) a jejich rizikového faktoru kouření v USA se věnují Loop a kol. (2017). Výsledkem jejich zjišťování je, že geografické rozdíly existují jak mezi jednotlivými státy, tak v jejich okresech a tyto vzorce se liší podle příslušnosti k rase. Cílem druhého příspěvku Marí-Dell'Olmo a kol. (2021) je analyzovat sociální nerovnosti ve výskytu COVID-19 mimo jiné stratifikované podle geografické oblasti. Jejich výsledky naznačují existenci sociálních nerovností ve výskytu COVID-19 i podle geografické oblasti.

6 Aplikace metod pro snížení rozměrnosti ukazatelů stavu zdraví a jeho determinantů

V této kapitole jsou naplňovány dílčí cíle disertační práce týkající se aplikací výše popsaných metod. Následně jsou jejich výsledky graficky vizualizovány. Výpočty a jejich vizualizace jsou provedeny převážně v programu R. Program R je volně dostupné softwarové prostředí pro statistické výpočty a grafiku (viz. R-project, 2022). V příloze 1 jsou uvedeny kódy pro provedení aplikovaných analýz v tomto programu v rámci použitých balíků.

Zde použitá agregovaná průřezová data týkající se stavu zdraví a determinantů stavu zdraví pro poslední dostupné roky pochází z databází a publikací Dyba a kol. (2021), Eurostat (2022a), OECD/European Union (2020), Timmis a kol. (2022), WHO (2022b), WHO (2022c) a WHO (2022d) pro země EU-27. Data jsou získána pro období před pandemií Covid-19, ve většině případů pro rok 2019 nebo nejbližší dostupný rok. V datech stavu zdraví a jeho determinantů jsou zahrnuty ukazatele jak pro všechny věkové kategorie dohromady, tak pro věkovou kategorii 65+. Všechny ukazatele jsou uvedeny bez rozlišení pohlaví.

První podkapitola 6.1 se věnuje použitým ukazatelům stavu zdraví pro státy EU-27. Ukazatele stavu zdraví jsou vybrány na základě PHI (indexu zdraví populace viz. kapitola 1.2), který zahrnuje jak ukazatele týkající se úmrtnosti, tak ukazatele týkající se nemocnosti. Dalšími podmínkami pro výběr ukazatelů do zde uvedených analýz je jejich dostupnost v rámci jednotlivých databází a publikací, jejich předchozí použití v souvislosti s měřením stavu zdraví, jejich využití v rámci indexů stavu zdraví, popř. se jedná o samotný index stavu zdraví (viz. kapitola 1).

Následně v podkapitole 6.2. jsou zjišťovány závislosti mezi dvojicemi ukazatelů stavu zdraví a je snížena rozměrnost těchto ukazatelů. Na základě zredukovaných datových matic na straně ukazatelů jsou v podkapitole 6.3 pomocí různých metod shlukové analýzy rozděleny země EU-27 do jednotlivých skupin (shluků) podle podobné situace ve stavu zdraví. Vzhledem k existenci odlehlých objektů (států) detekovaných pomocí metod pro snížení rozměrnosti ukazatelů jsou v podkapitole 6.4 dále identifikovány tyto odlehlé státy s podobnou úrovní stavu zdraví (i když s odlišným uspořádáním hodnot vstupních proměnných), které jsou v rámci hybridního přístupu lineárně uspořádány.

Podkapitoly 6.5 a 6.6 se věnují stejné problematice jako podkapitoly předchozí, nyní však pro případ datových souborů ohodnocujících země EU-27 pomocí determinantů stavu zdraví. Za účelem přehlednosti jsou výsledky analýz vizualizovány také pomocí geografických dat

v programu R. Podkapitola 6.5 se věnuje determinantům stavu zdraví pro státy EU-27. Determinanty stavu zdraví jsou vybrány na základě WHO (2017), Dahlgren, Whitehead (1991) a Golembilewski a kol. (2019). Dahlgren, Whitehead (1991) se zabývají determinanty stavu zdraví, které lze ovlivnit politickými zásahy, tzn. informace právě o těchto determinantech může být užitečná pro jejich řízení kompetentními orgány. Výběr determinantů stavu zdraví je také ovlivněn článkem Golembilewski a kol. (2019), kteří používají neklinické determinanty stavu zdraví. Dalšími podmínkami pro vstup ukazatelů do zde uvedených analýz je jejich dostupnost v rámci jednotlivých databází a publikací, jejich předchozí použití v souvislosti s měřením nerovností stavu zdraví a systémů zdravotní péče (viz. kapitola 1).

Podkapitoly 6.6 – 6.8 se věnují analyzování determinantů stavu zdraví v zemích EU-27 pomocí metod pro snížení rozměrnosti ukazatelů, metod shlukové analýzy a hybridního přístupu. Použité postupy jsou podobné jako při analyzování stavu zdraví s přihlédnutím k získaným výsledkům. Závěrečná podkapitola 6.9 se věnuje porovnání zemí EU-27 podle agregovaných měr stavu zdraví a determinantů stavu zdraví.

K nastavení hyperparametrů u jednotlivých metod, které tyto hyperparametry před samotným spuštěním analýz vyžadují, se využívají postupy založené na rešerši odborné literatury. Pro detaily viz. kapitoly 2 a 5.

6.1 Použité ukazatele stavu zdraví pro země EU-27

V tabulce 3 jsou uvedeny ukazatele (proměnné) použité pro splnění části stanovených cílů. Dvacet devět proměnných týkajících se stavu zdraví zahrnuje jak ukazatele úmrtnosti, tak ukazatele nemocnosti především týkajících se nejzávažnějších onemocnění ohrožujících evropskou populaci pro 27 evropských států. U některých úmrtností použitých pro nejzávažnější onemocnění (viz. kapitola 1) je v závorkách uvedena Mezinárodní klasifikace nemocí 10 (MKN-10). Pro detaily viz. podkapitoly 1.3 a 4.2.1 Použité ukazatele stavu zdraví.

Vzhledem k potřebě přehledných vizualizací je vhodné výsledky analýz vizualizovat sofistikovanějším způsobem, např. pomocí geografických dat. Geografická data týkající 27 evropských států jsou získána ze stránek Eurostatu (Eurostat, 2022a). Vzhledem k využití geografických dat a propojení těchto dat s atributy týkajícími se ukazatelů stavu zdraví jsou státy kódovány pomocí NUTS_ID získaného z těchto geografických dat (viz. balík *eurostat,* funkce *get_eurostat_geospatial*).

Tabulka 3: Vybrané proměnné stavu zdraví

Kódy	Názvy proměnných
HLY_0	zdravé roky života při narození (2019)
HLY_65	zdravé roky života ve věku 65 let (2019)
LE_0	střední délka života při narození (2019)
LE_65	střední délka života ve věku 65 let (2019)
HALE_0	očekávaná délka života ve zdraví při narození (2019)
HALE_60	očekávaná délka života ve zdraví ve věku 60 let (2019)
SDR1_0+	SDR – zhoubné novotvary všechny věkové kategorie (C00 – C97) (2019)
SDR1_65+	SDR – zhoubné novotvary 65+ (C00 – C97) (2019)
SDR2_0+	SDR – diabetes mellitus všechny věkové kategorie (2019)
SDR2_65+	SDR – diabetes mellitus 65+ (2019)
SDR3_0+	SDR – duševní poruchy a poruchy chování pro všechny věkové kategorie (F00 – F99) (2019)
SDR3_65+	SDR – duševní poruchy a poruchy chování pro 65+ (F00 – F99) (2019)
SDR4_0+	SDR – nemoci nervového systému a smyslových orgánů pro všechny věkové kategorie (G00 – H95) (2019)
SDR4_65+	SDR – nemoci nervového systému a smyslových orgánů pro 65+ (G00 – H95) (2019)
SDR5_0+	SDR – nemoci oběhového systému pro všechny věkové kategorie (I00 – I99) (2019)
SDR5_65+	SDR – nemoci oběhového systému pro 65+ (I00 – I99) (2019)
SDR6_0+	SDR – nemoci dýchacího systému pro všechny věkové kategorie (J00 – J99) (2019)
SDR6_65+	SDR – nemoci dýchacího systému pro 65+ (J00 – J99) (2019)
SDR7_P	SDR – zabranitelná úmrtnost (2019)
SDR7_T	SDR – léčitelná úmrtnost (2019)
KOJ_MOR	kojenecká úmrtnost (2019)
DALY	ztracená léta života v důsledku nemoci (2019)
PREV_CAD	prevalence kardiovaskulárních onemocnění, věkově standardizovaná (2019)
INC_CA	incidence rakovin (2019)
PREV_DIA	podíl dospělých s diabetem (2019)
PREV_PSYCH	prevalence symptomů psychické tísně (2018)
REP_AST	osoby hlásící astma (2019)
REP_CLRD	osoby hlásící chronické onemocnění dolních cest dýchacích (2019)
PER_HEALTH	osoby vnímajících svoje zdraví jako *„dobré“* nebo *„velmi dobré“* (2019)

Zdroj: Dyba a kol. (2021), Eurostat (2022a), OECD/European Union (2020), Timmis a kol. (2022), WHO (2022b), WHO (2022c) a WHO (2022d)

6.2 Snížení rozměrnosti ukazatelů stavu zdraví v EU-27

V této podkapitola jsou získaná (původní, nestandardizovaná) data nejprve vizuálně prozkoumána prostřednictvím jejich základních číselných charakteristik. Následně jsou ukazatele pomocí vztahu (8) *„z-skóre“* a vztahu (9) *„min-max“* uvedených v podkapitole 5.1.2 standardizovány. Dále jsou skrze Pearsonovy a Spearmanovy jednoduché korelační koeficienty zjišťovány závislosti mezi původními 29 ukazateli. Ukazatele stavu zdraví, které vykazují slabé

závislosti s ostatními proměnnými jsou ze dvou standardizovaných datových matic vyřazeny nejprve na základě Pearsonových korelačních koeficientů a následně dle Spearmanových korelačních koeficientů. Tímto vznikají čtyři datové matice pro dva způsoby standardizace dat. Čtyři datové matice jsou následně využity jako vstupní datové matice pro metody snížení rozměrnosti ukazatelů (PCA, SPCA a kPCA).

6.2.1 Základní charakteristiky a měření závislostí mezi proměnnými pro stav zdraví

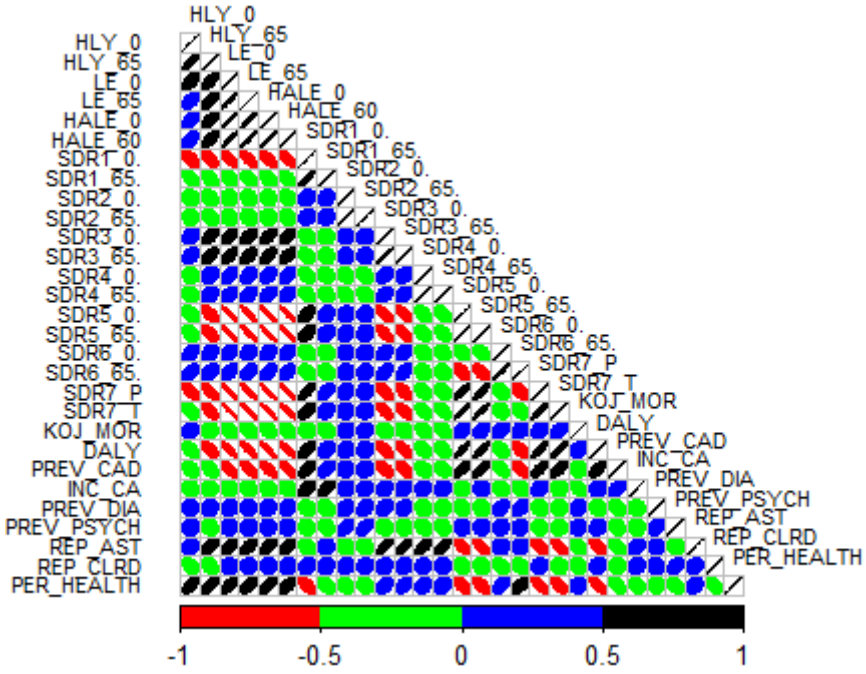
Podle základních charakteristik použitých proměnných lze odhalit odlehlé hodnoty ukazatelů, jejich maximální a minimální hodnoty, horní, dolní kvartil a medián. Největší rozptyl mají proměnné *SDR5_65* (nemoci oběhového systému 65+) a *PREV_CAD* (prevalence kardiovaskulárních onemocnění) oproti ostatním ukazatelům stavu zdraví. Tato skutečnost by vedla k potlačení vlivu ostatních proměnných v použitých analýzách. Existuje celá řada možností standardizace dat. Jednou z nich je standardizace pomocí „z-skóre“. To znamená, že standardizovaná data mají nulovou střední hodnotu a jednotkový rozptyl. Další zde použitou standardizací dat je „min-max“, kterou je vhodné použít v případě, kdy datový soubor obsahuje odlehlá pozorování. Následující analýzy pro snížení rozměrnosti ukazatelů stavu zdraví jsou aplikovány na standardizovaná data těmito dvěma způsoby.

Pozitivní nebo negativní korelaci mezi dvěma proměnnými lze vyjádřit jednoduchými korelačními koeficienty. Vzhledem k tomu, že všechny použité ukazatele stavu zdraví jsou vyjádřeny ve formě metrických (kvantitativních proměnných), je pro posouzení jejich závislosti primárně využit Pearsonův korelační koeficient (viz. podkapitola 5.1.1). Tento korelační koeficient ovšem není schopen postihnout případné nelineární vazby mezi proměnnými, z tohoto důvodu jsou závislosti mezi proměnnými posouzeny také Spearmanovým korelačním koeficientem (viz. podkapitola 5.1.1). Získané korelační matice jsou zde uvedené pouze dvě (obrázek 6 a 7), protože použité standardizace dat nemají vliv na změny v korelačních koeficientech. Pro přehlednost je použita dolní trojúhelníková korelační matice, kde silné pozitivní závislosti samozřejmě vykazují jednotlivé délky života uvedené pro všechny věkové kategorie a věkovou kategorii 65+ a jednotlivé úmrtnosti na nejzávažnější onemocnění opět uvedené pro všechny věkové kategorie a věkovou kategorii 65+. Další očekávané pozitivní závislosti mezi ukazateli stavu zdraví jsou zjištěny u některých incidencí a prevalencí závažných onemocnění a jejich úmrtností. Silné negativní závislosti jsou po dvojicích následně pozorovány u zabranitelné úmrtnosti a všemi délkami života a u léčitelné úmrtnosti a téměř všemi délkami života. Ukazatel ztracených let života v důsledku nemoci zahrnující nemocnost i úmrtnost vykazuje např. silné negativní korelace s ukazateli délky života a silné pozitivní

korelace s ukazateli zabranitelné a léčitelné úmrtnosti. Významnost korelačních koeficientů není v této práci testována, jelikož státy nepochází z náhodného výběru a výsledky nejsou zobecňovány na základní soubor.

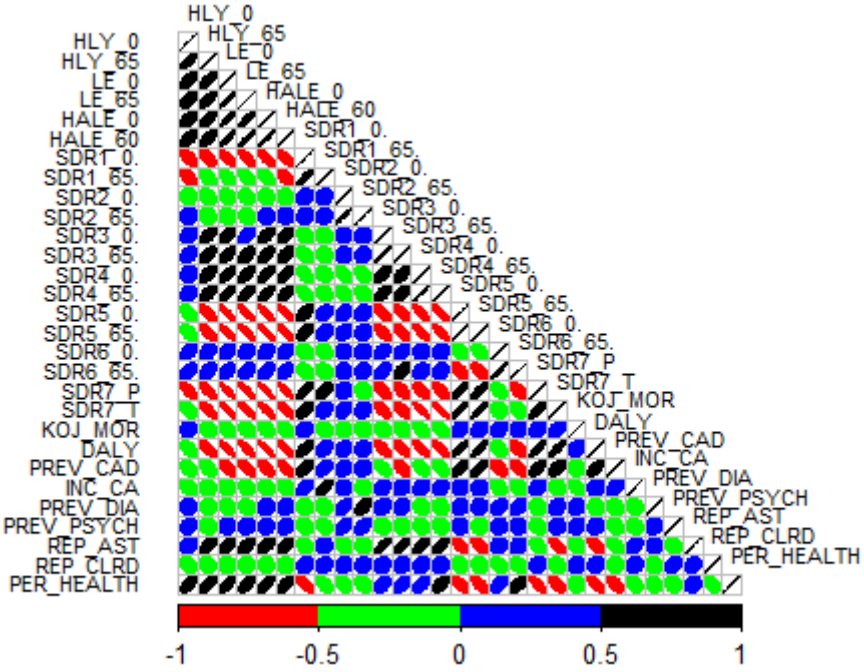
Vzhledem k tomu, že metody pro snížení rozměrnosti ukazatelů je vhodné využívat pro soubory vzájemně závislých proměnných, jsou proměnné, které nevykazují podstatnou závislost s žádnou jinou proměnnou z datových souborů vyřazeny. Pro posouzení těchto závislostí slouží vypočtené Pearsonovy (Spearmanovy) korelační koeficienty, kdy vyřazeny jsou ty proměnné, pro které žádná hodnota Pearsonova (Spearmanova) korelačního koeficientu nenaznačuje podstatnou závislost, tzn. není v absolutní hodnotě vyšší než 0,5 (viz. de Vaus, 2002). Podle Pearsonových korelačních koeficientů je z prvních dvou datových souborů standardizovaných pomocí „z-skóre“ a „min-max“ vyřazeno sedm následujících proměnných: *SDR2_0, SDR2_65, SDR6_0, KOJ_MOR, PREV_DIA, PREV_PSYCH* a *REP_CLRD*. Tudíž výsledný datový soubor obsahuje celkem 22 proměnných. Následně podle Spearmanových korelačních koeficientů jsou vytvořeny další dva standardizované datové soubory nyní pouze bez čtyřech původních proměnných: *SDR2_0, KOJ_MOR, PREV_PSYCH a REP_CLRD*. Nyní výsledný datový soubor obsahuje celkem 25 proměnných.

Obrázek 6: Dolní trojúhelníková matice Pearsonových korelačních koeficientů, stav zdraví



Zdroj: vlastní zpracování v programu R

Obrázek 7: Dolní trojúhelníková matice Spearmanových korelačních koeficientů, stav zdraví



Zdroj: vlastní zpracování v programu R

Jednoduché korelační koeficienty však jednoznačně nevypovídají o vhodnosti celé skupiny proměnných pro lineární metody snížení rozměrnosti ukazatelů. Vzájemnou závislost skupiny proměnných je proto vhodné hodnotit také pomocí KMO indexu (Kaiser–Meyer–Olkin index) (vztah (4) v podkapitole 5.1.1), který pro datové soubory obsahující 22 proměnných nabývá hodnoty 0,63 a datové soubory čítající 25 proměnných nabývá hodnoty 0,36. Na základě tabulky 1 v podkapitole 5.1.1 je adekvátnost dat pro lineární techniky snížení rozměrnosti ukazatelů, např. pro PCA, pro 22 použitých proměnných průměrná. Pokud je brán ohled při výběru proměnných pouze na Spearmanovy korelační koeficienty, stává se adekvátnost dat pro PCA nedostatečnou. Dále jsou v tabulkách 4 a 5 uvedeny dílčí míry adekvátnosti (MSA – Measure of Sampling Adequancy) pro jednotlivé proměnné nejprve pro datové matice s 22 proměnnými a následně pro datové matice s 25 proměnnými. Z porovnání tabulek 4 a 5 je zřejmé, že pro proměnné, které jsou vyřazeny na základě Pearsonova korelačního koeficientu, ale nejsou vyřazeny dle Spearmanova korelačního koeficientu (*SDR2_65*, *SDR6_0* a *PREV_DIA*), nabývá statistika MSA nízkých hodnot.

Tabulka 4: Hodnoty měr MSA pro 22 proměnných, stav zdraví

Index	***HLY_0***	***HLY_65***	***LE_0***	***LE_65***	***HALE_0***	***HALE_65***	***SDR1_0***
MSA	0,63	0,65	0,69	0,64	0,68	0,69	0,49
Index	***SDR1_65***	***SDR3_1***	***SDR3_65***	***SDR4_0***	***SDR4_65***	***SDR5_0***	***SDR5_65***
MSA	0,31	0,51	0,51	0,36	0,38	0,84	0,83
Index	***SDR6_65***	***SDR7_P***	***SDR7_T***	***DALY***	***PR_CAD***	***INC_CA***	***REP_AS***
MSA	0,63	0,78	0,86	0,70	0,58	0,48	0,52
Index	***PER_HE***						
MSA	0,64						

Zdroj: vlastní zpracování v programu R

Tabulka 5: Hodnoty měr MSA pro 25 proměnných, stav zdraví

Index	***HLY_0***	***HLY_65***	***LE_0***	***LE_65***	***HALE_0***	***HALE_65***	***SDR1_0***
MSA	0,24	0,35	0,50	0,48	0,46	0,42	0,39
Index	***SDR1_65***	***SDR2_65***	***SDR3_1***	***SDR3_65***	***SDR4_0***	***SDR4_65***	***SDR5_0***
MSA	0,26	0,03	0,30	0,31	0,24	0,25	0,59
Index	***SDR5_65***	***SDR6_0***	***SDR6_65***	***SDR7_P***	***SDR7_T***	***DALY***	***PR_CAD***
MSA	0,76	0,17	0,23	0,47	0,67	0,43	0,55
Index	***INC_CA***	***PR_DIA***	***REP_AS***	***PER_HE***			
MSA	0,11	0,05	0,36	0,27			

Zdroj: vlastní zpracování v programu R

6.2.2 Výsledky PCA a jejich interpretace pro stav zdraví

Analýza hlavních komponent (PCA) slouží pro snížení rozměrnosti ukazatelů, při kterém je hledán kompromis mezi zavedením co nejméně nových latentních proměnných a mezi co nejmenší ztrátou informace z původního datového souboru (vysvětleného rozptylu). V případě nedostatečných výsledků adekvátnosti dat pro PCA je možné skrze míry MSA odstranit z datového souboru proměnné, které nedostatečnou adekvátnost způsobují (viz. tabulky 4 a 5 v podkapitole 6.2.1). Na základě vlastních čísel z korelační matice je výběr následně extrahovaných PCs (hlavních komponent) omezen na čtyři pro standardizované datové matice pomocí „z-skóre“ s 22 proměnnými a 25 proměnnými. Pravidla ke stanovení počtu PCs jsou podrobněji uvedena v podkapitole 5.1.2 (viz. Hebák a kol., 2015; Stankovičová, Vojtková, 2007). Obrázky 21–22 v příloze 2 obsahují sutinové grafy (Scree plots) znázorňující body zlomu vždy u čtvrté dimenze, od které se pokles vlastních čísel výrazně snižuje.

Vzhledem k tomu, že k prvním zlomovým bodům dochází hned u druhé PC, je výběr čtyř komponent podpořen Kaiserovým pravidlem, tzn. čtyři vlastní čísla jsou větší než 1 a vysvětlenými rozptyly původních dat, které v obou případech přesahují 80 %. V případě 22 proměnných *PC1* vysvětluje 58,48 % variability, *PC2* 13,08 % variability, *PC3* 9,27 % variability a poslední *PC4* 5,74 % variability, tzn. celkový rozptyl vysvětlený všemi čtyřmi PCs činí 86,57 % variability původních dat. U 25 proměnných jsou výsledky PCA z hlediska

celkového vysvětleného rozptylu dat horší. *PC1* vysvětluje 52,22 % variability, *PC2* 12,68 %, *PC3* 10,07 % a *PC4* 5,83 % variability. Celkový rozptyl vysvětlený prvními čtyřmi PCs činí 80,80 % variability.

PCA aplikovaná na standardizované datové matice získané pomocí „min-max“ obsahující 22 a 25 proměnných přinesla v obou případech lepší výsledky z hlediska celkového vysvětleného rozptylu v původních datech než v předchozím případě. Na základě vlastních čísel z korelační matice je výběr omezen na tři PCs jak pro datovou matici s 22, tak pro datovou matici s 25 proměnnými. V případě 22 proměnných *PC1* vysvětluje 75,69 % variability, *PC2* 15,57 % variability a *PC3* 2,69 % variability, tzn. celkový rozptyl vysvětlený prvními třemi PCs činí 93,95 % variability původních dat. U 25 proměnných jsou výsledky PCA z hlediska celkového vysvětleného rozptylu dat opět horší. *PC1* vysvětluje 75,64 % variability, *PC2* 14,05 % a *PC3* 2,91 % variability. Celkový rozptyl vysvětlený prvními třemi PCs tedy činí 92,61 % variability. Opět obrázky 23–24 v příloze 2 vyobrazují zlomové body v sutinovém grafu, pomocí kterých jsou vybrány první tři PCs.

V příloze 3 v tabulce 19 jsou uvedeny komponentní zátěže po Varimax rotaci *RC1-RC4* pro data standardizovaná pomocí „z-skóre“. Komponentní zátěže bez rotace faktorů představují pouze korelace mezi původními proměnnými a extrahovanými PCs. Po rotaci těchto komponent je možné dojít k lepší interpretaci výsledků, jelikož se minimalizuje počet původních proměnných, které jsou silně korelované s více komponentami. V tuto chvíli se PCs mění na RCs (rotované komponenty).

Druhý až pátý sloupec v tabulce 19 v příloze 3 představuje rotované komponentní zátěže (*RC1 – RC4*), šestý sloupec vyjadřuje komunalitu pro jednotlivé proměnné (*h2*), v sedmém sloupci je vyobrazena jedinečnost rozptylu u jednotlivých proměnných (*u2*) a poslední sloupec představuje komplexitu (*com*). Pro detaily viz. podkapitola 5.1.2.

Na základě rotovaných komponentních zátěží (tabulka 19 v příloze 3) pro 22 proměnných je možné čtyři RCs interpretovat následovně (znaménka (+), resp. (-) značí kladné, resp. záporné hodnoty komponentních zátěží pro jednotlivé ukazatele stavu zdraví):

- *RC1* – rotovaná komponenta převážně střední délky života (+), očekávané délky života ve zdraví (+), úmrtností na nemoci oběhového systému (-), zabranitelné a léčitelné úmrtnosti (-), ztracených let života v důsledku nemoci (-) a prevalence kardiovaskulárních onemocnění (-) (vysvětluje 58,48 % celkové variability),

- *RC2* – rotovaná komponenta převážně úmrtí na zhoubné novotvary ve věku 65+ (+) a incidencí rakovin (+) (vysvětluje 13,08 % celkové variability),
- *RC3* – rotovaná komponenta převážně úmrtí na nemoci nervového systému a smyslových orgánů (+) (vysvětluje 9,27 % celkové variability),
- *RC4* – rotovaná komponenta převážně zdravých roků života (+) (vysvětluje 5,74 % celkové variability).

Nicméně z tabulky 19 v příloze 3 je zjevné, že existují proměnné, které mají významné komponentní zátěže s více než jednou RC v případě všech čtyř RCs (vyznačeno zeleně). Právě tento jev způsobuje nejednoznačnost v interpretaci RCs. Dále podle znamének komponentních zátěží lze konstatovat, že vysoké hodnoty *RC1* označují státy s vysokou střední délkou života a očekávanou délkou života ve zdraví. Na druhé straně nízké hodnoty *RC1* označují státy s vysokou mírou úmrtí na nemoci oběhového systému, vysokou mírou zabranitelné a léčitelné úmrtnosti, vysokými hodnotami ztracených let života v důsledku nemoci a vysokou prevalencí kardiovaskulárních chorob. Vysoké hodnoty *RC2* označují státy s vysokými mírami úmrtnosti na zhoubné novotvary a vysokou incidencí rakovin. Dále vysoké hodnoty *RC3* značí nízkou kvalitu stavu zdraví podle úmrtností na nemoci nervového systému a smyslových orgánů. Nakonec vysoké hodnoty *RC4* patří státům, které vykazují nejlepší situaci ve zdravých letech života v EU-27.

Komunality v pátém sloupci uvádí na kolik procent je původní proměnná vysvětlena v tomto případě prvními čtyřmi RCs. Například ukazatele *SDR5_0 a SDR5_65* jsou vysvětleny z 98 %, na druhou stranu ukazatel *PER_HEALTH* pouze z 61 %. Podle komplexity např. u ukazatelů *SDR5_0* a *SDR_65+* mají tyto proměnné vysokou komponentní zátěž pouze s jednou RC a s ostatními nízkou. Ve většině případů však komplexita výrazně přesahuje hodnotu 1, což indikuje významné zatížení s více než jednou RC.

Dále podle rotovaných komponentních zátěží (tabulka 19 v příloze 3) pro 25 proměnných lze čtyři RCs interpretovat následovně:

- *RC1* – rotovaná komponenta převážně střední délky života (+), očekávané délky života ve zdraví (+), úmrtností na duševní poruchy a poruchy chování (+), úmrtností na nemoci oběhového systému (-), zabranitelné a léčitelné úmrtnosti (-), ztracených let života v důsledku nemoci (-) a hlášení astmatu (+) (vysvětluje 52,22 % celkové variability),

- *RC2* – rotovaná komponenta převážně úmrtí na nemoci nervového systému a smyslových orgánů (-) a nemoci dýchacího systému (+) (vysvětluje 12,68 % celkové variability),
- *RC3* – rotovaná komponenta převážně úmrtí na zhoubné novotvary (+) a incidence rakovin (+) (vysvětluje 10,07 % celkové variability),
- *RC4* – rotovaná komponenta převážně úmrtí na diabetes mellitus (+) a prevalence diabetu (+) (vysvětluje 5,83 % celkové variability).

V tomto případě na základě komunality je např. proměnná *LE_0* vysvětlena prvními čtyřmi RCs z 94 %. Na druhé straně proměnná *SDR2_65* je vysvětlena pouze z 36 %. Z hlediska komplexity vykazují opět *SDR5_0* a *SDR5_65* vysokou komponentní zátěž pouze s jednou RC a s ostatními mizivou.

V příloze 3 v tabulce 20 jsou uvedeny komponentní zátěže po Varimax rotaci tří RCs pro data standardizovaná pomocí „min-max“. Na základě rotovaných komponentních zátěží pro 22 proměnných je možné tři RCs interpretovat následovně:

- *RC1* – rotovaná komponenta převážně střední délky života (+), očekávané délky života ve zdraví (+), úmrtností na duševní poruchy a poruchy chování (+), úmrtností na nemoci oběhového systému (-), zabranitelné a léčitelné úmrtnosti (-), ztracených let života v důsledku nemoci (-) a prevalence kardiovaskulárních onemocnění (-) (vysvětluje 75,69 % celkové variability),
- *RC2* – rotovaná komponenta převážně úmrtí na zhoubné novotvary ve věku 65+ (+) a incidencí rakovin (+) (vysvětluje 15,57 % celkové variability),
- *RC3* – rotovaná komponenta převážně úmrtí na nemoci nervového systému a smyslových orgánů (+) (vysvětluje 2,69 % celkové variability).

Například proměnné *HALE_0* a *SDR7_P* jsou vysvětleny prvními třemi RCs z 93 %, na druhou stranu proměnná *PER_HEALTH* pouze z 57 %. Podle ukazatele komplexity např. u proměnných *SDR1_65, SDR5_0, SDR_65+, DALY* a *INC_CA* existují vysoké komponentní zátěže pouze s jednou RC a s ostatními nízké. Ve většině případů však komplexita výrazně přesahuje hodnotu 1, což indikuje významné zatížení s více než jednou RC.

Dále podle rotovaných komponentních zátěží (tabulka 20 v příloze 3) pro 25 proměnných jsou tři RCs interpretované následovně:

- *RC1* – rotovaná komponenta převážně střední délky života (+), očekávané délky života ve zdraví (+), úmrtností na duševní poruchy a poruchy chování (+), úmrtností na nemoci oběhového systému (-), zabranitelné a léčitelné úmrtnosti (-), ztracených let života v důsledku nemoci (-) a hlášení astmatu (+) (vysvětluje 75,64 % celkové variability),
- *RC2* – rotovaná komponenta převážně úmrtí na nemoci nervového systému a smyslových orgánů (-), úmrtí na nemoci dýchacího systému (+) a podíl dospělých s diabetem (+) (vysvětluje 14,05 % celkové variability),
- *RC3* – rotovaná komponenta převážně úmrtí na zhoubné novotvary (+) a incidence rakovin (+) (vysvětluje 2,91 % celkové variability).

V tomto případě na základě komunality jsou např. proměnné *LE_0* a *HALE_0* vysvětleny prvními třemi RCs z 94 %. Na druhé straně proměnná *SDR2_65* je vysvětlena pouze z 21 %. Z hlediska komplexity vykazují *SDR1_65, SDR5_0, SDR5_65, SDR7_T* a *INC_CA* vysoké komponentní zátěže pouze s jednou RC a s ostatními mizivou.

Hodnoty extrahovaných RCs, tzv. *komponentní skóre*, jsou součástí výstupu funkce *principal* v programu R. Tyto skóre jsou vizualizovány na obrázku 8 v podkapitole 6.2.4 a na obrázcích 25–31 v příloze 5. Dále slouží jako vstupní datové matice pro shlukovou analýzu.

6.2.3 Výsledky SPCA a jejich interpretace pro stav zdraví

Jak již bylo zmíněno, SPCA slouží ke snadnější interpretaci PCA. Erichson a kol. (2018a) v rámci balíku *sparsepca* zavádějí funkci *spca* s následujícími argumenty:

- **X** – vstupní datová matice typu $n\ x\ p$,
- k – počet řídkých PCs, které mají být vypočítány,
- α – parametr způsobující řídkost PCs (čím vyšší je hodnota tohoto parametru, tím řidší jsou PCs),
- β – hodnota tohoto parametru (ridge shrinkage) se používá za účelem zlepšení podmíněnosti regresorů a má za následek posun zátěží směrem k nule.

Pro SPCA je důležité nastavení parametrů α a β, které způsobují řídkost PCs. U SPCA se jedná o kompromis mezi řídkostí a variabilitou zachycenou SPCs. Oba parametry jsou nastavovány po desetinásobcích od hodnoty 0,0001 do 1 včetně. Z hlediska řídkosti je podstatný parametr α. V případě nízkých hodnot tohoto parametru nedochází k řídkosti zátěží a tím pádem nedochází ke snadnější interpretaci PCs. Naopak, čím vyšší je hodnota parametru α, tím řidší zátěže jsou generovány a v extrémním případě nabývají u všech proměnných hodnoty nula.

Parametry ovlivňující řídkost jsou pro všechny použité datové matice nastaveny na $\alpha = 0{,}001$ a $\beta = 0{,}0001$. Takto nastavené parametry poskytují řídké komponentní zátěže (v některých případech nenulové zátěže původních proměnných existují pouze s jednou SPC) a zároveň vlastní čísla pro jednotlivé dimenze jsou porovnatelná s PCA (viz. tabulka 6 a tabulky 21 a 22 v příloze 4). Na základě výsledků PCA jsou extrahovány čtyři SPCs, pokud vstupní datové matice jsou standardizovány pomocí „z-skóre“ pro 22 a 25 proměnných. Následně tři SPCs jsou extrahovány, jestliže vstupní datové matice jsou standardizovány pomocí „min-max“ opět pro 22 a 25 proměnných. Pro všechny kombinace nastavovaných parametrů je stanoven ukazatel komplexity. Tento ukazatel přiřazuje každé původní proměnné s právě jednou nenulovou zátěží v rámci čtyřech nebo třech SPCs hodnotu jedna. Avšak pokud mají proměnné nenulové zátěže s více než jednou SPC, nabývá ukazatel komplexity hodnotu vyšší než 1. Jak již bylo uvedeno výše, SPCA může sloužit i k výběru proměnných. Například proměnná *SDR2_65* (úmrtnost na diabetes mellitus 65+) v případě standardizovaného datového souboru pomocí „min-max“ s 25 proměnnými stavu zdraví má se všemi třemi SPCs nulové komponentní zátěže (viz. tabulka 22 v příloze 4). To znamená, že je tato proměnná z SPCs vyřazena. V rámci prvních čtyř SPCs, získaných z datového souboru standardizovaného pomocí „z-skóre“, je zřejmé, že tyto komponenty jsou z hlediska interpretace (tabulka 21 v příloze 4) téměř totožné s RCs (tabulka 19 v příloze 3). Avšak v případě prvních tří SPCs a RCs, získaných z datových souborů standardizovaných pomocí „min-max“, jsou evidentní rozdíly v interpretaci jednotlivých komponent.

Tabulka 6: Vlastní čísla u RCs a SPCs pro datové soubory, stav zdraví

Datový soubor	RCs	SPCs
Standardizace „z-skóre“, 22 proměnných	12,866; 2,877; 2,038; 1,263	12,809; 2,831; 2,013; 1,232
Standardizace „z-skóre“, 25 proměnných	13,055; 3,171; 2,517; 1,457	12,993; 3,131; 2,472; 1,418
Standardizace „min-max“, 22 proměnných	4,709; 0,969; 0,167	4,695; 0,950; 0,149
Standardizace „min-max“, 25 proměnných	5,239; 0,973; 0,202	5,221; 0,952; 0,185

Zdroj: vlastní zpracování v programu R

Na základě řídkých komponentních zátěží, získaných z datového souboru standardizovaného pomocí „z-skóre“ (tabulka 21 v příloze 4) pro 22 proměnných je možné první čtyři SPCs interpretovat následovně:

- *SPC1* – řídká komponenta převážně střední délky života (-), očekávané délky života ve zdraví (-), úmrtností na nemoci oběhového systému (+), zabranitelné a léčitelné úmrtnosti (+), ztracených let života v důsledku nemoci (+) a prevalence kardiovaskulárních onemocnění (+),
- *SPC2* – řídká komponenta převážně úmrtí na zhoubné novotvary ve věku 65+ (+) a incidencí rakovin (+),
- *SPC3* – řídká komponenta převážně úmrtí na nemoci nervového systému a smyslových orgánů (+),
- *SPC4* – řídká komponenta převážně zdravích roků života (+).

Dále podle řídkých komponentních zátěží (tabulka 21 v příloze 4) pro 25 proměnných jsou čtyři SPCs interpretované následovně:

- *SPC1* – řídká komponenta převážně střední délky života (-), očekávané délky života ve zdraví (-), úmrtností na duševní poruchy a poruchy chování (-), úmrtností na nemoci oběhového systému (+), zabranitelné a léčitelné úmrtnosti (+), ztracených let života v důsledku nemoci (+) a hlášení astmatu (-),
- *SPC2* – řídká komponenta převážně úmrtí na nemoci nervového systému a smyslových orgánů (+) a nemoci dýchacího systému (-),
- *SPC3* – řídká komponenta převážně úmrtí na zhoubné novotvary (+) a incidence rakovin (+),
- *SPC4* – řídká komponenta převážně úmrtí na diabetes mellitus (-) a prevalence diabetu (-).

V příloze 4 v tabulce 22 jsou uvedeny řídké komponentní zátěže tří SPCs pro data standardizovaná pomocí „min-max“. Na základě řídkých komponentních zátěží pro 22 proměnných je možné tři SPCs interpretovat následovně:

- *SPC1* – řídká komponenta převážně zdravých roků života ve věku 65 (-), střední délky života (-), očekávané délky života ve zdraví (-), úmrtností na duševní poruchy a poruchy chování (-), úmrtností na nemoci dýchacího systému ve věku 65+ (-) a podílu osob vnímajících svoje zdraví jako dobré nebo velmi dobré (-),
- *SPC2* – řídká komponenta převážně úmrtí na zhoubné novotvary pro všechny věkové kategorie (-), úmrtností na nemoci oběhového systému (-), zabranitelné a léčitelné úmrtnosti (-), ztracených let života v důsledku nemoci (-) a prevalence kardiovaskulárních onemocnění (-),

- *SPC3* – rotovaná komponenta převážně zdravých let života při narození (+), úmrtí na zhoubné novotvary ve věku 65+ (-) a incidencí rakovin (-).

Nakonec podle řídkých komponentních zátěží (tabulka 22 v příloze 4) pro 25 proměnných jsou tři SPCs interpretované následovně:

- *SPC1* – řídká komponenta převážně zdravých roků života ve věku 65 (-), střední délky života (-), očekávané délky života ve zdraví (-), úmrtností na duševní poruchy a poruchy chování (-) a hlášení astmatu (-),
- *SPC2* – řídká komponenta převážně úmrtí na zhoubné novotvary (+) a nemoci oběhového systému (+), zabranitelné a léčitelné úmrtnosti (+), ztracených let života v důsledku nemoci (+) a prevalence kardiovaskulárních onemocnění (+),
- *SPC3* – řídká komponenta převážně zdravých roků života při narození (-), úmrtí na nemoci dýchacího systému (-), úmrtí na nemoci nervového systému a smyslových orgánů (+) a podíl dospělých s diabetem (-).

Hodnoty extrahovaných SPCs (komponentní skóre), jsou součástí výstupu funkce *spca* v programu R. Tyto skóre SPCs dále představují vstupní proměnné pro metody shlukové analýzy.

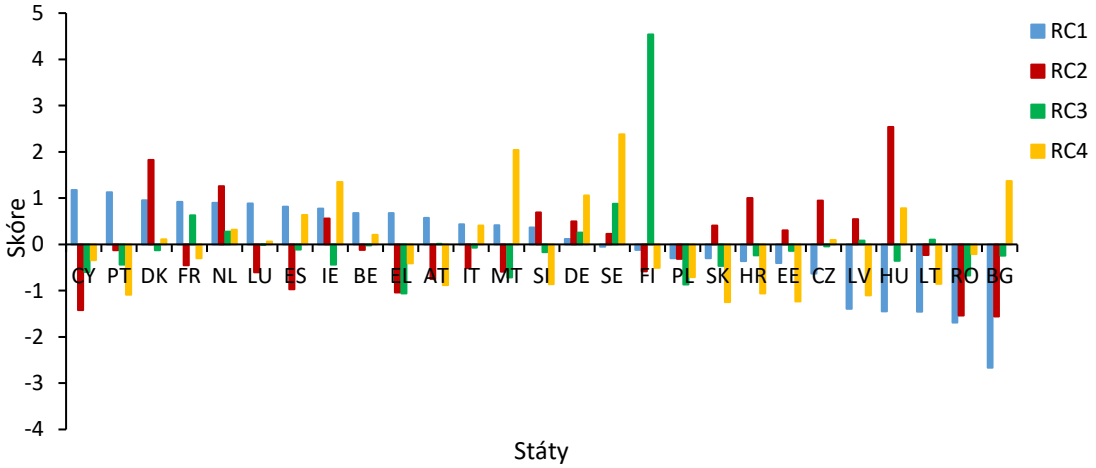
6.2.4 Vizualizace výsledků PCA a SPCA pro stav zdraví

V rámci této podkapitoly jsou vizualizovány a popsány hodnoty komponentních skóre získané pomocí PCA a SPCA. Na obrázku 8 a obrázcích 25–31 v příloze 5 jsou vizualizovány hodnoty komponentních skóre prvních čtyřech RCs a SPCs získaných z datových souborů s 22 a 25 proměnnými, standardizovaných pomocí „z-skóre“.

Na obrázku 8 jsou vizualizována rotovaná komponentní skóre získaná z datového souboru standardizovaného pomocí „z-skóre“. Pořadí států na obrázku 8 je určeno dle *RC1* od nejvyšších hodnot *RC1* k nejnižším. Hodnoty *RC1* zahrnují především ukazatele týkající se délky života (+), úmrtností na nemoci oběhového systému (-), zabranitelné a léčitelné úmrtnosti (-), ztracených let života v důsledku nemoci (-) a prevalence kardiovaskulárních onemocnění (-). Vysoké hodnoty *RC1* ukazují na dobrý stav zdraví, naopak jejich nízké hodnoty představují špatný stav zdraví na základě uvedených ukazatelů. V případě *RC2* komponenta zahrnuje především ukazatele týkající se úmrtností na zhoubné novotvary ve věku 65+ (+) a incidenci rakovin (+). Nízké hodnoty *RC2* ukazují na dobrý stav zdraví, naopak jejich vysoké hodnoty představují špatný stav zdraví na základě uvedených ukazatelů. Komponenta *RC3* představuje především úmrtí na nemoci nervového systému a smyslových orgánů (+).

Opět nízké hodnoty *RC3* ukazují na dobrý stav zdraví, naopak její vysoké hodnoty představují špatný stav zdraví. Nakonec *RC4* zahrnuje především zdravé roky života (+). Její vysoké hodnoty představují dobrý stav zdraví, a naopak nízké hodnoty špatný stav zdraví. Komponenta *RC1*, která vysvětluje nejvíce z variability původních dat, rozděluje státy EU-27 na post-socialistické státy s horším stavem zdraví podle uvedených ukazatelů a ostatní státy EU-27 s lepším stavem zdraví. Z hlediska onkologických onemocnění je v nejhorší situaci stavu zdraví Maďarsko. Vysokou úmrtnost na nemoci nervového systému a smyslových orgánů vykazuje především Finsko. Nakonec Malta a Švédsko patří mezi státy s nejlepší situací v délce života prožité ve zdraví.

Obrázek 8: Komponentní skóre RCs pro 22 proměnných, standardizace „z-skóre“, stav zdraví



Zdroj: vlastní zpracování v programu R a Excel

Na obrázku 25 v příloze 5 jsou opět prezentovány rotovaná komponentní skóre pro 25 původních proměnných standardizovaných pomocí „z-skóre“. Pořadí států je opět určeno dle *RC1* od nejvyšších hodnot *RC1* k nejnižším. Komponenta *RC1* zahrnuje především ukazatele týkající se délky života (+), úmrtností na duševní poruchy a poruchy chování (+), úmrtností na nemoci oběhového systému (-), zabranitelné a léčitelné úmrtnosti (-), ztracených let života v důsledku nemoci (-) a hlášení astmatu (+). I přes existenci kladných komponentních zátěží v případě úmrtností na duševní poruchy a poruchy chování a hlášení astmatu je možné vysoké hodnoty *RC1* interpretovat jako dobrý stav zdraví, jelikož zátěže patřící těmto dvěma ukazatelům jsou v porovnání se zátěžemi náležejícím délkám života nižší. Na druhou stranu její nízké hodnoty představují špatný stav zdraví na základě uvedených ukazatelů. V případě *RC1* je opět možné pozorovat dvě skupiny států s podobnou úrovní stavu zdraví, které si jsou podobné z hlediska obsahu jako na obrázku 8. V případě *RC2* komponenta zahrnuje především

ukazatele týkající se úmrtností na nemoci nervového systému a smyslových orgánů (-) a úmrtností na nemoci dýchacího systému (+). Vysoké hodnoty *RC2* ukazují na špatný stav zdraví z hlediska úmrtí na nemoci dýchacího systému, naopak její nízké hodnoty představují špatný stav zdraví na základě úmrtí na nemoci nervového systému a smyslových orgánů. *RC3* představuje především úmrtností na zhoubné novotvary ve věku 65+ (+) a incidenci rakovin (+). Nízké hodnoty *RC3* ukazují na dobrý stav zdraví, naopak její vysoké hodnoty představují špatný stav zdraví z hlediska onkologických onemocnění. Nakonec *RC4* zahrnuje především úmrtí na diabetes mellitus ve věku 65+ (+) a prevalenci diabetu (+). Její nízké hodnoty představují dobrý stav zdraví, a naopak vysoké hodnoty špatný stav zdraví. Na základě *RC2, RC3* a *RC4* není možné země EU-27 rozdělit na jednotlivé skupiny jako u *RC1*, kde lze pozorovat dvě skupiny států s horším stavem zdraví (především post-socialistické země) a lepším stavem zdraví (ostatní země EU-27).

Na obrázku 26 v příloze 5 jsou vizualizovány řídké komponentní skóre získané metodou SPCA z datového souboru s 22 proměnnými opět standardizovaného pomocí „z-skóre“. Pořadí států je určeno dle *SPC1* od nejnižších hodnot po nejvyšší. Řídké komponentní zátěže pro *SPC1* nabývají opačných znamének v porovnání s komponentními zátěžemi pro *RC1*, což se odráží na opačných pořadí komponentních skóre. S přihlédnutím k tomuto faktu je pořadí států z hlediska *SPC1* podobné jako pořadí států podle *RC1* na obrázku 8. V případě *SPC1* (s opačnými znaménky komponentních zátěží), *SPC2*, *SPC3* a *SPC4* lze tyto komponenty interpretovat podobně jako *RC1, RC2, RC3* a *RC4* na obrázku 8. Komponenty SPCs lze lépe interpretovat vzhledem k existenci řídkých komponentních zátěží v porovnání s RCs, jelikož jsou konstruovány na základě řídkých komponentních zátěží.

Na dalším obrázku 27 v příloze 5 jsou vizualizovány opět řídké komponentní skóre nyní získané z datového souboru s 25 proměnnými standardizovaného pomocí „z-skóre“. Pořadí států je opět určeno dle *SPC1* od nejnižších hodnot po nejvyšší. Řídké komponentní zátěže pro *SPC1, SPC2* a *SPC4* nabývají opačných znamének v porovnání s komponentními zátěžemi pro *RC1, RC2* a *RC4*, což se znovu odráží na opačných pořadí komponentních skóre. S přihlédnutím k tomuto faktu je i zde pořadí států z hlediska *SPC1* podobné pořadí států podle *RC1* na obrázku 25 v příloze 5. V případě *SPC1* (s opačnými znaménky komponentních zátěží oproti *RC1*), *SPC2* (s opačnými znaménky komponentních zátěží oproti *RC2*), *SPC3* (se stejnými znaménky komponentních zátěží jako u *RC3*) a *SPC4* (s opačnými znaménky komponentních zátěží oproti *RC4*) lze tyto komponenty interpretovat podobně jako *RC1, RC2, RC3* a *RC4*.

Dále jsou vizualizovány komponentní skóre prvních tří RCs a SPCs získaných z datových souborů s 22 a 25 proměnnými, standardizovaných pomocí „min-max“. Na obrázku 28 v příloze 5 jsou vizualizována rotovaná komponentní skóre získaná z datového souboru s 22 proměnnými. Pořadí států je určeno dle *RC1* od nejvyšších hodnot po nejnižší. *RC1* je interpretovaná především jako délka života (+), úmrtnosti na duševní poruchy a poruchy chování (+), úmrtnosti na nemoci oběhového systému (-), zabranitelné a vyhnutelné úmrtnosti (-), ztracené roky života v důsledku nemoci (-) a prevalence kardiovaskulárních onemocnění (-). Vysoké hodnoty *RC1* ukazují na státy s lepším stavem zdraví podle sledovaných ukazatelů. V případě *RC2* je možné tuto komponentu interpretovat jako úmrtí na zhoubné novotvary (+) a incidenci rakovin (+). Vysoké hodnoty *RC2* ukazují na státy s horším stavem zdraví převážně podle onkologických onemocnění. Poslední *RC3* poukazuje především na státy z hlediska úmrtností na nemoci nervového systému a smyslových orgánů (+). Státy, které jsou ohodnoceny vysokými hodnotami *RC3* vykazují horší zdravotní stav z hlediska nemocí nervového systému a smyslových orgánů. *RC1*, kterou je možné nazvat obecnou komponentou stavu zdraví, a která vysvětluje nejvíce z variability původních dat, rozděluje země EU-27 opět na státy především východní Evropy s nejhorším stavem zdraví, dále státy střední Evropy s průměrným stavem zdraví a na největší skupinu států, která se vyznačuje dobrým stavem zdraví. Z hlediska onemocnění na zhoubné novotvary patří k problémovým státům v rámci EU-27 Maďarsko a Dánsko. Onemocnění nervového systému a smyslových orgánů postihuje populaci opět především ve Finsku.

Obrázek 29 v příloze 5 vizualizuje opět rotovaná komponentní skóre nyní získaná z datového souboru s 25 proměnnými. Pořadí států je opět určeno dle *RC1* od nejvyšších hodnot po nejnižší. *RC1* je interpretovaná především jako délka života (+), úmrtnosti na duševní poruchy a poruchy chování (+), úmrtnosti na nemoci oběhového systému (-), zabranitelné a vyhnutelné úmrtnosti (-), ztracené roky života v důsledku nemoci (-) a hlášení astmatu (+). Vysoké hodnoty *RC1* ukazují na státy spíše s lepším stavem zdraví podle sledovaných ukazatelů. V případě *RC2* je možné tuto komponentu interpretovat jako úmrtí na nemoci nervového systému a smyslových orgánů (-), úmrtí na nemoci dýchacího systému (+) a podíl dospělých s diabetem (+). Nízké hodnoty *RC2* ukazují na státy s horším stavem zdraví v případě úmrtí na nemoci nervového systému a smyslových orgánů, vysoké hodnoty *RC2* patří naopak státům, jejichž populace trpí především nemocemi dýchacího systému. Poslední *RC3* poukazuje především na státy z hlediska úmrtností na zhoubné novotvary (+) a incidence rakovin (+). Státy, které jsou ohodnoceny vysokými hodnotami *RC3* vykazují horší zdravotní

stav z hlediska onkologických onemocnění. *RC1*, kterou je možné nazvat obecnou komponentou stavu zdraví, a která vysvětluje nejvíce z variability původních dat, rozděluje státy EU-27 opět na státy především východní Evropy s nejhorším stavem zdraví, dále státy střední Evropy s průměrným stavem zdraví a na největší skupinu států, která se vyznačuje dobrým stavem zdraví. Z hlediska onemocnění na zhoubné novotvary patří k problémovým státům v rámci EU-27 opět Maďarsko a Dánsko. Onemocnění nervového systému a smyslových orgánů postihuje populaci opět především ve Finsku.

Další obrázek 30 v příloze 5 ukazuje řídká komponentní skóre získaná z datového souboru s 22 proměnnými. Pořadí států je určeno dle *SPC1* od nejnižších hodnot po nejvyšší. *SPC1* je komponentou délky života (-), úmrtností na duševní poruchy a poruchy chování (-), úmrtností na nemoci dýchacího systému (-) a podílu osob vnímající svoje zdraví jako dobré nebo velmi dobré (-). Nízké hodnoty *SPC1* ukazují na státy spíše s lepším stavem zdraví z hlediska délky života, ale s horším stavem zdraví z hlediska duševních poruch a nemocí dýchacího ústrojí. V případě *SPC2* je tato komponenta interpretována jako úmrtí na nemoci oběhového systému (-), zabranitelná a léčitelná úmrtnost (-), ztracené roky života v důsledku nemoci (-) a prevalence kardiovaskulárních onemocnění (-). Nízké hodnoty *SPC2* ukazují na státy s horším stavem zdraví, naopak vysoké hodnoty *SPC2* patří státům s lepším stavem zdraví v této oblasti. Poslední *SPC3* poukazuje především na státy z hlediska zdravých roků života (+), úmrtností na zhoubné novotvary (-) a incidence rakovin (-). Státy, které jsou ohodnoceny vysokými hodnotami *SPC3* vykazují lepší zdravotní stav z hlediska onkologických onemocnění. Z obrázku 30 v příloze 5 je možné vidět, že státy především východní Evropy, které jsou nejvíce postiženy nejzávažnějšími onemocněními, tzn. kardiovaskulárními onemocněními (viz. *SPC2*), jsou v lepší situaci stavu zdraví z hlediska méně se vyskytujících onemocnění, tzn. např. nemocí dýchacího ústrojí (viz. *SPC1*).

Nakonec obrázek 31 v příloze 5 představuje řídká komponentní skóre získaná z datového souboru s 25 proměnnými. Pořadí států je opět určeno dle *SPC1* od nejnižších hodnot po nejvyšší. *SPC1* je možné interpretovat jako délku života (-), úmrtnosti na duševní poruchy a poruchy chování (-) a hlášení astmatu (-). Nízké hodnoty *SPC1* ukazují na státy s lepším stavem zdraví z hlediska délky života, ale horším stavem zdraví z hlediska duševních poruch a nemocí dýchacího ústrojí. V případě *SPC2* představuje tato komponenta úmrtí na zhoubné novotvary (+), úmrtí na nemoci oběhového systému (+), zabranitelnou a léčitelnou úmrtnost (+), ztracené roky života v důsledku nemoci (+) a prevalenci kardiovaskulárních onemocnění (+). Nízké hodnoty *SPC2* ukazují na státy s lepším stavem zdraví, naopak vysoké hodnoty *SPC2*

patří státům s horším stavem zdraví. Nakonec *SPC3* poukazuje především na státy z hlediska zdravých roků života (-), úmrtností na nemoci dýchacího systému (-) a podílu osob s diabetem (-). Státy, které jsou ohodnoceny vysokými hodnotami *SPC3* vykazují lepší zdravotní stav. Z obrázku 31 v příloze 5 je opět možné vidět, že státy především východní Evropy, které jsou nejvíce postiženy kardiovaskulárními onemocněními, jsou v lepší situaci stavu zdraví z hlediska méně se vyskytujících onemocnění (viz. opět *SPC1* a *SPC2*).

Hlavní výhodou SPCA je, že na rozdíl od PCA vytváří lépe interpretovatelné komponenty, kde se nepromítají, popř. promítají méně vlivy proměnných se slabými zátěžemi, což způsobuje rozdíly v komponentních skóre mezi RCs a SPCs. Obě tyto metody poskytují komponenty, které jsou lineárními kombinacemi původních dat. Předpokládá se tedy, že existují lineární závislosti mezi původními proměnnými, které mohou být zjištěny pomocí Pearsonových korelačních koeficientů. Výběr proměnných pro tyto analýzy na základě Pearsonových korelačních koeficientů podporuje tzv. KMO index, který ukazuje na průměrnou adekvátnost dat pro tyto analýzy při vynechání proměnných *SDR2_0+, SDR2_65+, SDR6+_0, KOJ_MOR, PREV_DIA, PREV_PSYCH* a *REP_CLRD* (viz. tabulka 3 v podkapitole 6.1).

Vzhledem k možným existencím nelineárních závislostí ve zkoumaných datových souborech je vhodné kromě lineárních technik pro snižování rozměrnosti ukazatelů použít i techniky nelineární. Mezi nelineární formy PCA patří KPCA, která předpokládá existenci nelineárních závislostí mezi původními proměnnými.

6.2.5 Výsledky KPCA pro stav zdraví

Karatzoglou a kol. (2019) popisují balík *kernlab* pro R, ve kterém uvádějí metody strojového učení založené na jádře jak pro klasifikaci, regresi a shlukování, tak pro redukci rozměrnosti. V rámci tohoto balíku je možné použít funkci *kpca* jako nelineární formu PCA. V rámci této funkce autoři zavádějí následující argumenty:

- *x* – datová matice indexovaná řádkem nebo vzorcem popisujícím model,
- *data* – volitelná data, která obsahují proměnné v modelu, pokud je použit vzorec,
- *kernel* – jedná se o funkci jádra, tento argument slouží k výpočtu skalárního součinu mezi dvěma vektorovými argumenty,
- *kpar* – jedná se o seznam hyperparametrů (parametrů jádra), seznam obsahuje parametry pro jednotlivé jádrové funkce,
- *features* – počet kPCs, které chceme získat,
- *th* – představuje hodnotu vlastního čísla, pod kterou jsou kPCs ignorovány.

Jako vstupní datové matice pro nelineární KPCA je použito 22 proměnných vybraných na základě Pearsonových korelačních koeficientů a 25 proměnných získaných podle Spearmanových korelačních koeficientů, standardizovaných jak pomocí „z-skóre“, tak pomocí „min-max“. Do programu R jsou tedy nahrány čtyři datové matice, dvě typu 27x22 a dvě typu 27x25.

Vzhledem k možnosti existence nelineárních vztahů mezi původními proměnnými (viz. Spearmanovy korelační koeficienty), je aplikována KPCA s nejčastěji používanými jádry: RBF (Radial Basis Function), polynomiální funkcí jádra (Polynomial Function) a funkce jádra hyperbolický tangens (Hyperbolic Tangent Function). V případě RBF jádra je parametr σ nastavován z intervalu $\langle 10^{-7}\,;10\rangle$ po desetinásobcích. U polynomiální funkce jádra je stupeň polynomu nastavován od 1 do 10 stupňů a v případě funkce hyperbolický tangens, který obsahuje škálový parametr (scale), je tento parametr nastavován z intervalu $\langle 10^{-7}\,;10\rangle$ opět po desetinásobcích jako u RBF. Rozmezí hodnot parametrů u jednotlivých funkcí jádra jsou nastavena na základě vysvětlených rozptylů dat prvními kPCs, kde při nižších (popř. vyšších) hodnotách parametrů nedochází k výrazným změnám v těchto rozptylech.

V příloze 6 jsou uvedeny obrázky 32–43, které vizualizují kumulativní vysvětlené rozptyly pro různé hodnoty parametrů a k nim odpovídající vlastní čísla získaná z jádrových matic NxN (vztah (24) viz. podkapitola 5.1.4). Ke vstupní datové matici s 22 proměnnými standardizované pomocí „z-skóre“, ze které jsou extrahovány čtyři kPCs, patří obrázky 32, 33 a 34 v příloze 5 pro parametry jader v tomto pořadí: RBF, polynomiální funkce a funkce hyperbolický tangens.

V případě RBF při nastavení parametru $\sigma = 0{,}0001$ první čtyři kPCs dohromady vysvětlují 86 % variability původních dat a zároveň u hodnoty tohoto parametru dochází k ustálení variability. Toto jsou důvody pro výběr parametru $\sigma = 0{,}0001$. Na základě hodnot vlastních čísel pro jednotlivé dimenze je možné si všimnout, že *kPC1* vysvětluje nejvíce variability z původních dat (58 %), a že ke zlomovému bodu opět dochází u čtvrté kPC stejně jako v případě PCA.

U polynomiální funkce jádra při výběru druhého stupně polynomu dochází ke snížení variability původních dat pro čtyři kPCs. Pro třetí stupeň polynomiální funkce je zaznamenán výrazný nárůst a dále ustálení této variability (na 85 %). Nicméně při nastavení vyšších stupňů polynomiální funkce není již KPCA schopná nacházet nelineární vzory.

U poslední funkce jádra hyperbolický tangens je škálový parametr nastaven na hodnotu 1 na základě obrázku 34 v příloze 6. Důvodem pro nastavení tohoto parametru na hodnotu 1 je

nalezení zlomového bodu u vysvětlené kumulativní variability původních dat pro čtyři kPCs, od jejíž hodnoty dále dochází k poklesu přírůstku této variability. V tomto případě první čtyři kPCs vysvětlují dohromady 93 % variability původních dat, což je lepší výsledek oproti vysvětlené variabilitě původních dat u PCA. Na základě hodnot vlastních čísel pro jednotlivé dimenze je možné si všimnout, že *kPC1* nyní opět vysvětluje nejvíce variability z původních dat (65 %), což je opět lepší výsledek než v případě PCA (58 %).

Tabulka 7 prezentuje vysvětlený kumulativní rozptyl a vysvětlený rozptyl *kPC1* při nastavení parametrů jednotlivých jader pro čtyři použité datové soubory. U všech čtyřech datových souborů nabývá vysvětlený kumulativní rozptyl nejvyšších hodnot v případě jádrové funkce hyperbolický tangens (při uvedeném nastavení parametrů) a překonává výsledky získané prostřednictvím PCA. Na druhou stranu při použití polynomiální funkce jádra dochází ke snížení hodnot kumulativního rozptylu oproti PCA. Použití jádra RBF vede ke získání téměř totožných výsledků s PCA, ale pouze v případě, kdy jsou datové matice standardizovány pomocí „z-skóre“. Naopak u „min-max“ dochází ke zhoršení výsledků vysvětleného kumulativního rozptylu. Výsledky vysvětlených rozptylů původních dat prostřednictvím PCA jsou uvedeny v podkapitole 6.2.2.

Tabulka 7: Parametry jader, vysvětlený kumulativní rozptyl, vysvětlený rozptyl kPC1 pro jednotlivé datové soubory, stav zdraví

Datové soubory	**RBF**	**Polynomiální**	**Hyperbolický tangens**
22 proměnných, standardizace „z-skóre“, 4 kPCs	sigma = 0,0001 kum. roz. = 0,86 kPC1 var. = 0,58	stupeň = 3 offset = 1 kum. roz. = 0,85 kPC1 var. = 0,36	scale = 1 offset = 0 kum. roz. = 0,93 kPC1 var. = 0,65
25 proměnných, standardizace „z-skóre“, 4 kPCs	sigma = 0,0001 kum. roz. = 0,81 kPC1 var. = 0,52	stupeň = 3 offset = 1 kum. roz. = 0,77 kPC1 var. = 0,30	scale = 0,1 offset = 0 kum. roz. = 0,87 kPC1 var. = 0,61
22 proměnných, standardizace „min-max“, 3 kPCs	sigma = 0,01 kum. roz. = 0,81 kPC1 var. = 0,63	stupeň = 1 offset = 1 kum. roz. = 0,82 kPC1 var. = 0,64	scale = 1 offset = 0 kum. roz. = 0,95 kPC1 var. = 0,73
25 proměnných, standardizace „min-max“, 3 kPCs	sigma = 0,01 kum. roz. = 0,77 kPC1 var. = 0,57	stupeň = 1 offset = 1 kum. roz. = 0,79 kPC1 var. = 0,58	scale = 1 offset = 0 kum. roz. = 0,95 kPC1 var. = 0,70

Zdroj: vlastní zpracování v programu R

Z tabulky 7 je zřejmé, že získané výsledky kumulativních vysvětlených rozptylů pomocí nelineární KPCA jsou téměř vždy lepší v případě datových souborů s 22 proměnnými

vybranými na základě Pearsonových korelačních koeficientů oproti kumulativním vysvětleným rozptylům získaným pomocí PCA (viz. podkapitola 6.2.2). Komponentní skóre jsou součástí výstupu funkce *kpca* v programu R. Tyto komponentní skóre (kPCs) získané pomocí jádrové funkce hyperbolický tangens, která vykazuje nejvyšší kumulativní vysvětlený rozptyl, představují vstupní proměnné pro shlukovou analýzu.

6.3 Rozdělení států EU-27 podle ukazatelů stavu zdraví

Dalším důležitým krokem zkoumání stavu zdraví populace v zemích EU-27 pro politiku EU zaměřenou na snížení nerovností ve stavu zdraví je stanovení skupin zemí s podobnou situací stavu zdraví. V podkapitole 6.2.4 bylo možné nejen některé skupiny států na základě jednotlivých komponentních skóre odhalit, ale také bylo možné pomocí nich detekovat odlehlé státy EU-27. Právě rozdělení zemí podle stavu zdraví, které se opírá o kvalitní data, může být užitečné např. pro rozhodování na úrovni EU. V případě existence ukazatelů stavu zdraví na regionální úrovni, mohou být tyto postupy využity např. pro rozhodování orgánů veřejné správy nebo veřejného zdravotnictví.

Pro zde aplikované metody shlukové analýzy jsou vybírány vstupní datové soubory z následujících v tomto pořadí:

- 1A – 22 standardizovaných ukazatelů stavu zdraví pomocí „z-skóre“,
- 1B – 25 standardizovaných ukazatelů stavu zdraví pomocí „z-skóre“,
- 1C – 22 standardizovaných ukazatelů stavu zdraví pomocí „min-max“,
- 1D – 25 standardizovaných ukazatelů stavu zdraví pomocí „min-max“,
- 2A – čtyři RCs z 22 ukazatelů stavu zdraví standardizovaných pomocí „z-skóre“,
- 2B – čtyři RCs z 25 ukazatelů stavu zdraví standardizovaných pomocí „z-skóre“,
- 2C – tři RCs z 22 ukazatelů stavu zdraví standardizovaných pomocí „min-max“,
- 2D – tři RCs z 25 ukazatelů stavu zdraví standardizovaných pomocí „min-max“,
- 3A – čtyři SPCs z 22 ukazatelů stavu zdraví standardizovaných pomocí „z-skóre“,
- 3B – čtyři SPCs z 25 ukazatelů stavu zdraví standardizovaných pomocí „z-skóre“,
- 3C – tři SPCs z 22 ukazatelů stavu zdraví standardizovaných pomocí „min-max“,
- 3D – tři SPCs z 25 ukazatelů stavu zdraví standardizovaných pomocí „min-max“,
- 4A – čtyři kPCs z 22 ukazatelů stavu zdraví standardizovaných pomocí „z-skóre“, (jádrová funkce hyperbolický tangens),
- 4B – čtyři kPCs z 25 ukazatelů stavu zdraví standardizovaných pomocí „z-skóre“, (jádrová funkce hyperbolický tangens),

- 4C – tři kPCs z 22 ukazatelů stavu zdraví standardizovaných pomocí „min-max“, (jádrová funkce hyperbolický tangens),
- 4D – tři kPCs z 25 ukazatelů stavu zdraví standardizovaných pomocí „min-max“, (jádrová funkce hyperbolický tangens).

6.3.1 Výsledky hierarchické aglomerativní shlukové analýzy pro stav zdraví

Na 16 datových souborů popsaných v předchozí podkapitole 6.3 je aplikována hierarchická aglomerativní metoda shlukové analýzy, konkrétně Wardova metoda. Tato metoda je z ostatních hierarchických metod shlukové analýzy vybrána kvůli výhodám, které z hlediska výsledných vytvořených shluků přináší. Jedná se především o tvorbu shluků pomocí vnitroshlukové homogenity a tvorbu shluků shodné velikosti. Dalším důvodem k výběru Wardovy metody je její oblíbenost. Pro detaily viz. podkapitoly 2.2.2 a 5.2.1. Pro konstrukci matice vzdáleností mezi státy je použita euklidovská vzdálenost. Před samotnou prezentací výsledků je třeba se podívat, jak dobře výsledný dendrogram modeluje skutečnost. Tabulka 8 obsahuje kofenetické korelační koeficienty vyjadřující závislost mezi maticí euklidovských vzdáleností použitých datových souborů a kofenetickou maticí.

Tabulka 8: Kofenetické korelační koeficienty, stav zdraví

Soubor	**1A**	**1B**	**1C**	**1D**	**2A**	**2B**	**2C**	**2D**
ρ	0,75	0,74	0,79	0,78	0,62	0,55	0,65	0,59
Soubor	**3A**	**3B**	**3C**	**3D**	**4A**	**4B**	**4C**	**4D**
ρ	0,76	0,76	0,85	0,84	0,93	0,87	0,85	0,91

Zdroj: vlastní zpracování v programu R

Pro prezentaci výsledných shluků získaných pomocí Wardovy metody a euklidovské vzdálenosti jsou na základě nejvyšších kofenetických koeficientů korelace vybrány čtyři datové soubory, tzn. každá metoda snižování rozměrnosti ukazatelů má svého zástupce. První datový soubor 1C představuje původních 22 proměnných standardizovaných za pomoci „min-max“, druhý a třetí datový soubor 2C a 3C jsou RCs a SPCs pocházející také z těchto původních 22 proměnných. Poslední datový soubor se týká kPCs, které jsou opět získané z 22 proměnných, nyní však standardizovaných pomocí „z-skóre“. Datové soubory obsahující 25 původních proměnných (datové soubory označené písmenem B a D, viz. podkapitola 6.3) nejsou v rámci Wardovy metody a euklidovské vzdálenosti použity.

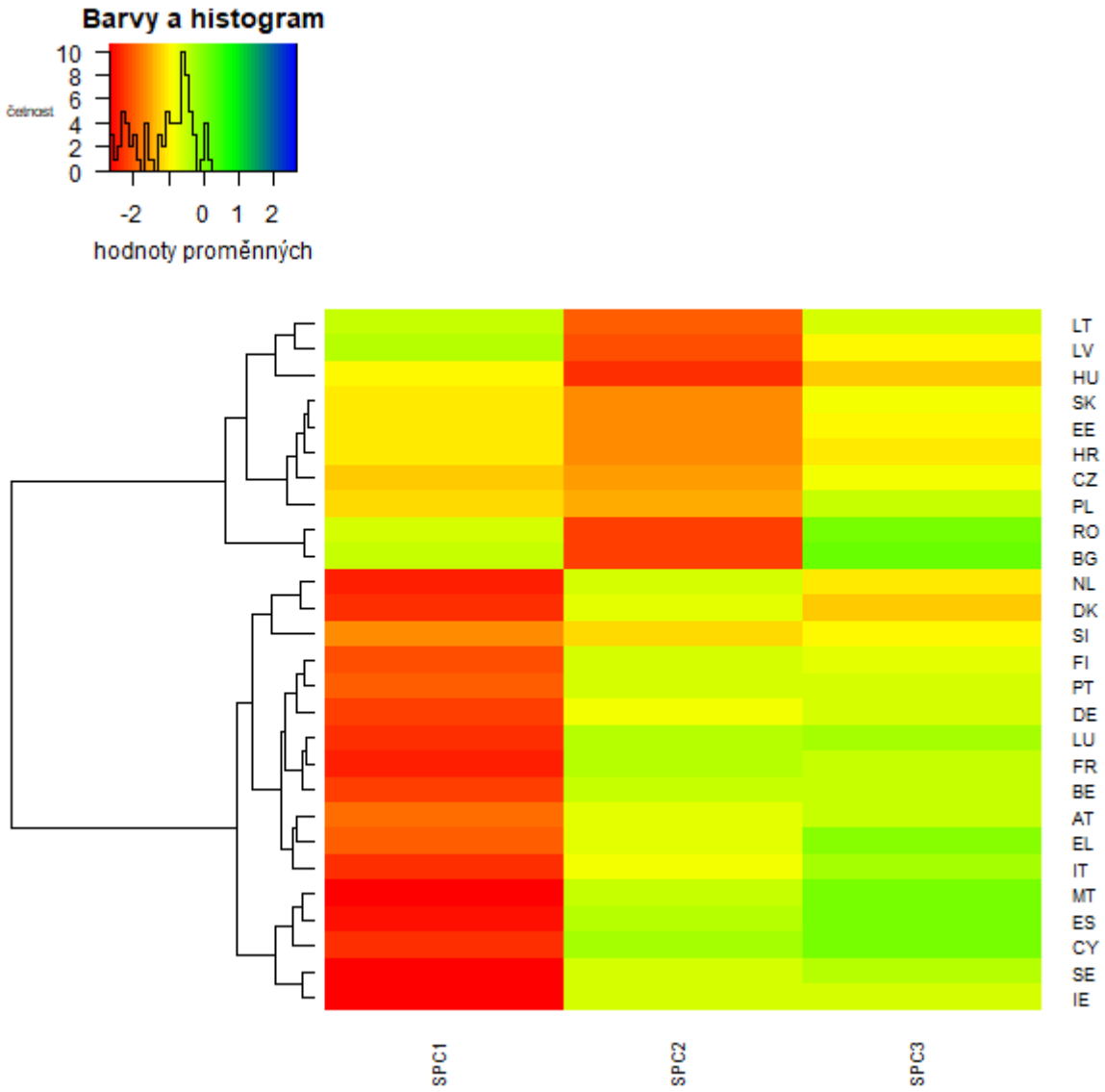
SPCA vytváří oproti PCA lépe interpretovatelné komponenty, což se pozitivně odráží i na výsledcích shlukové analýzy. Obrázek 9 prezentuje výsledný dendrogram získaný pomocí Wardovy metody a euklidovské vzdálenosti s heat mapou při použití datového souboru 3C (viz.

podkapitola 6.3). Heat mapa vizualizuje hodnoty proměnných, podle kterých ke shlukování států došlo. V legendě jsou vždy uvedeny také histogramy četností hodnot použitých proměnných vizualizovaných v heat mapě. V případě dvou shluků je výsledek shlukování srovnatelný s použitím datového souboru 1C až na případ Slovinska (viz. obrázek 44 v příloze 7). Při třech shlucích dochází k rozpadu post-socialistických států na shluk obsahující Bulharsko a Rumunsko a shluk tvořený zbylými zeměmi. Z hlediska hodnot proměnných lze rozdíly v uvedených shlucích popsat pomocí *SPC1,* tzn. zdravých roků života ve věku 65+, střední délky života, očekávané délky života ve zdraví, úmrtí na duševní poruchy a poruchy chování, úmrtí na nemoci dýchacího systému ve věku 65+ a podílu osob vnímající svoje zdraví jako dobré nebo velmi dobré (viz. podkapitola 6.2.3 a příloha 4). Je zřejmé, že horší situace stavu zdraví podle délky života a vlastního hodnocení je v post-socialistických zemích oproti zbytku Evropy. Na druhou stranu vyšší délka života v evropských zemích s sebou přináší vyšší úmrtnost na duševní poruchy a poruchy chování a úmrtnost na nemoci dýchacího systému. Rozdíly ve shlucích lze popsat dále podle *SPC2*, tzn. úmrtnosti na zhoubné novotvary pro všechny věkové kategorie, úmrtnosti na nemoci oběhového systému, zabranitelné a léčitelné úmrtnosti, ztracených let života v důsledku nemoci a prevalence kardiovaskulárních onemocnění. V tomto případě výsledky ukazují na nejhorší situaci stavu zdraví v post-socialistických zemích, především v Bulharsku, Litvě, Lotyšsku, Maďarsku a Rumunsku.

Jak již bylo uvedeno v podkapitole 2.1.4 v případě KPCA dochází při použití jádra ke ztíženému návratu k původnímu vstupnímu prostoru na rozdíl od lineárních metod pro snížení rozměrnosti ukazatelů, kde je možné se k původnímu vstupnímu prostoru vrátit prostřednictvím komponentních zátěží (viz. Shawn, 2006; Shiokawa, Kikuchi, 2018). Nelineární kPCs je obtížné interpretovat, protože neodpovídají vektorům ve vstupním prostoru. Z tohoto důvodu jsou získané výsledky shlukové analýzy vizualizované na obrázku 46 v příloze 7 porovnány s výsledky získanými pomocí vstupních datových souborů 1C a 3C, dále se zdravými roky života a zabranitelnou úmrtností patřící mezi indexy stavu zdraví, popř. jsou jejich součástí.

Obrázek 46 v příloze 7 prezentuje výsledky Wardovy metody a euklidovské vzdálenosti při použití datového souboru 4A (viz. podkapitola 6.3). Na rozdíl od ostatních použitých vstupních datových souborů jsou nyní ukazatele stavu zdraví standardizovány pomocí „z-skóre“. Ze všech použitých datových souborů nabývá kofenetický koeficient korelace pro Wardovu metodu a euklidovskou vzdálenost nejvyšší hodnoty.

Obrázek 9: Dendrogram a heat mapa (datový soubor 3C), stav zdraví



Zdroj: vlastní zpracování v programu R

Nyní na obrázku 46 v příloze 7 v případě nejvýraznějších dvou shluků je výsledek shlukování stejný s použitím datového souboru 1C (viz. obrázek 44 v příloze 7). Při třech shlucích dochází k rozpadu především post-socialistických států na shluk obsahující Slovinsko a shluk obsahující především post-socialistické země, čímž se na této úrovni shlukování liší od datového souboru 1C. Pokud jsou pro interpretaci dvou, třech, popř. čtyřech získaných shluků při použití datového souboru 4A použity zdravé roky života při narození patřící mezi indexy stavu zdraví (viz. obrázek 2 v podkapitole 1.3.2), je zjištěno, že se státy obsažené v konkrétních shlucích objevují mezi státy s vysokými hodnotami *HLY_0*, ale také s jejich nízkými, popř. průměrnými hodnotami. Další možností pro interpretaci výsledků shlukové analýzy je mimo jiné použití ukazatelů úmrtnosti, například ukazatele zabranitelné

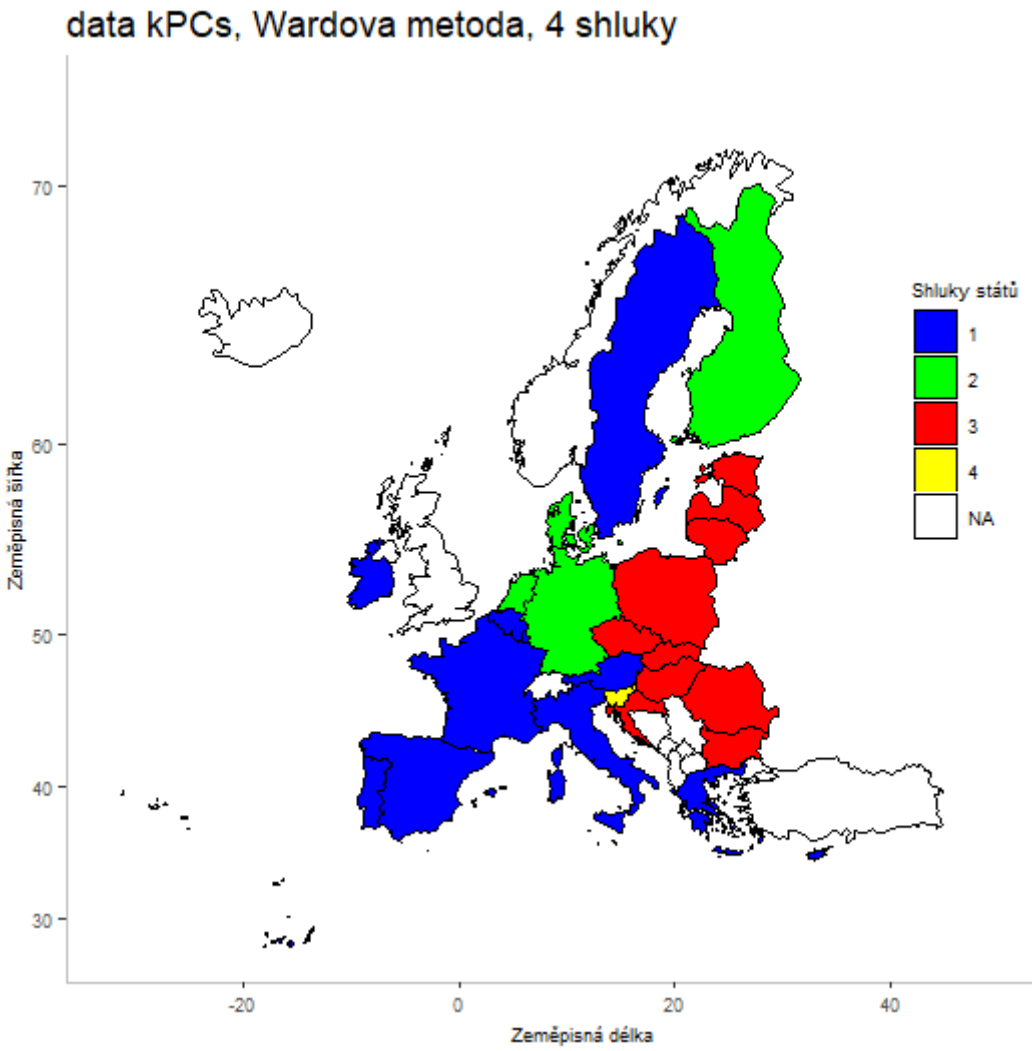
úmrtnosti uvedené v podkapitole 1.3.3 (viz. obrázek 3). Na základě tohoto ukazatele je možné získané výsledné shluky lépe interpretovat. Shluk převážně post-socialistických států lze interpretovat jako shluk států s vysokou mírou zabranitelné úmrtnosti, kde její nejnižší hodnoty nabývá Slovinsko. Na druhé straně hodnot tohoto ukazatele se nacházejí státy západní, jižní a severní Evropy, které se dále dělí na shluk obsahující Dánsko, Finsko, Německo a Nizozemsko, které nabývají z hlediska míry zabranitelné úmrtnosti průměrných hodnot tohoto ukazatele a shluk zbylých států s nejlepším stavem v rámci této úmrtnosti. Komponenta *kPC1* pomocí svých hodnot výrazně odlišuje dva shluky prezentované v uvedeném dendrogramu. Při porovnání heat map pro dva shluky na obrázcích 44 a 46 v příloze 7 je možné si všimnout podobností *kPC1* s proměnnými střední délky života, očekávané délky života ve zdraví, úmrtnosti na nemoci oběhového systému a zabranitelné a léčitelné úmrtnosti.

Již na zmíněném obrázku 44 v příloze 7 jsou uvedeny výsledky hierarchické shlukové analýzy, kde 1C (viz. podkapitola 6.3) je vstupním datovým souborem. Z dendrogramu jsou evidentní dva shluky, kde jeden obsahuje převážně post-socialistické státy a druhý zbylou část Evropy. V případě třech shluků se post-socialistické státy rozpadají na dva shluky, kde první z nich obsahuje Bulharsko, Litvu, Lotyšsko, Maďarsko a Rumunsko, druhý je tvořen CZ, Estonskem, Chorvatskem, Polskem, SK a Slovinskem. Z hlediska hodnot proměnných lze rozdíly v uvedených shlucích popsat především pomocí střední délky života, očekávané délky života ve zdraví, úmrtností na nemoci oběhového systému, zabranitelné a léčitelné úmrtnosti. Je zřejmé, že horší situace stavu zdraví podle těchto ukazatelů je vždy v post-socialistických zemích oproti zbytku Evropy. Opačná situace je detekována pouze v případě duševních poruch a poruch chování a nemocí nervového systém a smyslových orgánů.

Obrázek 45 v příloze 7 vizualizuje výsledky Wardovy metody a euklidovské vzdálenosti, kde 2C (viz. podkapitola 6.3) je vstupním datových souborem. Kofenetický koeficient korelace pro tento datový soubor je z vybraných čtyřech souborů nejnižší (viz. tabulka 8), což ukazuje na to, že zde státy v porovnání s ostatními datovými soubory tvoří spíše jeden shluk.

Vzhledem k nejlepšímu výsledku na základě kofenetického koeficientu korelace (viz. tabulka 8) a k nalezení interpretace výsledných shluků jsou pro vizualizaci pomocí geografických dat vybrány výsledky Wardovy metody a euklidovské vzdálenosti získané pomocí datového souboru 4A (viz. podkapitola 6.3). Na obrázku 10 jsou vizualizovány výsledné čtyři shluky zemí EU-27 získané z dendrogramu na obrázku 46 v příloze 7.

Obrázek 10: Vizualizace čtyř shluků států Wardovou metodou (kPCs)



Zdroj: vlastní zpracování v programu R

6.3.2 Výsledky metody k-průměrů pro stav zdraví

Pro porovnání výsledných shluků získaných pomocí Wardovy metody a euklidovské vzdálenosti je dále použita nehierarchická metoda shlukování, metoda *k*-průměrů. Vzhledem k tomu, že se jedná o iterační algoritmus, požaduje tato metoda specifikaci počáteční konfigurace. Problémem je, že funkce *kmeans* v programu R používá náhodný start, což znamená, že spuštění analýzy vícekrát za sebou způsobí pokaždé odlišné zařazení států do shluků. V tomto případě, je důležité vyzkoušet více náhodných počátečních nastavení a následně vybrat nejlepší z nich podle nejnižší hodnoty celkového vnitroshlukového součtu čtverců. Touto problematikou se zabývá např. Hand, Krzanowski (2005).

Optimální počet shluků je možné stanovit na základě podílu mezishlukového součtu čtverců s celkovým součtem čtverců. Čím vyšší je tato hodnota, tím větší jsou vzdálenosti mezi shluky.

Na obrázku 11 jsou graficky znázorněny hodnoty tohoto podílu až pro 26 shluků v pořadí datových souborů 1C, 2C, 3C a 4A. Následující tabulka 9 prezentuje tyto hodnoty pro dva až šest shluků všech šestnácti vstupních datových souborů popsaných v podkapitole 6.3.

Z tabulky 9 je zřejmé, že pro různá nastavení shluků, dosahují hodnoty podílu mezishlukového a celkového součtu čtverců vyšších hodnot u datových souborů 3C (zástupce lineární metody pro snížení rozměrnosti ukazatelů), 4A a 4C (zástupci nelineární metody pro snížení rozměrnosti ukazatelů). Vzhledem k tomu, že datové soubory 3C a 4A byly použity pro Wardovu metodu, jsou tyto dva datové soubory vstupními datovými soubory také pro metodu *k*-průměrů a jejich výsledky jsou následně porovnány. Na základě obrázku 11 je zřejmý zlomový bod při pěti shlucích u datových souborů 3C a 4A (červeně), nad kterým podíl mezishlukového a celkového součtu čtverců výrazně klesá.

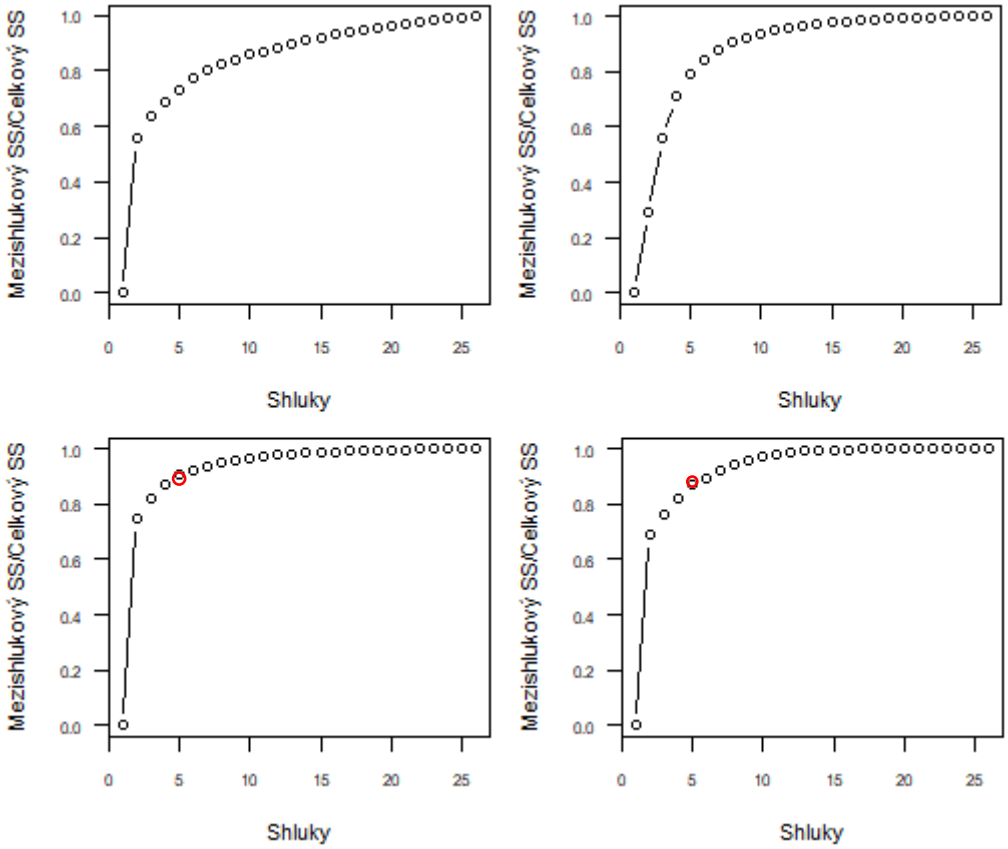
Výsledky pěti shluků získané metodou *k*-průměrů (obrázek 47 v příloze 8) jsou podobné výsledkům pěti shluků získaných pomocí Wardovy metody při použití vstupního datového souboru 3C (viz. dendrogram na obrázku 9). Tři shluky tvořené post-socialistickými zeměmi spadající nyní mezi země EU-27 tvoří v obou uvedených případech stejné shluky. K odlišnostem ve zbývajících dvou shlucích dochází v případě zemí západní a jižní Evropy, konkrétně se jedná o Belgii, Francii, Itálii, Lucembursko a Řecko.

Tabulka 9: Stanovení optimálního počtu shluků pro metodu k-průměrů, stav zdraví

Mezishlukový součet čtverců/celkový součet čtverců (v %)	**1A**	**1B**	**1C**	**1D**	**2A**	**2B**	**2C**	**2D**
2 shluky	50,9	45,4	55,9	50,6	21,8	22,0	29,1	29,1
3 shluky	59,6	54,0	63,6	58,1	42,7	38,8	55,7	48,5
4 shluky	67,6	61,5	68,6	63,8	57,2	52,3	71,0	64,7
5 shluků	72,4	66,3	73,0	67,9	70,0	64,4	79,1	73,4
6 shluků	76,5	70,1	77,1	71,6	77,8	70,9	83,9	79,4
Mezishlukový součet čtverců/celkový součet čtverců (v %)	**3A**	**3B**	**3C**	**3D**	**4A**	**4B**	**4C**	**4D**
2 shluky	58,9	56,6	74,5	71,6	68,8	66,2	43,7	40,8
3 shluky	68,7	65,6	82,1	79,4	75,9	73,2	66,7	61,3
4 shluky	76,8	73,4	87,3	84,4	82,2	78,7	78,1	74,8
5 shluků	81,8	78,0	90,3	87,2	86,8	82,3	87,2	85,4
6 shluků	85,5	81,5	92,2	89,4	89,6	85,3	90,6	89,1

Zdroj: vlastní zpracování v programu R

Obrázek 11: Grafické stanovení optimálního počtu shluků pro metodu k-průměrů a datové soubory 1C, 2C, 3C a 4A, stav zdraví



Zdroj: vlastní zpracování v programu R

Pět shluků vytvořených metodou *k*-průměrů je možné pomocí tří SPCs interpretovat následovně (viz. obrázek 9 a podkapitola 6.2.3):

- *Shluk 1:* Belgie, Irsko, Itálie, Francie, Kypr, Lucembursko, Malta, Řecko, Španělsko a Švédsko – nejlepší stav zdraví z hlediska střední délky života a očekávané délky života prožité ve zdraví, nejlepší stav zdraví z hlediska kardiovaskulárních onemocnění a lepší stav zdraví z hlediska rakovin v rámci zemí EU-27,
- *Shluk 2:* Dánsko, Finsko, Německo, Nizozemsko, Portugalsko, Rakousko a Slovinsko – lepší stav zdraví z hlediska střední délky života a očekávané délky života prožité ve zdraví, lepší stav zdraví z hlediska kardiovaskulárních onemocnění a horší stav zdraví z hlediska rakovin v rámci zemí EU-27,
- *Shluk 3*: Litva, Lotyšsko a Maďarsko – horší stav zdraví z hlediska střední délky života a očekávané délky života prožité ve zdraví, horší stav zdraví z hlediska kardiovaskulárních onemocnění, zabranitelné a léčitelné úmrtnosti a nejhorší stav zdraví z hlediska rakovin v rámci zemí EU-27,

- *Shluk 4:* Bulharsko a Rumunsko – horší stav zdraví z hlediska střední délky života a očekávané délky života prožité ve zdraví, nejhorší stav zdraví z hlediska kardiovaskulárních onemocnění, zabranitelné a léčitelné úmrtnosti a nejlepší stav zdraví z hlediska rakovin v rámci především post-socialistických zemí,
- *Shluk 5*: CZ, Estonsko, Chorvatsko, Polsko a SK – nejlepší stav zdraví z hlediska střední délky života a očekávané délky života prožité ve zdraví v rámci post-socialistických zemí, nejlepší stav zdraví z hlediska kardiovaskulárních onemocnění v post-socialistických zemích a horší stav zdraví z hlediska rakovin v rámci EU-27.

Výsledky pěti shluků získané metodou *k*-průměrů (obrázek 48 v příloze 8) jsou opět podobné výsledkům pěti shluků získaných pomocí Wardovy metody při použití vstupního datového souboru 4A (viz. obrázek 46 v příloze 7). K jedinému rozdílu v zařazení států v porovnání s Wardovou metodou dochází v případě Rakouska, které u metody *k*-průměrů tvoří shluk spolu s Řeckem. Již bylo uvedeno v podkapitole 6.3.1, že *kPC1* vykazuje podobnost s použitými ukazateli střední délky života, očekávané délky života prožité ve zdraví, úmrtnostmi na nemoci oběhového systému, zabranitelné a léčitelné úmrtnosti. Pět shluků vytvořených pomocí metody *k*-průměrů je možné pomocí tří vyjmenovaných proměnných interpretovat následovně (viz. obrázek 44 v příloze 7):

- *Shluk 1:* Dánsko, Finsko, Německo a Nizozemsko – dobrý stav zdraví podle uvedených ukazatelů, podle SPCs vykazují tyto státy horší stav zdraví z hlediska rakovin (podobnost *kPC2* na obrázku 46 v příloze 7 s *SPC3* na obrázku 9),
- *Shluk 2:* Slovinsko – nejlepší stav zdraví na základě střední délky života, očekávané délky života prožité ve zdraví, úmrtnosti na nemoci oběhového systému, zabranitelné a léčitelné úmrtnosti v rámci post-socialistických zemí,
- *Shluk 3:* především post-socialistické státy – nejhorší stav zdraví v rámci EU-27,
- *Shluk 4:* Belgie, Irsko, Itálie, Francie, Kypr, Lucembursko, Malta, Portugalsko, Řecko, Španělsko a Švédsko – nejlepší stav zdraví podle uvedených ukazatelů,
- *Shluk 5:* Rakousko a Řecko – vykazují horší stav zdraví podle uvedených ukazatelů v rámci zemí západní, jižní a severní Evropy.

6.3.3 Výsledky metody Fuzzy k-průměrů pro stav zdraví

Na základě výsledků metody *k*-průměrů bylo stanoveno pět shluků zemí EU-27 při použití vstupních datových souborů 3C a 4A (podkapitola 6.3). Nicméně existence odlehlých

pozorování (států) v těchto datových souborech (viz. hodnoty komponentních skóre v podkapitole 6.2.4) způsobuje problémy v rámci těchto metod.

Oproti předchozím metodám shlukových analýz má FCM algoritmus tu výhodu, že každý objekt je přiřazen do každého shluku s určitým stupněm příslušnosti. Použitá funkce *fcm* je obsažená v balíku *ppclust* v programu R (viz. příloha 1).

Na základě hodnot stupňů příslušnosti je možné zjistit, které země jsou s nízkou mírou příslušnosti přiřazené ke všem shlukům. Nezařazení států je určeno na základě hodnot stupňů příslušnosti, které pro všech pět shluků nabývá hodnot nižších než 0,5.

V následující tabulce 10 jsou uvedeny statistiky týkající se stanovení optimálního počtu shluků (viz. podkapitola 5.2.4), které se používají právě v případě FCM algoritmu. Na základě těchto statistik jsou porovnány výsledky pěti shluků pro datové soubory 3C a 4A.

Tabulka 10: Statistiky pro pět shluků u FCM algoritmu, stav zdraví

Statistiky	**3C**	**4A**
Mezishlukový součet čtverců/celkový součet čtverců (v %)	90,48	82,75
FSI	0,66	0,53
PE	0,67	0,60
PC	0,66	0,71
MPC	0,57	0,63

Zdroj: vlastní zpracování v programu R

Podle tří uvedených statistik v tabulce 10 dosahuje FCM algoritmus pro pět shluků lepších výsledků v případě vstupního datového souboru 4A. Avšak na základě podílu mezishlukového součtu čtverců a celkového součtu čtverců dosahuje FCM algoritmus lepších výsledků u datového souboru 3C. Výsledných pět shluků získaných pomocí FCM algoritmu je vizualizováno na obrázcích 49 a 50 v příloze 8.

Pět shluků prezentovaných na obrázku 49 v příloze 8 se v porovnání s pěti shluky získanými pomocí metody *k*-průměrů (obrázek 47 v příloze 8) liší pouze v odlišném zařazení Belgie a detekci dvou odlehlých objektů Řecka a Slovinska. U datového souboru 4A jsou za odlehlá pozorování považovány: Dánsko, Finsko, Rakousko, Řecko a Slovinsko. Mimo odlišnosti v existenci odlehlých pozorování se nyní převážně post-socialistické státy rozdělily na dva shluky, kde jeden ze shluků je tvořen Bulharskem a Rumunskem a druhý zbylými především post-socialistickými zeměmi oproti metodě *k*-průměrů. Státy západní, jižní a severní části

Evropy jsou nyní rozděleny do třech menších shluků v porovnání s metodou *k*-průměrů. První z nich obsahuje Německo a Nizozemsko stejně jako u *k*-průměrů. Další dva shluky tvoří státy, které v případě *k*-průměrů spadaly pod jeden shluk. První z nich obsahuje Francii a Itálii, druhý shluk zbylé evropské země. Na základě heat mapy pro datový soubor 4A vizualizované na obrázku 46 v příloze 7, je zřejmé, že odlehlé státy vykazují odlišné hodnoty nebo extrémní hodnoty čtyřech kPCs oproti ostatním hodnotám kPCs.

6.3.4 Výsledky DBSCAN algoritmu pro stav zdraví

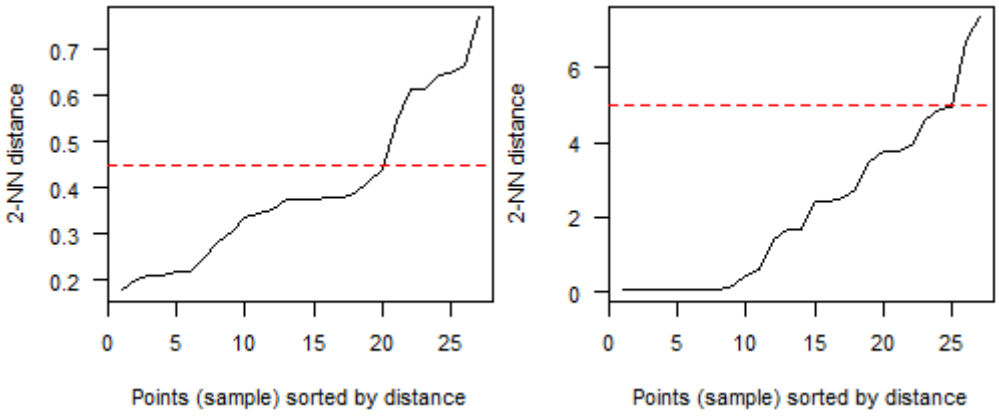
Problémem předchozích metod shlukové analýzy je jejich neschopnost najít shluky libovolného tvaru, proto je zde pro porovnání metod shlukové analýzy aplikován shlukovací algoritmus DBSCAN, který tento problém odstraňuje (Ester a kol., 1996). Program R nabízí DBSCAN algoritmus v rámci balíku *dbscan*, kde je tento algoritmus reimplementován pomocí *kd-tree* (*k*-dimensional tree), který celý proces urychlí a umožní tak práci s většími datovými soubory. Pro detaily viz Hahsler a kol. (2019).

Na základě Hahsler a kol. (2017) je pro stanovení shluků pomocí algoritmu DBSCAN nutné zadat vstupní datovou matici, velikost ε-ového okolí (*eps*) a minimální počet pozorování v oblasti určené velikostí ε-ového okolí (*minPts*). V rámci algoritmu DBSCAN dochází k odhadu hustoty kolem každého pozorování spočítáním počtu pozorování v ε-ovém okolí předem specifikovaném uživatelem. Minimální počet pozorování (*minPts*) slouží k identifikaci jádrových, hraničních a šumových pozorování. V případě *eps* a *minPts* se jedná o hyperparametry DBSCAN algoritmu. Inspirací pro volbu parametrů *eps* a *minPts* jsou články Sander a kol. (1998) a Rahmah, Sitanggang (2016) (viz. podkapitola 2.2.3). Vzhledem k malému rozsahu souboru (27 států) je *minPts* nastaven na hodnotu 2 pro oba použité datové soubory 3C a 4A (viz. podkapitola 6.3).

Parametr *eps* je nastaven pomocí výpočtů vzdáleností *k*-NN v matici pozorování. Tento výpočet nabízí v programu R funkce *kNNdist* a jeho grafické zobrazení *kNNdistplot*. Na základě grafu, který vizualizuje pozorování seřazená podle 2-NN vzdáleností v závislosti na těchto vzdálenostech (obrázek 12) je hodnota *eps* nastavena v bodě zlomu tzv. *„knee“*. Parametr *eps* je nastaven pro datové soubory 3C a 4A následujícím způsobem:

- pro 3C: *eps* = 0,45;
- pro 4A: *eps* = 5.

Obrázek 12: 2-NN vzdálenosti RCs, SPCs a kPCs, stav zdraví



Zdroj: vlastní zpracování v programu R

U algoritmu DBSCAN při nastavení počátečních parametrů došlo k nalezení čtyř shluků a skupiny šumových pozorování označených jako „0“ v případě datového souboru 3C (SPCs) a dvou shluků a skupiny šumových pozorování u datového souboru 4A. V případě datového souboru 3C (obrázek 51 v příloze 8) jsou ze shluků vyjmuty dvě země představující šumová pozorování. Jedná se o Maďarsko a Slovinsko. Shluk obsahující nejvíce zemí Evropy (modře) je tvořen zeměmi představující zástupce západní, severní a jižní části Evropy. Druhý shluk obsahuje pouze dvě země (červeně), mezi které patří Bulharsko a Rumunsko, třetí shluk (zeleně) je tvořen CZ, Estonskem, Chorvatskem, Polskem a SK. Poslední čtvrtý shluk (žlutě) tvoří Litva a Lotyšsko.

V porovnání s výslednými shluky získanými prostřednictvím FCM algoritmu v případě datového souboru 3C (obrázky 49 a 51 v příloze 8) je zřejmé, že vlivem konkrétního nastavení parametrů *eps* a *minPts* došlo k získání odlišného počtu shluků. Právě nastavení hodnot parametrů ovlivňuje výsledný počet shluků. I přes tuto skutečnost se získané výsledné shluky v obou případech příliš neliší. Při použití DBSCAN algoritmu vznikl jeden větší shluk obsahující země západní, jižní a severní Evropy oproti použití FCM algoritmu, kde byl tento shluk rozdělen na dva shluky. Především u post-socialistických států nedošlo k vytvoření rozdílných shluků, vyjma rozdílů u odlehlých států.

Čtyři shluky vytvořené pomocí DBSCAN algoritmu je možné podle tří SPCs interpretovat následovně (viz. obrázek 9 v podkapitole 6.3.1 a podkapitola 6.2.3):

- *Shluk 1*: Belgie, Dánsko, Finsko, Francie, Irsko, Itálie, Kypr, Lucembursko, Malta, Německo, Nizozemsko, Portugalsko, Rakousko, Řecko, Španělsko a Švédsko – lepší stav zdraví z hlediska střední délky života a očekávané délky života prožité ve zdraví,

lepší stav zdraví z hlediska kardiovaskulárních onemocnění a lepší stav zdraví z hlediska rakovin,

- *Shluk 2:* Bulharsko a Rumunsko – horší stav zdraví z hlediska střední délky života a očekávané délky života prožité ve zdraví, nejhorší stav zdraví z hlediska kardiovaskulárních onemocnění, zabranitelné a léčitelné úmrtnosti a nejlepší stav zdraví z hlediska rakovin v rámci především post-socialistických zemí,
- *Shluk 3:* CZ, Estonsko, Chorvatsko, Polsko a SK – nejlepší stav zdraví z hlediska střední délky života a očekávané délky života prožité ve zdraví v rámci post-socialistických zemí, nejlepší stav zdraví z hlediska kardiovaskulárních onemocnění v post-socialistických zemích a lepší stav zdraví z hlediska rakovin v rámci EU-27,
- *Shluk 4:* Litva a Lotyšsko – horší stav zdraví z hlediska střední délky života a očekávané délky života prožité ve zdraví, horší stav zdraví z hlediska kardiovaskulárních onemocnění, zabranitelné a léčitelné úmrtnosti a nejhorší stav zdraví z hlediska rakovin.

Jak již bylo zmíněno, oproti výsledným shlukům získaným pomocí FCM algoritmu nejsou státy západní, jižní a severní Evropy rozděleny do dvou shluků, kde Belgie, Dánsko, Finsko, Německo, Nizozemsko, Portugalsko, Rakousko vykazují horší situaci z hlediska onkologických onemocnění oproti zbylým státům západní, jižní a severní Evropy.

Na obrázku 52 v příloze 8 jsou vizualizované shluky států získané pomocí DBSCAN algoritmu pro datový soubor 4A. V tomto případě jsou ze shluků vyjmuty opět dvě země, konkrétně Řecko a Slovinsko. Vlivem nastavení parametrů *eps* a *minPts* jsou vytvořeny dva shluky obsahující v prvním z nich státy západní, jižní a severní Evropy (modře), dále je vytvořen shluk se zástupci především post-socialistických zemí (červeně). Vzhledem k tomu, že *kPC1* vykazuje podobnost s použitými ukazateli střední délky života, očekávané délky života prožité ve zdraví, úmrtností na nemoci oběhového systému, zabranitelné a léčitelné úmrtnosti, je možné říci, že státy jsou zařazeny do dvou shluků (obrázek 52 v příloze 8) na základě nejhoršího stavu zdraví a nejlepšího stavu zdraví podle *kPC1* (viz. také obrázek 9 v podkapitole 6.3.1).

Na nutnost snížení dimenze původních ukazatelů stavu zdraví ukazují nejen statistiky pro stanovení optimálního počtu shluků používané v rámci jednotlivých metod shlukové analýzy, ale také potřeba interpretace výsledných shluků. Právě schopnost interpretovat jednotlivé shluky je užitečná pro přijímání rozhodnutí na úrovni nadnárodní, celostátní nebo regionální.

6.4 Identifikace států s podobnou celkovou úrovní stavu zdraví a jejich lineární uspořádání

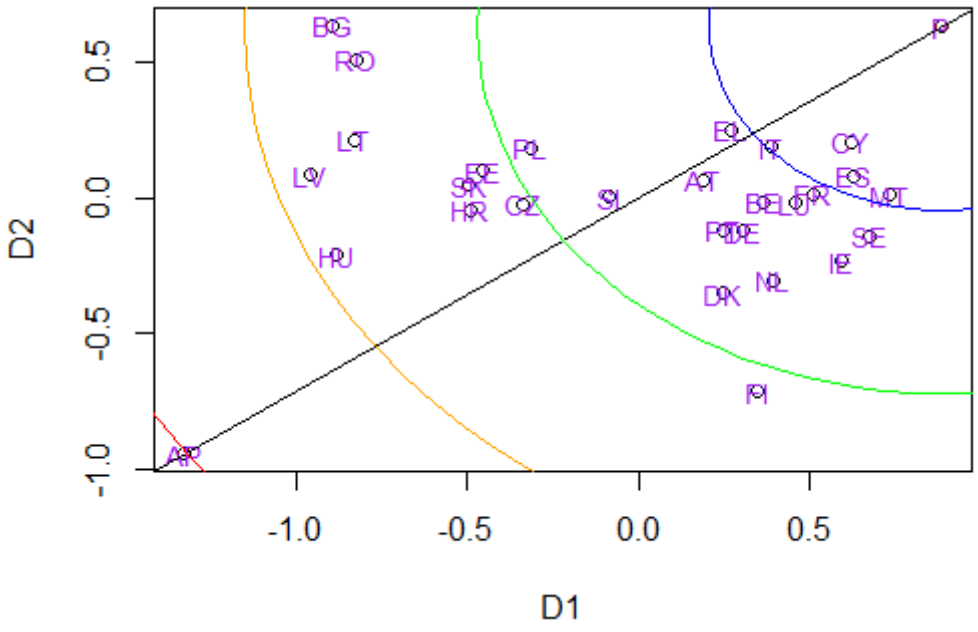
Kromě pouhého zařazování států do shluků pomocí algoritmů shlukové analýzy jsou dále zjišťovány země s podobnou celkovou úrovní stavu zdraví, i když s odlišnou konfigurací (uspořádáním) hodnot vstupních proměnných pomocí MDS (vícerozměrného škálování). Následně jsou tyto země lineárně uspořádány podle vzdálenosti od P objektu (ideálního vzorového objektu). Toho lze dosáhnout hybridním přístupem (viz. podkapitola 5.3) popsaném také ve článcích Walesiak (2016); Walesiak, Dehnel (2018); Walesiak, Dehnel (2019). Odlehlé země, které jsou v použitých datových souborech detekovány pomocí některých metod shlukové analýzy, konkrétně FCM a DBSCAN algoritmu, lze nyní s ostatními zeměmi porovnat pomocí ideálního objektu P.

Hybridní přístup je aplikován na dva datové soubory původních 22 proměnných standardizovaných pomocí „min-max“ (1C) a „z-skóre“ (1A), viz. podkapitola 6.3. MDS slouží k redukci rozměrnosti ukazatelů stejně jako PCA, SPCA nebo kPCA, a proto jsou použity datové soubory, které nevznikly pomocí metod pro snížení rozměrnosti ukazatelů. Důvodem pro aplikování hybridního přístupu na datové soubory s 22 proměnnými jsou lepší výsledky metod shlukové analýzy oproti 25 proměnným (viz. tabulka 8 v podkapitole 6.3.1). Nejprve je důležité u datových souborů 1C a 1A určit, které z 22 proměnných jsou stimulanty (jsou žádoucí jejich vysoké hodnoty), a které z nich jsou destimulanty (jsou žádoucí jejich nízké hodnoty). Na základě určení, které z proměnných jsou stimulanty a destimulanty jsou k původním 27 státům připojeny další dva objekty, jedná se o objekt vzor (P) a objekt anti-vzor (AP). Z datových matic typu 29x22 jsou dále počítány euklidovské vzdálenosti, jelikož vstupem do MDS může být mimo jiné matice vzdáleností.

Vstupní maticí pro MDS je tedy matice euklidovských vzdáleností, která má nyní ve sloupcích i řádcích 29 objektů (včetně objektu P a AP). Matice euklidovských vzdáleností je určena podle vztahu (29) v podkapitole 5.2.1 ze dvou již zmíněných datových souborů. Odlišnosti mezi evropskými státy jsou vizualizovány ve 2D grafech na základě nových souřadnic získaných pro každý stát metodou MDS (obrázky 13 a 14). Před samotnou interpretací výsledků je však nutné ověřit kvalitu modelu MDS pomocí hodnot *stress*. Na základě tabulky 2 v podkapitole 5.3 je kvalita modelů dobrá. V případě datového souboru 1C je hodnota *stress* 0,075 a u datového souboru 1A je tato hodnota 0,067.

Na obrázcích 13 a 14 je možné vidět, jaké jsou pozice zemí Evropy vůči objektům P a AP, které jsou spojené tzv. *osou souboru*. Dále je zřejmá podobná, popř. stejná úroveň evropských zemí ve zdravotním stavu obyvatelstva dle polohy zemí k vyobrazeným tzv. *izokvantám rozvoje* (kružnicím), jejichž střed představuje objekt P. Na těchto obrázcích jsou znázorněny čtyři izokvanty rozvoje, které dělí 2D souřadnicový systém získaný pomocí MDS mezi objektem P a AP na čtyři části, kde každá představuje podobnou celkovou úroveň stavu zdraví vzhledem k objektu P. Země, které leží ve stejném mezikruží, mají podobnou celkovou úroveň stavu zdraví vzhledem k objektu vzor. Například na obrázku 13 se jedná především o většinu post-socialistických zemí, které leží ve stejném mezikruží (ohraničeno oranžovou a zelenou izokvantou rozvoje), a které jsou ze všech zemí EU-27 nejblíže k objektu AP. Například Itálie a Malta leží téměř na stejné izokvantě rozvoje (modře), díky čemuž mohou být interpretovány jako země s téměř stejnou celkovou úrovní zdravotního stavu obyvatelstva vzhledem k objektu P, ale s odlišným uspořádáním hodnot původních proměnných v datové matici 1C. Pomocí metod shlukové analýzy byly také detekovány odlehlé státy, mezi kterými se vyskytovaly např. Finsko a Slovinsko. I v 2D souřadnicovém systému získaném pomocí MDS se jeví tyto dva státy jako odlehlé. Avšak díky možnosti porovnání těchto států s objektem P a možnosti konstrukce izokvant rozvoje je evidentní, že i když metody shlukové analýzy zařazovaly Finsko ke státům západní Evropy, jeho celková úroveň ve stavu zdraví je spíše na úrovni CZ nebo Estonska.

Obrázek 13: Výsledky MDS ve 2D souřadnicovém systému, datový soubor 1C, stav zdraví

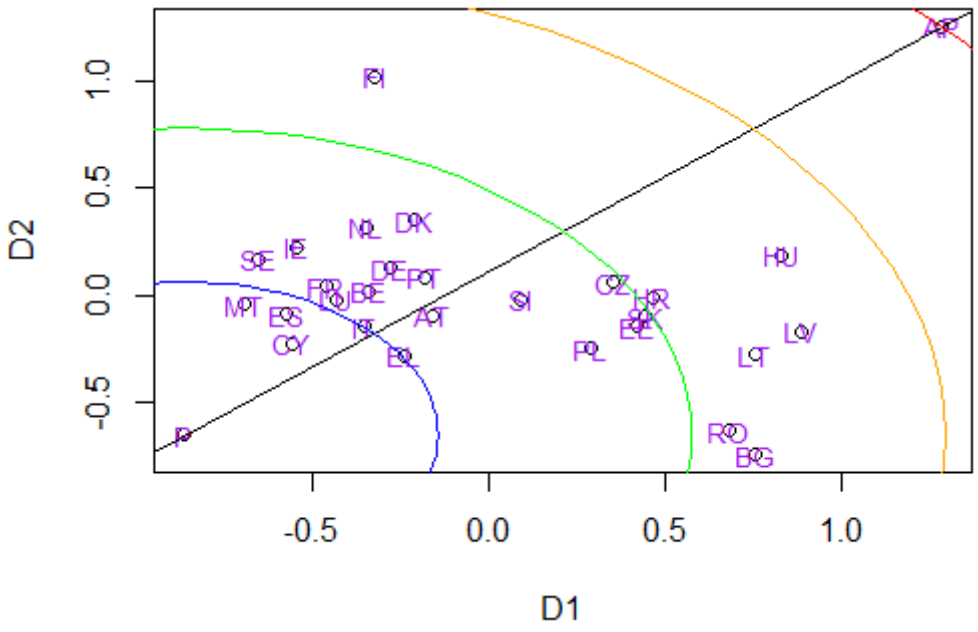


Zdroj: vlastní zpracování v programu R

Na obrázku 14, kde je vstupní matice euklidovských vzdáleností pro MDS získána z datového souboru 1A, je možné si všimnout, že získané výsledky 2D souřadnicového systému jsou téměř totožné s 2D souřadnicovým systémem vyobrazeným na obrázku 13. Tato podobnost je způsobena tím, že oproti datovému souboru 1C je rozdíl pouze v použité standardizaci původních dat. Například u Finska na obrázku 14 je zřejmý pokles celkové úrovně stavu zdraví na podobnou celkovou úroveň Litvy nebo Lotyšska oproti obrázku 13.

V rámci MDS jsou získány nové 2D souřadnice pro každý stát. Na základě agregované míry d_i (viz. vztah (45) v podkapitole 5.3) jsou státy následně lineárně uspořádány podle vzdálenosti od objektu P. V příloze 9 na obrázcích 53 a 54 jsou k dispozici grafy takto lineárně uspořádaných států Evropy bez objektů P a AP. Na obrázcích 55 a 56 v příloze 9 jsou vyobrazeny agregované míry d_i za využití geografických dat. Z těchto obrázků je možné lépe porovnat vliv použité standardizace na výsledky získané pomocí hybridního přístupu. Je evidentní, že pokud je použita standardizace 22 proměnných prostřednictvím „z-skóre“ (obrázek 56 v příloze 9), čili všechny proměnné vstupují do analýzy se stejnou vahou (jednotkovým rozptylem), vykazují např. CZ, Finsko zhoršení celkové úrovně stavu zdraví oproti některým post-socialistickým zemím. Na druhou stranu zlepšení celkové úrovně stavu zdraví zaznamenalo Řecko oproti Francii.

Obrázek 14: Výsledky MDS ve 2D souřadnicovém systému, datový soubor 1A, stav zdraví



Zdroj: vlastní zpracování v programu R

V porovnání s algoritmy shlukové analýzy, které pouze zařazují země do jednotlivých shluků podle jejich podobností v rámci stavu zdraví, hybridní přístup výsledky shlukových analýz doplňuje o posouzení celkové úrovně stavu zdraví. Dále je získáno porovnání celkové úrovně stavu zdraví odlehlých zemí, které pomocí algoritmů shlukových analýz tvoří samostatné shluky, s ostatními zeměmi a objektem P. Nakonec jsou detekovány rozdíly ve výsledcích hybridního přístupu způsobené použitím odlišné standardizace původních 22 proměnných.

6.5 Použité determinanty stavu zdraví pro země EU-27

V tabulce 11 jsou nyní uvedeny zvolené determinanty stavu zdraví. Dvacet sedm těchto proměnných zahrnuje ukazatele, které je možné do určité míry ovlivňovat pomocí různých politik (viz. podkapitola 1.4). Vybrané determinanty stavu zdraví spadají pod následující skupiny: *životní prostředí* (environment conditions), *individuální faktory životního stylu* (individual lifestyle factors), *zdravotnické služby* (health services), *sociálně ekonomické podmínky* (socio-economics conditions), *rodinné podmínky* (family conditions) a *vzdělání* (education) Pro detaily viz. podkapitola 4.2.2 Použité determinanty stavu zdraví.

Tabulka 11: Vybrané determinanty stavu zdraví

Kódy	Názvy proměnných
EC1	emise jemných částic na osobu (2018)
ILF1	míra dospělých kouřících denně v % (2018)
ILF2	nadměrná konzumace alkoholu dospělými v litrech na osobu (2018)
ILF3	hlášení míry obezity mezi dospělými v % (2018)
HS1	výdaje do zdraví na osobu (EUR) při zohlednění parity kupní síly (2019)
HS2	výdaje do zdraví jako podíl na HDP (vládní/povinné) (2019)
HS3	výdaje do zdraví jako podíl na HDP (dobrovolné/vlastní) (2019)
HS4	výdaje do zdraví *„z vlastní kapsy“* vyjádřené jako podíl na konečné spotřebě domácností (2018)
HS5	výdaje do zdraví na dlouhodobou péči na osobu (EUR) (2019)
HS6	výdaje do zdraví na dlouhodobou péči % z HDP (2019)
HS7	lékaři na 1 000 obyvatel (2018)
HS8	zdravotní sestry na 1 000 obyvatel (2018)
HS9	nemocniční lůžka na 1 000 obyvatel (2018)
HS10	průměrná délka pobytu v nemocnici (dny) (2018)
HS11	hlášení neuspokojených potřeb lékařské prohlídky osob 16+ (příliš drahé, příliš daleko, dlouhé čekací doby) v % (2019)
HS12	hlášení neuspokojených potřeb lékařské prohlídky osob 65+ (příliš drahé, příliš daleko, dlouhé čekací doby) v % (2019)
HS13	hlášení používání služeb domácí péče při těžkých úrovních obtíží (2019)
SEC1	celková míra nezaměstnanosti % z pracovní síly (2019)
SEC2	medián ekvivalizovaného disponibilního příjmu (PPS) (2019)
SEC3	Giniho koeficient ekvivalizovaného disponibilního příjmu (2019)
SEC4	osoby ohrožené chudobou nebo sociálním vyloučením % (2019)

Kódy	Názvy proměnných
SEC5	míra hmotné a sociální deprivace % (2019)
SEC6	průměrná velikost domácnosti (průměrný počet osob v domácnosti) (2019)
FC1	hrubá míra uzavření sňatků na 1000 obyvatel (2019)
FC2	hrubá míra rozvodovosti na 1000 obyvatel (2019)
ED1	terciální vzdělání osob ve věku 15–64 let v % (2019)
ED2	méně než základní, základní a nižší sekundární vzdělání osob ve věku 15–64 let (2019)

Zdroj: Eurostat (2022a), OECD/European Union (2020)

6.6 Snížení rozměrnosti determinantů stavu zdraví v EU-27

V této a v následujících podkapitolách se analyzují determinanty stavu zdraví pomocí metod pro snížení rozměrnosti ukazatelů, metod shlukové analýzy a hybridního přístupu. Dále jsou získané výsledky pro determinanty stavu zdraví porovnány s výsledky získanými pro stav zdraví v zemích EU-27. Provedené analýzy v programu R jsou k dispozici opět v příloze 1.

Před samotným snížením rozměrnosti ukazatelů jsou stejně jako v případě ukazatelů stavu zdraví provedeny standardizace proměnných pomocí „z-skóre“ a „min-max“. Následně jsou měřeny závislosti mezi dvojicemi proměnných pomocí Pearsonových a Spearmanových korelačních koeficientů. Vzhledem k dosažení podobných výsledků u obou použitých korelačních koeficientů jsou z datového souboru vyřazeny následující proměnné: *ILF3, HS7, HS10, SEC1, FC1, FC2* a *ED2* (viz. tabulka 11 v podkapitole 6.5). Oproti ukazatelům stavu zdraví jsou vytvořeny dva datové soubory, které se liší pouze použitou standardizací.

KMO index (Kaiser–Meyer–Olkin index) nyní pro 20 původních proměnných nabývající hodnotu 0,59 ukazuje na slabou adekvátnost dat pro metody snížení rozměrnosti ukazatelů (viz. podkapitola 5.1.1). Z tohoto důvodu jsou dále ze dvou standardizovaných datových souborů vyřazeny proměnné s nejnižšími mírami MSA (Measure of Sampling Adequancy) (viz. podkapitola 5.1.1). Jedná se o proměnné *ILF2, HS3* a *HS4* s mírami MSA v tomto pořadí: 0,25; 0,32 a 0,25. Po vyřazení těchto proměnných vykazuje 17 determinantů stavu zdraví KMO index 0,66; což značí průměrnou adekvátnost použitých ukazatelů pro metody snížení rozměrnosti.

6.6.1 Výsledky PCA a jejich interpretace pro determinanty stavu zdraví

V případě datového souboru standardizovaného pomocí „z-skóre“ je výběr extrahovaných PCs omezen na čtyři. Pro čtyři PCs čtyři vlastní čísla přesahují hodnotu 1 v tomto pořadí: 8,76; 2,25; 1,70 a 1,28 a dohromady vysvětlují 82,42 % variability původních dat. U datového souboru standardizovaného pomocí „min-max“ je výběr PCs omezen na dva z důvodu 87,49 %

vysvětlené variability původních dat. Vlastní čísla pro první dvě PC nabývají hodnot 2,90 a 0,65.

V příloze 10 v tabulce 23 jsou uvedeny komponentní zátěže po Varimax rotaci *RC1-RC4* pro data standardizovaná pomocí „z-skóre“. Druhý až pátý sloupec v tabulce 23 v příloze 10 představuje opět rotované komponentní zátěže (*RC1* – *RC4*), šestý sloupec vyjadřuje komunality pro jednotlivé proměnné (*h2*), v sedmém sloupci je vyobrazena jedinečnost rozptylu u jednotlivých proměnných (*u2*) a poslední sloupec představuje komplexitu (*com*). Pro detaily viz. podkapitola 5.1.2.

Na základě rotovaných komponentních zátěží (tabulka 23 v příloze 10) pro 17 proměnných je možné čtyři RCs interpretovat následovně:

- *RC1* – rotovaná komponenta převážně výdajů do zdraví: vládní, povinné a na dlouhodobou péči (+), počtů zdravotních sester (+), hlášení používání služeb domácí péče při těžkých úrovních obtíží (+) a průměrného počtu osob v domácnosti (-),
- *RC2* – rotovaná komponenta převážně hlášení neuspokojených potřeb zdravotní péče (-),
- *RC3* – rotovaná komponenta obsahující převážně Giniho koeficient (+), osoby ohrožené chudobou (+), hmotnou a sociální deprivaci (+),
- *RC4* – rotovaná komponenta převážně počtu nemocničních lůžek (-) a terciálního vzdělání (+).

Nicméně z tabulky 23 v příloze 10 je opět zjevné, že existují proměnné, které mají významné komponentní zátěže s více než jednou RC v případě všech čtyř RCs (vyznačeno zeleně). Podle znamének komponentních zátěží lze konstatovat, že vysoké hodnoty *RC1* označují státy s vysokými výdaji do zdraví, vysokými počty sester a značným využíváním služeb domácí péče. Na druhé straně vysoké hodnoty *RC1* označují státy s nízkým průměrným počtem osob v domácnosti. Nízké hodnoty *RC2* označují státy s vysokou mírou nespokojenosti s uspokojováním zdravotní péče. Dále vysoké hodnoty *RC3* označují státy s vysokou mírou chudoby, a nakonec vysoké hodnoty *RC4* patří státům, které vykazují především nejvyšší počty osob s terciálním vzděláním a nejnižší počty nemocničních lůžek v rámci EU-27.

Podle hodnot komunalit v tabulce 23 v příloze 10 je například ukazatel *HS12* vysvětlen z 95 %, na druhou stranu ukazatel *ILF1* pouze z 60 %. Podle komplexity např. u ukazatele *HS9* a *HS11* mají tyto proměnné vysokou komponentní zátěž pouze s jednou RC a s ostatními nízkou.

Ve většině případů však komplexita výrazně přesahuje hodnotu 1, což indikuje významné zatížení s více než jednou RC.

V příloze 10 v tabulce 24 jsou uvedeny komponentní zátěže po Varimax rotaci dvou RCs pro data standardizovaná pomocí „min-max“. Na základě rotovaných komponentních zátěží pro 17 proměnných je možné dvě RCs interpretovat následovně:

- *RC1* – rotovaná komponenta převážně emise jemných částic (-), míry dospělých kouřících denně (-), počtu nemocničních lůžek (-), průměrného počtu osob v domácnosti (-) a osob s terciálním vzděláním (+),
- *RC2* – rotovaná komponenta převážně hlášení neuspokojených potřeb lékařské prohlídky (-) a míry chudoby (-).

Například proměnná *HS1* je vysvětlena dvěma RCs z 89 %, na druhou stranu proměnné *SEC6* a *ED1* pouze z 49 %. Podle ukazatele komplexity např. u proměnných *EC1*, *HS9*, *HS12* a *ED1* existují vysoké komponentní zátěže pouze s jednou RC a s ostatními nízké. Komplexita výrazně přesahuje hodnotu 1 v případě proměnných *HS1*, *HS6*, *SEC2* a *SEC5* (viz. podkapitola 6.5).

6.6.2 Výsledky SPCA a jejich interpretace pro determinanty stavu zdraví

Stejně jako v případě analyzování stavu zdraví pomocí SPCA jsou i zde nejprve nastaveny parametry způsobující řídkost (viz. příloha 1). Parametry α a β jsou pro datový soubor se sedmnácti původními proměnnými standardizovanými pomocí „z-skóre“ nastaveny na hodnoty $\alpha = 0{,}01$ a $\beta = 0{,}0001$. Pro datový soubor se sedmnácti proměnnými standardizovaný pomocí „min-max“ jsou tyto parametry nastaveny na hodnoty $\alpha = 0{,}001$ a $\beta = 0{,}0001$.

Takto nastavené parametry poskytují řídké komponentní zátěže (v některých případech nenulové zátěže původních proměnných existují pouze s jednou SPC) a zároveň vlastní čísla pro jednotlivé dimenze jsou porovnatelná s PCA (viz. tabulka 12 a tabulky 23, 24, 25 a 26 v přílohách 10 a 11). Stejně jako v případě RCs jsou extrahovány čtyři SPCs, pokud vstupní datová matice sedmnácti proměnných je standardizována pomocí „z-skóre“. Následně dvě SPCs jsou extrahovány stejně jako v případě RCs, jestliže vstupní datová matice je standardizována pomocí „min-max“ opět pro 17 proměnných.

Tabulka 12: Vlastní čísla u RCs a SPCs pro datové soubory, determinanty stavu zdraví

Datový soubor	RCs	SPCs
Standardizace „z-skóre“, 17 proměnných	8,78; 2,25; 1,70; 1,28	8,34; 2,07; 1,61; 1,11
Standardizace „min-max“, 17 proměnných	2,90; 0,65	2,89; 0,64

Zdroj: vlastní zpracování v programu R

Na základě SPCs, získaných z datového souboru standardizovaného pomocí „z-skóre“ (tabulka 25 v příloze 11) pro 17 proměnných je možné první čtyři SPCs interpretovat následovně:

- *SPC1* – řídká komponenta převážně výdajů do zdraví: vládní, povinné a na dlouhodobou péči (-), počtů zdravotních sester (-), hlášení používání služeb domácí péče při těžkých úrovních obtíží (-) a průměrného počtu osob v domácnosti (+),
- *SPC2* – řídká komponenta převážně hlášení neuspokojených potřeb zdravotní péče (-),
- *SPC3* – řídká komponenta obsahující převážně Giniho koeficient (-), osoby ohrožené chudobou (-), hmotnou a sociální deprivaci (-),
- *SPC4* – řídká komponenta převážně emise jemných částic (-), míry dospělých kouřících denně (-), počtu nemocničních lůžek (-) a terciálního vzdělání (+).

V příloze 11 v tabulce 26 jsou uvedeny řídké komponentní zátěže dvou SPCs pro data standardizovaná pomocí „min-max“. Na základě řídkých komponentních zátěží pro 17 proměnných je možné dvě SPCs interpretovat následovně:

- *SPC1* – řídká komponenta převážně výdajů do zdraví: vládní, povinné a na dlouhodobou péči (-), počtu zdravotních sester (-), hlášení používání služeb domácí péče při těžkých úrovních obtíží (-), mediánu ekvivalizovaného čistého příjmu (-) a terciálního vzdělání (-),
- *SPC2* – řídká komponenta převážně emise jemných částic (-), míry dospělých kouřících denně (-), počtu nemocničních lůžek (-), hlášení neuspokojených potřeb lékařské prohlídky (-), míry chudoby (-) a průměrného počtu osob v domácnosti (-).

6.6.3 Vizualizace výsledků PCA a SPCA pro determinanty stavu zdraví

V této podkapitole jsou vizualizovány komponentní skóre získané pomocí PCA a SPCA. Na obrázku 57 v příloze 12 jsou vizualizována RCs získaná z datového souboru standardizovaného pomocí „z-skóre“. Pořadí států je určeno dle *RC1* od nejvyšších hodnot po nejnižší. Vysoké hodnoty *RC1* ukazují na státy s většími zdroji v systému zdravotní péče

(např. výdaji do zdraví, počty zdravotních sester). Nízké hodnoty *RC2* ukazují na státy s vyšším počtem hlášení neuspokojených potřeb lékařských prohlídek. Dále státy, které jsou ohodnoceny vysokými hodnotami *RC3* vykazují vyšší míru chudoby, a nakonec vysoké hodnoty *RC4* prezentují státy s vysokou mírou terciálního vzdělání a nízkým počtem nemocničních lůžek. Komponenta *RC1,* která vysvětluje nejvíce z variability původních dat, rozděluje státy EU-27 na státy západní a severní Evropy s největšími zdroji v systému zdravotní péče a dále post-socialistické státy a státy jižní Evropy s nízkými zdroji plynoucími do systému zdravotní péče. Z hlediska neuspokojených potřeb lékařského vyšetření je v nejhorší situaci Estonsko, Finsko, Rumunsko a Řecko. Vysoká míra chudoby ohrožuje populaci především v Bulharsku a Rumunsku.

Obrázek 58 v příloze 12 prezentuje RCs získaná z datového souboru standardizovaného pomocí „min-max“. Pořadí států je určeno opět podle *RC1*. Nízké hodnoty *RC1* ukazují na státy s větším znečištěním ovzduší, větším počtem denních kuřáků a s nižší mírou terciálního vzdělání (viz. podkapitola 6.6.1). Dále nízké hodnoty *RC2* ukazují na státy s vyšší mírou chudoby a vyšším počtem hlášení neuspokojených potřeb lékařských prohlídek. Z hlediska *RC1* se mezi státy s nejmenším znečištěním, nejmenším počtem denních kuřáků řadí především země severní Evropy. Opět např. Estonsko, Rumunsko a Řecko patří mezi státy s nejvyšší mírou chudoby a nejvyšším počtem hlášených neuspokojených potřeb lékařského vyšetření.

Na obrázku 59 v příloze 12 jsou vizualizovány SPCs získané z datového souboru standardizovaného pomocí „z-skóre“. Pořadí států je určeno dle *SPC1* od nejnižších hodnot po nejvyšší. Pořadí států z hlediska SPCs je podobné pořadí států podle RCs na obrázku 57 v příloze 12. V případě *SPC1, SPC2, SPC3* a *SPC4* lze tyto komponenty interpretovat podobně jako *RC1, RC2, RC3* a *RC4*. Rozdíly jsou pouze u znamének komponentních zátěží v případě první a třetí komponenty.

Obrázek 60 v příloze 12 vizualizuje SPCs získané z datového souboru standardizovaného pomocí „min-max“. Pořadí států je opět určeno podle *SPC1*. Nyní však pořadí států z hlediska dvou SPCs neodpovídá pořadí států podle RCs na obrázku 58 v příloze 12. Důvodem je odlišná interpretace získaných SPCs. Nízké hodnoty *SPC1* představují větší zdroje plynoucí do systému zdravotní péče, vyšší medián ekvivalizovaného čistého příjmu a vyšší míru terciálního vzdělání. Nízké hodnoty jsou zaznamenány v případě zemí západní a severní Evropy. Na druhou stranu nízkých hodnot *SPC2* nabývají státy s vysokou mírou chudoby, větším znečištěním, vyšším počtem denních kuřáků a vyšší mírou hlášení neuspokojených potřeb lékařského vyšetření. S těmito problémy se potýkají především post-socialistické země, ale také země jižní Evropy.

6.6.4 Výsledky kPCA pro determinanty stavu zdraví

Vstupní datové matice pro nelineární KPCA (Kernel analýzu hlavních komponent) jsou stejné jako pro PCA a SPCA. Sedmnáct proměnných vybraných na základě jednoduchých korelačních koeficientů a KMO indexu je standardizováno pomocí „z-skóre“ a „min-max“. Do programu R jsou tedy nahrány dvě datové matice typu 27x17.

Vzhledem k možnosti existence nelineárních vztahů mezi původními proměnnými je opět použita KPCA s nejčastěji používanými jádry: RBF, polynomiální funkcí jádra a funkcí jádra hyperbolický tangens. V případě RBF jádra je parametr σ nastavován opět z intervalu $\langle 10^{-7}\,;10\rangle$ po desetinásobcích. U polynomiální funkce jádra je stupeň polynomu nastavován od 1 do 10 stupňů a v případě funkce hyperbolický tangens, který obsahuje škálový parametr, je tento parametr nastavován z intervalu $\langle 10^{-7}\,;10\rangle$ opět po desetinásobních. Tabulka 13 prezentuje vysvětlený kumulativní rozptyl a vysvětlený rozptyl *kPC1* při nastavení parametrů jednotlivých jader pro dva použité datové soubory.

Tabulka 13: Parametry jader, vysvětlený kumulativní rozptyl, vysvětlený rozptyl kPC1 pro jednotlivé datové soubory, determinanty stavu zdraví

Datové soubory	RBF	Polynomiální	Hyperbolický tangens
17 proměnných, standardizace „z-skóre“, 4 kPCs	sigma = 0,001 kum. roz. = 0,81 kPC1 var. = 0,51	stupeň = 1 offset = 1 kum. roz. = 0,82 kPC1 var. = 0,52	scale = 0,1 offset = 0 kum. roz. = 0,86 kPC1 var. = 0,54
17 proměnných, standardizace „min-max“, 2 kPCs	sigma = 0,01 kum. roz. = 0,67 kPC1 var. = 0,55	stupeň = 1 offset = 1 kum. roz. = 0,68 kPC1 var. = 0,56	scale = 1 offset = 0 kum. roz. = 0,73 kPC1 var. = 0,53

Zdroj: vlastní zpracování v programu R

Z tabulky 13 je zřejmé, že získané výsledky kumulativního vysvětleného rozptylu pomocí nelineární KPCA (jádra hyperbolické tangens) jsou nejlepší v případě datového souboru standardizovaného pomocí „z-skóre“ v porovnání s výsledky kumulativního rozptylu získanými pomocí PCA (viz. podkapitola 6.6.1). Z tohoto důvodu jsou čtyři kPCs získané pomocí jádrové funkce hyperbolický tangens použity jako vstupní proměnné pro shlukovou analýzu.

6.7 Rozdělení států EU-27 podle determinantů stavu zdraví

V této podkapitole jsou opět uvedeny shluky zemí EU-27 nyní s podobnou situací v rámci determinantů stavu zdraví, což může být důležitou informací pro řízení těchto determinantů

kompetentními institucemi. Pro zde aplikované metody shlukové analýzy jsou vybrány vstupní datové soubory z následujících v tomto pořadí:

- 1A – 17 standardizovaných determinantů stavu zdraví pomocí „z-skóre“,
- 1B – 17 standardizovaných determinantů stavu zdraví pomocí „min-max“,
- 2A – čtyři RCs ze 17 determinantů stavu zdraví standardizovaných pomocí „z-skóre“,
- 2B – dvě RCs ze 17 determinantů stavu zdraví standardizovaných pomocí „min-max“,
- 3A – čtyři SPCs ze 17 determinantů stavu zdraví standardizovaných pomocí „z-skóre“,
- 3B – dvě SPCs ze 17 determinantů stavu zdraví standardizovaných pomocí „min-max“,
- 4A – čtyři kPCs ze 17 determinantů stavu zdraví standardizovaných pomocí „z-skóre“, (jádrová funkce hyperbolický tangens),

6.7.1 Výsledky hierarchické aglomerativní shlukové analýzy pro determinanty stavu zdraví

Na sedm datových souborů popsaných v předchozí podkapitole je aplikována hierarchická aglomerativní metoda shlukové analýzy, konkrétně Wardova metoda. Pro konstrukci matice vzdáleností mezi státy je opět použita euklidovská vzdálenost. Tabulka 14 obsahuje kofenetické korelační koeficienty.

Tabulka 14: Kofenetické korelační koeficienty, determinanty stavu zdraví

Soubor	**1A**	**1B**	**2A**	**2B**	**3A**	**3B**	**4A**
ρ	0,68	0,73	0,58	0,62	0,69	0,82	0,79

Zdroj: vlastní zpracování v programu R

Pro vizualizaci výsledných shluků získaných pomocí Wardovy metody a euklidovské vzdálenosti je na základě nejvyššího kofenetického koeficientu korelace vybrán datový soubor 3B představující dvě SPCs vycházející ze standardizovaných sedmnácti původních proměnných pomocí „min-max“.

Na obrázku 61 v příloze 13 je vizualizován dendrogram získaný pomocí Wardovy metody a euklidovské vzdálenosti s heat mapou prezentující hodnoty SPCs použitých pro shlukování. V legendě je opět uveden histogram četností hodnot SPCs vizualizovaných v heat mapě.

Vzhledem k existenci řídkých komponentních zátěží u SPCA lze jednotlivé shluky na obrázku 61 v příloze 13 podle prvních dvou SPCs snadněji interpretovat. Na základě vzdáleností v dendrogramu jsou nejvýraznější tři shluky obsahující následující země EU-27, které je možné pomocí heat mapy interpretovat následovně:

- *Shluk 1:* CZ, Estonsko, Itálie, Kypr, Litva, Malta, Portugalsko, Slovinsko a Španělsko – dosahují spíše průměrných hodnot uvedených ukazatelů v rámci zemí EU-27,
- *Shluk 2:* Bulharsko, Chorvatsko, Lotyšsko, Maďarsko, Polsko, Rumunsko, Řecko a SK - nízké výdaje do zdraví: vládní, povinné a na dlouhodobou péči, nízké počty zdravotních sester, nízká míra hlášení používání služeb domácí péče při těžkých úrovních obtíží, nízký medián ekvivalizovaného čistého příjmu, nízká míra terciálního vzdělání, vysoká míra emise jemných částic, vysoký počet dospělých kouřících denně, vyšší počet nemocničních lůžek, vysoká míra hlášení neuspokojených potřeb lékařské prohlídky, vysoká míra chudoby a vysoký průměrný počet osob v domácnosti v rámci zemí EU-27,
- *Shluk 3:* Belgie, Dánsko, Finsko, Francie, Irsko, Lucembursko, Německo, Nizozemsko, Rakousko a Švédsko, – vysoké výdaje do zdraví: vládní, povinné a na dlouhodobou péči, vysoké počty zdravotních sester, vysoká míra hlášení používání služeb domácí péče při těžkých úrovních obtíží, vysoký medián ekvivalizovaného čistého příjmu, vysoká míra terciálního vzdělání, nízká míra emise jemných částic, nízký počet dospělých kouřících denně, nižší počet nemocničních lůžek, nízká míra hlášení neuspokojených potřeb lékařské prohlídky, nízká míra chudoby a nízký průměrný počet osob v domácnosti v rámci zemí EU-27.

Na obrázku 62 v příloze 14 jsou vizualizovány výsledné tři shluky pomocí geografických dat v programu R, což k získaným interpretacím pro jednotlivé shluky přidává informaci o poloze zařazených států. Z obrázku 62 v příloze 14 je evidentní, že jednotlivé shluky jsou tvořeny především zeměmi, které spolu až na zelený shluk sousedí.

6.7.2 Výsledky metody k-průměrů pro determinanty stavu zdraví

Opět pro porovnání výsledných shluků získaných pomocí Wardovy metody je použita metoda *k*-průměrů. Optimální počet shluků je i nyní nastaven na základě podílu mezishlukového součtu čtverců s celkovým součtem čtverců. Na obrázku 15 jsou graficky znázorněny hodnoty tohoto podílu až pro 26 shluků pro datový soubor 3B. Tabulka 15 prezentuje tyto hodnoty pro dva až šest shluků všech sedmi vstupních datových souborů popsaných v podkapitole 6.7. Z tabulky 15 je zřejmé, že pro různá nastavení shluků, dosahují hodnoty podílu mezishlukového a celkového součtu čtverců vyšších hodnot u datového souboru 3B (zástupce lineární metody pro snížení rozměrnosti ukazatelů). Vzhledem k tomu, že datový soubor 3B byl také použit pro Wardovu metodu, je nyní vstupním datovým souborem pro metodu *k*-průměrů. Výsledky těchto dvou metod shlukové analýzy jsou následně porovnány. Na základě obrázku 15 je zřejmý

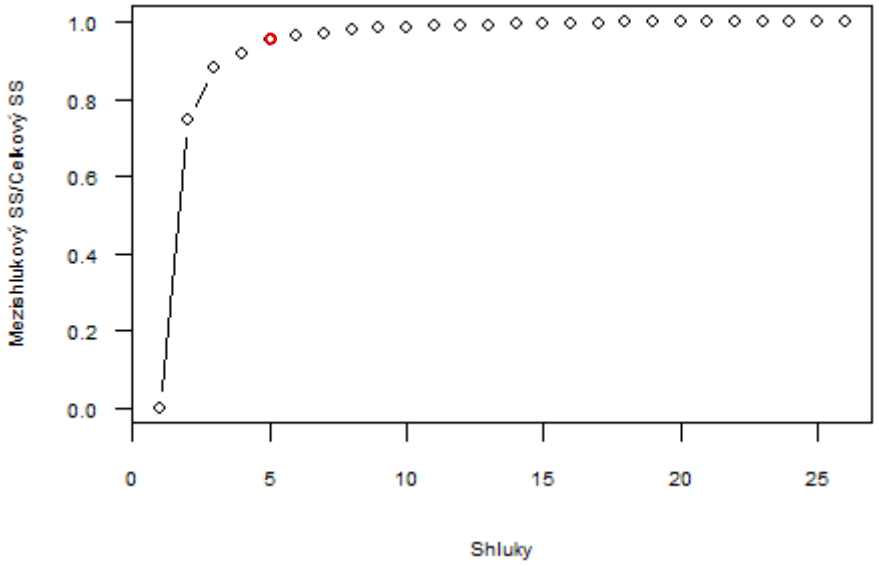
zlomový bod při pěti shlucích (červeně), nad kterým podíl mezishlukového a celkového součtu čtverců výrazně klesá.

Tabulka 15: Stanovení optimálního počtu shluků pro metodu k-průměrů, determinanty stavu zdraví

Mezishlukový součet čtverců/celkový součet čtverců (v %)	1A	1B	2A	2B	3A	3B	4A
2 shluky	40,1	45,1	20,7	38,1	49,5	74,5	53,4
3 shluky	52,6	55,3	40,0	62,7	63,6	87,9	65,5
4 shluky	59,7	61,5	56,2	74,4	71,3	91,9	72,1
5 shluků	66,1	66,9	69,2	81,7	78,2	95,1	77,2
6 shluků	71,7	71,9	74,5	88,5	82,4	96,2	81,4

Zdroj: vlastní zpracování v programu R

Obrázek 15: Grafické stanovení optimálního počtu shluků pro metodu k-průměrů a datový soubor 3B, determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R

Výsledky pěti shluků získané metodou *k*-průměrů (obrázek 63 v příloze 14) jsou podobné výsledkům pěti shluků získaných pomocí Wardovy metody při použití vstupního datového souboru 3B (viz. obrázek 61 v příloze 13). Jediný rozdíl je zaznamenán u Litvy. V případě Wardovy metody je Litva zařazena mezi státy především jižní Evropy. U metody *k*-průměrů patří Litva do shluku spolu s Chorvatskem, Lotyšskem, Maďarskem, Polskem a SK. Pět shluků vytvořených metodou *k*-průměrů je možné pomocí dvou SPCs interpretovat následovně (viz. obrázek 61 v příloze 13):

- *Shluk 1:* Belgie, Francie, Irsko, Lucembursko, Německo a Rakousko – vysoké výdaje do zdraví: vládní, povinné a na dlouhodobou péči, vysoké počty zdravotních sester,

vysoká míra hlášení používání služeb domácí péče při těžkých úrovních obtíží, vysoký medián ekvivalizovaného čistého příjmu, vysoká míra terciálního vzdělání, nižší míra emise jemných částic, nižší počet dospělých kouřících denně, nižší počet nemocničních lůžek, nižší míra hlášení neuspokojených potřeb lékařské prohlídky, nižší míra chudoby a nižší průměrný počet osob v domácnosti v rámci zemí EU-27,

- *Shluk 2:* CZ, Estonsko, Itálie, Kypr, Malta, Portugalsko, Slovinsko a Španělsko – dosahují spíše průměrných hodnot uvedených ukazatelů,
- *Shluk 3:* Bulharsko, Rumunsko a Řecko – nejnižší výdaje do zdraví: vládní, povinné a na dlouhodobou péči, nízké počty zdravotních sester, nízká míra hlášení používání služeb domácí péče při těžkých úrovních obtíží, nízký medián ekvivalizovaného čistého příjmu, nízká míra terciálního vzdělání, vysoká míra emise jemných částic, vysoký počet dospělých kouřících denně, vyšší počet nemocničních lůžek, vysoká míra hlášení neuspokojených potřeb lékařské prohlídky, vysoká míra chudoby a vysoký průměrný počet osob v domácnosti v rámci zemí EU-27,
- *Shluk 4:* Chorvatsko, Litva, Lotyšsko a Maďarsko, Polsko a SK – nižší výdaje do zdraví: vládní, povinné a na dlouhodobou péči, nízké počty zdravotních sester, nízká míra hlášení používání služeb domácí péče při těžkých úrovních obtíží, nízký medián ekvivalizovaného čistého příjmu, nízká míra terciálního vzdělání, vyšší míra emise jemných částic, vyšší počet dospělých kouřících denně, vyšší počet nemocničních lůžek, vyšší míra hlášení neuspokojených potřeb lékařské prohlídky, vyšší míra chudoby a vyšší průměrný počet osob v domácnosti v rámci zemí EU-27,
- *Shluk 5:* Dánsko, Finsko, Nizozemsko a Švédsko – nejvyšší výdaje do zdraví: vládní, povinné a na dlouhodobou péči, nejvyšší počty zdravotních sester, nejvyšší míra hlášení používání služeb domácí péče při těžkých úrovních obtíží, nejvyšší medián ekvivalizovaného čistého příjmu, nejvyšší míra terciálního vzdělání, nejnižší míra emise jemných částic, nejnižší počet dospělých kouřících denně, nízký počet nemocničních lůžek, nejnižší míra hlášení neuspokojených potřeb lékařské prohlídky, nejnižší míra chudoby a nejnižší průměrný počet osob v domácnosti v rámci zemí EU-27.

6.7.3 Výsledky metody Fuzzy k-průměrů pro determinanty stavu zdraví

Vzhledem k použití jednoho vstupního datového souboru 3B u metody *k*-průměrů na základě nejlepších výsledků podílu mezishlukového a celkového součtu čtverců jsou v následující tabulce 16 uvedeny statistiky pro stanovení optimálního počtu shluků, konkrétně pro 3 až 6

shluků. Na základě těchto statistik jsou následně vizualizovány výsledky čtyř shluků pro datový soubor 3B, kde nejvíce těchto statistik dosahuje nejlepších hodnot (viz. podkapitola 5.2.4).

Tabulka 16: Statistiky pro datový soubor 3B u FCM algoritmu, determinanty stavu zdraví

Statistiky pro 3B	3 shluky	4 shluky	5 shluků	6 shluků
Mezishlukový součet čtverců/celkový součet čtverců (v %)	88,19	92,15	95,08	96,23
FSI	0,79	0,82	0,79	0,74
PE	0,40	0,42	0,47	0,56
PC	0,79	0,79	0,78	0,74
MPC	0,69	0,72	0,71	0,68

Zdroj: vlastní zpracování v programu R

Výsledné čtyři shluky získané pomocí FCM algoritmu jsou vizualizovány na obrázku 64 v příloze 14. Tyto čtyři shluky se v porovnání s pěti shluky získanými pomocí metody *k*-průměrů (obrázek 63 v příloze 14) liší pouze v existenci jednoho shluku zemí západní a severní Evropy a detekci Kypru jako odlehlého pozorování. Shluk států západní a severní Evropy obecně představuje země s dobrou situací v rámci determinantů stavu zdraví.

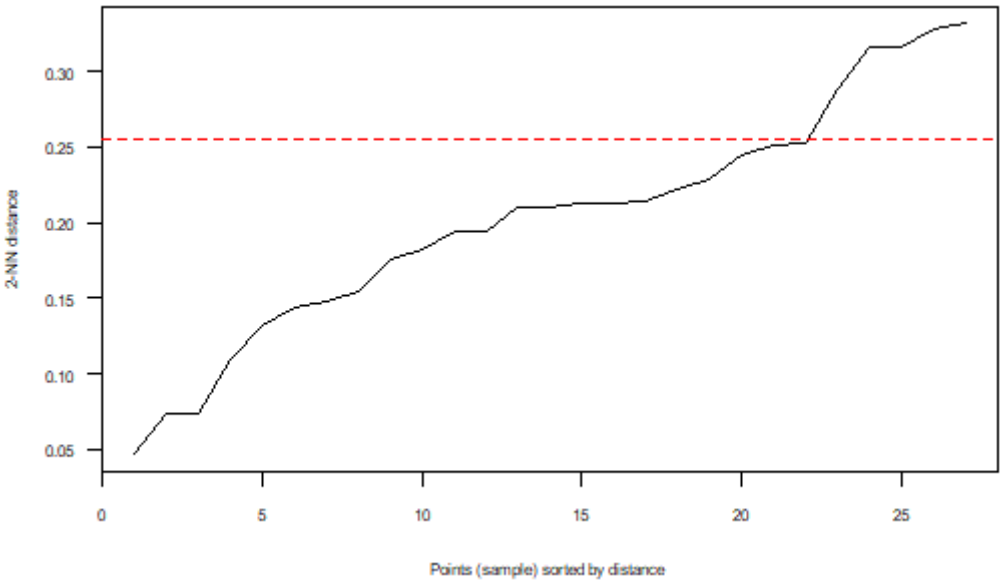
6.7.4 Výsledky DBSCAN algoritmu pro determinanty stavu zdraví

Stejně jako při analyzování stavu zdraví v zemích EU-27 je i zde použit pro vytvoření shluků zemí DBSCAN algoritmus, který je schopen najít shluky libovolného tvaru na rozdíl od předchozích použitých metod shlukové analýzy. Pro shlukování zemí EU-27 na základě determinantů stavu zdraví je opět nastaven parametr *eps* pomocí výčtů vzdáleností *k*-NN v matici pozorování. Podle grafu znázorňujícího státy seřazené podle 2-NN vzdáleností v závislosti na těchto vzdálenostech (obrázek 16) je hodnota *eps* nastavena v bodě zlomu tzv. *„knee“*. Parametr *eps* je nastaven pro vstupní datový soubor 3B na hodnotu 0,255.

U algoritmu DBSCAN při nastavení parametru *eps* na hodnotu 0,255 dochází k nalezení pěti shluků a skupiny šumových pozorování „0“. Na obrázku 65 v příloze 14 je z těchto pěti shluků vyjmuto Finsko. Shluk obsahující nejvíce zemí Evropy (oranžově) je tvořen post-socialistickými zeměmi a zeměmi jižní Evropy. Další shluk obsahuje tři země (červeně), mezi které patří Bulharsko a Rumunsko a Řecko, shluk označen zeleně je tvořen Francií, Irskem, Lucemburskem a Rakouskem. Shluk označen žlutě tvoří Belgie a Německo. Poslední shluk (modře) obsahuje státy severní Evropy v rámci EU-27.

V porovnání s výslednými pěti shluky získanými prostřednictvím metody *k*-průměrů (viz. obrázek 63 v příloze 14) je zřejmé, že vlivem konkrétního nastavení parametrů *eps* a *minPts* došlo k získání stejného počtu shluků zemí EU-27. Nicméně i přes tuto skutečnost se získané výsledné shluky v obou případech liší. Při použití DBSCAN algoritmu vznikl jeden větší shluk obsahující především post-socialistické země a země jižní Evropy oproti metodě *k*-průměrů, u které byl tento shluk rozdělen na dva shluky. Další odlišností je vytvoření dvou výsledných shluků v případě zemí západní Evropy. Obě použité metody shlukové analýzy vytvořily stejný shluk obsahující Bulharsko, Rumunsko a Řecko.

Obrázek 16: 2-NN vzdálenosti SPCs, determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R

Pět shluků vytvořených pomocí DBSCAN algoritmu je možné pomocí dvou SPCs interpretovat následovně (viz. obrázek 61 v příloze 13):

- *Shluk 1:* Belgie a Německo - vysoké výdaje do zdraví: vládní, povinné a na dlouhodobou péči, vysoké počty zdravotních sester, vysoká míra hlášení používání služeb domácí péče při těžkých úrovních obtíží, vysoký medián ekvivalizovaného čistého příjmu, vysoká míra terciálního vzdělání, nižší míra emise jemných částic, nižší počet dospělých kouřících denně, nižší počet nemocničních lůžek, nižší míra hlášení neuspokojených potřeb lékařské prohlídky, nižší míra chudoby a nižší průměrný počet osob v domácnosti v rámci zemí EU-27,
- *Shluk 2:* Bulharsko, Rumunsko a Řecko - nejnižší výdaje do zdraví: vládní, povinné a na dlouhodobou péči, nízké počty zdravotních sester, nízká míra hlášení používání

služeb domácí péče při těžkých úrovních obtíží, nízký medián ekvivalizovaného čistého příjmu, nízká míra terciálního vzdělání, nejvyšší míra emise jemných částic, vysoký počet dospělých kouřících denně, vyšší počet nemocničních lůžek, vysoká míra hlášení neuspokojených potřeb lékařské prohlídky, vysoká míra chudoby a vysoký průměrný počet osob v domácnosti v rámci zemí EU-27,

- *Shluk 3:* CZ, Estonsko, Chorvatsko, Itálie, Kypr, Litva, Lotyšsko, Maďarsko, Malta, Polsko, Portugalsko, Slovinsko, SK a Španělsko – dosahují lepších hodnot uvedených ukazatelů v rámci post-socialistických zemí a zemí jižní Evropy,
- *Shluk 4:* Dánsko, Nizozemsko a Švédsko - vysoké výdaje do zdraví: vládní, povinné a na dlouhodobou péči, vysoké počty zdravotních sester, vysoká míra hlášení používání služeb domácí péče při těžkých úrovních obtíží, vysoký medián ekvivalizovaného čistého příjmu, vysoká míra terciálního vzdělání, nejnižší míra emise jemných částic, nejnižší počet dospělých kouřících denně, nízký počet nemocničních lůžek, nejnižší míra hlášení neuspokojených potřeb lékařské prohlídky, nejnižší míra chudoby a nejnižší průměrný počet osob v domácnosti v rámci zemí EU-27,
- *Shluk 5:* Francie, Irsko, Lucembursko a Rakousko - vysoké výdaje do zdraví: vládní, povinné a na dlouhodobou péči, vysoké počty zdravotních sester, vysoká míra hlášení používání služeb domácí péče při těžkých úrovních obtíží, vysoký medián ekvivalizovaného čistého příjmu, vysoká míra terciálního vzdělání v porovnání s ostatními státy západní a severní Evropy, vysoká míra emise jemných částic, vyšší počet dospělých kouřících denně, vyšší počet nemocničních lůžek, vyšší míra hlášení neuspokojených potřeb lékařské prohlídky, vyšší míra chudoby a vyšší průměrný počet osob v domácnosti v rámci zemí EU-27.

6.8 Identifikace států s podobnou celkovou úrovní determinantů stavu zdraví a jejich lineární uspořádání

I v případě determinantů stavu zdraví jsou zjišťovány země EU-27 s podobnou úrovní determinantů stavu zdraví, i když s odlišným uspořádáním hodnot vstupních proměnných vzhledem k objektu P. Následně jsou tyto země opět lineárně uspořádány podle vzdálenosti od P objektu (ideálního vzorového objektu). Hybridní přístup je aplikován na dva datové soubory původních 17 proměnných standardizovaných pomocí „z-skóre“ (1A) a „min-max“ (1B), viz. podkapitola 6.7. Vstupní maticí pro MDS (vícerozměrné škálování) je matice euklidovských vzdáleností, která má nyní ve sloupcích i řádcích 27 evropských států a dva uměle vytvořené objekty P (vzor) a AP (anti-vzor). Odlišnosti mezi evropskými státy jsou

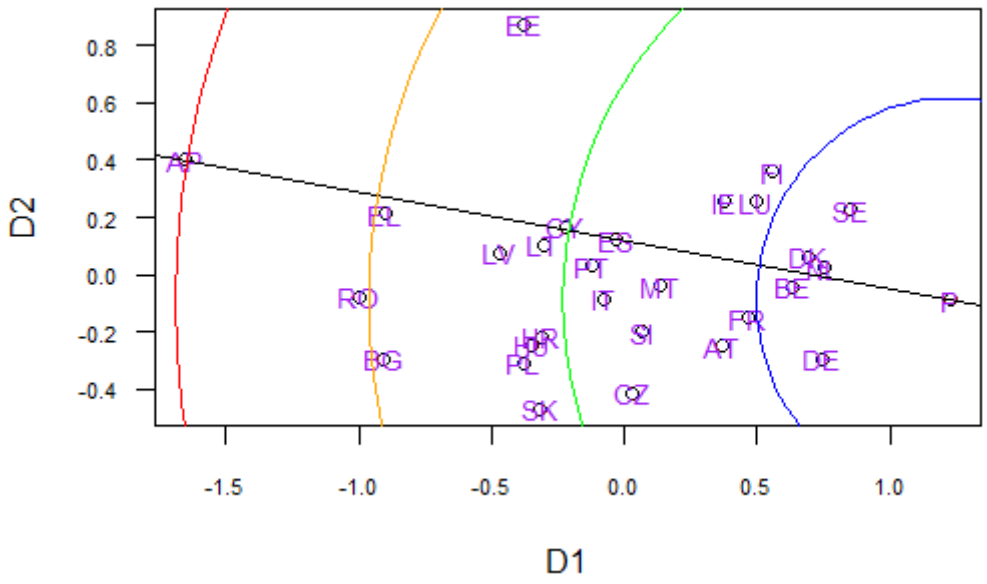
vizualizovány ve 2D grafech na základě nových souřadnic získaných pro každý stát metodou MDS (obrázky 17 a 18). Před samotnou interpretací výsledků je však nutné ověřit kvalitu modelu MDS pomocí hodnot *stress*. Na základě tabulky 2 v podkapitole 5.3 je kvalita modelů dobrá. V případě datového souboru 1A je hodnota *stress* 0,082 a u datového souboru 1B je tato hodnota 0,079.

Na obrázcích 17 a 18 jsou opět vyobrazeny čtyři izokvanty rozvoje, které dělí 2D souřadnicový systém mezi objektem P a AP získaný pomocí MDS na čtyři stejně velké části, kde každá představuje podobnou celkovou úroveň determinantů stavu zdraví vzhledem k objektu P. Podobnou celkovou úroveň determinantů stavu zdraví (nejhorší) vykazují například Bulharsko, Rumunsko a Řecko, které pomocí aplikovaných metod shlukové analýzy při vizualizovaných konkrétních počtech shluků spadaly vždy pod jeden shluk. Nejblíže k objektu P mají Belgie, Dánsko, Německo, Nizozemsko a Švédsko. Tyto státy vykazují podobnou celkovou úroveň determinantů stavu zdraví, i když při odlišném uspořádání hodnot vstupních proměnných, což se odráží i na výsledcích shlukové analýzy. U pěti a čtyř shluků získaných pomocí metod *k*-průměrů a FCM algoritmu se post-socialistické státy zařazují do třech odlišných shluků. CZ, Estonsko tvoří shluk společně se státy jižní Evropy, dalším shlukem je již zmiňovaný shluk obsahující Bulharsko a Rumunsko a poslední shluk tvoří zbylé post-socialistické země. Výsledky hybridního přístupu však ukazují např. na podobnou celkovou úroveň determinantů stavu zdraví ve státech jižní Evropy a CZ na základě porovnání s objektem P. Na druhou stranu Estonsko, které s CZ tvořilo již zmiňovaný shluk, patří ke státům s horší celkovou úrovní determinantů stavu zdraví než CZ.

V rámci MDS jsou získány nové 2D souřadnice pro každý stát. Na základě agregované míry d_i dle vztahu (45) jsou státy následně lineárně uspořádány podle vzdálenosti od objektu P. V příloze 15 na obrázcích 66 a 67 jsou opět k dispozici grafy takto lineárně uspořádaných států Evropy bez objektů P a AP.

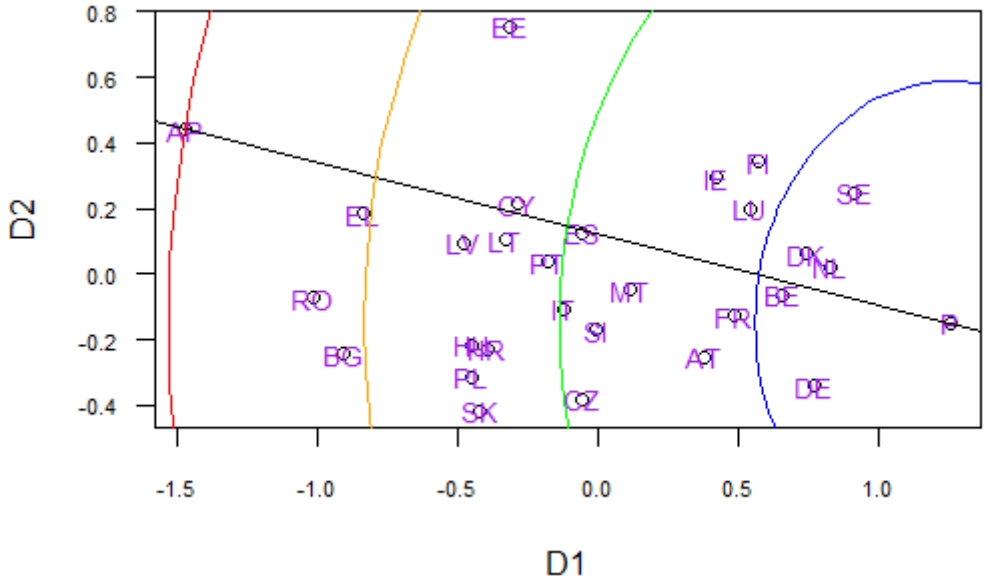
Na obrázcích 68 a 69 v příloze 15 jsou také vizualizovány agregované míry d_i za využití geografických dat. Z těchto obrázků je možné lépe porovnat vliv použité standardizace na výsledky získané pomocí hybridního přístupu. Je evidentní, že pokud je použita standardizace 17 proměnných prostřednictvím „z-skóre“ (obrázek 68 v příloze 15), čili všechny proměnné vstupují do analýzy se stejnou vahou (jednotkovým rozptylem), vykazují např. Bulharsko, Rumunsko a Řecko téměř stejnou úroveň determinantů stavu zdraví. Celkově je možné říci, že při použití standardizace pomocí „z-skóre“ se státy nacházejí blíže objektu P.

Obrázek 17: Výsledky MDS ve 2D souřadnicovém systému, datový soubor 1A, determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R

Obrázek 18: Výsledky MDS ve 2D souřadnicovém systému, datový soubor 1B, determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R

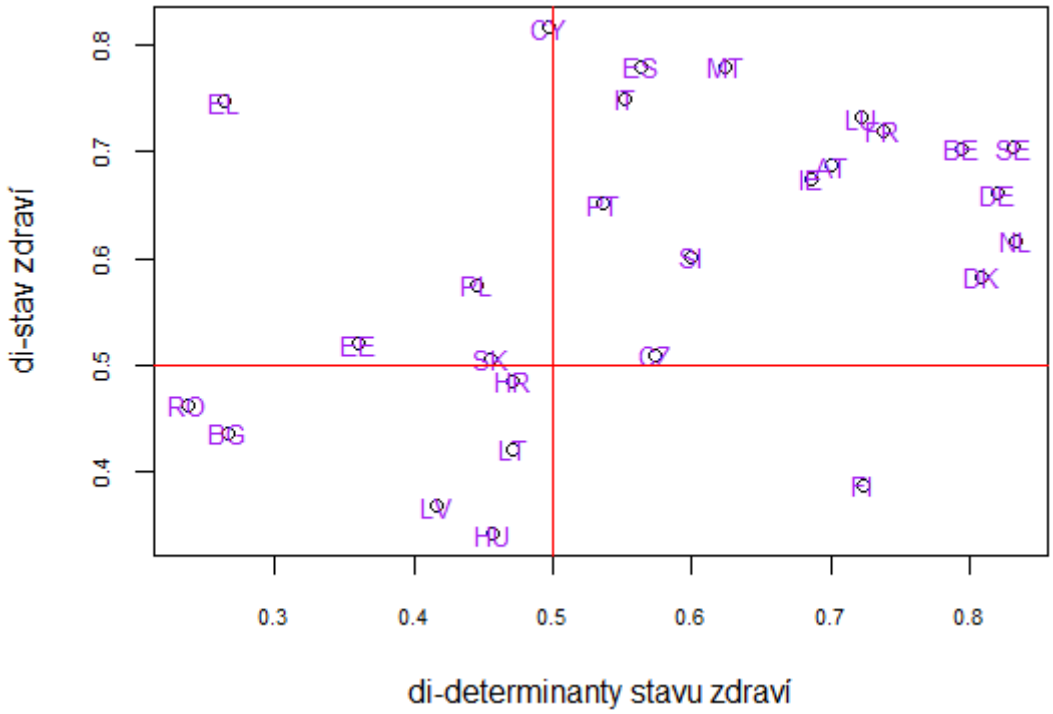
6.9 Porovnání zemí EU-27 z hlediska stavu zdraví a jeho determinantů

Existuje celá řada proměnných stavu zdraví a jeho determinantů, které stav zdraví do jisté míry ovlivňují. Aby bylo možné porovnat země EU-27 na základě těchto dvou vzájemně se ovlivňujících skupin ukazatelů, je třeba nejen snížit rozměrnost těchto ukazatelů, ale také

vytvořit pro každou skupinu ukazatelů jeden agregovaný ukazatel, který postihne nejdůležitější část informace z původních ukazatelů. Vzhledem k dobré kvalitě modelů na základě hodnoty *stress* získaných pomocí MDS, jsou vytvořeny dvě agregované míry d_i pro stav zdraví v zemích EU-27 a pro determinanty stavu zdraví, z kterých je zřejmé, že většina zemí leží blíže objektu P.

Na obrázcích 19 a 20 jsou vizualizovány agregované míry stavu zdraví v závislosti na agregovaných mírách determinantů stavu zdraví nejprve pro ukazatele standardizované pomocí „z-skóre" a následně pro ukazatele standardizované pomocí „min-max". V obou grafech na obrázcích 19 a 20 jsou vizualizované dělící čáry, které dělí 2D souřadnicový systém agregovaných měr na čtyři kvadranty.

Obrázek 19: Porovnání úrovně stavu zdraví a determinantů stavu zdraví, standardizace „z-skóre"

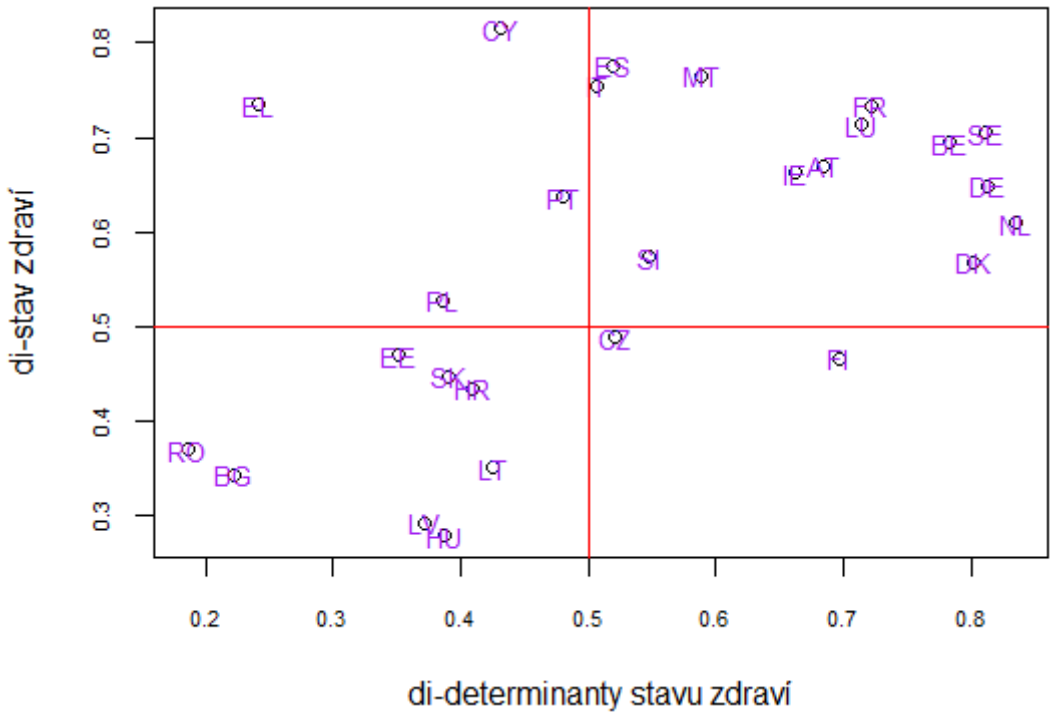


Zdroj: vlastní zpracování v programu R

První kvadrant, kde agregované míry determinantů stavu zdraví a také stavu zdraví nabývají hodnot od 0,5 do 1, obsahuje země, které vykazují dobrou celkovou úroveň stavu zdraví a zároveň dobrou celkovou úroveň determinantů stavu zdraví. Jedná se o státy, které jsou v obou případech blíže objektu P. V případě této skupiny zemí, by bylo vhodné dále posoudit, zda zdroje plynoucí do zdravotní péče (výdaje na zdravotní péči, počty zdravotnického personálu, materiální zdroje aj.) jsou využívány efektivně. Na základě výsledků metod

pro snížení rozměrnosti ukazatelů právě tyto státy vynakládají největší prostředky na zdravotní péči. Měřením efektivnosti systémů zdravotní péče v zemích OECD se zabývali např. Pilyavskyy, Kopecká (2018). Na základě tohoto článku je zřejmé, že velkým problémem pro měření efektivnosti systémů zdravotní péče pomocí DEA modelů (Data Envelopment Analysis) je právě výběr porovnávaných zemí.

Obrázek 20: Porovnání úrovně stavu zdraví a determinantů stavu zdraví, standardizace „min-max“



Zdroj: vlastní zpracování v programu R

Pro druhý kvadrant agregovaná míra determinantů stavu zdraví nabývá hodnot od 0 do 0,5 a agregovaná míra stavu zdraví hodnot od 0,5 do 1. Jedná se o státy, jejichž populace vykazuje dobrý stav zdraví, avšak řízení determinantů stavu zdraví z hlediska množství zdrojů zdravotní péče, sociálního zabezpečení, prevence, životního prostředí je nedostatečné. Druhý kvadrant obsahuje několik zemí EU-27. Nejvýraznější zástupce tohoto kvadrantu je Řecko, jehož polohu v rámci 2D souřadnicového systému mohou vysvětlovat především úsporná opatření. Právě u těchto zemí hrozí do budoucna zhoršování zdravotního stavu populace.

Třetí kvadrant obsahuje státy, které dosahují nedostatečné kvality jak determinantů stavu zdraví, tak stavu zdraví a mají blíže k objektu AP. Hodnoty agregovaných měr determinantů stavu zdraví a stavu zdraví se pohybují od 0 do 0,5. Jedná se především o bývalé post-socialistické země a zároveň novější členy EU. Problémem těchto zemí je to, že i přes zlepšující

se situaci ve stavu zdraví, popř. v řízení determinantů stavu zdraví existuje velká propast mezi těmito státy a ostatními státy Evropy. Pro detaily viz. článek Pacáková, Kopecká (2018c).

Poslední čtvrtý kvadrant se jeví jako nejproblematičtější. Agregovaná míra pro determinanty stavu zdraví dosahuje hodnot od 0,5 do 1, čili obsahuje státy, které jsou v tomto ohledu blíže objektu P, avšak agregovaná míra pro stav zdraví dosahuje pouze hodnot od 0 do 0,5; což ukazuje na špatný stav zdraví populace. Tento kvadrant obsahuje pouze Finsko v případě použití 22 proměnných standardizovaných pomocí „z-skóre“, viz. obrázek 19. Na obrázku 20 se v tomto kvadrantu spolu s Finskem nachází i CZ v případě, že 22 proměnných je standardizováno pomocí „min-max“. Nejvýraznějším zástupcem čtvrtého kvadrantu je tedy Finsko, které se nacházelo v pozici odlehlých objektů detekovaných skrze metody shlukové analýzy. Polohu Finska v 2D souřadnicovém systému agregovaných měr stavu zdraví a jeho determinantů lze přisuzovat špatné situaci stavu zdraví hlavně vlivem nemocí nervového systému a smyslových orgánů (viz. podkapitola 6.2.4). Problematická je v tomto ohledu i CZ, která i přes vynakládání značných zdrojů do systému zdravotní péče vykazuje spíše průměrné výsledky stavu zdraví.

Skutečnost, že na obrázcích 19 a 20 většina zemí EU-27 leží v prvním a třetím kvadrantu ukazuje na pozitivní závislost mezi celkovou úrovní stavu zdraví a celkovou úrovní jeho determinantů, což je důležitá zpráva pro kompetentní řídící orgány. V případě vstupních dat pro MDS standardizovaných pomocí „z-skóre“ (obrázek 19) nabývá Spearmanův korelační koeficient mezi agregovanou mírou pro stav zdraví a agregovanou mírou pro determinanty stavu zdraví hodnoty 0,50. Pokud jsou vstupní data pro MDS standardizovaná pomocí „min-max“, dosahuje tento koeficient mezi zmiňovanými agregovanými mírami hodnoty 0,69. Podle de Vaus (2002) takové hodnoty korelačního koeficientu naznačují podstatnou závislost mezi zkoumanými veličinami. Pomocí hybridního přístupu jsou doplněny výsledky shlukové analýzy o celkovou úroveň stavu zdraví a celkovou úroveň jeho determinantů na základě vzdálenosti od ideálního objektu, což chybí např. v článcích Balçık a kol. (2021) nebo Pacáková a kol. (2019), zabývající se nerovnostmi ve zdravotním stavu evropské populace (viz. podkapitola 1.5). S přihlédnutím k rozdílům ve volbě ukazatelů stavu zdraví a jeho determinantů a k rozdílům ve výběru shlukovaných evropských zemí ve článcích Balçık a kol. (2021) a Pacáková a kol. (2019) nejsou zaznamenány extrémní rozdíly v jejich výsledných shlucích a zde vizualizovaných skupin států na obrázcích 19 a 20. Oba uvedené články identifikují shluky podle ukazatelů stavu zdraví a jeho determinantů dohromady, avšak z těchto výsledných shluků není zřejmý vliv determinantů stavu zdraví na stav zdraví.

7 Diskuze

Všechny výsledky získané pomocí zde použitých analýz v kapitole 6 vedly k naplnění hlavního cíle disertační práce skrze stanovené dílčí cíle (viz. kapitola 3). Veřejně dostupné datové soubory v databázích Eurostatu, OECD a WHO poskytly data o stavu zdraví a jeho determinantech, která bez hlubšího zkoumání prostřednictvím vhodně zvolených metod nepřináší dostatek informací k příjímání rozhodování v oblasti veřejného zdraví. Existuje celá řada publikací týkající se stavu zdraví a jeho determinantů evropské populace, které tyto ukazatele prezentují separovaně v jednotlivých grafech. Jedná se např. o publikace OECD (2019a), OECD (2019b), OECD (2021a), OECD/European Observatory on Health Systems and Policies (2021) nebo OECD/European Union (2020). Takto jednotlivě vizualizovaná data způsobují nepřehlednost ve zkoumané problematice.

Metody snižování rozměrnosti ukazatelů aplikované na ukazatele stavu zdraví v rámci podkapitoly 6.2 přinesly výsledky naplňující dílčí cíl *C1:* Vhodnými metodami snížit rozměrnost ukazatelů stavu zdraví pro země EU-27. Lineární techniky snižování rozměrnosti ukazatelů (PCA při Varimax rotaci a SPCA) požadovaly lineární závislosti mezi původními proměnnými. Vhodnost použitých datových souborů pro lineární techniky snižování rozměrnosti ukazatelů byla ověřena pomocí KMO indexu (Kaiser–Meyer–Olkin index). Mimo lineární techniky snižování rozměrnosti ukazatelů byla aplikována i jedna nelineární technika, konkrétně KPCA, u které byly porovnávány výsledky získané pomocí třech jader: RBF jádro, polynomiální jádro a jádro hyperbolický tangens. Tabulka 17 poskytuje přehled koncepčního porovnání použitých metod pro snižování rozměrnosti ukazatelů.

Důvody nezbytnosti snižování rozměrnosti ukazatelů pro další analýzy (shlukovou analýzu, klasifikační nebo regresní modely atd.) řešili např. autoři Pramoditha (2021) nebo Terek a kol. (2010), kteří uváděli jejich výhody. Techniky pro snižování rozměrnosti ukazatelů v této práci přinesly nové latentní proměnné, které byly následně aplikovány jako vstupní datové matice pro metody shlukové analýzy. Výsledné datové matice získané pomocí vyjmenovaných metod pro snížení rozměrnosti ukazatelů byly dále pro metody shlukové analýzy vybrány především na základě celkové vysvětlené variability původních dat (vysoké). Kofenetický koeficient korelace a podíl mezishlukového a celkového součtu čtverců extrahované v rámci Wardovy metody a metody *k*-průměrů pro jednotlivé použité datové matice ukázaly na získání lepších výsledků metod shlukové analýzy, pokud došlo ke snížení rozměrnosti ukazatelů ve vstupní datové matici. Jedná se především o datové soubory 3C a 4A (viz. podkapitola 6.3).

Tabulka 17: Koncepční porovnání metod pro snižování rozměrnosti ukazatelů

Porovnání	PCA při Varimax rotaci	SPCA	KPCA
Cíl metody	maximalizace rozptylu	řídké komponentní zátěže	lineárně oddělitelná data
Použitý balík a funkce v R	*psych* *principal()*	*sparsepca* *spca()*	*kernlab* *kpca()*
Učení	učení bez učitele	učení bez učitele	učení bez učitele
Vztahy mezi proměnnými	lineární	lineární	nelineární
Hyperparametry	počet hlavních komponent	počet hlavních komponent, parametry způsobující řídkost	počet hlavních komponent, jádro, parametry jádra
Způsob nastavení hyperparametrů	Kaiserovo pravidlo: vl. č. > 1, dostatečně vysvětlený celkový rozptyl původních dat, zlom v sutinovém grafu	převzato z PCA, prostřednictvím „for cyclu“ (viz. příloha 1)	převzato z PCA, nejpoužívanější, prostřednictvím „for cyclu“ (viz. příloha 1)
Výhody	odstranění silných korelací mezi proměnnými	vytvoření řídkých komponent a jejich snadnější interpretace	možnost použití při nelineárních vztazích mezi proměnnými
Nevýhody	silné komponentní zátěže s více než jednou komponentou, funguje dobře pouze na lineárně oddělitelná data	výsledek se odvíjí od nastavení parametrů způsobujících řídkost, funguje dobře pouze na lineárně oddělitelná data	obtížná interpretace nových latentních proměnných, výsledek se odvíjí od výběru jádra a jeho parametrů

Zdroj: vlastní zpracování na základě Anowar a kol. (2021); Erichson a kol. (2018b); Vats, Sharma (2020); Zou a kol. (2006)

V podkapitole 6.3 dochází k naplnění dílčího cíle *C2:* Vybranými metodami posoudit nerovnosti ve stavu zdraví v 27 evropských státech a identifikovat skupiny států s podobnou situací ve stavu zdraví. Byly použity následující metody shlukové analýzy: Wardova metoda, metoda *k*-průměrů, FCM algoritmus a DBSCAN algoritmus. Důvodem k aplikování těchto metod byly vhodnost použití těchto metod vzhledem k dostupným agregovaným ukazatelům a postupné odstraňování jejich nevýhod. Tabulka 18 opět poskytuje přehled koncepčního porovnání použitých metod, nyní však pro případ shlukové analýzy.

Tabulka 18: Koncepční porovnání metod shlukové analýzy

Porovnání	**Wardova metoda**	**Metoda k-průměrů**	**FCM algoritmus**	**DBSCAN algoritmus**
Cíl metody	maximalizace vnitroshlukové homogenity	zařazení objektů k nejbližšímu centroidu	zařazení objektů do shluků pomocí stupně příslušnosti	spojení objektů do shluků na základě četností a vzdáleností objektů v sousedství
Zařazení metody	hierarchická	nehierarchická – pevné shlukování	nehierarchická – fuzzy shlukování	metoda založená na hustotě
Použitý balík a funkce v R	*stats* *hclust()*	*stats* *kmeans()*	*ppclust* *fcm()*	*dbscan* *dbscan()*
Hyperparametry	-	počet shluků	počet shluků	parametr *minPts*, parametr *eps*
Způsob nastavení hyperparametrů	-	prostřednictvím „for cycle“ (viz. příloha 1)	převzato z k-průměrů, statistiky pro stanovení optimálního počtu shluků	*minPts* = 2, *eps*: výpočet vzdáleností *k*-NN v matici pozorování
Výhody	vizualizace pomocí dendrogramu	již zařazený objekt lze v dalším kroku přeřadit do jiného shluku	fuzzy shlukování, zařazení objektů do hluků dle stupňů příslušnosti	nepožaduje apriorní informaci o počtu shluků před spuštěním algoritmu
Nevýhody	již zařazený objekt nelze přeřadit do jiného shluku	pevné shlukování, citlivost vůči odlehlým objektům, vyžaduje apriorní zadání počtu shluků před spuštěním algoritmu	vyžaduje apriorní zadání počtu shluků před spuštěním algoritmu, citlivost vůči odlehlým objektům	vyžaduje nastavení parametrů před spuštěním algoritmu

Zdroj: vlastní zpracování na základě Hair a kol. (1992), Hebák a kol. (2015), Řezanková a kol. (2009), Stankovičová, Vojtková (2007)

Pro pět shluků poskytovaly Wardova metoda, metoda *k*-průměrů a FCM algoritmus podobné výsledné shluky. Aby bylo možné pro takto identifikované shluky států posoudit nerovnosti ve stavu zdraví, bylo třeba vrátit se k hodnotám ukazatelů vstupujícím do shlukové analýzy. Právě díky snížení rozměrnosti původních ukazatelů došlo ke snadnějšímu posouzení těchto nerovností (viz. podkapitoly 6.2.4 a 6.3.1). Obecně lze říci, že především populace v post-

socialistických zemích nacházející se v EU-27, která v rámci použitých metod shlukové analýzy tvoří 1 až 3 shluky, je nejvíce zatížena kardiovaskulárními onemocněními a nízkou délkou života. Větší nerovnosti v rámci post-socialistických zemí byly zjištěny v případě onkologických onemocnění. Výskyty a úmrtností na onkologická onemocnění trpí nejvíce populace v Litvě, Lotyšsku a Maďarsku ze všech zemí EU-27. Naopak nejlepší stav zdraví na základě onkologických onemocnění vykazuje Bulharsko a Rumunsko.

Ze zemí západní, severní a jižní Evropy se nejčastěji v jednom shluku objevovalo Dánsko, Finsko, Německo a Nizozemsko. Pro tyto země je typická vyšší délka života, nižší incidence a úmrtnost na kardiovaskulární onemocnění, ale vyšší riziko výskytu a úmrtí na onkologická onemocnění ze zemí západní, severní a jižní Evropy. Nejlepší situaci ve zdravotním stavu populace však vykazují jihoevropské státy.

Metody shlukové analýzy zařazovaly státy do shluků, kde si byly tyto státy podobné podle stavu zdraví. Nicméně existence odlehlých států ovlivňovala získané výsledné shluky. Aby bylo možné i tyto státy porovnat s ostatními alespoň podle celkové úrovně stavu zdraví vzhledem k ideálnímu objektu P, byl dále aplikován tzv. hybridní přístup kombinující MDS s lineárním uspořádáním. Hybridní přístup zde sloužil k naplnění dílčího cíle *C3:* Nalézt státy s podobnou celkovou úrovní stavu zdraví, ale s odlišným uspořádáním (konfigurací) hodnot původních proměnných a následně provést jejich uspořádání pomocí hybridního přístupu.

Finsko, Řecko a Slovinsko byly nejčastěji detekovanými odlehlými státy pomocí FCM nebo DBSCAN algoritmu. V závislosti na použité standardizaci 22 vstupních proměnných (datový soubor 1A, 1C, viz. podkapitola 6.3) vykazovaly státy EU-27 následující podobné celkové úrovně stavu zdraví vzhledem k objektu P (viz. obrázky 13 a 14 v podkapitole 6.4):

- Finsko s CZ a Estonskem v případě datového souboru 1C,
- Finsko s Litvou a Lotyšskem v případě datového souboru 1A,
- Řecko s Francií v případě datového souboru 1C,
- Řecko s Itálií v případě datového souboru 1A,
- Slovinsko s Dánskem v případě datových souborů 1A a 1C.

Součástí hybridního přístupu bylo získání agregované míry d_i na základě vzdálenosti od objektu P. Tato agregovaná míra slouží pro lineární uspořádání států podle celkové úrovně stavu zdraví, tzn. jednoho ukazatele stavu zdraví.

Stav zdraví evropské populace je do značné míry ovlivňován jeho determinanty (viz. podkapitoly 1.4 a 1.5). Detekce těchto determinantů ovlivňujících zdraví populace je proto důležitá pro kompetentní řídící složky. Z tohoto důvodu je stanoven čtvrtý dílčí cíl disertační práce *C4*: Rozšířit aplikaci zvolených metod na identifikaci hlavních determinantů stavu zdraví pro zkvalitnění politik veřejného zdraví. Tento dílčí cíl je naplňován v podkapitolách 6.5 až 6.8.

Po snížení rozměrnosti původních ukazatelů determinantů stavu zdraví vybraných na základě předchozí rešerše byly aplikovány metody shlukové analýzy jak na původní standardizované datové matice, tak na datové matice nových latentních proměnných (viz. podkapitola 6.7). Opět kofenetický koeficient korelace a podíl mezishlukového a celkového součtu čtverců extrahované v rámci Wardovy metody a metody *k*-průměrů pro jednotlivé použité datové matice ukázaly na získání lepších výsledků metod shlukové analýzy, pokud došlo ke snížení rozměrnosti ukazatelů ve vstupní datové matici. Nejlepšího výsledku dosáhl datový soubor 3B.

Pro různé metody shlukové analýzy Wardovu metodu, metodu k-průměrů, FCM algoritmus a DBSCAN algoritmus byl zobrazen různý počet shluků, kvůli odlišným výsledkům statistik pro stanovení optimálního počtu shluků. Nicméně i přes tuto skutečnost se některé státy vyskytovaly spolu v rámci jednoho shluku. Opět i zde byly posuzovány nerovnosti u identifikovaných shluků pomocí determinantů stavu zdraví na základě vstupního datového souboru 3B získaného pomocí SPCA (viz. podkapitoly 6.6.3 a 6.7.1).

Opět lze říci, že především post-socialistické země a Řecko, které se v rámci použitých metod shlukové analýzy nacházely v 1 až 3 shlucích, jsou v nejhorší situaci týkající se determinantů stavu zdraví. Tyto země vykazují nízké výdaje na zdravotní péči, nízké zdroje plynoucí do systému zdravotní péče, horší sociálně ekonomické podmínky, nižší míru vzdělání, horší stav životního prostředí a horší přístup k životnímu stylu.

Výjimku z post-socialistických zemí představovaly CZ a Estonsko, které se často vyskytovaly v jednom shluku spolu se zeměmi jižní Evropy, které vykazovaly průměrné hodnoty sledovaných ukazatelů. Severní část Evropy vykazovala nejlepší situaci v determinantech stavu zdraví, avšak Finsko bylo pomocí DBSCAN algoritmu detekováno jako šumové pozorování (tzn. odlehlý objekt). Na základě výsledků hybridního přístupu vykazovalo Finsko podobnou celkovou úroveň stavu zdraví s Francií a Lucemburskem podle vzdálenosti od objektu P. I pro determinanty stavu zdraví byla konstruována agregovaná míra d_i.

Průběžně v rámci naplňování jednotlivých cílů byl naplňován i dílčí cíl *C5:* Vizualizovat výsledky analýz v rámci států pomocí různých možností vizualizace včetně využití

geografických dat. Komponentní skóre získaná pomocí lineárních metod pro snižování rozměrnosti ukazatelů byla vyobrazována pomocí sloupcových grafů (viz. podkapitoly 6.2.4 a 6.6.3). Dále výsledné shluky získané pomocí Wardovy metody byly vizualizovány běžně používanými dendrogramy, které byly doplněny o heat mapy vyobrazující hodnoty použitých vstupních proměnných (viz. podkapitoly 6.3.1 a 6.7.1). U hybridního přístupu byla využita vizualizace celkové úrovně stavu zdraví a jeho determinantů v 2D souřadnicovém systému získaném pomocí MDS. Následně byly státy lineárně uspořádány a v přílohách 9 a 15 jsou na obrázcích 53, 54, 66 a 67 znázorněny d_i v závislosti na jejich vzdálenosti od objektu P. Vzhledem k tomu, že i geografická poloha zemí EU-27 hraje významnou roli v situaci ve stavu zdraví a jeho determinantech (tzn. některé sousední státy se nachází ve stejném shluku), byly státy vizualizovány také pomocí geografických dat, což výsledky uvedených analýz činí přehlednějšími. Použité balíky a jejich funkce v programu R jsou uvedeny v příloze 1.

Již na základě získaných výsledných shluků pomocí metod shlukové analýzy bylo zřejmé, že se některé shluky států pro stav zdraví překrývají se shluky států pro determinanty stavu zdraví. Z hybridního přístupu byly extrahovány agregované míry jak pro celkovou úroveň stavu zdraví, tak pro celkovou úroveň jeho determinantů, které jsou v podkapitola 6.9 porovnány v 2D souřadnicovém systému. Na základě obrázků 19 a 20 byl naplněn dílčí cíl *C6:* Propojit získané výsledky stavu zdraví a jeho determinantů a následně je porovnat s již publikovanými. Takto propojené výsledky stavu zdraví a determinantů stavu zdraví byly porovnány s výslednými shluky publikovanými v článcích Balçık a kol. (2021) a Pacáková a kol. (2019). Analyzování stavu zdraví a determinantů stavu zdraví zvlášť je užitečné z hlediska možnosti posouzení vlivu použitých determinantů na stav zdraví.

8 Vědecké a praktické přínosy disertační práce

Cíle disertační práce byly stanoveny v souladu se současným stavem poznání v oblasti stavu zdraví evropské populace a jeho determinantů, v souladu se získanými výsledky v rámci vlastní vědecko-výzkumné činnosti a vzhledem k veřejně dostupným datům. Hlavním cílem disertační práce bylo porovnání a vyhodnocení výsledků lineárních a nelineárních technik pro snížení rozměrnosti ukazatelů stavu zdraví a jejich determinantů v zemích EU-27 a využití takto předzpracovaných dat k posouzení nerovností ve stavu zdraví a identifikování skupin států s podobnou, resp. rozdílnou úrovní stavu zdraví.

V souvislosti se stanovenými cíli lze identifikovat následující vědecké přínosy disertační práce:

- *Předzpracování získaných dat z veřejných databází Eurostatu, OECD a WHO prostřednictvím lineárních a nelineárních metod pro snižování rozměrnosti ukazatelů stavu zdraví a jeho determinantů a následné porovnání takto nově vytvořených datových souborů skrze metody shlukové analýzy*. Statistiky pro stanovení optimálního počtu shluků, ukázaly na nutnost snižování rozměrnosti ukazatelů stavu zdraví. V tomto smyslu SPCs překonaly nedostatky RCs díky vyřazení nadbytečných ukazatelů v RCs zahrnutých. Nelineární technika KPCA, stejně jako uvedené lineární techniky pro snížení rozměrnosti ukazatelů, poskytla vstupní datové soubory pro metody shlukové analýzy, tzn. Wardovy metody, metody *k*-průměrů, FCM a DBSCAN algoritmu. Výsledky, které metody shlukové analýzy přinesly při použití různých vstupních datových matic (původní standardizovaná data, RCs, SPCs a kPCs), byly dle statistik pro stanovení optimálního počtu shluků nejlepší pro vstupní datové soubory SPCs a kPCs.
- *Nastavení hodnot důležitých hyperparametrů u metod pro snižování rozměrnosti ukazatelů a metod shlukové analýzy*. V programu R byly vytvořeny algoritmy sloužící pro výběr optimálních hyperparametrů některých zde použitých metod. V případě parametrů způsobujících řídkost u SPCA byly tyto parametry nastavovány pomocí ukazatele komplexity. Parametry σ pro RBF jádro, stupeň polynomiální funkce jádra a škálový parametr funkce jádra hyperbolický tangens u KPCA byly vybrány na základě kumulativního vysvětleného rozptylu získaného z vlastních čísel pro konkrétní počet kPCs a pro různé hodnoty tohoto parametru. Hyperparametr představující počet shluků pro metodu *k*-průměrů byl nastaven prostřednictvím podílu mezishlukového a celkového součtu čtverců, a nakonec u DBSCAN algoritmu došlo k nastavení

parametru *eps* pomocí výčtů vzdáleností *k*-NN v matici pozorování. Pro detaily viz. příloha 1.

- *Nalezení zemí EU-27 s podobnou celkovou úrovní stavu zdraví a celkovou úrovní determinantů stavu zdraví pomocí hybridního přístupu.* Hybridní přístup byl aplikován prof. Markem Walesiakem a kol. pro porovnávací analýzy v oblastech ekonomické efektivnosti a sociální soudržnosti. Disertační práce rozšířila tuto oblast o oblast veřejného zdraví. Prostřednictvím hybridního přístupu byly nalezeny evropské státy s podobnou celkovou úrovní stavu zdraví a jeho determinantů na základě vzdálenosti od ideálního objektu P, ale s odlišným uspořádáním hodnot vstupních proměnných. Tímto byly doplněny výsledky získané metodami shlukové analýzy, u nichž došlo k posouzení nerovností ve stavu zdraví a jeho determinantů. Výsledky hybridního přístupu ukázaly, že podobnou úroveň stavu zdraví s odlišným uspořádáním hodnot vstupních proměnných mohou mít země, které se spolu nenacházejí v jednom shluku.
- *Vytvoření agregovaných měr celkové úrovně stavu zdraví a celkové úrovně determinantů stavu zdraví.* Na základě agregovaných měr získaných pomocí vzdálenosti od ideálního objektu zvlášť pro stav zdraví a zvlášť pro jeho determinanty je možné změřit jejich závislost.
- *Výběr vhodných vizualizačních technik v programu R*. Disertační práce ukázala možnosti vizualizace vícerozměrných dat stavu zdraví, popř. jeho determinantů nejen v rovině, ale také prostřednictvím vizualizací pomocí geografických dat. Jestliže vstupní datová matice obsahuje velké množství objektů, je nutné provést vizualizaci s prostorovou složkou, bez které by byla tato vizualizace nepřehledná.

Kromě vědeckých přínosů, má práce značné praktické využití, zejména v oblastech:

- *Veřejného zdraví*. Disertační práce je v této oblasti přínosná zejména možnostmi, které poskytuje pro identifikaci determinantů stavu zdraví, které mohou sloužit pro zkvalitnění politik veřejného zdraví a přijímání rozhodování v této oblasti.
- *Sběru dat souvisejících se stavem zdraví a jeho determinantů*. Veřejně dostupné datové soubory o stavu zdraví a jeho determinantech např. v databázích Eurostatu, OECD a WHO, které byly v této práci použity, neposkytují dostatek těchto ukazatelů. Zejména se jedná o ukazatele týkající se dlouhodobé péče o seniory, na které má právě veřejná správa zásadní vliv. Disertační práce upozorňuje na nedostatečnou dostupnost především determinantů stavu zdraví pro NUTS2 regiony, které z tohoto důvodu nebyly v práci analyzovány.

- *Poskytování informací o stavu zdraví a jeho determinantech široké veřejnosti*. Především publikace Eurostatu a OECD poskytují data o stavu zdraví a jeho determinantech jednotlivě v grafech, což bez hlubšího zkoumání prostřednictvím vhodně zvolených metod nepřináší dostatek komplexních informací. Zde jsou uvedeny metody a interpretace jejich výsledků, které mohou široké veřejnosti poskytnout více informací než sloupcové grafy jednotlivých ukazatelů.
- *Vytvoření indexu celkové úrovně stavu zdraví a jeho determinantů prostřednictvím hybridního přístupu*. Agregované míry získané prostřednictvím hybridního přístupu mohou představovat další indexy stavu zdraví a jeho determinantů, podle kterých lze posuzovat vývoj stavu zdraví a jeho determinantů v průběhu let.
- *Možnost použití balíků v programu R pro potřeby veřejné správy*. Práce poskytuje přehled použitých balíků v programu R nejen pro jednotlivé analýzy, ale také pro vizualizace jejich výsledků. Tyto balíky jsou využívané a podložené řadou odborných publikací a mohou tedy být dále použity pro praktické účely orgány veřejné správy.

Závěr

Disertační práce s názvem *Metody snížení dimenze pro modelování stavu zdraví* měla za hlavní cíl porovnat a vyhodnotit výsledky lineárních a nelineárních technik pro snížení rozměrnosti ukazatelů stavu zdraví a jeho determinantů v zemích EU-27 a využít takto předzpracovaná data k posouzení nerovností ve stavu zdraví a identifikování skupin států s podobnou, resp. rozdílnou úrovní stavu zdraví.

V kapitole 1 a 2 je věnována pozornost stavu zdraví evropské populace, jeho determinantům, možnostem snižování rozměrnosti ukazatelů, shlukování objektů a využití hybridního přístupu. Stav zdraví evropské populace stejně jako jeho determinanty jsou vícerozměrné kategorie. Veřejně dostupné datové soubory z databází Eurostatu, OECD nebo WHO obsahují velké množství ukazatelů stavu zdraví a jeho determinantů, což bez použití vhodných analýz pro snižování rozměrnosti ukazatelů vede k jejich nepřehlednosti. Publikace, které vycházejí ze zde uvedených databází, poskytují ukazatele stavu zdraví a jeho determinanty jednotlivě v grafech, což bez hlubšího zkoumání prostřednictvím vhodně zvolených analýz nepřináší dostatek informací pro širokou veřejnost, popř. orgány veřejné správy aj. Mimo jiné tyto databáze poskytují ukazatele považované za indexy stavu zdraví HLY, HALE, DALY atd.

Metody pro snižování rozměrnosti ukazatelů PCA, SPCA a KPCA zde byly použity k vytvoření nových vstupních datových souborů pro metody shlukové analýzy. Avšak tyto nové datové soubory mohou také sloužit jako vstupy např. do klasifikačních nebo regresních modelů. Využitím datových souborů s nižší dimenzí než původní datový soubor pro metody shlukové analýzy nebo klasifikační modely se zabývaly např. články Aungkulanon a kol. (2017), Chang a kol. (2015), Du a kol. (2016), Kallas a kol. (2012), Latifoğlu a kol (2008) nebo Wang a kol. (2017). Výsledky získané pomocí metod shlukové analýzy v této práci ukázaly na nutnost snížit rozměrnost ukazatelů stavu zdraví a jeho determinantů. Následně byly metody shlukové analýzy doplněny o výsledky hybridního přístupu aplikovaného prof. Walesiakem a kol. Hybridní přístup doplnil výsledky shlukové analýzy o nalezení států s podobnou celkovou úrovní stavu zdraví, popř. s podobnou celkovou úrovní determinantů stavu zdraví, ale s odlišným uspořádáním jejich původních hodnot. I přes existenci odlehlých států v rámci zemí EU-27 byly nalezeny jiné státy, se kterými mají tyto odlehlé státy podobnou celkovou úroveň stavu zdraví nebo podobnou celkovou úroveň determinantů stavu zdraví.

Vícerozměrná data stavu zdraví, jeho determinantů a výsledky získané prostřednictvím použitých analýz jsou vizualizovány nejen v rovině, ale také pomocí geografických dat. Takto

vizualizované výsledky mohou přinášet cenné informace např. jak pro orgány veřejné správy, tak pro širokou veřejnost.

Disertační práce byla zpracována na základě rozsáhlé české a zahraniční odborné literatury. Všechny výpočty, většina 2D grafických znázornění a všechny uvedené vizualizace pomocí geografických dat byly vytvořeny v programu R. Použité balíky v rámci programu R jsou využívané a podložené řadou odborných publikací a mohou tedy být dále použity pro uvedené praktické účely.

Prostřednictvím aplikovaných metod byly posouzeny nerovnosti ve stavu zdraví na celostátní úrovni v zemích EU-27 v období let 2018-2019. Dále došlo také k identifikaci determinantů stavu zdraví ve stejném období v těchto zemích Evropy a následně byly posouzeny nerovnosti ve stavu zdraví a jeho determinantech, což může být klíčem pro zkvalitnění politik veřejného zdraví. Porovnání agregovaných měr stavu zdraví a jeho determinantů získaných pomocí hybridního přístupu pro 27 evropských států poskytlo souhrnný přehled o situaci ve stavu zdraví a jeho determinantech. Vzhledem k naplnění všech sedmi dílčích cílů disertační práce je hlavní cíl práce splněn.

Disertační práce využívá data z období těsně před pandemií Covid-19, a proto získané výsledky lze využít jako základ pro porovnání stavu zdraví a determinantů stavu zdraví před a po pandemii v rámci další vědecko-výzkumné a publikační činnosti.

Seznam použité literatury

Abrahart, R. J., Openshaw, S., See, L. M. (2000). *GeoComputation*. CRC Press. ISBN 0-203-30580-9

Alam, M. A., Fukumizu, K. (2014). Hyperparameter Selection in Kernel Principal Component Analysis. *Journal of Computer Science 10*(7), 1139-1150. DOI:10.3844/jcssp.2014.1139.1150

Alptekin, N. (2014). Comparison of Turkey and European Union Countries' Health Indicators by Using Fuzzy Clustering Analysis. *International Journal of Business and Social Research, 10*(4), 68-74.

Al-Shammari, A., Zhou, R., Naseriparsaa, M., Liu, C. (2019). An Effective Density-Based Clustering and Dynamic Maintenance Framework for Evolving Medical Data Streams. *International Journal of Medical Informatics, 126*, 176-186. DOI: 10.1016/j.ijmedinf.2019.03.016

Anowar, F., Sadaoui, S., Selim, B. (2021). Conceptual and Empirical Comparison of Dimensionality Reduction Algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review, 40*, 13. DOI: 10.1016/j.cosrev.2021.100378

Araújo, D., Neto, A. D., Martins, A., Melo, J. (2011). Comparative Study on Dimension Reduction Techniques for Cluster Analysis of Microarray Data. *In The 2011 international Joint conference on neural networks*, 1835-1842. DOI: 10.1109/IJCNN.2011.6033447

Arya, R. (2014). Emblematic Fuzzy C-means Clustering for Demographic Dataset. *International Journal of Computer Science & Engineering Technology (IJCSET), 5*(08). 835-847.

Ashraf, K., Ng, C. J., Teo, C. H., Goh, K. L. (2019). Population Indices Measuring Health Outcomes: A Scoping Review. *Journal of Global Health, 9*(1), 1-14. DOI: 10.7189/jogh.09.010405

Aungkulanon, S., Tangcharoensathien, V., Shibuya, K., Bundhamcharoen, K., Chongsuvivatwong, V. (2017). Area-Level Socioeconomic Deprivation and Mortality Differentials in Thailand: Results from Principal Component Analysis and Cluster Analysis. *International journal for equity in health, 16*(1), 117. DOI 10.1186/s12939-017-0613-z

Azevedo, A. I. R. L., Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a Parallel Overview. *IADS-DM*, 6.

Badash, I., Kleinman, N. P., Barr, S., Jang, J., Rahman, S., Wu, B. W. (2017). Redefining Health: the Evolution of Health Ideas from Antiquity to the Era of Value-Based Care. *Cureus, 9*(2).

Balçik, P., Y., Demirci, Ş., Konca, M. (2021). Determinants of Health System Performance in Europe: A Study Based on Clustering Analysis. *Acıbadem Üniversitesi Sağlık Bilimleri Dergisi, 12*(3), 682-689. DOI: 10.31067/acusaglik.851235

Bambra, C., Riordan, R., Ford, J., Matthews, F. (2020). The COVID-19 Pandemic and Health Inequalities. *J Epidemiol Community Health, 74*(11), 964-968. DOI:10.1136/jech-2020-214401

Beaujot, R., Niu, J. (2005). Aggregate Level Comunity Characteristics and Health. *Discussion Paper no*. 05-14, 1-10. Dostupné z: https://www.researchgate.net/profile/Roderic-Beaujot/publication/41660896_Aggregate_Level_Community_Characteristics_and_Health/links/543fc7570cf21227a11b7bf0/Aggregate-Level-Community-Characteristics-and-Health.pdf

Beckfield, J., Krieger, N. (2009). Epi + Demos + Cracy: Linking Political Systems and Priorities to the Magnitude of Health Inequities – Evidence, Gaps, and a Research Agenda. *Epidemiol Rev, 31,* 152–177. DOI: 10.1093/epirev/mxp002.

Berkhin, P. (2006). *Survey of Clustering Data Mining Techniques*. In: Kogan, J., Nicholas, C., Teboulle, M. (eds.), Grouping Multidimensional Data: Recent Advances in Clustering. Berlin, Springer, 25-71.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press. ISBN 978-1-4757-0452-5

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., Pebesma, E. J. (2008). *Applied Spatial Data Analysis with R*. New York: Springer. ISBN 978-0-387-78170-9

Bountziouka, V., Panagiotakos, D. B. (2012). The Role of Rotation Type Used to Extract Dietary Patterns Through Principal Component Analysis, on Their Short-Term Repeatability. *J Data Sci, 10*, 19-36.

Bro, R., Smilde, A. K. (2014). Principal Component Analysis. *Analytical Methods, 6,* 2812-2831. DOI: 10.1039/c3ay41907j

Byung, S. L. (2018). Varimax Rotation and Thereafter: Tutorial on PCA Using Linear Algebra, Visualization, and Python Programming for R and Q analysis. *Journal of Research Methodology, 3*(1), 79-130. DOI: 10.21487/jrm.2018.5.3.1.79

Cao, L. J., Chua, K. S., Chong, W. K., Lee, H. P., Gu, Q. M. (2003). A Comparison of PCA, KPCA and ICA for Dimensionality Reduction in Support Vector Machine. *Neurocomputing, 55*(1-2), 321-336.

Carr, D. B., Young, C. J., Aster, R. C., Zhang, X. (1999). *Cluster Analysis for CTBT Seismic Event Monitoring*. (No. SAND99-1406C). Sandia National Lab. (SNL-NM), Albuquerque, NM (United States, Sanda National Lab. (SNL-CA, Livermore CA (United States).

Carrilero, N., García-Altés, A., Mendicuti, V. M., Ruiz García, B. (2021). Do Governments Care about Socioeconomic Inequalities in Health? Narrative Review of Reports of EU-15 Countries. *European Policy Analysis, 7*(2), 521-536. DOI: 10.1002/epa2.1124

Chakraborty, S., Nagwani, N. K. (2014). Analysis and Study of Incremental DBSCAN Clustering Algorithm. *International Journal of Enterprise Computing and Business Systems Systems*, 15, arXiv preprint arXiv:1406.4754

Chang, T. S., Gangnon, R. E., Page, C. D., Buckingham, W. R., Tandias, A., Cowan, K. J., et al. (2015). Sparse Modeling of Spatial Environmental Variables Associated with Asthma. *Journal of biomedical informatics, 53*, 320-329.

Chuang, K. S., Tzeng, H. L., Chen, S., Wu, J., Chen, T. J. (2006). Fuzzy C-Means Clustering with Spatial Information for Image Segmentation. *Computerized Medical Imaging and Graphics, 30*, 9–15. DOI: 10.1016/j.compmedimag.2005.10.001

Costa, C., Freitas, Â., Stefanik, I., Krafft, T., Pilot, E., Morrison, J., Santana, P. (2019). Evaluation of Data Availability on Population Health Indicators at the Regional Level Across the European Union. *Population Health Metrics, 17*(11), 1-15. DOI: 10.1186/s12963-019-0188-6

Coste, J., Bouée, S., Ecosse, E., Leplège, A., Pouchot, J. (2005). Methodological Issues in Determining the Dimensionality of Composite Health Measures Using Principal Component Analysis: Case Illustration and Suggestions for Practice. *Quality of Life Research, 14*(3), 641-654.

Čermáková, I., Sedlák, P., Komárková, J. (2016). Pardubice Region Health Data Visualization: Steps of Geoinformatics Project. *In Computer Science Research Notes CSRN 2603: poster´s proceedings. Plzeň: Václav Skála - UNION Agency*, 5-8.

Dahlgren, G., Whitehead, M. (1991). Policies and Strategies to Promote Social Equity in Health. *Background Document to WHO – Strategy Paper for Europe. Institute for Futures Studies, Arbetsrapport, 14*, 1-70.

Dahlgren, G., Whitehead, M. (2021). The Dahlgren-Whitehead Model of Health Determinants: 30 Years on and Still Chasing Rainbows. *Public Health, 199*, 20-24. DOI: 10.1016/j.puhe.2021.08.009

Dash, M., Liu, H., Yao, J. (1997). Dimensionality Reduction of Unsupervised Data. *In Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, 532-539.

Dash, Y., Dubey, S. K. (2012). Application of Principal Component Analysis in Software Quality Improvement. *IJARCSSE, 2*(4), 202-205.

Deaton, A. (2003). Health, Inequality, and Economic Development. *Journal of Economic Literature, 41*(1), 113–158. DOI: 10.1257/002205103321544710

Dehnel, G., Walesiak, M. (2019). A Comparative Analysis of Economic Efficiency of Medium-Sized Manufacturing Enterprises in Districts of Wielkopolska Province Using the Hybrid Approach with Metric and Interval-Valued Data. *Statistics in Transition New Series, 20*(2), 49-68. DOI 10.21307/stattrans-2019-014

Dehnel, G., Walesiak, M., Obrębalski, M. (2019). Comparative Analysis of the Ordering of Polish Provinces in Terms of Social Cohesion. *Argumenta Oeconomica Cracoviensia, 1*(20), 71-85. DOI: https://doi.org/10.15678/AOC.2019.2005

Dickes, P., Valentová, M. (2013). Construction, Validation and Application of the Measurement of Social Cohesion in 47 European Countries and Regions. *Social Indicators Research, 113*(3), 827-846. DOI 10.1007/s11205-012-0116-7

Du, K. L., Swamy, M. N. (2013). *Neural Networks and Statistical Learning.* Springer Science & Business Media, 881. DOI: 10.1007/978-1-4471-5571-3

Du, J., Wang, L., Jie, B., Zhang, D. (2016). Network-Based Classification of ADHD Patients Using Discriminative Subnetwork Selection and Graph Kernel PCA. *Computerized Medical Imaging and Graphics, 52*, 82-88.

Dudik, J. M., Kurosu, A., Coyle, J. L., Sejdić, E. (2015). A Comparative Analysis of DBSCAN, K-Means, and Quadratic Variation Algorithms for Automatic Identification of Swallows from

Swallowing Accelerometry Signals. *Computers in Biology and Medicine, 59*, 10-18. DOI: 10.1016/j.compbiomed.2015.01.007

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics, 3*, 32-57.

Dyba T, Randi G, Bray F, Martos C, Giusti F, Nicholson N, Gavin A, Flego M, Neamtiu L, Dimitrova N, Negrão Carvalho R, Ferlay J, Bettio M. (2021). The European cancer burden in 2020: Incidence and mortality estimates for 40 countries and 25 major cancers. *Eur J Cancer. 157*, 308-347. DOI: 10.1016/j.ejca.2021.07.039

Elgeldawi, E., Sayed, A., Galal, A. R., Zaki, A. M. (2021). Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. *Informatics, 8*(79), 1-21. DOI: 10.3390/informatics8040079

Erichson N. B., Zheng P., Aravkin, S. (2018a). *Package "sparsepca"*. Verze 0.1.2. [online 2020-09-26]. Dostupné z: https://cran.r-project.org/web/packages/sparsepca/sparsepca.pdf

Erichson N. B., Zheng P., Manohar K., Brunton S., Kutz J. N., Aravkin A. Y. (2018b). Sparse Principal Component Analysis via Variable Projection. *Submitted to IEEE Journal of Selected Topics on Signal Processing*, arXiv preprint arXiv:1804.00341

Eslava-Schmalbach, J., Alfonso, H., Oliveros, H., Gaitan, H., Agudelo, C. (2008). A new Inequity-in-Health Index based on Millenium Development Goals: Methodology and Validation. *Journal of Clinical Epidemiology*, *61*(2), 142-150. DOI: 10.1016/j.jclinepi.2007.05.001

Ester, M., Kriegel, H. P., Sander, J., Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In Kdd, 96*(34), 226-231.

Eszergár-Kiss, D., Caesar, B. (2017). Definition of user groups applying Ward's method. *Transportation Research Procedia, 22*, 25-34. DOI: 10.1016/j.trpro.2017.03.004

European Commission. (2018a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: on Enabling the Digital Transformation of Health and Care in the Digital Single Market; Empowering Citizens and Building a Healthier Society*. Document 52018DC0233. [online 2022-06-09]. Dostupné z: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:233:FIN

European Commission. (2018b). *The 2018 Ageing Report: Economics and Budgetary Projections for the 28 EU Member States (2016 – 2070)*. [2019-02-20]. DOI: 10.2765/615631

European Commission. (2022). *ECHI Data Tool*. [online 2022-06-09]. Dostupné z: https://webgate.ec.europa.eu/dyna/echi/

Eurostat. (2020). Eurostat Statistics Explained: Self-Perceived Health Statistics. [online 2022-02-26]. Dostupné z: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Self-perceived_health_statistics

Eurostat. (2021). Eurostat Statistics Explained: *Preventable and Treatable Mortality Statistics*. [online 2022-02-22]. Dostupné z: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Preventable_and_treatable_mortality_statistics

Eurostat. (2022a). Eurostat Database. *European Commission* [online 2022-10-13]. Dostupné z: https://ec.europa.eu/eurostat/data/database

Eurostat. (2022b). GISCO: Geografical Information and Maps: *Administrative Units/Statistical Uni*ts [online 2022-01-22]. Dostupné z: https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units

Everitt, S. B., Dunn, G. (2001). *Applied Multivariate Data Analysis*. 2. ed. London: Hodder Arnold. ISBN 0-340-74122-8

Ezukwoke, K., Zareian, S. (2019). Kernel Methods for Principal Component Analysis (PCA). *A Comparative Study of Classical and Kernel PCA*, 1-9. DOI: 10.13140/RG.2.2.17763.09760

Fiscella, K., Franks, P. (2000). Individual Income, Income Inequality, Health and Mortality: What Are the Relationships? *Health Services Research*, *35*(1 Pt 2), 307-318.

Fitzmaurice, C., Allen, C., Barber, R. M., Barregard, L., Bhutta, Z. A., Brenner, H. et al. (2017). Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life-Years for 32 Cancer Groups, 1990 to 2015: a Systematic Analysis for the Global Burden of Disease Study. *JAMA oncology*, *3*(4), 524-548.

Fu, Y.; Kruger, U.; Li, Z.; Xie, L.; Thompson, J.; Rooney, D.; Hahn, J.; Yang, H. (2017). Cross-Validatory Framework for Optimal Parameter Estimation of KPCA and KPLS Models. *Chemom. Intell. Lab. Syst., 167,* 196–207.

Gajjar, S., Kulahci, M., Palazoglu, A. (2016). Use of Sparse Principal Component Analysis (SPCA) for Fault Detection. *IFAC-PapersOnLine, 49*(7), 693-698. DOI: 10.1016/j.ifacol.2016.07.259

Gan, G., Ma CH., Wu J. (2007). *Data Clustering Theory, Algorithms, and Applications*. ASA, Philadelphia.

Gracia, A., González, S., Robles, V., Menasalvas, E. (2014). A Methodology to Compare Dimensionality Reduction Algorithms in Terms of Loss of Quality. *Information Sciences, 270,* 1-27. DOI: 10.1016/j.ins.2014.02.068

García-Escudero, L. A., Greselin, F., Iscar, A. M. (2016). Fuzzy Clustering through Robust Factor Analyzers. *In International Conference on Soft Methods in Probability and Statistics*, 229-235.

García-Escudero, L. A., Greselin, F., Iscar, A. M. (2018). Robust, Fuzzy, and Parsimonious Clustering, Based on Mixtures of Factor Analyzers. *International Journal of Approximate Reasoning, 94*, 60-75. DOI: 10.1016/j.ijar.2018.01.001

Gavurova, B., Ivankova, V., Rigelsky, M., Kmecova, I. (2020). How Do Gender Inequalities in Health Relate to the Competitiveness of Developed Countries? An Empirical Study. *Journal of Competitiveness, 12*(3), 99. DOI: 10.7441/joc.2020.03.06

Genton, M. G. (2001). Classes of Kernels for Machine Learning: A Statistics Perspective. *Journal of Machine Learning Research, 2*(12), 299-312.

Göpfertová, D., Šmerhovský, Z. (2015). *Výkladový slovník termínů v epidemiologii*. Praha: Institut postgraduálního vzdělávání ve zdravotnictví.

Gerdtham, U. G., Johanesson, M. (2004) Absolute Income, Relative Income, Income Inequality, and Mortality. *Journal of Human Resources, 39*(1), 228-247.

Godoy, J. L., Zumoffen, D. A., Vega, J. R., Marchetti, J. L. (2014). New Contributions to Non-Linear Process Monitoring through Kernel Partial Least Squares. *Chemometrics and Intelligent Laboratory Systems, 135,* 76-89. DOI: 10.1016/j.chemolab.2014.04.001

Golembiewski, E., Allen, K. S., Blackmon, A. M., Hinrichs, R. J., Vest, J. R. (2019). Combining Nonclinical Determinants of Health and Clinical Data for Research and Evaluation: Rapid Review. *JMIR public health and surveillance, 5*(4), e12846.

Grabiński, T. (1984). *Wielowymiarowa Analiza Porównawcza w Badaniach Dynamiki Zjawisk Ekonomicznych*. Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, Seria Specjalna – Monografie, 61.

Grabiński, T. (1992). *Metody Taksonometrii*. Akademia Ekonomiczna.

Grabiński, T., Wydymus, S., Zeliaś, A. (1983). *Metody Prognozowania Rozwoju Społeczno-Gospodarczego*. Warszawa: Państwowe Wydawnictwo Ekonomiczne.

Gutmann, M. (2017). *Data Mining and Exploration*. Lecture Notes, University of Edinburg.

Hahsler, M., Piekenbrock, M., Arya, S., Mount, D. (2017). *Package "dbscan"*. Verze 1.1-8. Online [2021-04-26]. Dostupné z: https://cran.r-project.org/web/packages/dbscan/dbscan.pdf

Hahsler, M., Piekenbrock, M., Doran, D. (2019). dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software, 91*(1), 30. DOI: 10.18637/jss.v091.i01

Hair, J. F., Anderson, R. E., Tatham, R. L, Black, W. C. (1992). *Multivariate Data Analysis*. 3. ed. Macmillan Publishing Company

Hamplová, L. (2019). *Veřejné zdravotnictví a výchova ke zdraví.* Grada Publishing, a.s. ISBN 978-80-271-2826-6

Han, J., Ge, Z. (2020). Effect of Dimensionality Reduction on Stock Selection with Cluster Analysis in Different Market Situation. *Expert Systems with Applications, 147*, 1-15. DOI: 10.1016/j.eswa.2020.113226

Han, L., Embrechts, M. J., Szymanski, B. K., Sternickel, K., Ross, A. (2011). Sigma Tuning of Gaussian Kernels Detection of Ischemia from Magnetocardiograms. *In Computational Modeling and Simulation of Intellect: Current State and Future Perspectives*, 206-223.

Hand, D. J. (1998). Data Mining: Statistics and More? *The American Statistician, 52*(2), 112-118.

Hand, D. J., Krzanowski, W. J. (2005). Optimising K-Means Clustering Results with Standard Software Packages. *Computational Statistics & Data Analysis, 49*(4), 969-973. DOI 10.1016/j.csda.2004.06.017

Hartigan, J. A., Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics, 28*, 100–108. DOI 10.2307/2346830

Hathaway, R. J., Bezdek, J. C. (2001). Fuzzy C-Means Clustering of Incomplete Data. *In IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 31*(5), 735-744. DOI: 10.1109/3477.956035

Hebák, P. Hustopecký, J., Pecáková, I., Průša, M., Řezanková, H., Svobodová, A., Vlach, P. (2007). *Vícerozměrné statistické metody [3]*. 2. vyd. Praha: Informatorium. ISBN 978-80-7333-052-1

Hebák, P a kolektiv. (2015). *Statistické myšlení a nástroje analýzy dat*. 2. vyd. Praha: Informatorium. ISBN 978-80-7333-118-4

Hellwig, Z. (1968). Zastosowanie Metody Taksonomicznej do Typologicznego Podziału Krajów ze Względu na Poziom Rozwoju Oraz Zasady i Strukturę Wykwalifikowanych Kadr. *Przegląd Statystyczny, 4*, 307-326.

Hellwig, Z. (1972). Procedure of Evaluating High-Level Manpower Data and Typology of Countries by Means of the Taxonomic Method. Towards a System of Human Resources Indicators for Less Developed Countries. *Papers prepared for a UNESCO Research Project, Ossolineum, The Polish Academy of Sciences, Wrocław*, 131-134.

Hernandez, J. B., & Kim, P. (2022). Epidemiology morbidity and mortality. *StatPearls*. Dostupné z: https://www.statpearls.com/ArticleLibrary/viewarticle/21202

Hlaváček, M., Lakotová, L. (2019). Délka života ve zdraví. *Úřad národní rozpočtové rady: Informační studie, sekce makroekonomických a fiskálních analýz*. [online 2022-02-19]. Dostupné z: https://unrr.cz/wp-content/uploads/2020/08/D%C3%A9lka-%C5%BEivota-ve-zdrav%C3%AD.pdf

Hoffmann, H. (2007). Kernel PCA for Novelty Detection. *Pattern Recognition, 40*(3), 863-874.

Holčík, J., Komenda, M. a kol. (2015). *Matematická biologie: e-learningová učebnice*. 1. vyd. Brno: Masarykova univerzita, [online 2020-03-19]. Dostupné z: https://portal.matematickabiologie.cz/

Holgersson M. (1977). The Limited Value of Cophenetic Correlation as a Clustering Criterion. *Pattern Recognition, 10*, 287-295.

d'Hombres, B., Elia, L., Weber, A. (2013) Multivariate Analysis of the Effect of Income Inequality on Health, Social Capital, and Happiness. *Joint Research Centre, European Commission.* DOI: 10.2788/68427

Hout, M. C., Papesh, M. H., Goldinger, S. D. (2013). Multidimensional Scaling. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(1), 93-103.

Hu, G., Kaur, M., Hewage, K., Sadig, R. (2019). Fuzzy Clustering Analysis of Hydraulic Fracturing Additives for Environmental and Human Health Risk Mitigation. *Clean Technologies and Environmental Policy, 21*, 39–53. DOI: 10.1007/s10098-018-1614-3

Hudrliková, L. (2013). Composite Indicators as a Useful Tool for International Comparison: the Europe 2020 Example. *Prague Economic Papers, 4*, 459-473. DOI: 10.18267/j.pep.462

Institute for Health Metrics and Evaluation. (2022). *About GBD*. [online 2022-10-14]. Dostupné z: https://www.healthdata.org/gbd/about

Jayasinghe, S. (2015). Social Determinants of Health Inequalities: Towards a Theoretical Perspective Using Systems Science. *International Journal for Equity in Health, 14*(71). DOI 10.1186/s12939-015-0205-8

Jindrová, P., Labudová, V. (2020). The Impact of Socio-Economic and Demographic Determinants on Self-Perceived Health. *E&M Economics and Management, 23*(4), 68–88. DOI: 10.15240/tul/001/2020-4-005

Jolliffe, I., T. (2002). *Principal Component Analysis*. 2. vyd., Springer Science & Business Media. ISBN 0-387-95442-2

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., Daszak, P. (2008). Global Trends in Emerging Infectious Diseases. *Nature, 451*(7181), 990-993.

Kahle, D. J., Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *R Journal, 5*(1), 144-161.

Kallas, M., Francis, C., Kanaan, L., Merheb, D., Honeine, P., Amoud, H. (2012). Multi-Class SVM Classification Combined with Kernel PCA Feature Extraction of ECG Signals. *In 2012 19th International Conference on Telecommunications (ICT). IEEE*, 1-5.

Kallas, M.; Mourot, G.; Maquin, D.; Ragot, J. (2014). Diagnosis of Nonlinear Systems Using Kernel Principal Component Analysis. *J. Phys. Conf. Ser., 570*(7). DOI: 10.1088/1742-6596/570/7/072004

Kaltenthaler, E., Maheswaran, R., Beverley, C. (2004). Population-Based Health Indexes: A Systematc Review. *Health Policy, 68*(2), 245-255. DOI: 10.1016/j.healthpol.2003.10.005

Karatzoglou, A., Smola, A., Hornik, K. (2019). *Package "kernlab"*. Verze 0.9-29. [online 2020-10-10]. Dostupné z: https://cran.r-project.org/web/packages/kernlab/kernlab.pdf

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A. (2004). kernlab-an S4 Package for Kernel Methods in R. *Journal of Statistical Software, 11*(9), 1-20.

Khanteymoori, A., Kumar, A. (2021). *Clustering in Machine Learning (Galaxy Training Materials)*. [online 2022-01-08]. Dostupné z: https://training.galaxyproject.org/archive/2021-07-01/topics/statistics/tutorials/clustering_machinelearning/tutorial.html

Kossarova, L., Holland, W., Mossialos, E. (2013). Avoidable Mortality: a Measure of Health System Performance in the Czech Republic and Slovakia between 1971 and 2008. *Health Policy and Planning, 28*(5), 508-525. DOI: 10.1093/heapol/czs093

Košťál, J. (2013). *Vybrané metody vícerozměrné statistiky (se zvláštním zaměřením na kriminologický výzkum)*. 4. svazek. Praha: Institut pro kriminologii a sociální prevenci, 114. ISBN 978-80-7338-128-8

Kruk, M. E., Gage, A. D., Joseph, N. T., Danaei, G., García-Saisó, S., Salomon, J. A. (2018). Mortality due to Low-Quality Health Systems in the Universal Health Coverage Era: a Systematic Analysis of Amenable Deaths in 137 Countries. *The Lancet, 392*(10160), 2203-2212. DOI: 10.1016/S0140-6736(18)31668-4

Kříž, K. (2016). Ztracené roky zdravého života v České republice. *Hygiena, 61*(2), 88-91. DOI: 10.21101/hygiena.a1451

Kuc, M. (2012). The Implementation of Synthetic Variable for Constructing the Standard of Living Measure in European Union Countries. *Oeconomia Copernicana. 3*(3), 5-19. DOI: 10.12775/OeC.2012.012

Lahti, L., Huovari, J., Kainu, M., Biecek, P., Antal, D., Hernangomez, D., Lehtomaki, J., Briatte, F., Stauffer, R., Rougieux, P., Vasylytsya, A., Reiter, O., Kantanen, P. (2022). *Package „eurostat".* Verze 3.7.10. [online 2022-03-11]. Dostupné z: https://cran.r-project.org/web/packages/eurostat/eurostat.pdf

Lakićević, M. (2021). Creating Maps in R (Case Study: National Park „Fruška Gora"). *Contemporary Agriculture, 70*(1-2). 41-45. DOI: 10.2478/contagri-2021-0008

Latifoğlu, F., Polat, K., Kara, S., Güneş, S. (2008). Medical Diagnosis of Atherosclerosis from Carotid Artery Doppler Signals Using Principal Component Analysis (PCA), k-NN Based

Weighting Pre-Processing and Artificial Immune Recognition System (AIRS). *Journal of Biomedical Informatics, 41*(1), 15-23.

Lebano, A., Hamed, S., Bradby, H., Gil-Salmerón, A., Durá-Ferrandis, E., Garcés-Ferrer, J. a kol. (2020). Migrants' and Refugees' Health Status and Healthcare in Europe: a Scoping Literature Review. *BMC Public Health, 20*(1), 1-22. DOI: 10.1186/s12889-020-08749-8

Lee, J. M., Yoo, C., Choi, S. W., Vanrolleghem, P. A., Lee, I. B. (2004). Nonlinear Process Monitoring Using Kernel Principal Component Analysis. *Chemical Engineering Science, 59*, 223-234. DOI: 10.1016/j.ces.2003.09.012

Lee, J. M., Qin, S. J., Lee, I. B. (2008). Fault Detection of Non-Linear Processes Uning Kernel Independent Component Analysis. *Can. J. Chem. Eng., 85*, 526-536.

di Lego, V. (2021). Health Expectancy Indicators: What Do They Measure? *Cad Saúde Colet, 2021*; *Ahead of Print*. DOI: 10.1590/1414-462X202199010376

Lemenkova, P. (2020). Using R Packages „tmap", „raster" and „ggmap" for Dartographic Visualization: An Example of Dem-Based Terrain Modelling of Italy, Apennine Peninsula. *Zbornik radova-Geografski fakultet Univerziteta u Beogradu, 68*, 99-116. DOI: 10.5937/zrgfub2068099L

Leon-Gonzalez, R., Tseng, F. M. (2011). Socio-Economic Determinants of Mortality in Taiwan: Combining Individual and Aggregate Data. *Health policy, 99*(1), 23-36. DOI: 10.1016/j.healthpol.2010.07.005

Lin, H. T., Lin, C. J. (2003). A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type methods. *Neural Computation, 3*, 1-28.

Livesley, W. J., Jang, K. L., Vernon, P. A. (1998). Phenotypic and Genetic Structure of Traits Delineating Personality Disorder. *Archives of General Psychiatry, 55*(10), 941-948.

Loop, M. S., Howard, G., de Los Campos, G., Al-Hamdan, M. Z., Safford, M. M., Levitan, E. B., McClure, L. A. (2017). Heat Maps of Hypertension, Diabetes Mellitus, and Smoking in the Continental United States. *Circulation: Cardiovascular Quality and Outcomes, 10*(1), 6. DOI: 10.1161/CIRCOUTCOMES.116.003350.

Löster, T. (2016). Determining the Optimal Number of Clusters in Cluster Analysis. *Proceedings of the10th International Days of Statistics and Economics, Prague*, 1078-1090.

Lovelace, R., Nowosad, J., Muenchow, J. (2019). *Geocomputation with R.* 1. vyd. Chapman and Hall/CRC. [online 2022-04-30]. Dostupné z: https://geocompr.robinlovelace.net/index.html

Lundberg, O., Dahl, E., Fritzell, J., Palme, J., Sjöberg, O. (2016). *Social Protection, Income and Health Inequities.* WHO Regional Office for Europe. [cit. 2017-09-03]. Dostupné z: http://www.euro.who.int/__data/assets/pdf_file/0006/302874/TG-GDP-taxes-income-welfare-final-report.pdf?ua=1

Luss, R., d'Aspremont, A. (2010). Clustering and Feature Selection Using Sparse Principal Component Analysis. *Optimization and Engineering, 11*(1), 145-157.

Mallik, S., Zhao, Z. (2019). Multi-Objective Optimized Fuzzy Clustering for Detecting Cell Clusters from Single-Cell Expression Profiles. *Genes 2019, 10*(611), 22. DOI: 10.3390/genes10080611

Marí-Dell'Olmo, M., Gotsens, M., Pasarín, M. I., Rodríguez-Sanz, M., Artazcoz, L., Garcia de Olalla, P., Rius, C., Borrell, C. (2021). Socioeconomic Inequalities in COVID-19 in a European Urban Area: Two Waves, Two Patterns. *International Journal of Environmental Research and Public Health, 18*(3), 1256-1267.

Marmot, M. G., Shipley, M., Rose, G. (1984). Inequalities in Death-Specific Explanations of a General Pattern. *Lancet, 1*(8384), 1003–1006.

Marmot, M. G. (2002). The Influence of Income on Health: Views of an Epidemiologist. *Health Affairs, 21*(2), 31–46.

Marra, C. A., Lynd, L. D., Harvard, S. S., Grubisic, M. (2011). Agreement between Aggregate and Individual-Level Measures of Income and Education: a Comparison across Three Patient Groups. *BMC Health Services Research, 11*(1), 1-7. DOI: 10.1186/1472-6963-11-69

Millstein, S. G., Irwin, C. E. Jr. (1987). Concepts of Health and Illness: Different Constructs of Variations on a Theme? *Health Psychol, 6*, 515-524.

Młodak A. (2006). *Analiza Taksonomiczna w Statystyce Regionalnej*. Centrum Doradztwa i Informacji Difin.

Mohamed, A. A. (2020). An Effective Dimension Reduction Algorithm for Clustering Arabic Text. *Egyptian Informatics Journal, 21*(1), 1-5. DOI: 10.1016/j.eij.2019.05.002

Molnár, Z., Mildeová, S., Řezanková, H., Brixí, R., Kalina, J. (2012). *Pokročilé metody vědecké práce.* 1. vyd., Profess Consulting s.r.o. ISBN 978-80-7259-064-3

Morissette, L., Chartier, S. (2013). The K-Means Clustering Technique: General Considerations and Implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology, 9*(1), 15-24. DOI 10.20982/tqmp.09.1.p015

Murtagh, F., Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification 31,* 274-295. DOI: 10.1007/s00357-014-9161-z

Niskanen, M., Silvén, O. (2003). Comparison of Dimensionality Reduction Methods for Wood Surface Inspection. *In Sixth International Conference on Quality Control by Artificial Vision, 5132,* 178-188.

Nykiforuk, C. I., Flaman, L. M. (2011). Geographic Information Systems (GIS) for Health Promotion and Public Health: a Review. *Health Promotion Practice, 12*(1), 63-73.

NZIP. (2022). *Veřejné zdraví.* Praha: Ministerstvo zdravotnictví České republiky a Ústav zdravotnických informací a statistiky České republiky. [online 2022-06-09]. Dostupné z: https://www.nzip.cz/a /rejstrikovy-pojem/1724

OECD. (2019a). *Health at a Glance 2019: OECD Indicators*, OECD Publishing, Paris. DOI: 10.1787/4dd50c09-en

OECD. (2019b). *Society at a Glance 2019: OECD Social Indicators*, OECD Publishing, Paris. DOI: 10.1787/soc_glance-2019-en

OECD. (2021a). *Health at a Glance 2021: OECD Indicators*, OECD Publishing. Paris. DOI: 10.1787/ae3016b9-en.

OECD. (2022). *List of Variables in OECD Health Statistics 2021*. [online 2022-06-09]. Dostupné z: https://www.oecd.org/els/health-systems/List-of-variables-OECD-Health-Statistics-2021.pdf

OECD/European Observatory on Health Systems and Policies. (2021). *Country Health Profiles 2021, State of Health in the EU.* OECD Publishing, Paris/European Observatory on Health Systems and Policies, Brussels. [online 2022-06-24]. Dostupné z: https://www.oecd.org/health/country-health-profiles-eu.htm

OECD/European Union. (2020). *Health at a Glance: Europe 2020: State of Health in the EU Cycle*, OECD Publishing, Paris. DOI: 10.1787/82129230-en

Pacáková, V. a kolektiv. (2015). *Štatistická indukcia pre ekonómov a manažérov*. Wolters Kluwer. ISBN 978-80-8168-081-6

Pacáková, V., Jindrová, P., Zapletal, D. (2016). Comparison of Health Care Results in Public Health Systems of European Countries. *European Financial Systems 2016, 13*, 534-541.

Pacáková, V., Žáková, N. (2019). Základní rizika předčasných úmrtí v Evropských zemích. *Slovenská štatistika a demografia, 29*(3), 1-13.

Pasin, O., Ankarali, H. (2015). Usage of Kernel K-Means and DBSCAN Cluster Algorıthms in Health Studies: an Application. *Clin Res Trials 1*. DOI: 10.15761/CRT.1000116

Pawełek B. (2008). *Metody Normalizacji Zmiennych w Badaniach Porównawczych Złożonych Zjawisk Ekonomicznych*. Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.

Pebesma, E. J. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *R Journal, 10*(1), 439-446.

Petruševski, A. B. (2013). History of Infectious Diseases Development in the Old and the Middle Ages with the Emphasis on the Plague and Leprosy. *Vojnosanitetski Pregled, 70*(7), 704-708.

Pettersson E., Turkheimer E. (2010). Item Selection, Evaluation, and Simple Structure in Personality Data. *Journal of Research in Personality, 44*, 407-420. DOI: 10.1016/j.jrp.2010.03.002

Pilario, K. E., Cao, Y., Shafiee, M. (2019). Mixed Kernel Canonical Variate Dissimilarity Analysis for Incipient Fault Monitoring in Nonlinear Dynamic Processes. *Computers and Chemical Engineering, 123*, 143-154. DOI: 10.1016/j.compchemeng.2018.12.027

Pilario, K. E., Shafiee, M., Cao, Y., Lao, L., Yang, S. H. (2020). A Review of Kernel Methods for Feature Extraction in Nonlinear Process Monitoring. *Processes, 8*(1), 24.

Porter, D. (2005). *Health, Civilization and the State: a History of Public Health from Ancient to Modern Times*. Routledge. ISBN 0-203-98057-3

Pramoditha, R. (2021). 11 Dimensionality Reduction Techniques You Should Know in 2021. *Towards Data Science*. [online 2022-02-24]. Dostupné z: https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b

Preston, S. H. (2007). The Changing Relation between Mortality and Level of Economic Development. *International Journal of Epidemiology*, *36*(3), 484–490. DOI: 10.1093/ije/dym075

Probst, P., Boulesteix, A. L., Bischl, B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *The Journal of Machine Learning Research, 20*(1), 1934-1965.

Qiu, J., Wang, H., Lu, J., Zhang, B., Du, K. L. (2012). Neural Network Implementations for PCA and its Extensions. *ISRN Artificial Intelligence*, 2012, 19. DOI: 10.5402/2012/847305

Rahmah, N., Sitanggang, I. S. (2016). Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra. *IOP Conf. Ser.: Earth Environ. Sci. 31*(012012), 1-4. DOI: 10.1088/1755-1315/31/1/012012

Raschka, S., Linear, P., Gaussian, R., & LLE, L. L. E. (2014). Kernel Tricks and Nonlinear Dimensionality Reduction via RBF Kernel PCA. *Blog, September, 24*. [online 2020-10-25]. Dostupné z: https://www.academia.edu/22711914/PAC_RBF_In_detail

Rathi, Y., Dambreville, S., Tannenbaum, A. (2006). Statistical Shape Analysis Using Kernel PCA. *In Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning, International Society for Optics and Photonics, 6064*.

Rawashdeh, M., Ralescu, A. (2012). Fuzzy Cluster Validity with Generalized Silhouettes. *MAICS, 2012*, 8.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., Baker, T. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access, 8*, 54776-54788. DOI: 10.1109/ACCESS.2020.2980942

Rehm, J., Probst, C. (2018). Decreases of Life Expectancy despite Decreases in Non-Communicable Diseases Mortality: The Role of Substance Use and Socio-Economic Status. *Eur Addict Res 2018, 24*, 53–59. DOI: 10.1159/000488328

Revelle, W. (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois. R package version 2.0.7. [online 2020-07-26]. Dostupné z: https://cran.r-project.org/web/packages/psych/psych.pdf

Reverter, F., Vegas, E., Oller, J. M. (2014). Kernel-PCA Data Integration with Enhanced Interpretability. *BMC systems biology, 8*(S2). DOI: 10.1186/1752-0509-8-S2-S6

Richman, M. B. (1986). Rotation of Principal Components. *Journal of Climatology, 6,* 293-335.

Robine, J. M., Cambois, E., Nusselder, W., Jeune, B., Oyen, H. V., Jagger, C. (2013). The joint action on healthy life years (JA: EHLEIS). *Archives of Public Health, 71*(1), 1-5.

Rodgers, G. B. (1979). "Income and Inequality as Determinants of Mortality: an International Cross-Section Analysis." *Population Stud, 33*(2), 343–351.

Rodrigues, V., Mota-Pinto, A., de Sousa, B., Botelho, A., Alves, C., de Oliveira, C. R. (2014). The Aging Profile of the Portuguese Population: a Principal Component Analysis. *Journal of Community Health, 39*(4), 747-752. DOI 10.1007/s10900-014-9821-2

Romdhani, S., Gong, S., Psarrou, A. (1999). A Multi-View Nonlinear Active Shape Model Using Kernel PCA. *In BMVC, 10,* 483-492.

Rosicova, K., Reijneveld, S. A., Geckova, A. M., Stewart, R. E., Rosic, M., Groothoff, J. W., van Dijk, J. P. (2015). Inequalities in Mortality by Socioeconomic Factors and Roma Ethnicity in the Two Biggest Cities in Slovakia: a Multilevel Analysis. *International Journal for Equity in Health, 14*(1), 1-10. DOI: 10.1186/s12939-015-0262-z

R-project. (2022). *The R Project for Statistical Computing.* [online 2022-03-12]. Dostupné z: https://www.r-project.org/

Rumel, D., Contanzo, G. (1992). General Index of Health. *Revue Canadienne de Sante Publique, 83*(1), 82-83.

Řezanková, H., Húsek, D., Snášel, V. (2009). *Shluková analýza dat*. 2. vyd. Praha: Professional Publishing. ISBN 978-80-86946-81-8

Říhová, E., Říha, D. (2019). Validation Approaches for FCM Algorithm. *The 13th International Days of Statistics and Economics, Prague*, 1281-1289.

Sander, J., Ester, M., Kriegel, H. P., Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery, 2*, 169–194.

Santana, P., Freitas, Â., Costa, C., Stefanik, I., Santinha, G., Krafft, T., Pilot, E. (2020a). The Role of Cohesion Policy Funds in Decreasing the Health Gaps Measured by the EURO-HEALTHY Population Health Index. *International Journal of Environmental Research and Public Health, 17*(5), 1-21. DOI: 10.3390/ijerph17051567

Santana, P., Freitas, Â., Stefanik, I., Costa, C., Oliveira, M., Rodrigues, T. C., Vieira, A., Ferreira, P. L., Borrell, C., Dimitroulopoulou, S., Rican, S., Mitsakou, C., Marí-Dell'Olmo, M., Schweikart, J., Corman, D., Bana e Costa, C. A. (2020b). Advancing Tools to Promote Health Equity across European Union Regions: the EURO-HEALTHY Project. *Health Research Policy and Systems, 18*(1), 1-14.

Satyanarayana, L., Indrayan, A., Sachdev, H. P. S., Gupta, S. M. (1995). A Comprehensive Index for Longitudinal Monitoring of Child Health Status. *Indian Pediatrics, 32*, 443-452.

Schölkopf, B., Smola, A., Müller, K. R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation, 10*, 1299-1319.

Sharma, A., Gupta, R. K., Tiwari, A. (2016). Improved Density Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data. *Mathematical Problems in Engineering, 2016*, 9. DOI: 10.1155/2016/1564516

Shawe-Taylor, J., Sun, S. (2014). Kernel Methods and Support Vector Machines. *In Academic Press Library in Signal Processing, 1*, 857-881. DOI: 10.1016/B978-0-12-396502-8.00016-4

Shawn, M. (2006). An Approximate Version of Kernel PCA. *In 2006 5th International Conference on Machine Learning and Applications (ICMLA'06),* 239-244.

Sheng, J., Dong, S., Liu, Z., Gao, H. (2016). Fault Feature Extraction Method Based on Local Mean Decomposition Shannon Entropy and Improved Kernel Principal Component Analysis Model. *Advances in Mechanical Engineering, 8*(8), 1-8. DOI: 10.1177/1687814016661087

Shiokawa, Y., Kikuchi, J. (2018). Application of Kernel Principal Component Analysis and Computational Machine Learning to Exploration of Metabolites Strongly Associated with Diet. *Scientific reports, 8*(1), 1-8. DOI: 10.1038/s41598-018-20121-w

Siddique, M., Bakr, A., Arif, R. B., Khan, M. M. R., Ashrafi, Z. (2018). Implementation of Fuzzy C-Means and Possibilistic C-Means Clustering Algorithms, Cluster Tendency Analysis and Cluster Validation. *arXiv preprint arXiv:1809.08417*. 8. DOI: 10.20944/preprints201811.0581.v1

Sing, T., Svicher, V., Beerenwinkel, N., Ceccherini-Silberstein, F., Däumer, M., Kaiser, R et al. (2005). Characterization of Novel HIV Drug Resistance Mutations Using Clustering, Multidimensional Scaling and SVM-Based Feature Ranking. *In European Conference on Principles of Data Mining and Knowledge Discovery, Springer, Berlin, Heidelberg*, 285-296.

Sjöstrand, K., Stegmann, M. B., Larsen, R. (2006). Sparse Principal Component Analysis in Medical Shape *Modeling. In Proceedings of SPIE - The International Society for Optical Engineering*, 6144. DOI: 10.1117/12.651658

Sokal, R. R., Rohlf, F. J. (1962). The Comparison of Dendrograms by Objective Methods. *Taxon 11*(2), 33-40.

Souto, T., S., Ramires, A., Leite, Â., Santos, V., Santo, R., E. (2018). Health Perception: Validation of a Scale for the Portuguese Population. *Temas em Psicologia, 26*(4), 2167-2183.

Spruyt, K., O'Brien, L. M., Coxon, A. M., Cluydts, R., Verleye, G., Ferri, R. (2006). Multidimensional Scaling of Pediatric Sleep Breathing Problems and Bio-Behavioral Correlates. *Sleep Medicine, 7*(3), 269-280.

Staníčková M., Melecký L. (2018). Understanding of Resilience in the Context of Regional Development Using Composite Index Approach: the Case of European Union NUTS-2 Regions, *Regional Studies, Regional Science, 5*(1), 231-254. DOI: 10.1080/21681376.2018.1470939

Stankovičová, I., Vojtková, M. (2007). *Viacrozmerné štatistické metódy s aplikáciami*. Iura Edition, Bratislava. ISBN 978-80-8078-152-1

Subbalakshmi, C., Krishna, G. R., Rao, S. K. M., Rao P. V. (2015). A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set. *Procedia Computer Science, 46*, 346–353. DOI: 10.1016/j.procs.2015.02.030

Svalastog, A. L., Donev, D., Kristoffersen, N. J., Gajović, S. (2017). Concepts and Definitions of Health and Health-Related Values in the Knowledge Landscapes of the Digital Society. *Croatian Medical Journal, 58*(6), 431-435.

Székely G. J., Rizzo, M. L. (2005). Hierarchical Clustering via Joint between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification, 22*, 151-183. DOI 10.1007/s00357-005-0012-9

Škaloudová, A. (2010). *Faktorová analýza, základní pojmy*. Univerzita Karlova, Pedagogická fakulta. [online]. [2020-08-07]. Dostupné z: http://kps.pedf.cuni.cz/skalouda/fa/zakl_pojmy.htm

Tabachnick, B. G., Fidell, F. S. (2007). *Using Multivariate Statistics*. Pearson, Boston. ISBN 0-205-45939-2

Terek, M., Horníková, A., Labudová, V. (2010). *Hĺbková analýza údajov*. Bratislava: IURA EDITION. ISBN 978-80-8078-336-5

Thorpe, M. G., Milte, C. M., Crawford, D., McNaughton, S. A. (2016). A Comparison of the Dietary Patterns Derived by Principal Component Analysis and Cluster Analysis in Older Australians. *International Journal of Behavioral Nutrition and Physical Activity, 13*(1), 30. DOI 10.1186/s12966-016-0353-2

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B, 58,* 267–288.

Timmis, A., Vardas, P., Townsend, N., Torbica, A., Katus, H., De Smedt, D., ..., Atlas Writing Group, European Society of Cardiology. (2022). European Society of Cardiology: cardiovascular disease statistics 2021. *European heart journal, 43*(8), 716-799. DOI: 10.1093/eurheartj/ehab892

Tango, T. (2010). *Statistical Methods for Disease Clustering*. Springer Science & Business Media.

ÚZIS. (2020). *Registry a sběr dat: Mezinárodní klasifikace nemocí.* [online 2020-07-08]. Dostupné z: https://www.uzis.cz/index.php?pg=registry-sber-dat--klasifikace--mezinarodni-klasifikace-nemoci

Van Der Maaten, L., Postma, E., Van den Herik, J. (2009). Dimensionality Reduction: a Comparative Review. *J. Mach. Learn. Res., 10*(66-71), 13.

Vats, D., Sharma, A. (2020). Dimensionality Reduction Techniques: Comparative Analysis. *Journal of Computational and Theoretical Nanoscience, 17,* 2687-2691. DOI: 10.1166/jctn.2020.8967

de Vaus, D. (2002). *Analyzing Social Science Data. 50 Key Problems in Data Analysis.* Sage, London. ISBN-13: 978-0761959373

Walesiak, M. (2016). Visualization of Linear Ordering Results for Metric Data with the Application of Multidimensional Scaling. *Ekonometria*, (52), 9-21. DOI: 10.15611/ekt.2016.2.01

Walesiak, M., Dehnel, G. (2018). Evaluation of Economic Efficiency of Small Manufacturing Enterprises in Districts of Wielkopolska Province Using Interval-Valued Symbolic Data and the Hybrid Approach. *In Proceedings of the 12th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Cracow: Foundation of the Cracow University of Economics*, 563-572.

Walesiak, M., Dehnel, G. (2019). A Comparative Analysis of Rankings of Polish Provinces in Terms of Social Cohesion for Metric and Interval-Valued Data. *In Proceedings of the 12th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Cracow: Foundation of the Cracow University of Economics*, 250-258.

Walesiak, M., Dehnel, G. (2020). The Measurement of Social Cohesion at Province Level in Poland Using Metric and Interval-Valued Data. *Sustainability, 12*(18), 19. DOI: 10.3390/su12187664

Walesiak, M., Dudek, A. (2022). Package *"clusterSim"*. Verze 0.50-1. [online 2022-06-16]. Dostupné z: https://cran.r-project.org/web/packages/clusterSim/clusterSim.pdf

Wang, F. (2020). Why Public Health Needs GIS: a Methodological Overview. *Annals of GIS, 26*(1), 1-12. DOI: 10.1080/19475683.2019.1702099

Wang, W., Zhang, M., Wang, D., & Jiang, Y. (2017). Kernel PCA Feature Extraction and the SVM Classification Algorithm for Multiple-Status, Through-Wall, Human Being Detection. *EURASIP Journal on Wireless Communications and Networking*, (1), 1-7. DOI 10.1186/s13638-017-0931-2

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58*(301), 236-244.

Weide, A. C., Beauducel, A. (2019). Varimax Rotation Based on Gradient Projection Is a Feasible Alternative to SPSS. *Front. Psychol., 10*(645). DOI: 10.3389/fpsyg.2019.00645

WHO. (1995). *Constitution of the world health organization*. [online 2020-03-15]. Dostupné z: https://apps.who.int/iris/bitstream/handle/10665/121457/em_rc42_cwho_en.pdf

WHO. (2017). *Determinants of Health*. [online 2022-05-07]. Dostupné z: https://www.who.int/news-room/questions-and-answers/item/determinants-of-health

WHO. (2020a). Coronavirus. *World Health Organization*. [online 2020-03-16]. Dostupné z: https://www.who.int/health-topics/coronavirus

WHO. (2020b). *WHO Methods and Data Sources for Global Burden of Disease Estimates 2000-2019*. Global Health Estimates Technical Paper WHO/ DDI/DNA/GHE/2020.3. [online 2022-09-10] Dostupné z: https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2019_daly-methods.pdf?sfvrsn=31b25009_7

WHO. (2022a). *Age-Standardized DALY's*. [online 2022-02-23]. Dostupné z: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/158

WHO. (2022b). *European Health for All database (HFA-DB).* [online 2022-06-13]. Dostupné z: https://gateway.euro.who.int/en/datasets/european-health-for-all-database/

WHO. (2022c). *Global Health Estimates 2019: Disease Burden by Cause, Age, Sex, by Country and by Region, 2000-2019*. Geneva. [online 2022-06-18]. Dostupné z: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/global-health-estimates-leading-causes-of-dalys

WHO. (2022d). *Healthy Life Expectancy (HALE) at Birth (Years).* [online 2022-02-19]. Dostupné z: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/gho-ghe-hale-healthy-life-expectancy-at-birth

Wickham, H., Chang, W., Henry, L., Pedersen T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D. (2021). *Package „ggplot2". Create Elegant Data Visualisations Using the Grammar of Graphics*. R Package Version 3.3.5. [online 2021-06-25]. Dostupné z: https://cloud.r-project.org/web/packages/ggplot2/ggplot2.pdf

Wickham, H., François, R., Henry, L., Müller, K. (2022). *Package „dplyr". A Grammar of Data Manipulation.* R Package Version 1.0.9. [online 2022-04-27]. Dostupné z: https://cloud.r-project.org/web/packages/dplyr/dplyr.pdf

Wilkins E., Wilson L., Wickramasinghe K., Bhatnagar P., Leal J., Luengo-Fernandez R., Burns R., Rayner M., Townsend N. (2017). *European Cardiovascular Disease Statistics 2017*. European Heart Network, Brussels.

Wilkinson, R. (1992). Income Distribution and Life Expectancy. *British Medical Journal, 304*(6824), 165–168.

Wilkinson R., Pickett K. (2006). Income Inequality and Population Health: a Review and Explanation of the Evidence. *Soc Sci Med., 62*(7), 1768–1784. DOI: 10.1016/j.socscimed.2005.08.036

Williams, C. K. I. (2002). On a Connection between Kernel PCA and Metric Multidimensional Scaling. *Machine Learning, 46*, 11-19.

Wilson, A. G., Hu, Z., Salakhutdinov, R. R., Xing, E. P. (2016). Stochastic Variational Deep Kernel Learning. *In Advances in Neural Information Processing Systems*, 2586-2594.

Wolfson, M. C. (1996). Health-Adjusted Life Expectancy. *Health Reports, 8*(1), 41-46.

Woodside, J. M. (2016). Bemo: A Parsimonious Big Data Mining Methodology. *AJIT-e, 7*(24), 113-123. DOI: 10.5824/1309-1581.2016.3.007.x

Xia, J., Zhang, Y., Song, J., Chen, Y., Wang, Y., Liu, S. (2021). Revisiting Dimensionality Reduction Techniques for Visual Cluster Analysis: an Empirical Study. *IEEE Transactions on Visualization and Computer Graphics, 28*(1), 529-539.

Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., Chen, X. (2021). A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Frontiers in genetics, 12*, 12. DOI: 10.3389/fgene.2021.646936

Xie, H., Li, J., Zhang, Q., Wang, Y. (2016). Comparison among Dimensionality Reduction Techniques Based on Random Projection for Cancer Classification. *Computational Biology and Chemistry*, *65*, 165-172. DOI: 10.1016/j.compbiolchem.2016.09.010

Zahra, A., Lee E. W., Sun L. Y., Park, J. H. (2015). Perception of Lay People Regarding Determinants of Health and Factors Affecting It: an Aggregate Analysis from 29 Countries. *Iran J Public Health, 44*(12), 1620-1631.

Zákon č. 258/2000 Sb., o ochraně veřejného zdraví a o změně některých souvisejících zákonů. *Sbírka zákonů České republiky*. [online 2022-06-09]. Dostupné z: https://www.zakonyprolidi.cz/cs/2000-258

Zeliaś A., Malina A. (1997). O Budowie Taksonomicznej Miary Jakości Życia. Syntetyczna Miara Rozwoju Jest Narzędziem Statystycznej Analizy Porównawczej, *Taksonomia*, *4*, 238-262.

Zheng, X., Xu, X., Liu, Y., Xu, Y., Pan, S., Zeng, X., Yi, Q., Xiao, N., Lin, L. (2020). Age-Standardized Mortality, Disability-Adjusted Life-Years and Healthy Life Expectancy in Different Cultural Regions of Guangdong, China: a Population-Based Study of 2005-2015. *BMC Public Health, 20*(858), 1-20. DOI: 10.1186/s12889-020-8420-7

Zou, H., Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B, 67*, 301–320.

Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics, 15*(2), 265-286.

Přehled publikační činnosti autora

Gogola, J., Kopecká, L. (2017). Multiple State Models for Critical Illness Policy. *In European Financial Systems 2017: Proceedings of the 14th International Scientific Conference. Brno: Masaryk University*, 159-165.

Gogola, J., Kopecká, L. (2018). Actuarial Model for Pricing Disability Insurance Policy. *In European Financial Systems 2018: Proceedings of the 15th International Scientific Conference. Brno: Masaryk University*, 120-126.

Jindrová, P., Kopecká, L. (2017a). Assessment of Risk Factors of Serious Diseases in OECD Countries. *In Proceedings of the 11th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Cracow: Foundation of the Cracow University of Economics*, 123-132.

Jindrová, P., Kopecká, L. (2017b). Empirical Bayes Credibility Models for Economic Catastrophic Losses by Regions. *In ITM Web of Conferences. Les Ulis: EDP Sciences, 9*(01006), 1-6. DOI: 10.1051/itmconf/20170901006

Jindrová, P., Kopecká, L. (2017c). Kvantifikace rizik pro úrazové pojištění. *Scientific Papers of the University of Pardubice - Series D, Faculty of Economics and Administration, 24*(39), 75-86.

Kopecká, L. (2018a) Bayesian Estimates of Probability of Incidences and Mortalities of Selected Serious Diseases. *Herald of Lviv University of Trade and Economics. Economic Sciences, 54*, 59-65.

Kopecká, L. (2018b). Comparison of Mortality Caused by Serious Diseases within the Czech Republic. *Scientific Papers of the University of Pardubice – Series D, Faculty of Economics and Administration, 43*(2), 112 – 122.

Kopecká, L. (2018c). Vliv věkové struktury obyvatelstva na výskyt a léčbu chronických onemocnění v rámci Evropy. *Slovak Statistical and Demographic Society: Forum Statisticum Slovacum, 14*(1), 19 – 28.

Kopecká, L. (2019). Posouzení stavu zdraví a nerovností ve stavu zdraví populace v zemích EU-28. *Slovenská štatistika a demografia, 29*(3), 28-40.

Kopecká, L., Jindrová, P. (2017). Comparison of Mortality due to Critical Illnesses in the EU Countries. *In Proceedings of the 11th Professor Aleksander Zelias International Conference*

on Modelling and Forecasting of Socio-Economic Phenomena. Cracow: Foundation of the Cracow University of Economics, 133-142.

Kopecká, L., Pacáková, V. (2017). Bayesian Estimation of Probability of Incidences of the Most Serious Oncological Diseases in the Czech Republic. *In Proceedings of the 11th International Scientific Conference Financial Management of Firms and Financial Institutions, Ostrava*, 407-414.

Kopecká, L. Zapletal, D. Pacáková, V. (2020). Comparison of Incidences of Serious Diseases within Regions in the Czech Republic. *Ekonomický časopis SAV, 68*(4), 409-428.

Pacáková, V., Jindrová, P., Kopecká, L. (2019). Statistical Analysis of Health Inequalities in European Countries. *In ITM Web of Conferences, International Conference on Applied Mathematics, Computational Science and Systems Engineering, 24*(02002), 1-9. DOI: 10.1051/itmconf/20192402002

Pacáková, V., Kopecká, L. (2018a). Comparing Inequalities in Health Outcomes in European Countries. *Journal of International Studies, 11*(4), 215-227. DOI: 10.14254/2071-8330.2018/11-4/15

Pacáková, V., Kopecká, L. (2018b). Health and Economic Risks of Longevity in European Countries. *In Proceedings of the 9th International Scientific Conference – Managing and Modelling of Financial Risks. Ostrava*, 380-387.

Pacáková, V., Kopecká, L. (2018c). Inequalities in Health Status Depending on Socio-Economic Situation in the European Countries. *E+M Economics and Management, 21*(2), 4-20. DOI: 10.15240/tul/001/2018-2-001.

Pacáková, V. Kopecká, L. (2019a). Assessment of the Impact of Socio-Economic Situation on Health Status of Inhabitants in the European Union Countries. *In Proceedings of the 13th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena: Conference Proceedings, Warsaw: C.H. Beck*, 169-177.

Pacáková, V. Kopecká, L. (2019b). Comparison of Serious Diseases Mortality in Regions of V4. *In 12th Scientific Meeting Clasification and Data Analysis: Book of Short Papers. Cassino: Università di Cassino e del Lazio Meridionale*, 365-368.

Pacáková V., Šild, P., Zapletalová, L. (2021). Demographics and Social Factors of Unmet Health Care Needs and Avoidable Mortality in European Union Countries. *In RELIK 2021:*

reprodukce lidského kapitálu – vzájemné vazby a souvislosti. Praha: Vysoká škola ekonomická v Praze, 569-578.

Pacáková, V. Zapletalová, L. Šild, P. (2020). Úmrtnost na závažné nemoci podle demografických charakteristik v zemích EU-28. *In RELIK 2020: reprodukce lidského kapitálu – vzájemné vazby a souvislosti. Praha: Vysoká škola ekonomická v Praze*, 435-444.

Pilyavskyy, A. I., Kopecká, L. (2018). The Efficiency of Health Care Systems in OECD Countries. Does it make a difference? *In European Financial Systems 2018: proceedings of the 15th International Scientific Conference. Brno: Masaryk University,* 530-535.

Zapletal, D. Kopecká, L. (2019). Application of Survival Analysis to Critical Illness Insurance Data. *In 12th Scientific Meeting Clasification and Data Analysis: book of short papers. Cassino: Università di Cassino e del Lazio Meridionale*, 472-475.

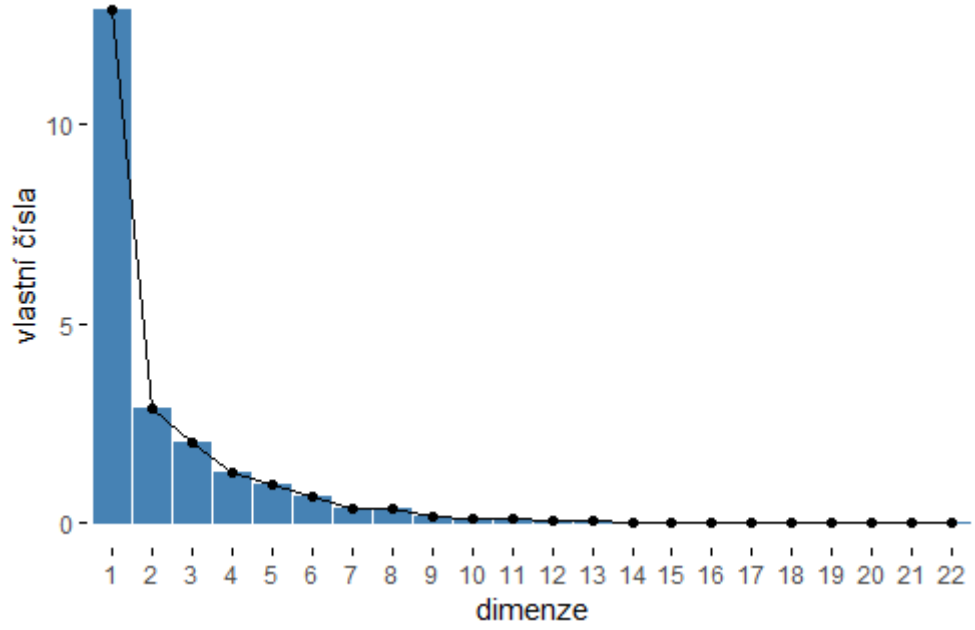
Seznam příloh

Příloha 1: Kódy v programu R

viz. přiložené CD

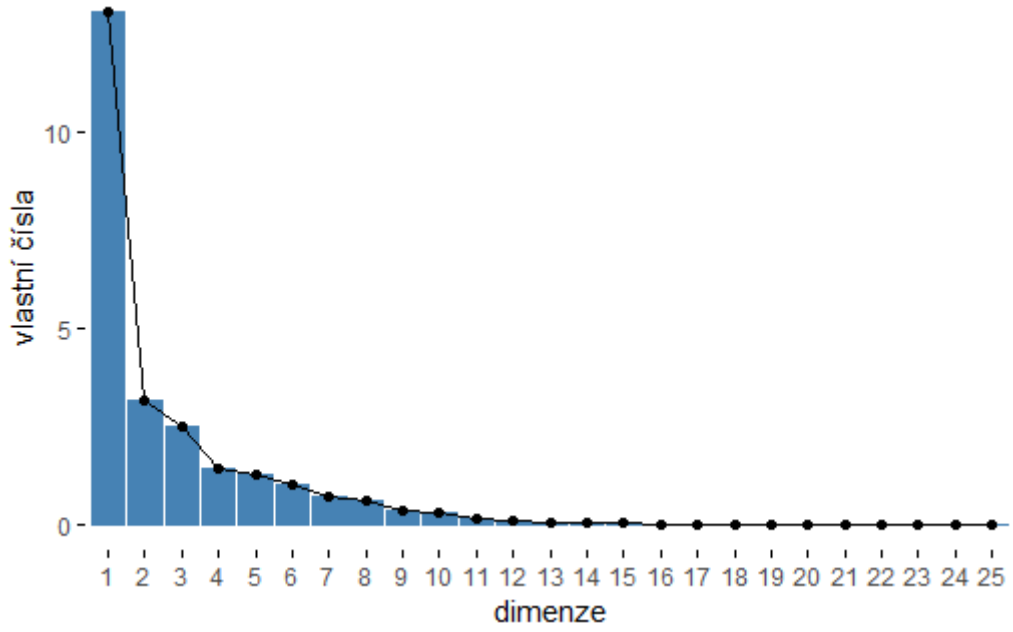
Příloha 2: Sutinové grafy pro stav zdraví

Obrázek 21: Sutinový graf, 22 proměnných, standardizovaná data „z-skóre“, stav zdraví



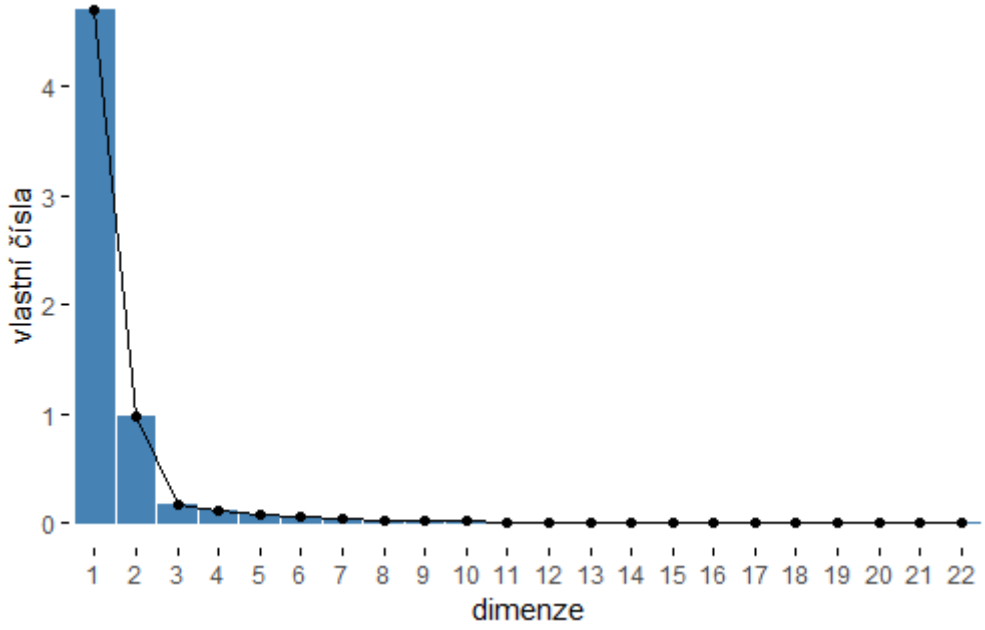
Zdroj: vlastní zpracování v programu

Obrázek 22: Sutinový graf, 25 proměnných, standardizovaná data „z-skóre“, stav zdraví



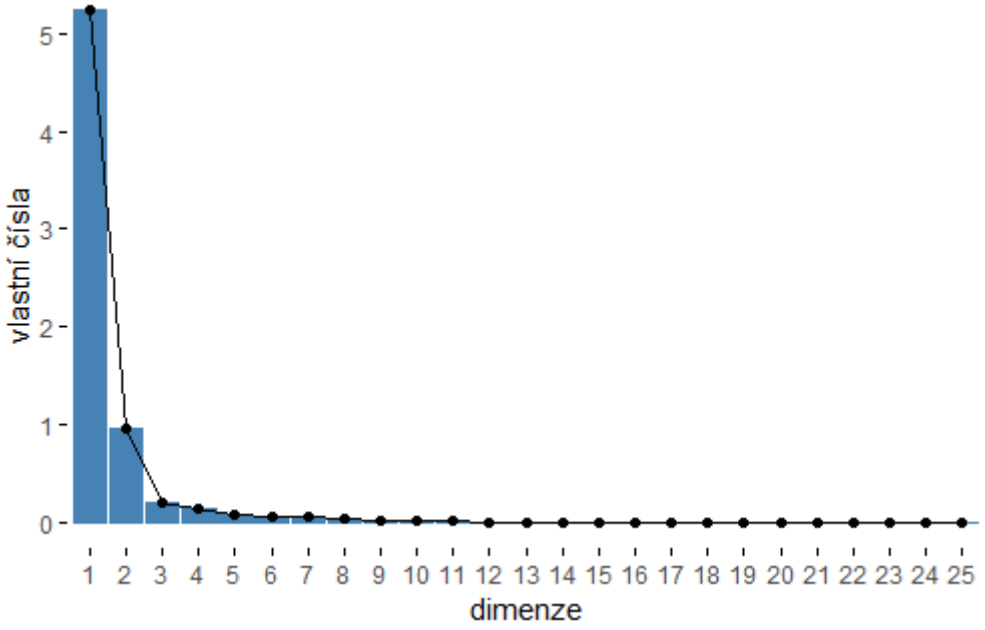
Zdroj: vlastní zpracování v programu R

Obrázek 23: Sutinový graf, 22 proměnných, standardizovaná data „min-max“, stav zdraví



Zdroj: vlastní zpracování v programu R

Obrázek 24: Sutinový graf, 25 proměnných, standardizovaná data „min-max“, stav zdraví



Zdroj: vlastní zpracování v programu R

Příloha 3: Rotované komponentní zátěže pro stav zdraví

Tabulka 19: Komponentní zátěže po varimax rotaci, standardizace „z-skóre“, stav zdraví

22 proměnných	RC1	RC2	RC3	RC4	h2	u2	com
HLY_0	0,24	-0,24	-0,18	0,86	0,89	0,11	1,4
HLY_65	0,39	-0,08	0,15	0,85	0,90	0,10	1,5
LE_0	0,90	-0,11	0,20	0,30	0,95	0,05	1,4
LE_65	0,88	-0,19	0,21	0,24	0,91	0,09	1,4
HALE_0	0,92	-0,13	0,17	0,27	0,96	0,04	1,3
HALE_60	0,90	-0,22	0,19	0,24	0,95	0,05	1,4
SDR1_0	-0,45	0,71	-0,25	-0,40	0,92	0,08	2,7
SDR1_65	-0,16	0,83	-0,15	-0,35	0,87	0,13	1,5
SDR3_0	0,60	0,43	0,21	0,46	0,81	0,19	3,1
SDR3_65	0,60	0,38	0,20	0,50	0,79	0,21	3,0
SDR4_0	0,26	-0,03	0,93	-0,01	0,94	0,06	1,2
SDR4_65	0,27	-0,04	0,93	0,01	0,93	0,07	1,2
SDR5_0	-0,96	-0,03	-0,17	-0,17	0,98	0,02	1,1
SDR5_65	-0,96	-0,02	-0,17	-0,17	0,98	0,02	1,1
SDR6_65	0,58	-0,01	-0,47	0,34	0,68	0,32	2,6
SDR7_P	-0,84	0,25	-0,13	-0,38	0,93	0,07	1,7
SDR7_T	-0,90	0,01	-0,25	-0,25	0,93	0,07	1,3
DALY	-0,89	0,08	-0,10	-0,25	0,87	0,13	1,2
PREV_CAD	-0,75	0,27	-0,08	-0,15	0,67	0,33	1,4
INC_CA	-0,09	0,82	0,17	0,05	0,71	0,29	1,1
REP_AST	0,54	0,35	0,60	0,31	0,86	0,14	3,2
PER_HEALTH	0,49	-0,29	-0,01	0,54	0,61	0,39	2,5
25 proměnných	**RC1**	**RC2**	**RC3**	**RC4**	**h2**	**u2**	**com**
HLY_0	0,47	0,59	-0,26	-0,37	0,78	0,23	3,1
HLY_65	0,70	0,29	-0,14	-0,40	0,76	0,24	2,1
LE_0	0,94	0,00	-0,22	-0,08	0,94	0,06	1,1
LE_65	0,89	-0,04	-0,29	-0,14	0,90	0,10	1,3
HALE_0	0,93	0,01	-0,25	-0,01	0,93	0,07	1,1
HALE_60	0,90	-0,03	-0,34	-0,05	0,92	0,08	1,3
SDR1_0	-0,51	0,03	0,78	0,11	0,89	0,11	1,8
SDR1_65	-0,20	-0,02	0,89	0,06	0,83	0,17	1,1
SDR2_65	-0,07	0,13	0,13	0,57	0,36	0,64	1,2
SDR3_0	0,81	0,03	0,28	0,21	0,78	0,22	1,4
SDR3_65	0,81	0,06	0,22	0,20	0,76	0,25	1,3
SDR4_0	0,47	-0,77	-0,07	-0,12	0,83	0,18	1,7
SDR4_65	0,48	-0,75	-0,08	-0,12	0,82	0,18	1,8
SDR5_0	-0,95	0,03	0,10	-0,05	0,92	0,08	1,0
SDR5_65	-0,95	0,02	0,11	-0,05	0,92	0,08	1,0
SDR6_0	0,37	0,70	-0,11	0,39	0,79	0,21	2,2
SDR6_65	0,52	0,65	-0,14	0,38	0,85	0,15	2,7
SDR7_P	-0,88	-0,10	0,37	0,04	0,92	0,08	1,4

25 proměnných	RC1	RC2	RC3	RC4	h2	u2	com
SDR7_T	-0,95	0,07	0,12	0,05	0,92	0,08	1,0
DALY	-0,89	-0,09	0,20	-0,03	0,84	0,16	1,1
PREV_CAD	-0,70	-0,03	0,42	-0,31	0,76	0,24	2,1
INC_CA	0,11	-0,10	0,84	-0,07	0,72	0,28	1,1
PREV_DIA	0,13	0,13	-0,14	0,74	0,60	0,40	1,2
REP_AST	0,79	-0,35	0,24	-0,02	0,81	0,19	1,6
PER_HEALTH	0,60	0,34	-0,33	-0,26	0,66	0,34	2,7

Zdroj: vlastní zpracování v programu R

Tabulka 20: Komponentní zátěže po varimax rotaci, standardizace min-max, stav zdraví

22 proměnných	RC1	RC2	RC3	h2	u2	com
HLY_0	0,56	-0,38	-0,39	0,60	0,40	2,6
HLY_65	0,72	-0,24	-0,05	0,58	0,42	1,2
LE_0	0,92	-0,23	0,21	0,94	0,06	1,2
LE_65	0,86	-0,31	0,24	0,89	0,11	1,4
HALE_0	0,91	-0,25	0,19	0,93	0,07	1,2
HALE_60	0,87	-0,33	0,22	0,92	0,08	1,4
SDR1_0	-0,48	0,80	-0,19	0,90	0,10	1,8
SDR1_65	-0,18	0,89	-0,08	0,83	0,17	1,1
SDR3_0	0,81	0,30	0,13	0,76	0,24	1,3
SDR3_65	0,81	0,25	0,11	0,73	0,27	1,2
SDR4_0	0,26	-0,06	0,92	0,91	0,09	1,2
SDR4_65	0,27	-0,07	0,91	0,91	0,09	1,2
SDR5_0	-0,93	0,08	-0,22	0,92	0,08	1,1
SDR5_65	-0,93	0,09	-0,22	0,92	0,08	1,1
SDR6_65	0,64	-0,11	-0,49	0,66	0,34	1,9
SDR7_P	-0,88	0,38	-0,11	0,93	0,07	1,4
SDR7_T	-0,91	0,13	-0,26	0,92	0,08	1,2
DALY	-0,89	0,20	-0,12	0,84	0,16	1,1
PREV_CAD	-0,70	0,35	-0,12	0,62	0,38	1,5
INC_CA	0,08	0,80	0,12	0,67	0,33	1,1
REP_AST	0,69	0,24	0,53	0,81	0,19	2,2
PER_HEALTH	0,63	-0,41	-0,11	0,57	0,43	1,8
25 proměnných	**RC1**	**RC2**	**RC3**	**h2**	**u2**	**com**
HLY_0	0,38	0,35	-0,51	0,53	0,47	2,7
HLY_65	0,65	0,1	-0,35	0,55	0,45	1,6
LE_0	0,92	0,01	-0,29	0,93	0,07	1,2
LE_65	0,87	-0,06	-0,36	0,89	0,11	1,3
HALE_0	0,91	0,05	-0,3	0,93	0,07	1,2
HALE_60	0,88	-0,01	-0,38	0,92	0,08	1,4
SDR1_0	-0,48	0,11	0,79	0,87	0,13	1,7
SDR1_65	-0,17	0,07	0,86	0,77	0,23	1,1
SDR2_65	-0,06	0,39	0,25	0,21	0,79	1,8

25 proměnných	RC1	RC2	RC3	h2	u2	com
SDR3_0	0,82	0,21	0,25	0,78	0,22	1,3
SDR3_65	0,82	0,23	0,19	0,76	0,25	1,3
SDR4_0	0,54	-0,7	0,05	0,78	0,22	1,9
SDR4_65	0,55	-0,68	0,03	0,77	0,23	1,9
SDR5_0	-0,95	-0,06	0,14	0,92	0,08	1,1
SDR5_65	-0,94	-0,06	0,15	0,92	0,08	1,1
SDR6_0	0,30	0,81	-0,19	0,79	0,21	1,4
SDR6_65	0,45	0,77	-0,21	0,85	0,15	1,8
SDR7_P	-0,85	-0,1	0,44	0,92	0,08	1,5
SDR7_T	-0,94	0,02	0,17	0,92	0,08	1,1
DALY	-0,87	-0,14	0,26	0,84	0,16	1,2
PREV_CAD	-0,69	-0,19	0,36	0,64	0,36	1,7
INC_CA	0,14	-0,04	0,78	0,63	0,37	1,1
PREV_DIA	0,14	0,46	0,03	0,24	0,76	1,2
REP_AST	0,83	-0,24	0,25	0,81	0,19	1,3
PER_HEALTH	0,54	0,19	-0,5	0,58	0,42	2,2

Zdroj: vlastní zpracování v programu R

Příloha 4: Řídké komponentní zátěže pro stav zdraví

Tabulka 21: Řídké komponentní zátěže, standardizace „z-skóre“, stav zdraví

22 proměnných	SPC1	SPC2	SPC3	SPC4	com
HLY_0	0,04	-0,02	-0,01	0,64	1,01
HLY_65	0	0	0	0,59	1,00
LE_0	-0,28	0	0	0	1,00
LE_65	-0,21	0	0	0	1,00
HALE_0	-0,31	0	0	0	1,00
HALE_60	-0,31	0	0	0	1,00
SDR1_0	0,07	0,24	-0,20	-0,16	2,89
SDR1_65	0	0,55	-0,16	-0,22	1,51
SDR3_0	-0,07	0,39	0	0,16	1,42
SDR3_65	-0,10	0,27	0	0,30	2,20
SDR4_0	-0,03	0,05	0,61	0	1,02
SDR4_65	-0,06	0,03	0,58	0	1,03
SDR5_0	0,37	0	0	0	1,00
SDR5_65	0,37	0	0	0	1,00
SDR6_65	-0,13	0	-0,41	0,09	1,32
SDR7_P	0,26	0	0	-0,01	1,01
SDR7_T	0,29	0	0	0	1,00
DALY	0,36	0	0	0	1,00
PREV_CAD	0,32	0,04	0	0	1,03
INC_CA	0,07	0,56	0	0	1,03
REP_AST	0,00	0,36	0,24	0	1,75
PER_HEALTH	-0,02	-0,06	-0,03	0,26	1,13
25 proměnných	**SPC1**	**SPC2**	**SPC3**	**SPC4**	**com**
HLY_0	-0,04	-0,29	0	0,51	1,62
HLY_65	-0,19	0	0	0,39	1,46
LE_0	-0,28	0	0	0	1,00
LE_65	-0,23	0	0	0	1,00
HALE_0	-0,30	0	0	0	1,00
HALE_60	-0,28	0	0	0	1,00
SDR1_0	0,22	0	0,38	0	1,58
SDR1_65	0,08	0,01	0,60	0	1,04
SDR2_65	0	-0,15	0,08	-0,40	1,36
SDR3_0	-0,19	-0,03	0,35	-0,01	1,56
SDR3_65	-0,21	-0,05	0,26	-0,03	2,02
SDR4_0	-0,14	0,45	0	-0,06	1,23
SDR4_65	-0,17	0,42	0	-0,05	1,36
SDR5_0	0,26	0	0	0	1,00
SDR5_65	0,26	0	0	0	1,00
SDR6_0	-0,07	-0,48	0,08	-0,08	1,16

25 proměnných	SPC1	SPC2	SPC3	SPC4	com
SDR6_65	-0,12	-0,45	0,01	-0,02	1,16
SDR7_P	0,28	0	0	0	1,00
SDR7_T	0,23	0	0	0	1,00
DALY	0,25	0	0	0	1,00
PREV_CAD	0,26	0,01	0,06	0,27	2,10
INC_CA	0	0,10	0,53	0,08	1,11
PREV_DIA	-0,05	-0,20	0	-0,52	1,31
REP_AST	-0,21	0,11	0,17	0	2,52
PER_HEALTH	-0,15	-0,15	-0,08	0,24	2,76

Zdroj: vlastní zpracování v programu R

Tabulka 22: Řídké komponentní zátěže, standardizace „min-max“, stav zdraví

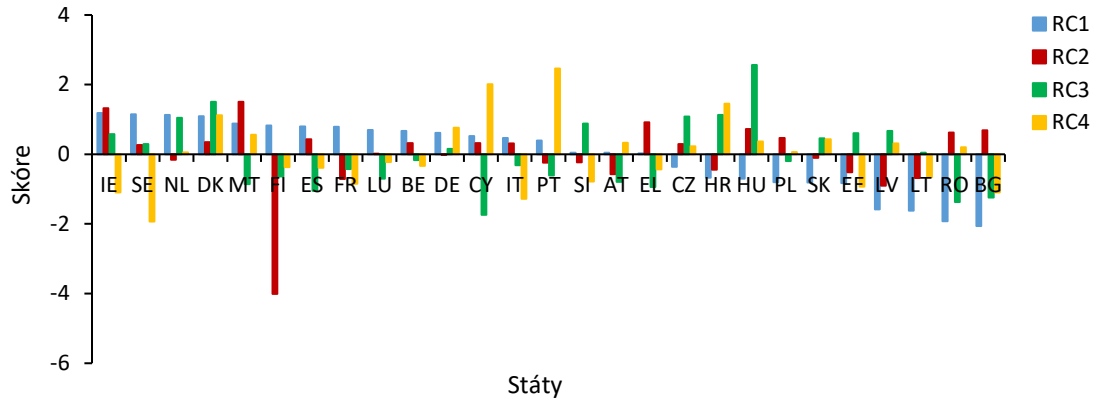
22 proměnných	SPC1	SPC2	SPC3	com
HLY_0	-0,20	-0,12	0,51	1,42
HLY_65	-0,24	0	0	1,00
LE_0	-0,49	0	0	1,00
LE_65	-0,34	0	0	1,00
HALE_0	-0,45	0	0	1,00
HALE_60	-0,27	0	0	1,00
SDR1_0	0	-0,25	-0,11	1,39
SDR1_65	0	-0,10	-0,71	1,04
SDR3_0	-0,21	0	-0,20	1,99
SDR3_65	-0,22	0	-0,10	1,44
SDR4_0	0	0	-0,02	1,00
SDR4_65	0	0	0,00	1,00
SDR5_0	0	-0,34	0	1,00
SDR5_65	0	-0,34	0	1,00
SDR6_65	-0,22	-0,04	0,06	1,23
SDR7_P	0	-0,33	0	1,00
SDR7_T	0	-0,53	0,02	1,00
DALY	0	-0,33	0	1,00
PREV_CAD	0	-0,44	0	1,00
INC_CA	-0,03	0	-0,37	1,01
REP_AST	-0,28	0	-0,32	1,97
PER_HEALTH	-0,32	0	0,11	1,23
25 proměnných	**SPC1**	**SPC2**	**SPC3**	**com**
HLY_0	-0,01	0	-0,37	1,00
HLY_65	-0,15	0	0	1,00
LE_0	-0,52	0	0	1,00
LE_65	-0,34	0	0	1,00
HALE_0	-0,50	0	0	1,00

25 proměnných	SPC1	SPC2	SPC3	com
HALE_60	-0,28	-0,03	0	1,02
SDR1_0	0	0,37	0	1,00
SDR1_65	-0,11	0,30	0,14	1,71
SDR2_65	0	0	0	-
SDR3_0	-0,26	0	0	1,00
SDR3_65	-0,24	0	0	1,00
SDR4_0	-0,01	0	0,07	1,02
SDR4_65	-0,02	0	0,05	1,16
SDR5_0	0	0,25	0	1,00
SDR5_65	0	0,26	0	1,00
SDR6_0	0	0	-0,69	1,00
SDR6_65	0	0	-0,54	1,00
SDR7_P	0	0,43	0	1,00
SDR7_T	0,01	0,43	-0,07	1,05
DALY	0	0,34	0	1,00
PREV_CAD	0	0,41	0	1,00
INC_CA	-0,18	0,08	0,08	1,74
PREV_DIA	0	0,04	-0,22	1,08
REP_AST	-0,45	0	0,05	1,02
PER_HEALTH	-0,20	0	-0,13	1,73

Zdroj: vlastní zpracování v programu R

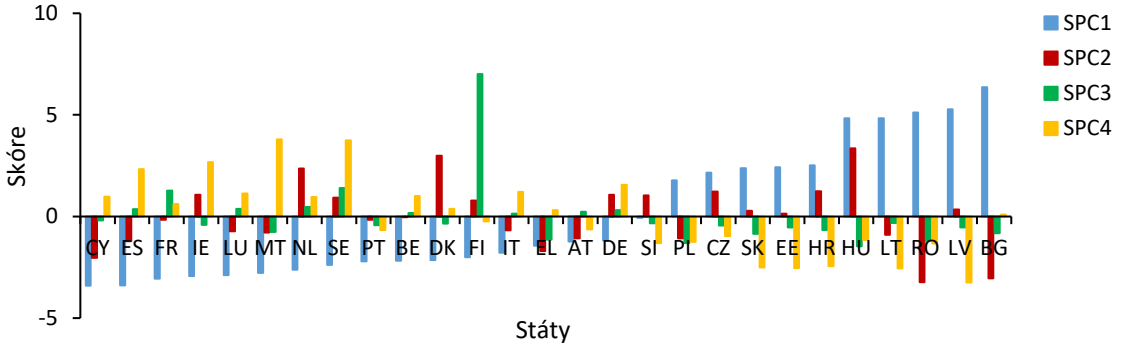
Příloha 5: Vizualizované hodnoty komponentních skóre pro stav zdraví

Obrázek 25: Komponentní skóre RCs pro 25 proměnných, standardizace „z-skóre“, stav zdraví



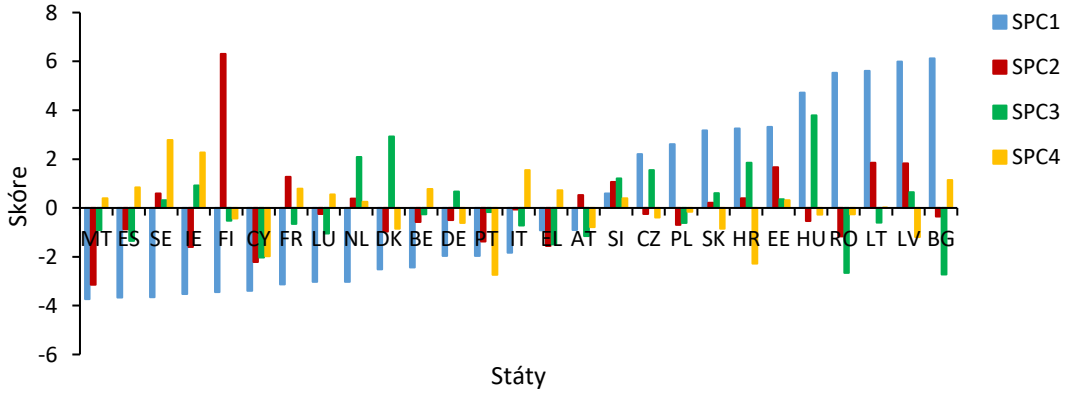
Zdroj: vlastní zpracování v programu R a Excel

Obrázek 26: Komponentní skóre SPCs pro 22 proměnných, standardizace „z-skóre“, stav zdraví



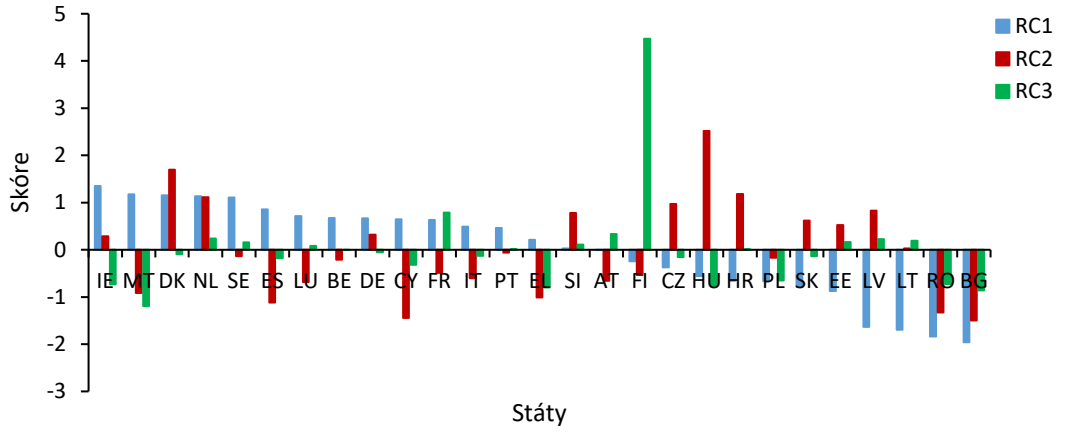
Zdroj: vlastní zpracování v programu R a Excel

Obrázek 27: Komponentní skóre SPCs pro 25 proměnných, standardizace „z-skóre“, stav zdraví



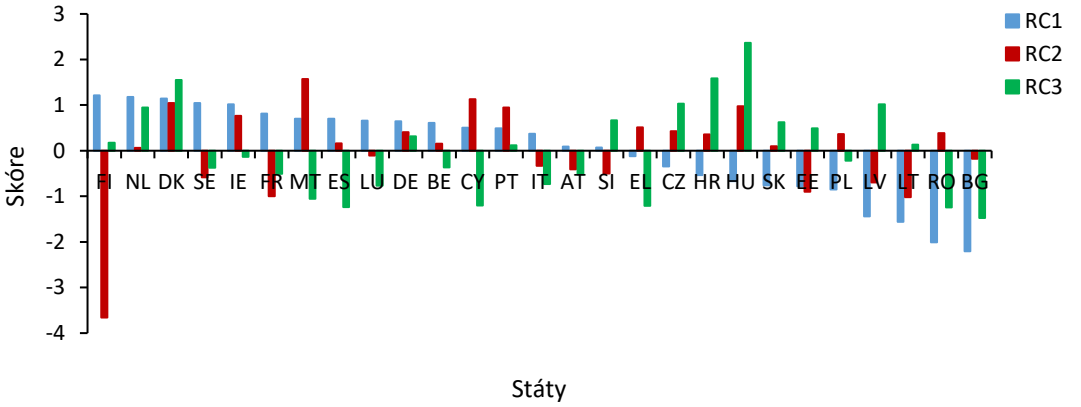
Zdroj: vlastní zpracování v programu R a Excel

Obrázek 28: Vizualizace výsledků RCs (22 proměnných, standardizace „min-max“), stav zdraví



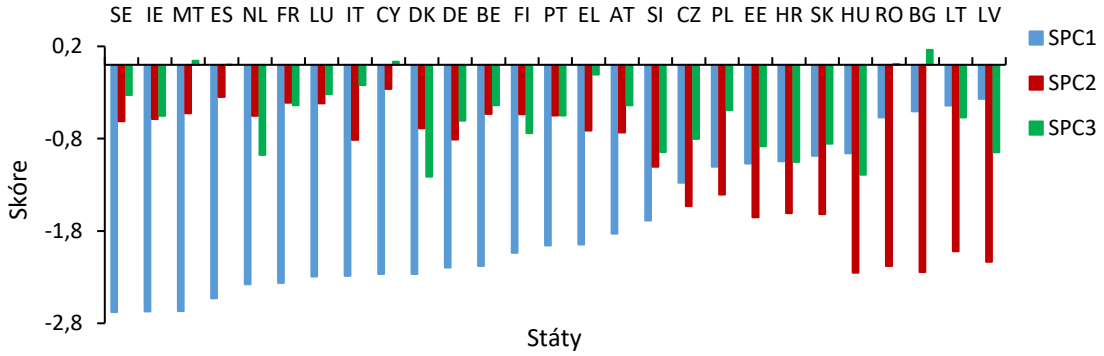
Zdroj: vlastní zpracování v programu R a Excel

Obrázek 29: Vizualizace výsledků RCs (25 proměnných, standardizace „min-max“), stav zdraví



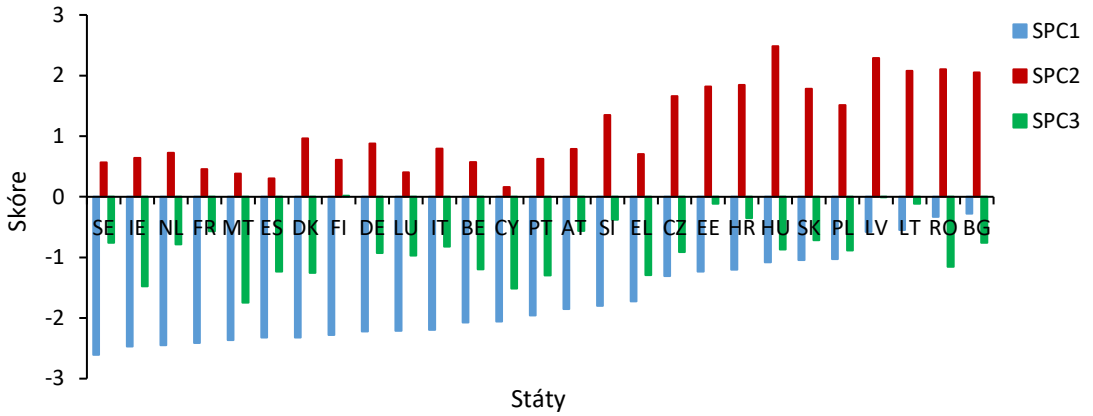
Zdroj: vlastní zpracování v programu R a Excel

Obrázek 30: Vizualizace výsledků SPCs (22 proměnných, standardizace „min-max“), stav zdraví



Zdroj: vlastní zpracování v programu R a Excel

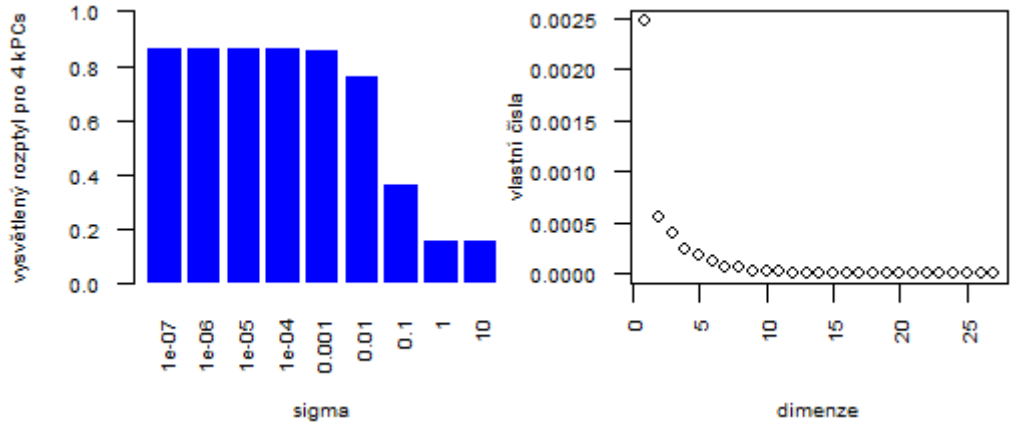
Obrázek 31: Vizualizace výsledků SPCs (25 proměnných, standardizace „min-max“), stav zdraví



Zdroj: vlastní zpracování v programu R a Excel

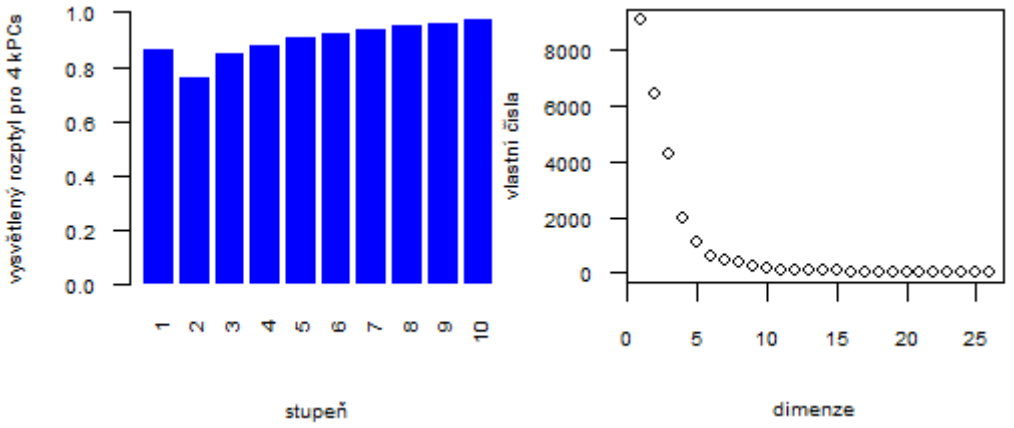
Příloha 6: Kumulativní vysvětlené rozptyly pro různé hodnoty parametrů a vlastní čísla, stav zdraví

Obrázek 32: Výběr parametru σ na základě vysvětleného rozptylu pro 4 kPCs a vlastní čísla získaná z jádrové matice 27x27 (22 proměnných, standardizace „z-skóre“)



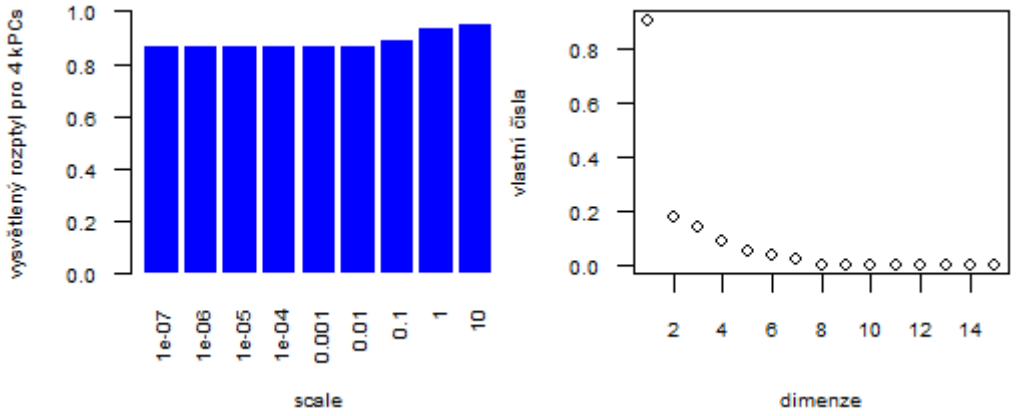
Zdroj: vlastní zpracování v programu R

Obrázek 33:Výběr stupně na základě vysvětleného rozptylu pro 4 kPCs a vlastní čísla získaná z jádrové matice 27x27 (22 proměnných, standardizace „z-skóre“)



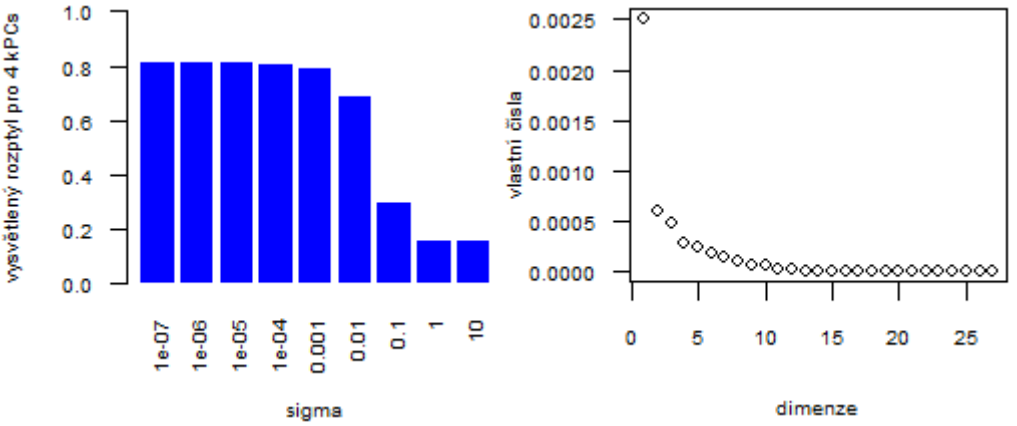
Zdroj: vlastní zpracování v programu R

Obrázek 34: Výběr škálového parametru na základě vysvětleného rozptylu pro 4 kPCs a vlastní čísla získaná z jádrové matice 27x27 (22 proměnných, standardizace „z-skóre“)



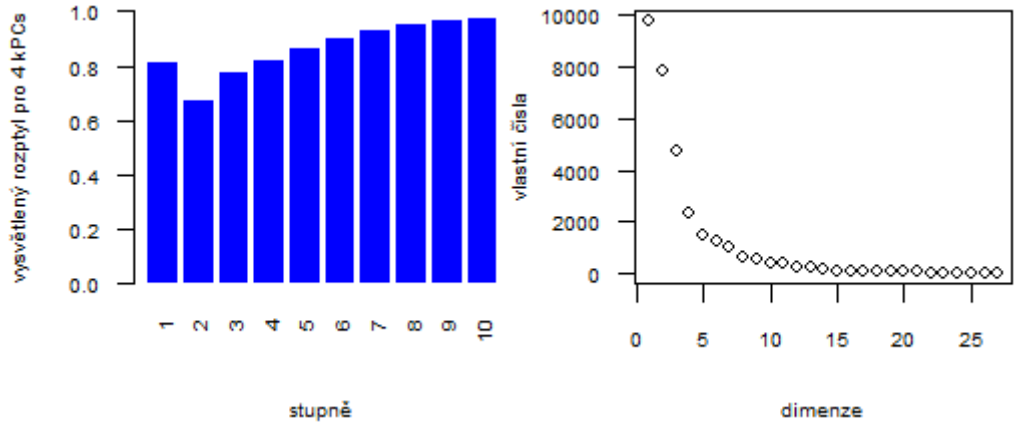
Zdroj: vlastní zpracování v programu R

Obrázek 35: Výběr parametru σ na základě vysvětleného rozptylu pro 4 kPCs a vlastní čísla získaná z jádrové matice 27x27 (25 proměnných, standardizace „z-skóre“)



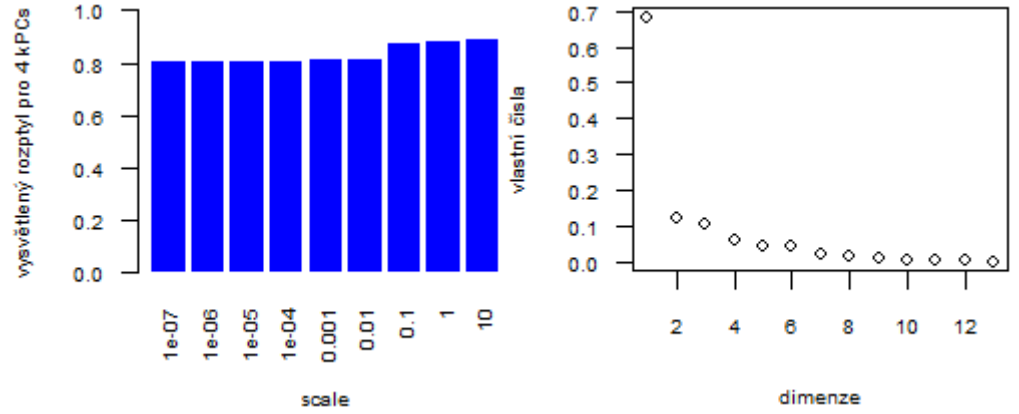
Zdroj: vlastní zpracování v programu R

Obrázek 36: Výběr stupně na základě vysvětleného rozptylu pro 4 kPCs a vlastní čísla získaná z jádrové matice 27x27 (25 proměnných, standardizace „z-skóre“)



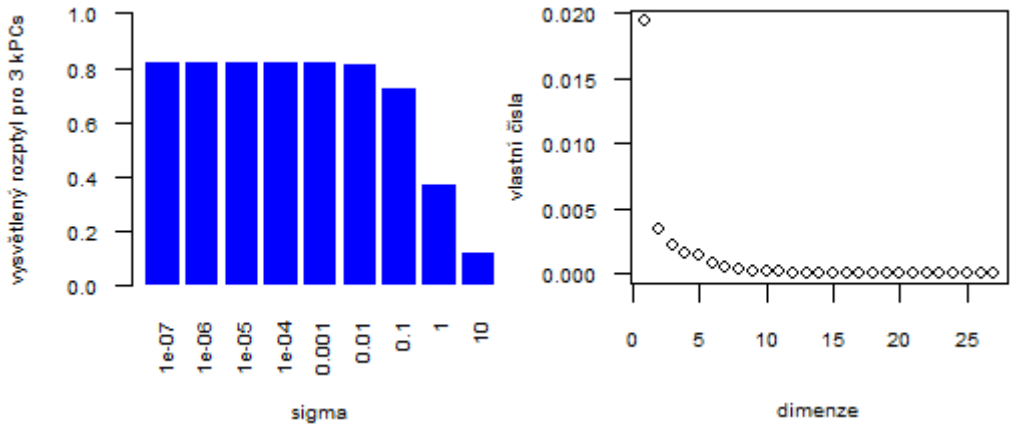
Zdroj: vlastní zpracování v programu R

Obrázek 37: Výběr škálového parametru na základě vysvětleného rozptylu pro 4 kPCs a vlastní čísla získaná z jádrové matice 27x27 (25 proměnných, standardizace „z-skóre“)



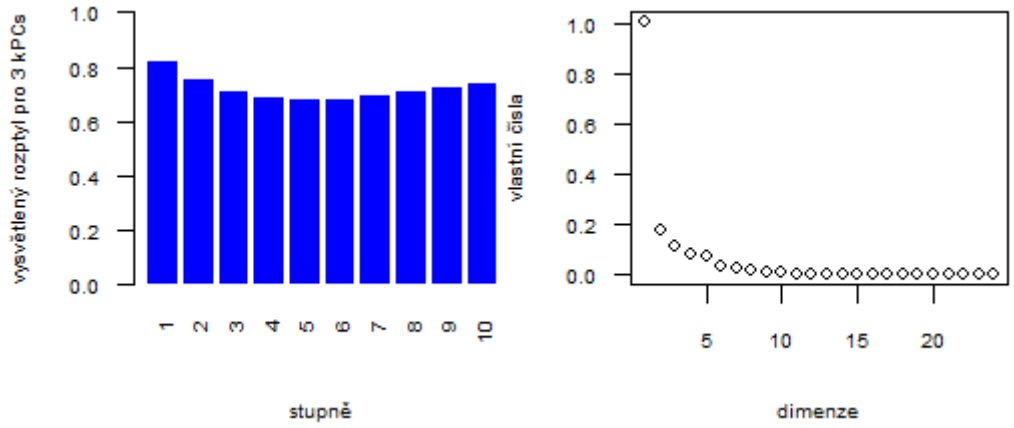
Zdroj: vlastní zpracování v programu R

Obrázek 38: Výběr parametru σ na základě vysvětleného rozptylu pro 3 kPCs a vlastní čísla získaná z jádrové matice 27x27 (22 proměnných, standardizace „min-max“)



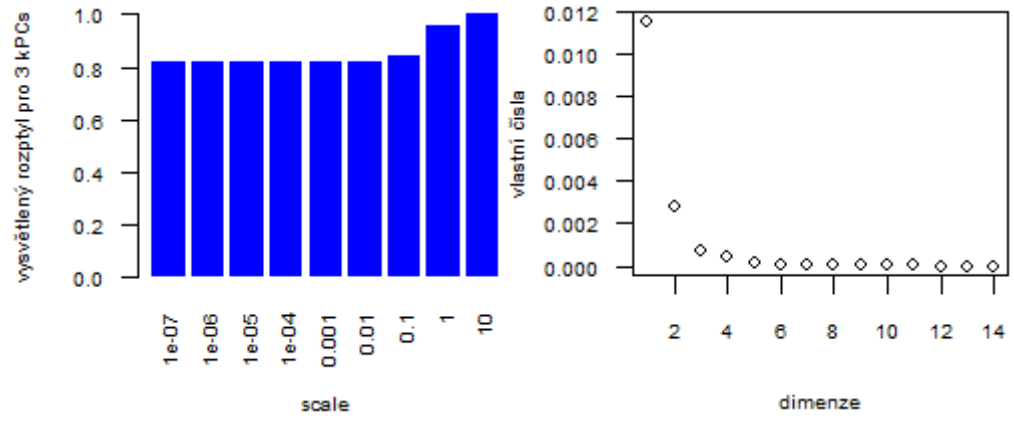
Zdroj: vlastní zpracování v programu R

Obrázek 39: Výběr stupně na základě vysvětleného rozptylu pro 3 kPCs a vlastní čísla získaná z jádrové matice 27x27 (22 proměnných, standardizace „min-max“)



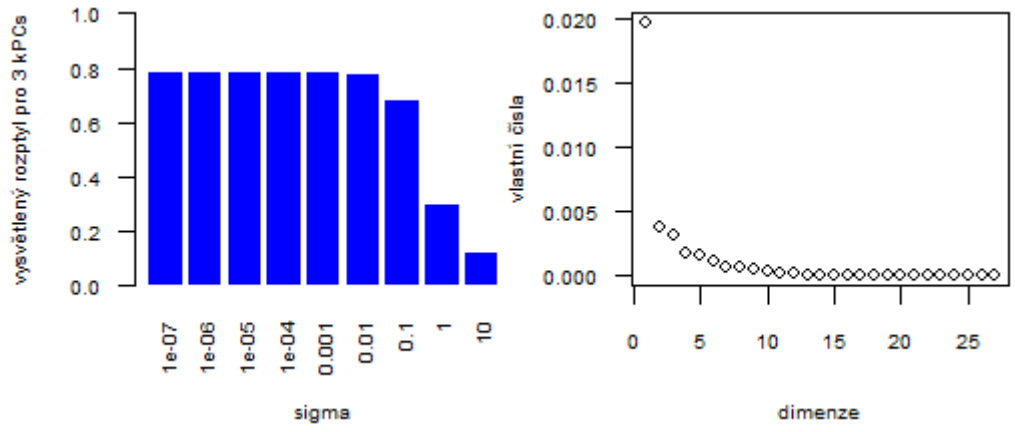
Zdroj: vlastní zpracování v programu R

Obrázek 40: Výběr škálového parametru na základě vysvětleného rozptylu pro 3 kPCs a vlastní čísla získaná z jádrové matice 27x27 (22 proměnných, standardizace „min-max“)



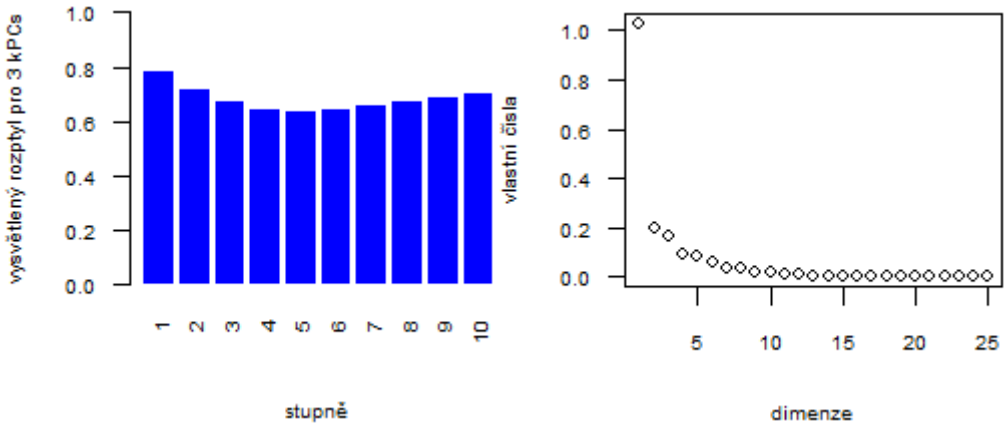
Zdroj: vlastní zpracování v programu R

Obrázek 41: Výběr parametru σ na základě vysvětleného rozptylu pro 3 kPCs a vlastní čísla získaná z jádrové matice 27x27 (25 proměnných, standardizace „min-max“)



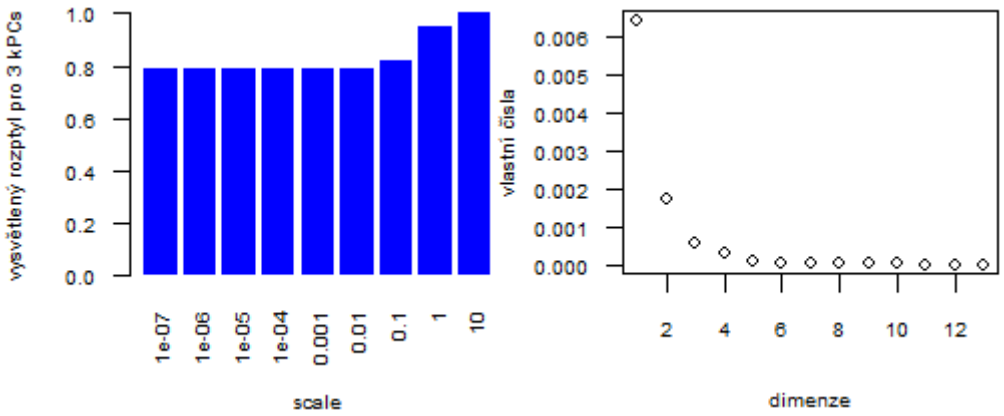
Zdroj: vlastní zpracování v programu R

Obrázek 42: Výběr stupně na základě vysvětleného rozptylu pro 3 kPCs a vlastní čísla získaná z jádrové matice 27x27 (25 proměnných, standardizace „min-max“)



Zdroj: vlastní zpracování v programu R

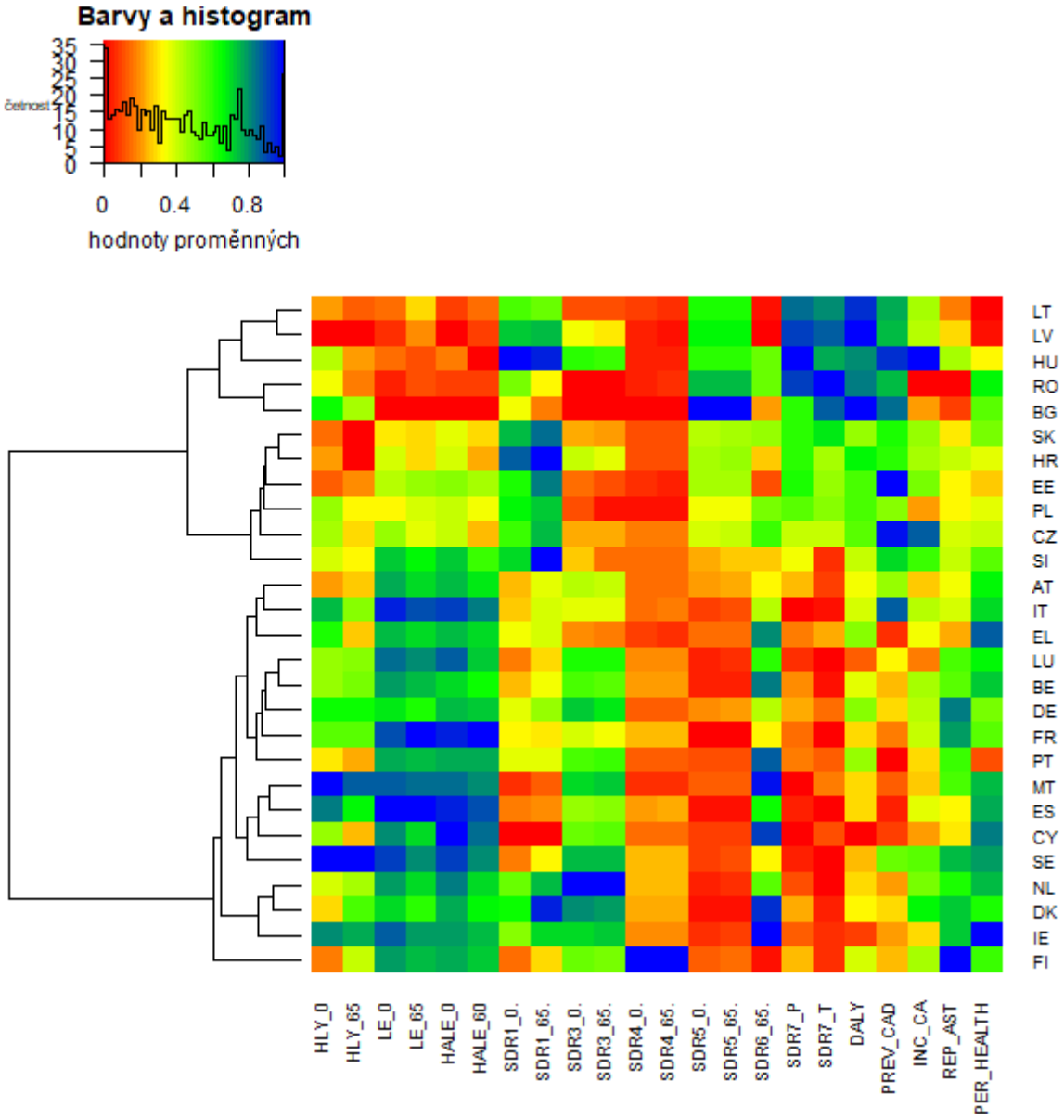
Obrázek 43: Výběr škálového parametru na základě vysvětleného rozptylu pro 3 kPCs a vlastní čísla získaná z jádrové matice 27x27 (25 proměnných, standardizace „min-max“)



Zdroj: vlastní zpracování v programu R

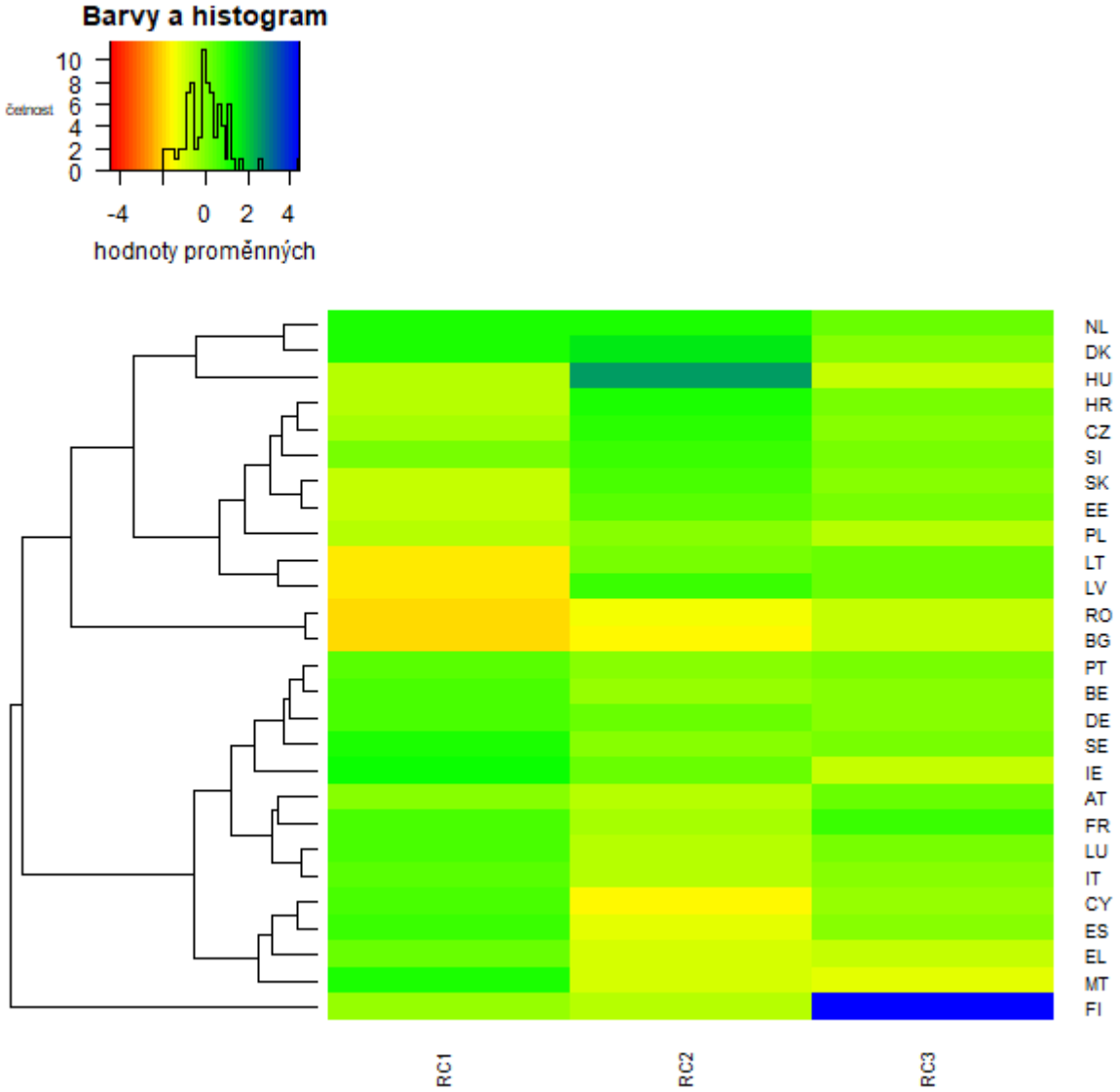
Příloha 7: Dendrogramy a heat mapy pro stav zdraví

Obrázek 44: Dendrogram a heat mapa (datový soubor 1C), stav zdraví



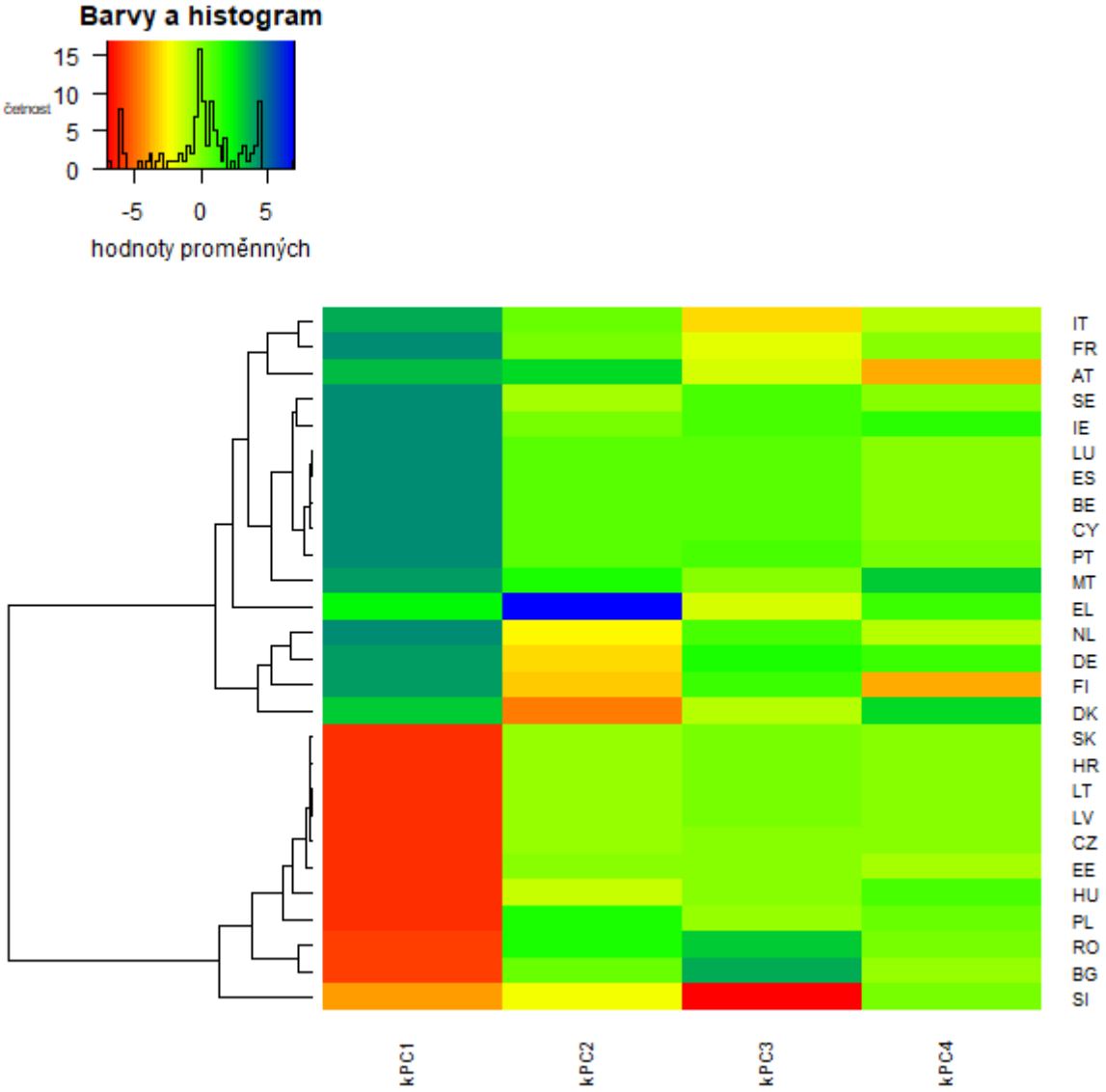
Zdroj: vlastní zpracování v programu R

Obrázek 45: Dendrogram a heat mapa (datový soubor 2C), stav zdraví



Zdroj: vlastní zpracování v programu R

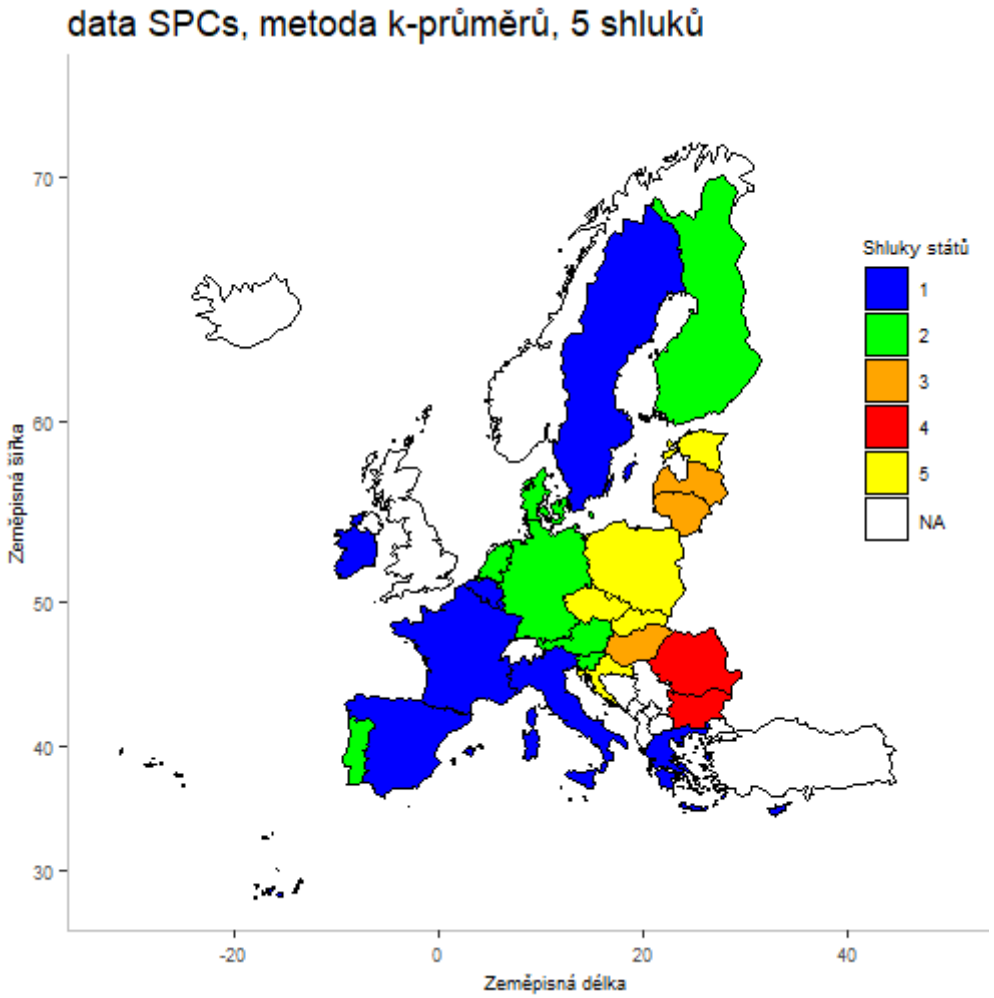
Obrázek 46: Dendrogram a heat mapa (datový soubor 4A), stav zdraví



Zdroj: vlastní zpracování v programu R

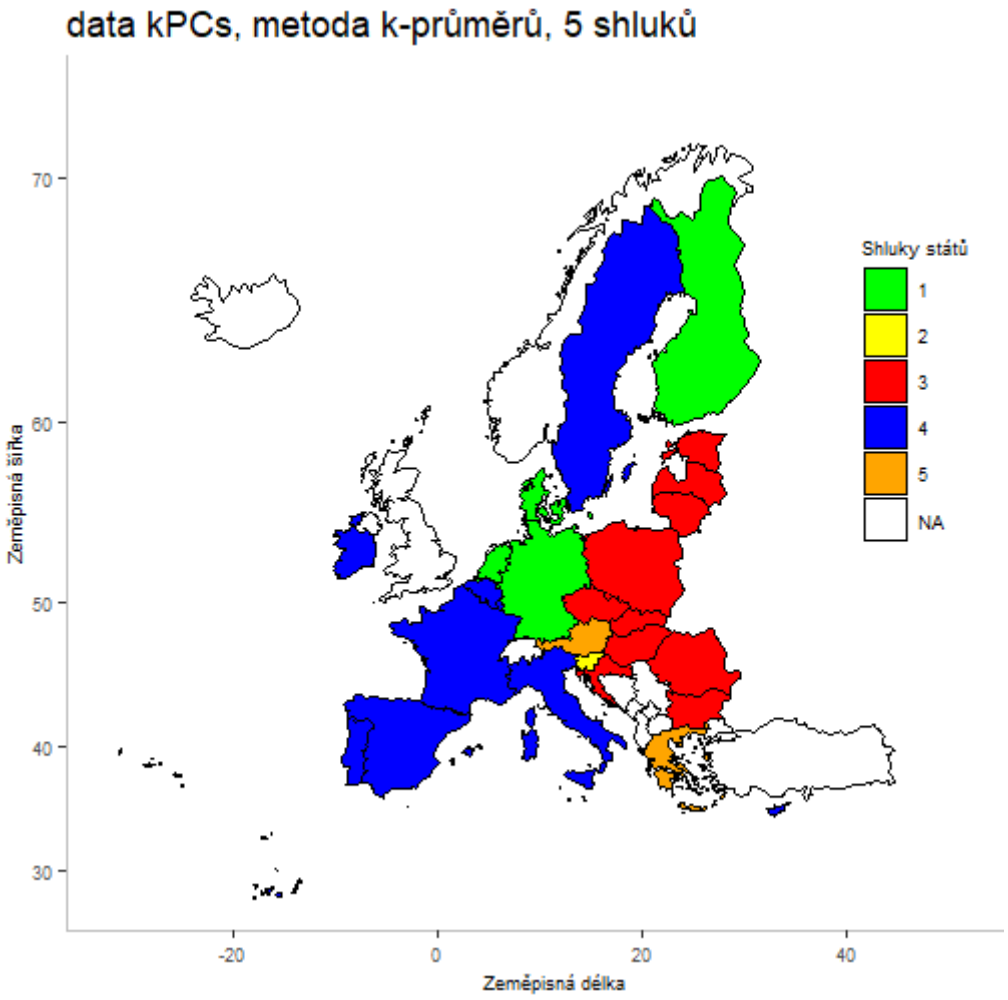
Příloha 8: Vizualizace výsledků shlukové analýzy pomocí geografických dat pro stav zdraví

Obrázek 47: Vizualizace pěti shluků států metodou k-průměrů (SPCs), stav zdraví



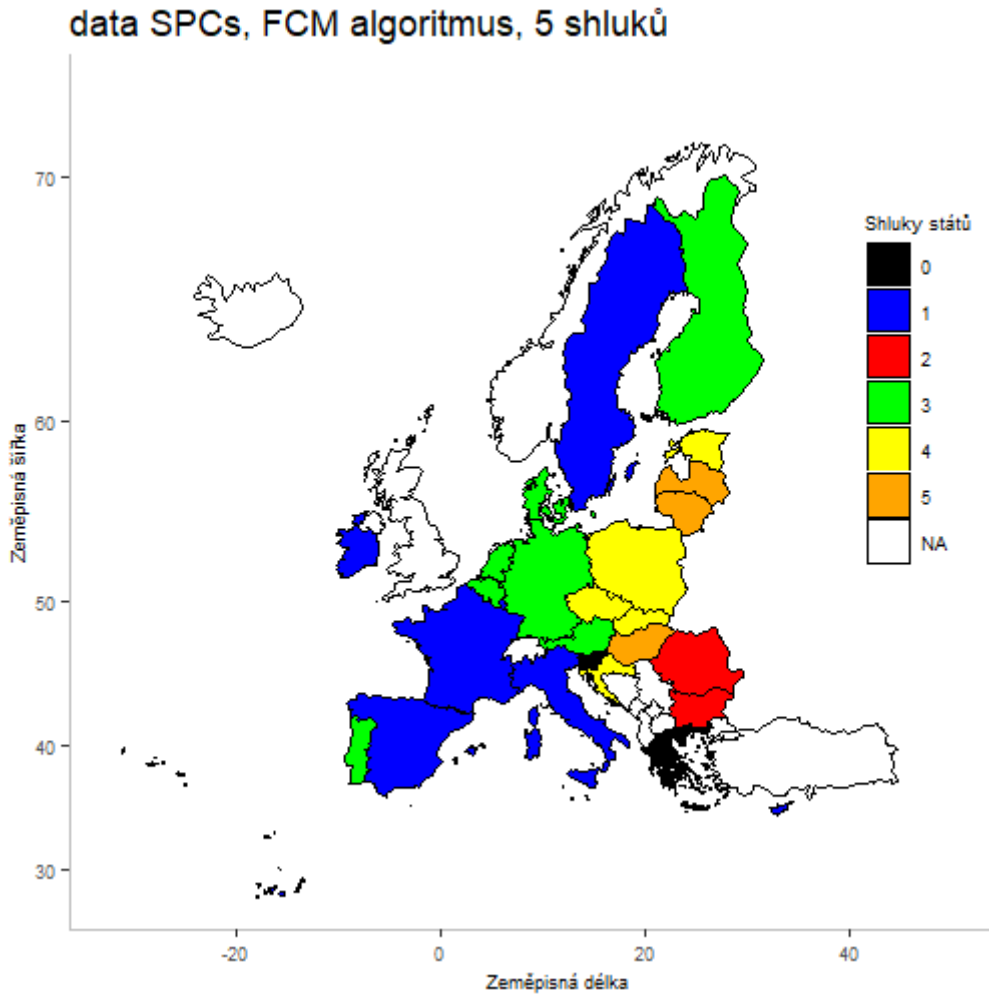
Zdroj: vlastní zpracování v programu R

Obrázek 48: Vizualizace pěti shluků států metodou k-průměrů (kPCs)



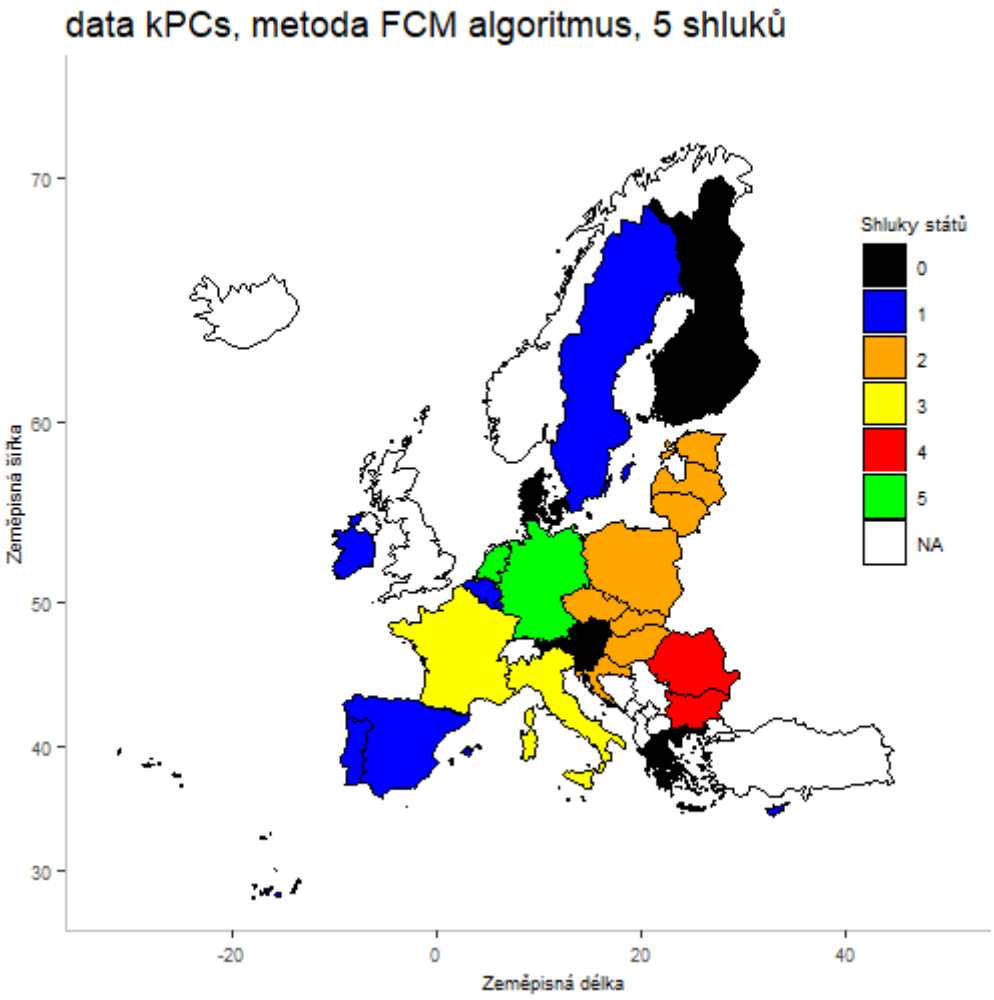
Zdroj: vlastní zpracování v programu R

Obrázek 49: Vizualizace pěti shluků států FCM algoritmem (SPCs), stav zdraví



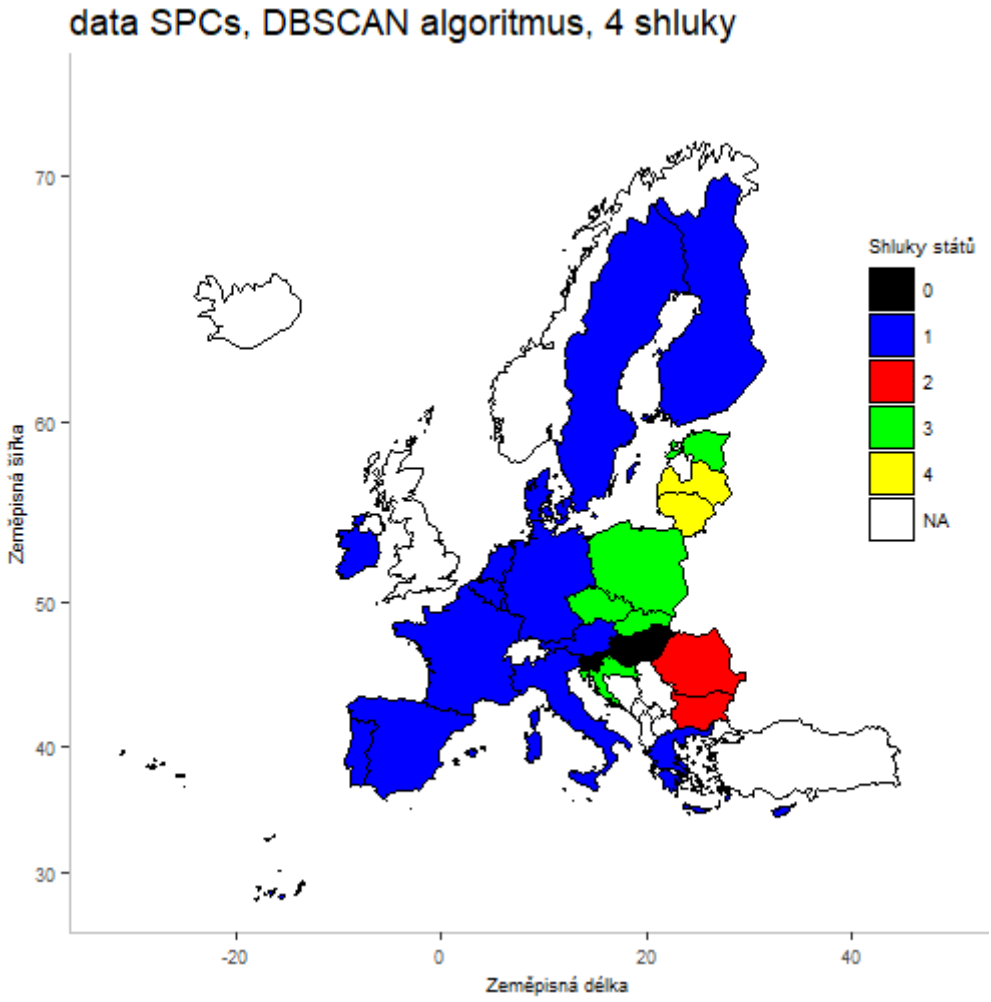
Zdroj: vlastní zpracování v programu R

Obrázek 50: Vizualizace pěti shluků států FCM algoritmem (kPCs), stav zdraví



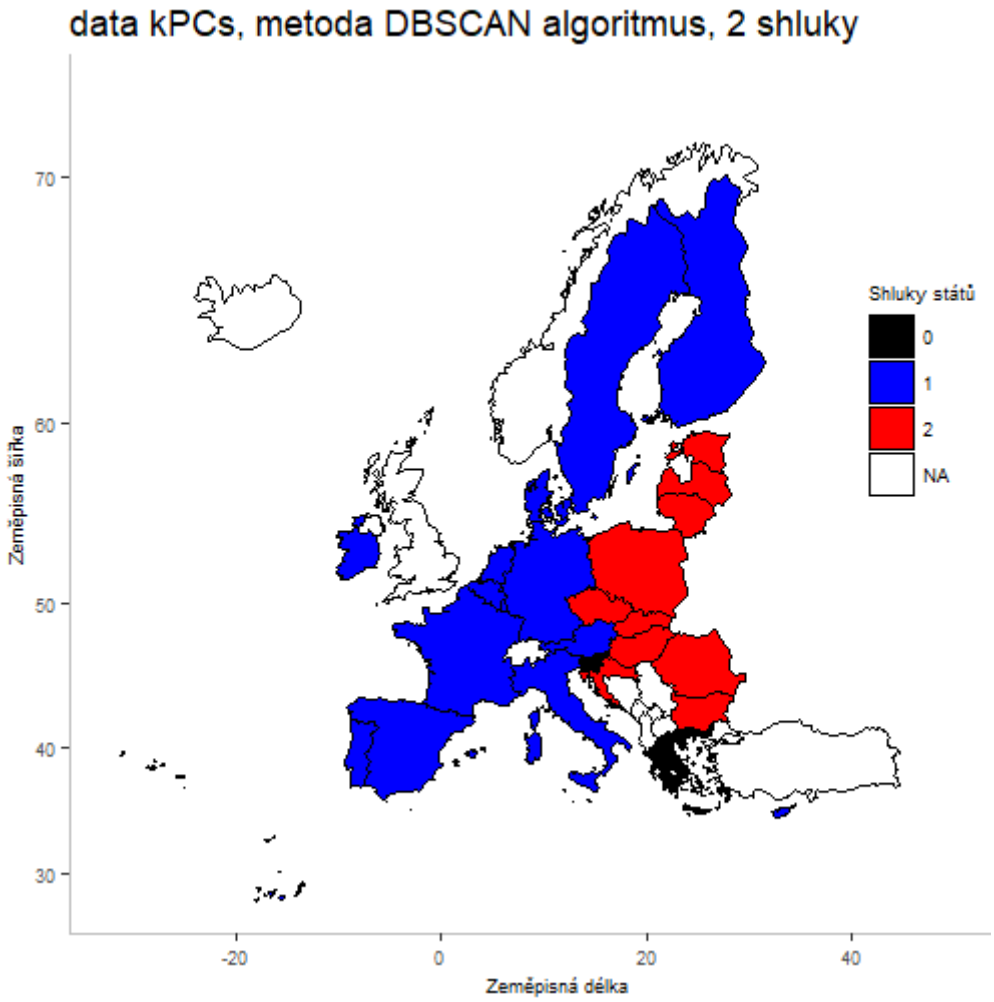
Zdroj: vlastní zpracování v programu R

Obrázek 51: Vizualizace čtyř shluků států DBSCAN algoritmem (SPCs), stav zdraví



Zdroj: vlastní zpracování v programu R

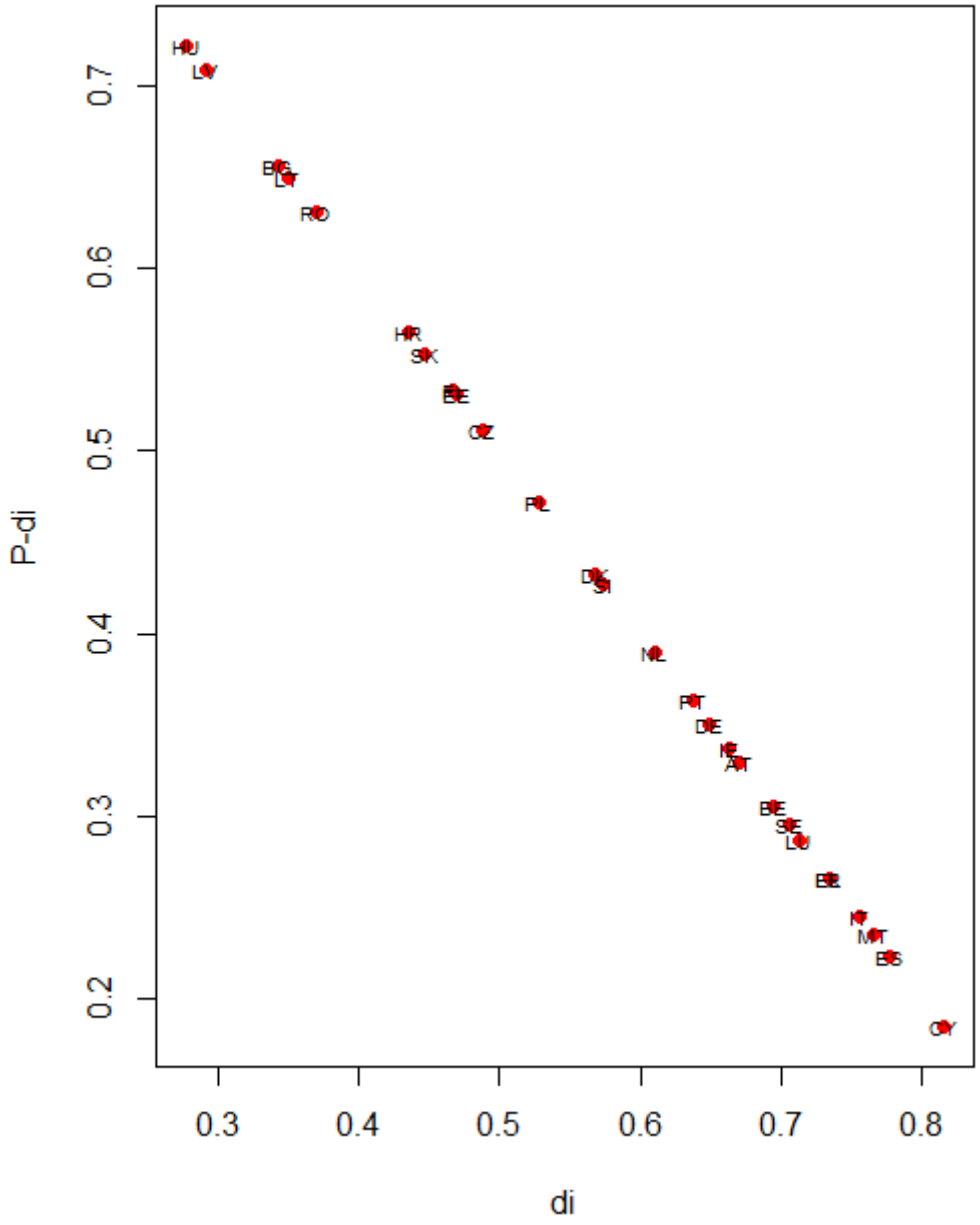
Obrázek 52: Vizualizace dvou shluků států DBSCAN algoritmem (kPCs)



Zdroj: vlastní zpracování v programu R

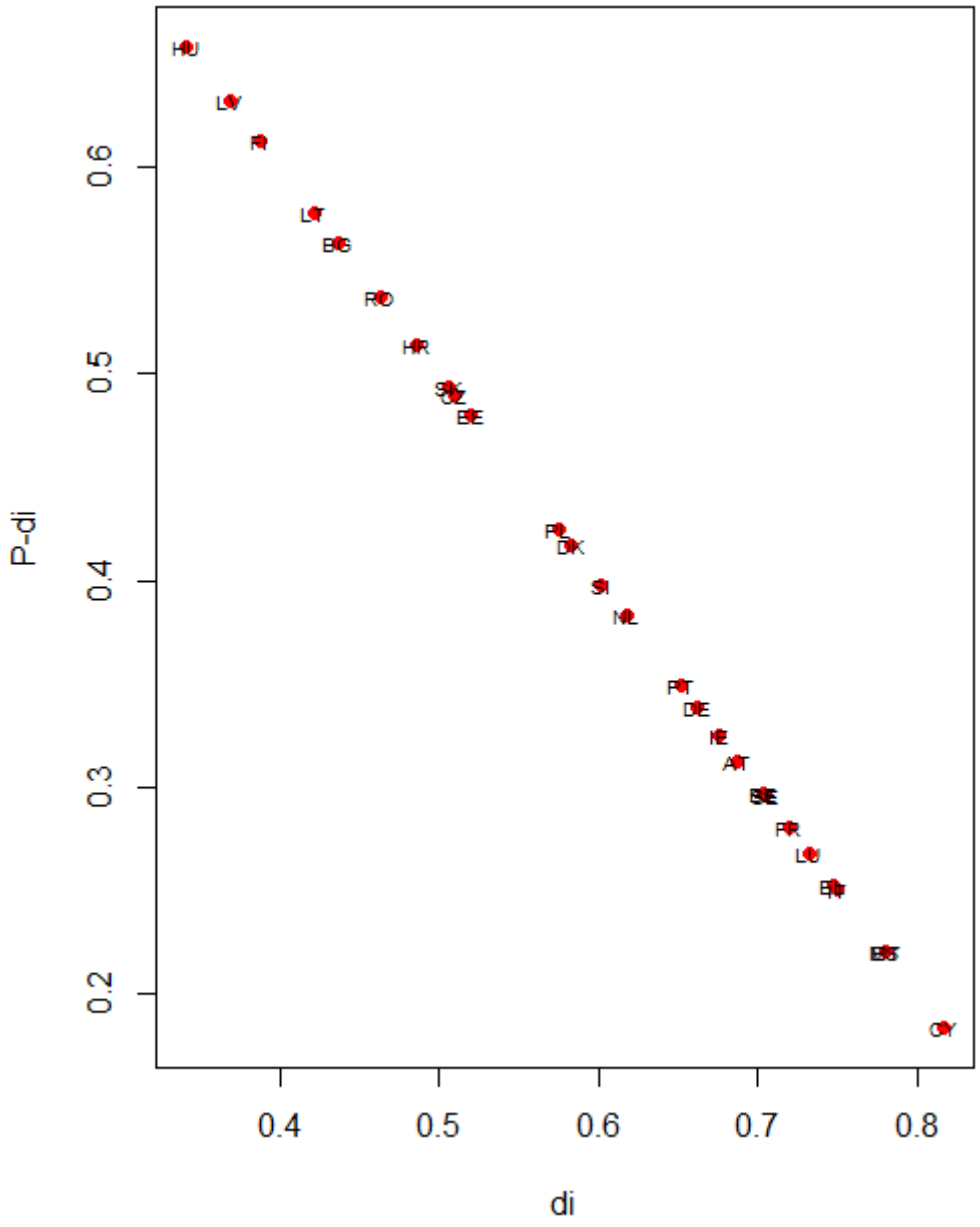
Příloha 9: Uspořádání států pomocí hybridního přístupu podle stavu zdraví

Obrázek 53: Lineární uspořádání států, datový soubor 1C, stav zdraví



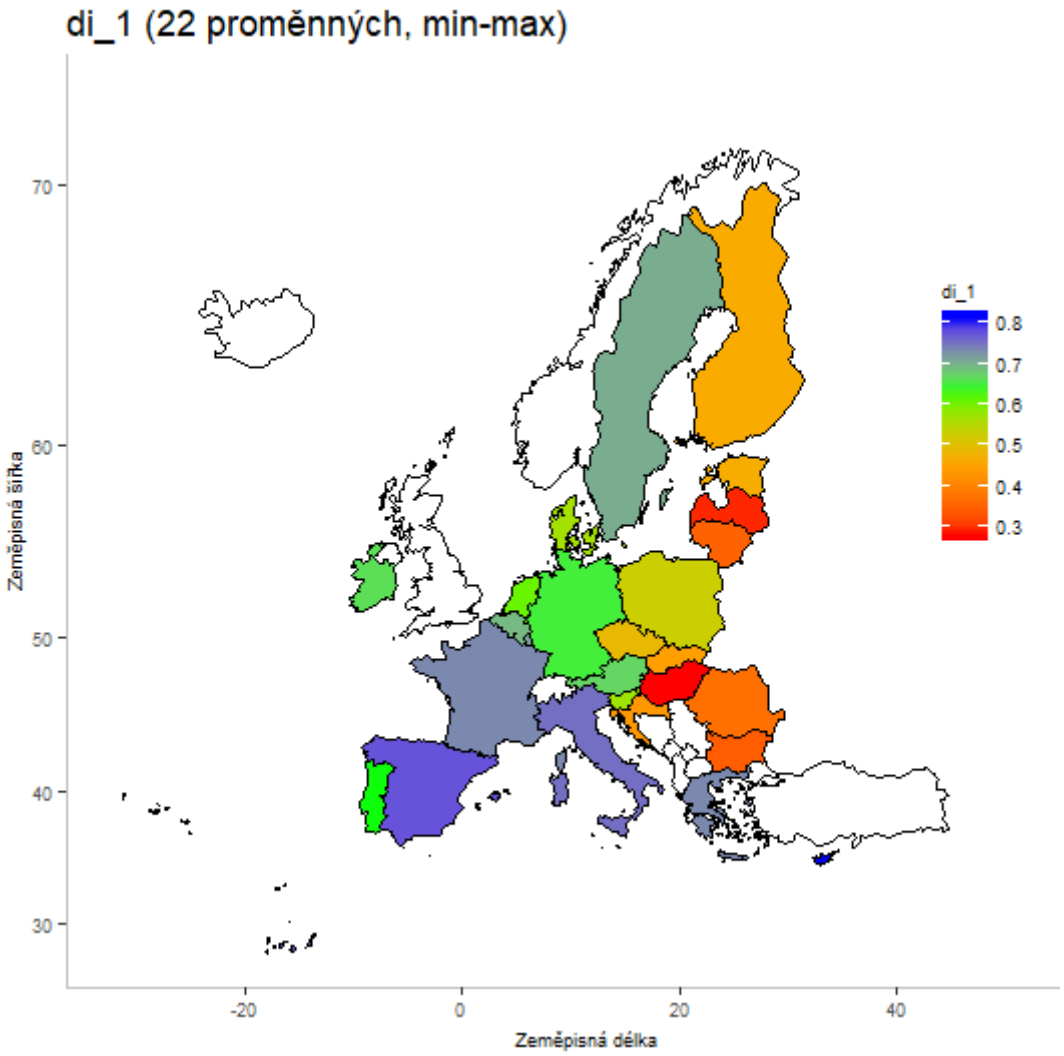
Zdroj: vlastní zpracování v programu R

Obrázek 54: Lineární uspořádání států, datový soubor 1A, stav zdraví



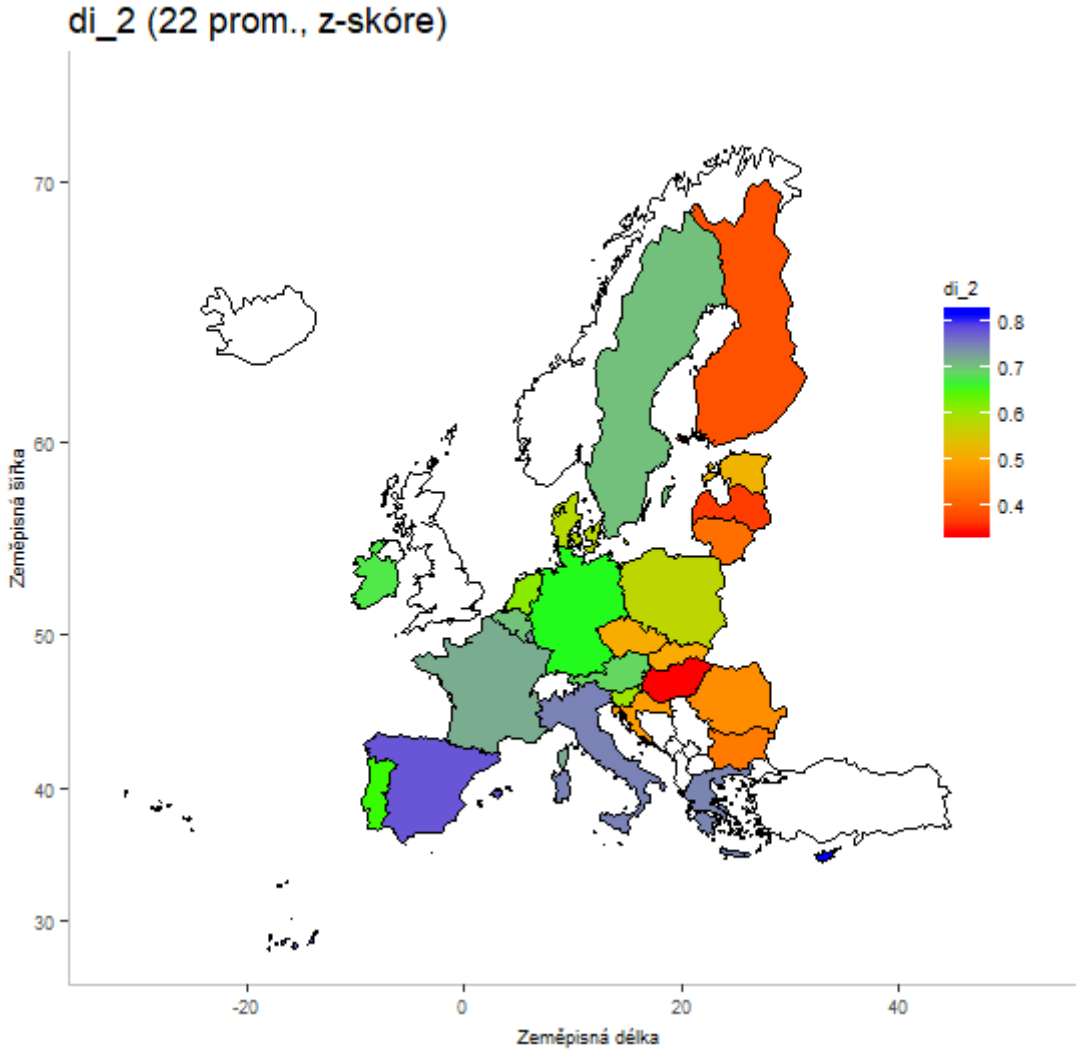
Zdroj: vlastní zpracování v programu R

Obrázek 55: Vizualizace agregované míry (1C), stav zdraví



Zdroj: vlastní zpracování v programu R

Obrázek 56: Vizualizace agregované míry (1A), stav zdraví



Zdroj: vlastní zpracování v programu R

Příloha 10: Rotované komponentní zátěže pro determinanty stavu zdraví

Tabulka 23: Komponentní zátěže po varimax rotaci, standardizace „z-skóre“, determinanty stavu zdraví

17 proměnných	RC1	RC2	RC3	RC4	h2	u2	com
EC1	-0,52	0,17	0,03	-0,74	0,85	0,15	1,9
ILF1	-0,53	0,04	0,25	-0,51	0,60	0,40	2,4
HS1	0,83	0,34	-0,24	0,29	0,93	0,07	1,8
HS2	0,85	0,16	-0,29	-0,10	0,84	0,16	1,3
HS5	0,80	0,26	-0,18	0,40	0,90	0,10	1,8
HS6	0,86	0,20	-0,25	0,24	0,91	0,10	1,5
HS8	0,73	0,13	-0,30	0,30	0,74	0,26	1,8
HS9	-0,02	0,01	0,13	-0,80	0,66	0,34	1,1
HS11	-0,13	-0,95	0,06	0,08	0,93	0,07	1,1
HS12	-0,21	-0,94	0,15	-0,04	0,95	0,05	1,2
HS13	0,71	0,40	-0,20	0,22	0,76	0,24	2,0
SEC2	0,64	0,34	-0,30	0,47	0,83	0,17	3,0
SEC3	-0,18	0,01	0,93	0,02	0,90	0,11	1,1
SEC4	-0,20	-0,22	0,92	-0,11	0,94	0,06	1,2
SEC5	-0,31	-0,13	0,76	-0,37	0,83	0,17	1,9
SEC6	-0,84	0,14	-0,05	-0,12	0,74	0,26	1,1
ED1	0,24	0,08	-0,04	0,80	0,70	0,30	1,2

Zdroj: vlastní zpracování v programu R

Tabulka 24: Komponentní zátěže po varimax rotaci, standardizace „min-max“, determinanty stavu zdraví

17 proměnných	RC1	RC2	h2	u2	com
EC1	-0,91	0,00	0,82	0,18	1,00
ILF1	-0,73	-0,24	0,60	0,40	1,20
HS1	0,66	0,68	0,89	0,11	2,00
HS2	0,46	0,64	0,62	0,38	1,80
HS5	0,74	0,57	0,86	0,14	1,90
HS6	0,69	0,61	0,85	0,15	2,00
HS8	0,67	0,51	0,72	0,28	1,90
HS9	-0,59	0,07	0,36	0,64	1,00
HS11	0,24	-0,80	0,70	0,30	1,20
HS12	0,09	-0,85	0,73	0,27	1,00
HS13	0,52	0,67	0,72	0,28	1,90
SEC2	0,67	0,60	0,81	0,19	2,00
SEC3	-0,18	-0,54	0,32	0,68	1,20
SEC4	-0,22	-0,69	0,52	0,48	1,20
SEC5	-0,48	-0,55	0,54	0,46	2,00
SEC6	-0,66	-0,22	0,49	0,51	1,20
ED1	0,70	0,06	0,49	0,51	1,00

Zdroj: vlastní zpracování v programu R

Příloha 11: Řídké komponentní zátěže pro determinanty stavu zdraví

Tabulka 25: Řídké komponentní zátěže, standardizace „z-skóre“, determinanty stavu zdraví

17 proměnných	SPC1	SPC2	SPC3	SPC4	com
EC1	0,00	0,00	0,00	-0,36	1,00
ILF1	0,07	0,00	0,00	-0,27	1,15
HS1	-0,46	0,00	0,00	0,00	1,00
HS2	-0,35	0,00	0,00	0,00	1,00
HS5	-0,30	0,00	0,00	0,00	1,00
HS6	-0,44	0,00	0,00	0,00	1,00
HS8	-0,36	0,00	0,00	0,00	1,00
HS9	0,00	0,00	0,00	-0,58	1,00
HS11	0,00	-0,66	0,00	0,00	1,00
HS12	0,00	-0,72	0,00	0,00	1,00
HS13	-0,18	0,05	0,00	0,00	1,14
SEC2	-0,18	0,01	0,00	0,10	1,53
SEC3	0,00	0,00	-0,66	0,00	1,00
SEC4	0,00	0,00	-0,58	0,00	1,00
SEC5	0,00	0,00	-0,40	-0,01	1,00
SEC6	0,41	0,00	0,00	0,00	1,00
ED1	0,00	0,00	0,00	0,63	1,00

Zdroj: vlastní zpracování v programu R

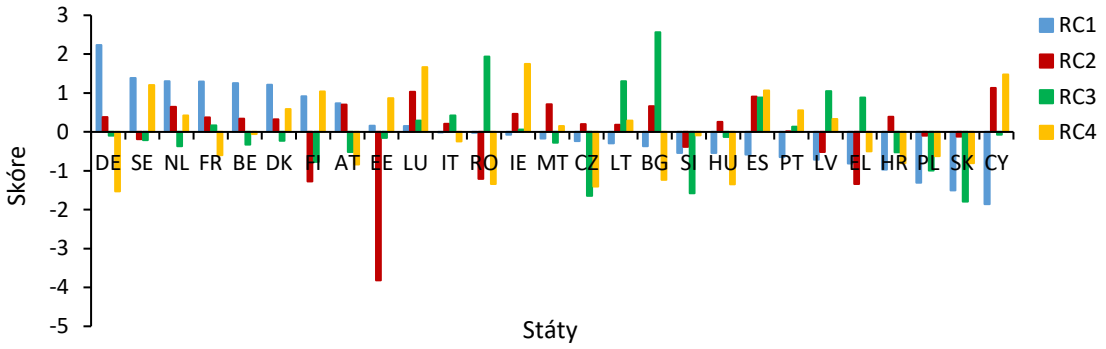
Tabulka 26: Řídké komponentní zátěže, standardizace „min-max“, determinanty stavu zdraví

17 proměnných	SPC1	SPC2	com
EC1	0,00	-0,44	1,00
ILF1	0,00	-0,37	1,00
HS1	-0,42	0,02	1,01
HS2	-0,34	0,00	1,00
HS5	-0,34	0,11	1,20
HS6	-0,33	0,00	1,00
HS8	-0,33	0,00	1,00
HS9	-0,05	-0,37	1,04
HS11	0,00	-0,06	1,00
HS12	0,00	-0,26	1,00
HS13	-0,36	0,00	1,00
SEC2	-0,33	0,00	1,00
SEC3	0,00	-0,29	1,00
SEC4	0,00	-0,35	1,00
SEC5	0,00	-0,29	1,00
SEC6	0,00	-0,38	1,00
ED1	-0,36	-0,11	1,19

Zdroj: vlastní zpracování v programu R

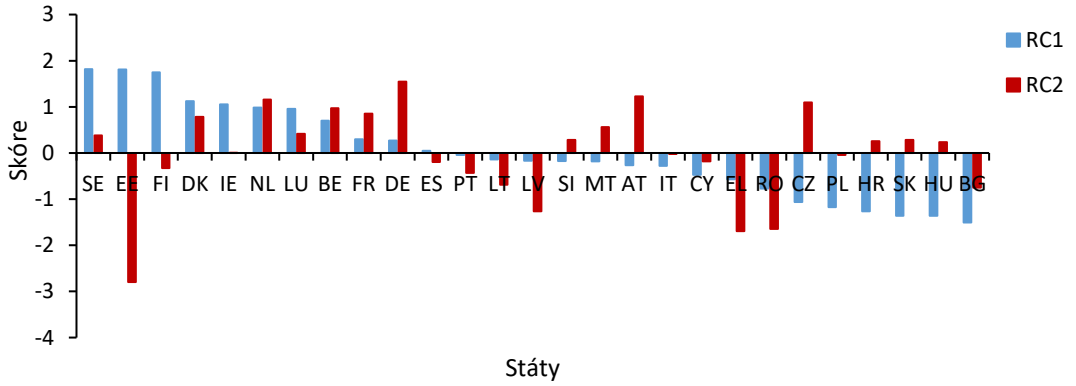
Příloha 12: Vizualizované hodnoty komponentních skóre pro determinanty stavu zdraví

Obrázek 57: Komponentní skóre RCs pro 17 proměnných, standardizace „z-skóre“, determinanty stavu zdraví



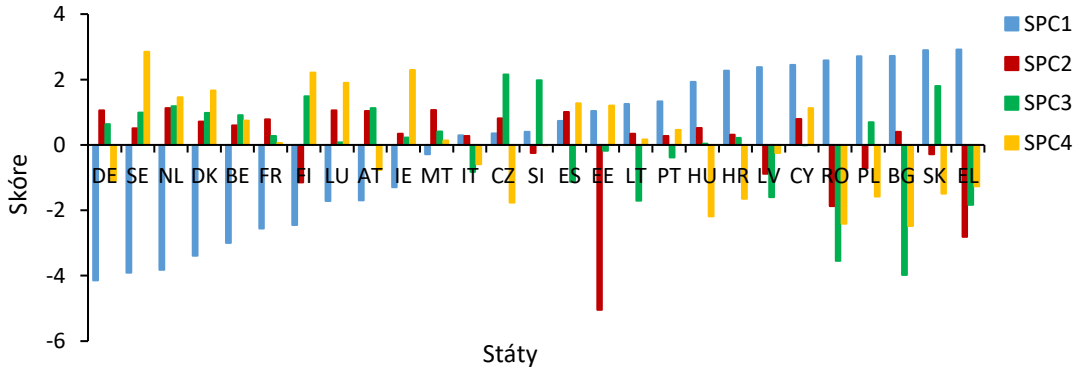
Zdroj: vlastní zpracování v programu R a Excel

Obrázek 58: Komponentní skóre RCs pro 17 proměnných, standardizace „min-max“, determinanty stavu zdraví



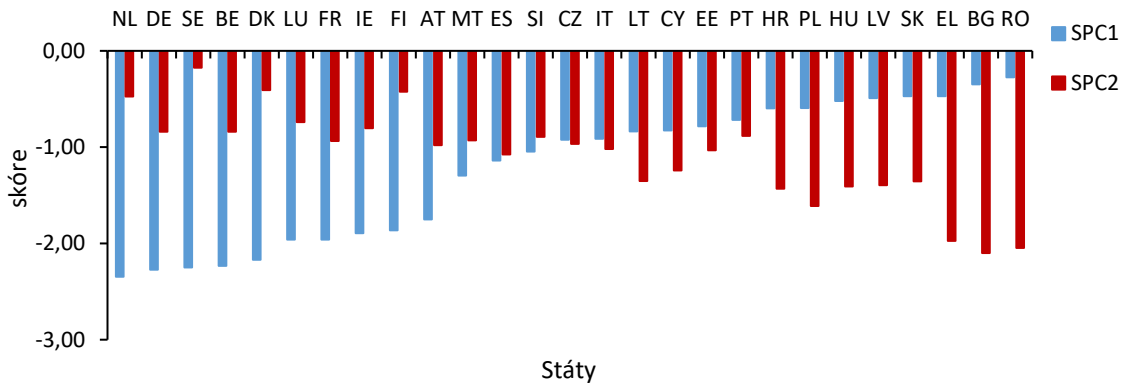
Zdroj: vlastní zpracování v programu R a Excel

Obrázek 59: Řídká komponentní skóre SPCs pro 17 proměnných, standardizace „z-skóre“, determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R a Excel

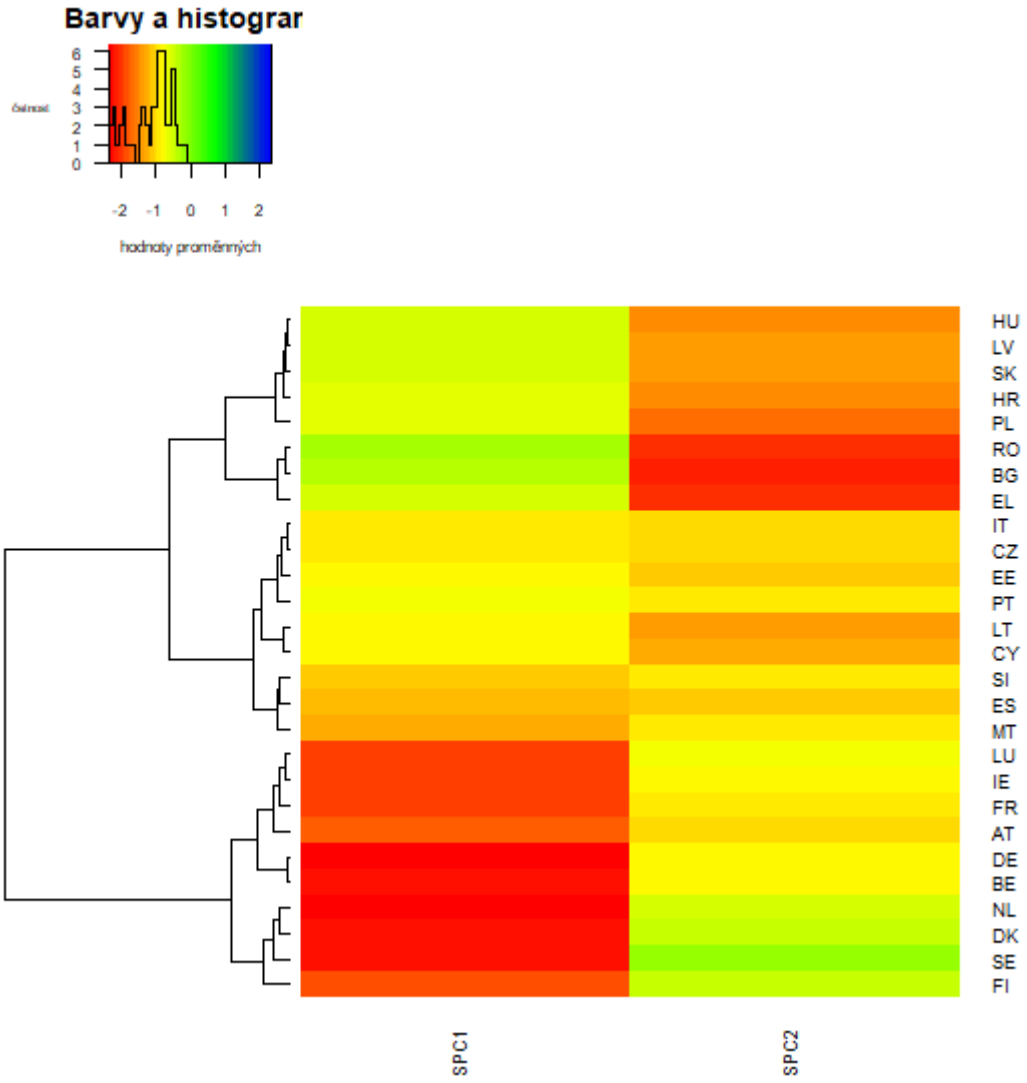
Obrázek 60: Řídká komponentní skóre SPCs pro 17 proměnných, standardizace „min-max“, determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R a Excel

Příloha 13: Dendrogram a heat mapa pro determinanty stavu zdraví

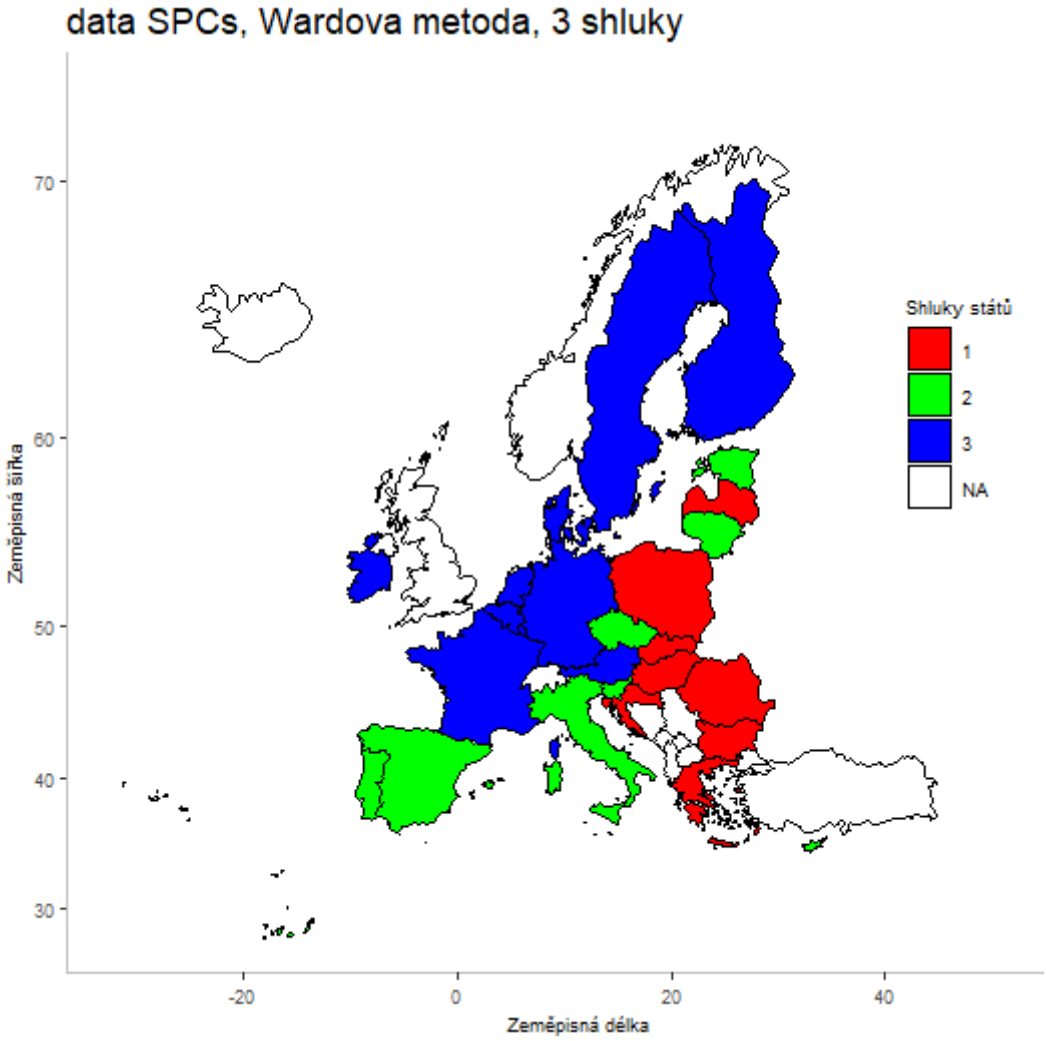
Obrázek 61: Dendrogram a heat mapa (datový soubor 3B), determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R

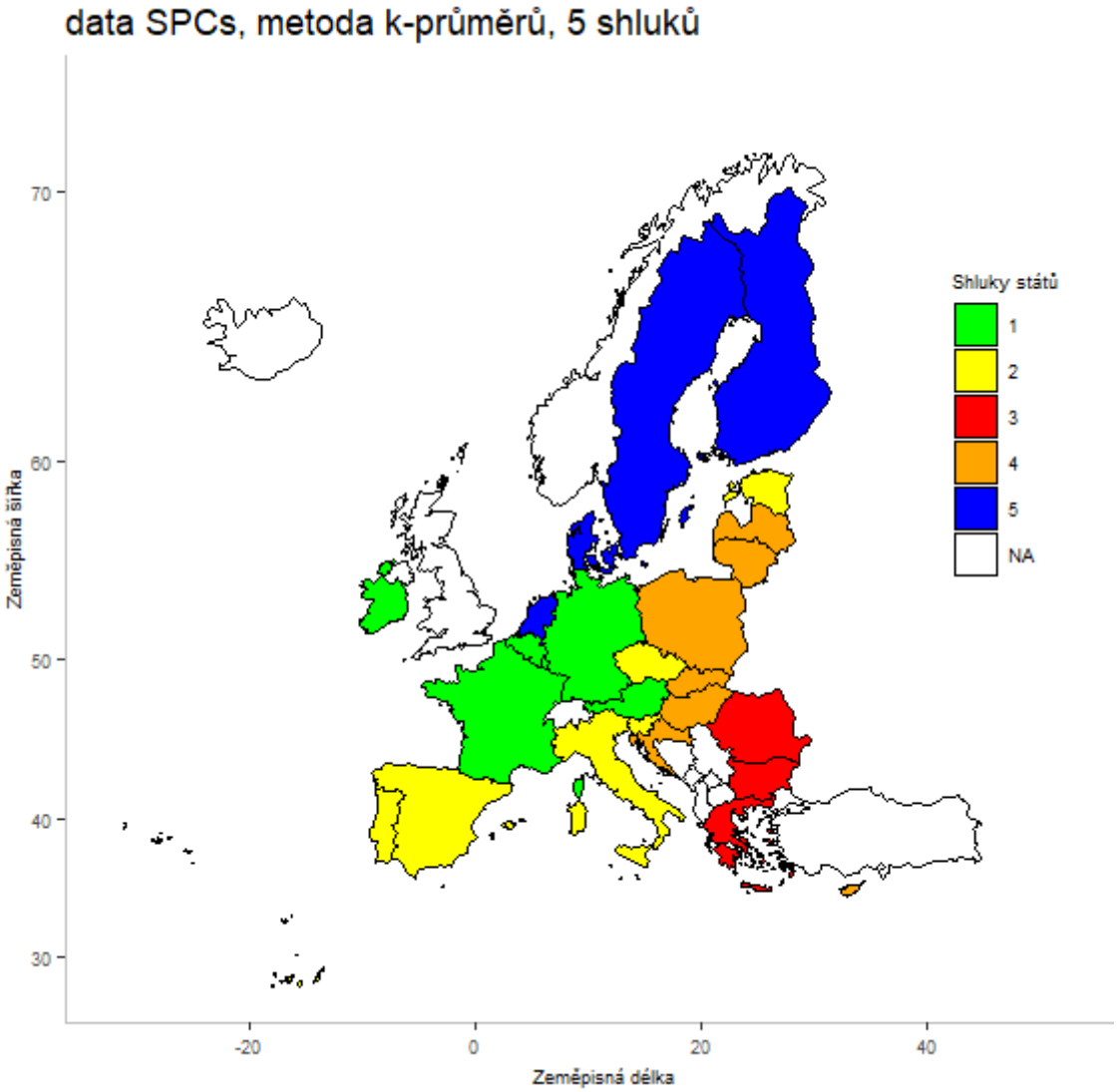
Příloha 14: Vizualizace výsledků shlukové analýzy pomocí geografických dat pro determinanty stavu zdraví

Obrázek 62: Vizualizace třech shluků států Wardovou metodou (SPCs)



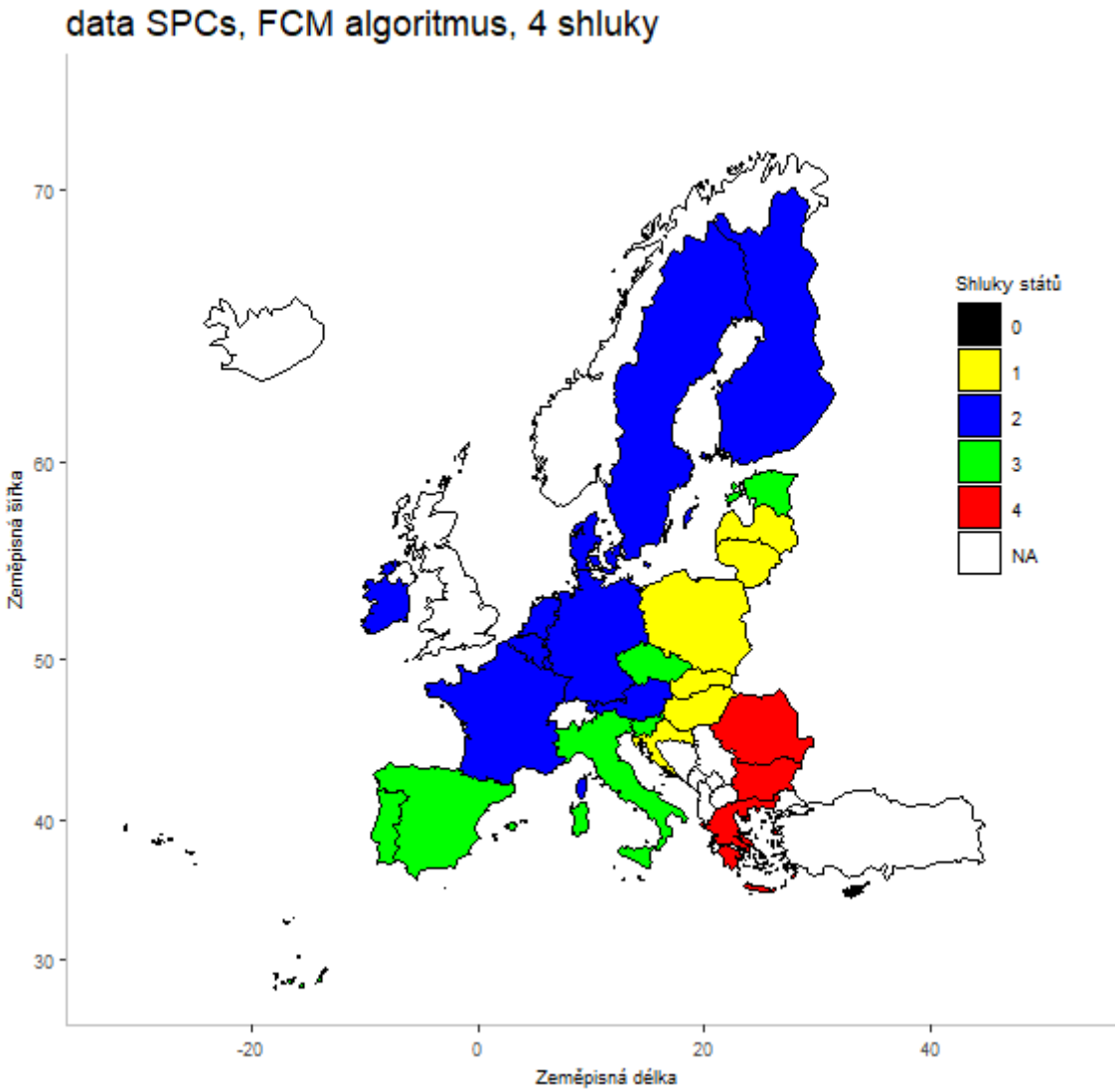
Zdroj: vlastní zpracování v programu R

Obrázek 63: Vizualizace pěti shluků států metodou k-průměrů (SPCs), determinanty stavu zdraví



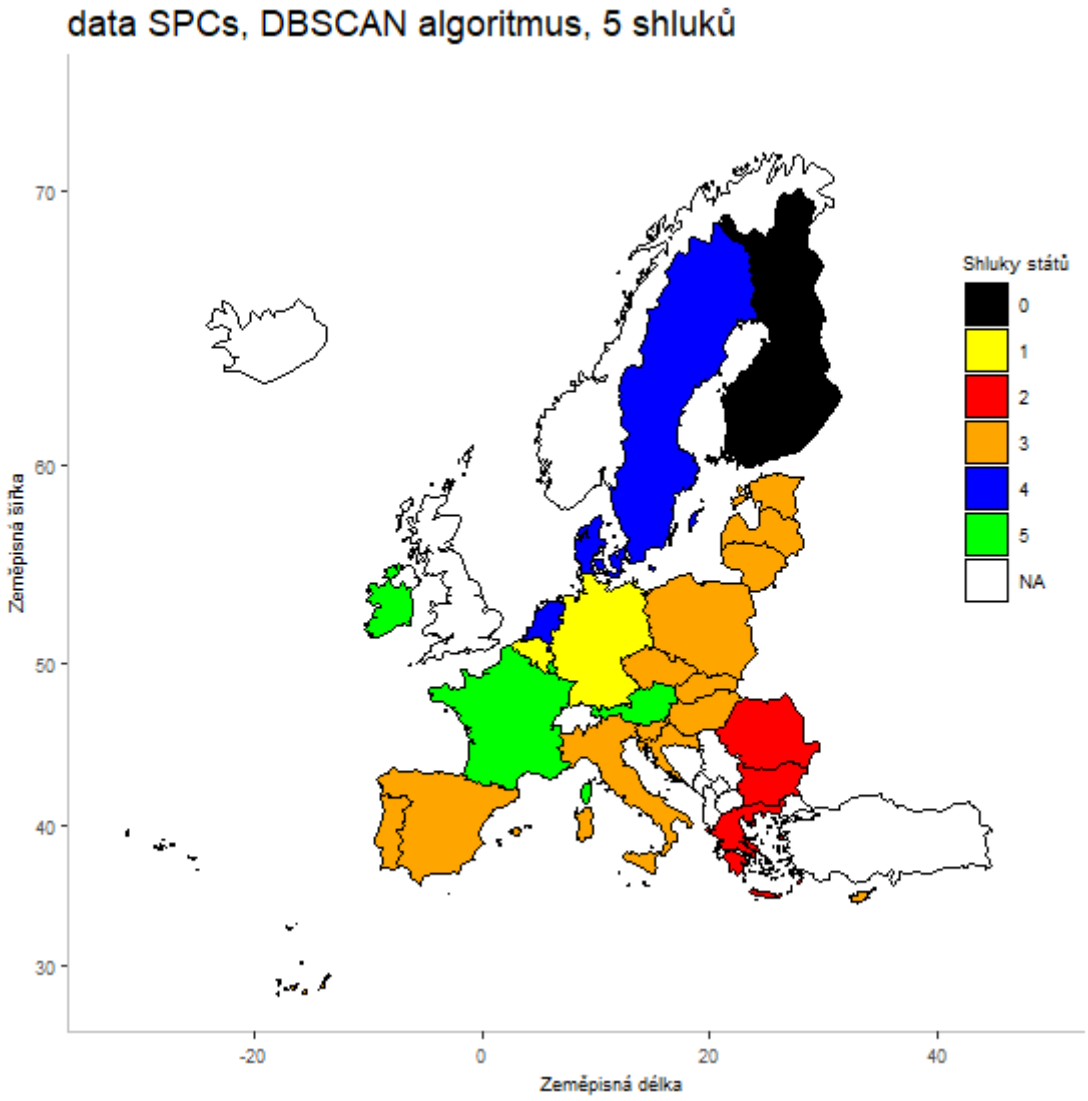
Zdroj: vlastní zpracování v programu R

Obrázek 64: Vizualizace čtyř shluků států FCM algoritmem (SPCs), determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R

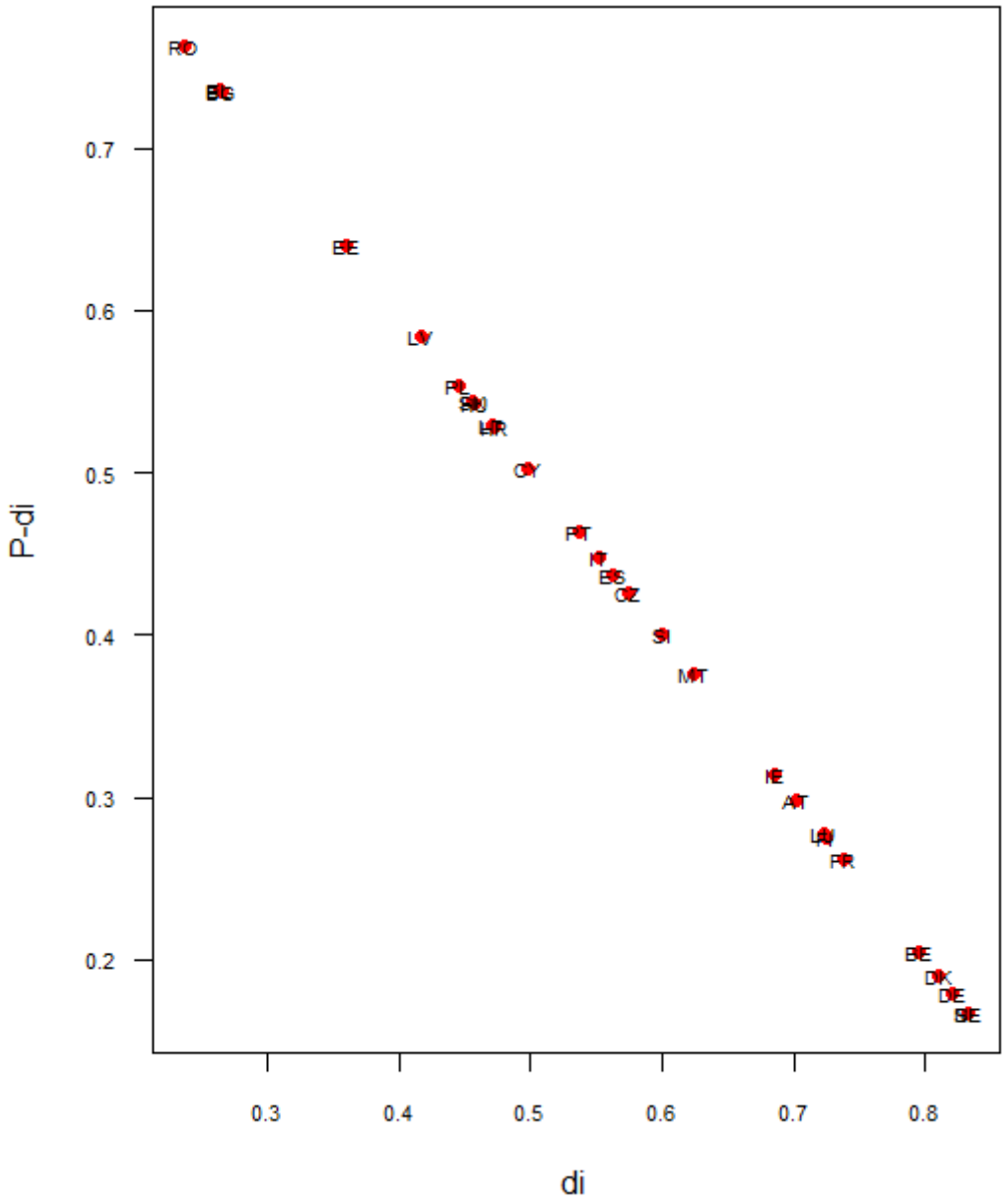
Obrázek 65: Vizualizace pěti shluků států DBSCAN algoritmem (SPCs), determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R

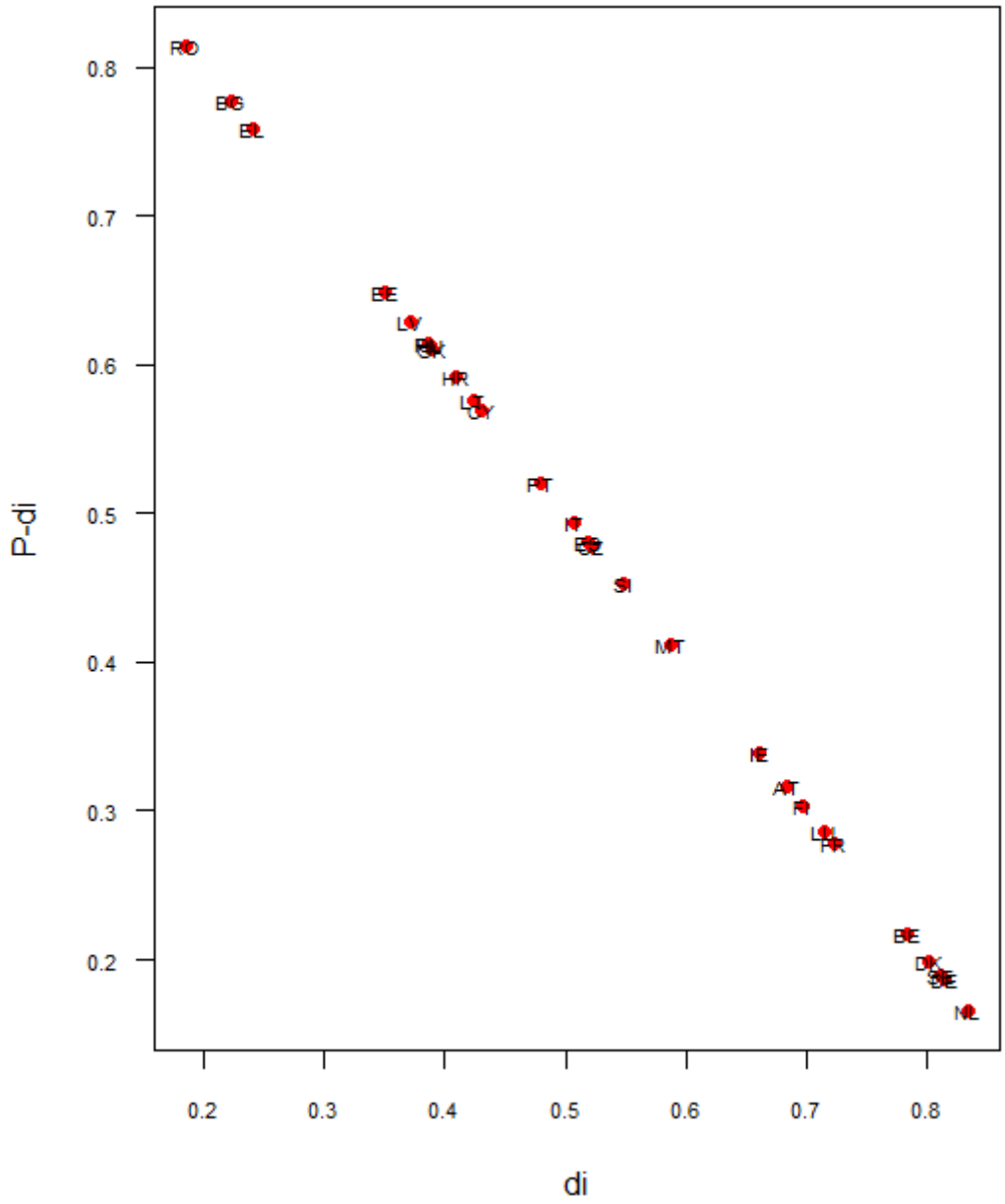
Příloha 15: Uspořádání států pomocí hybridního přístupu podle determinantů stavu zdraví

Obrázek 66: Lineární uspořádání států, datový soubor 1A, determinanty stavu zdraví



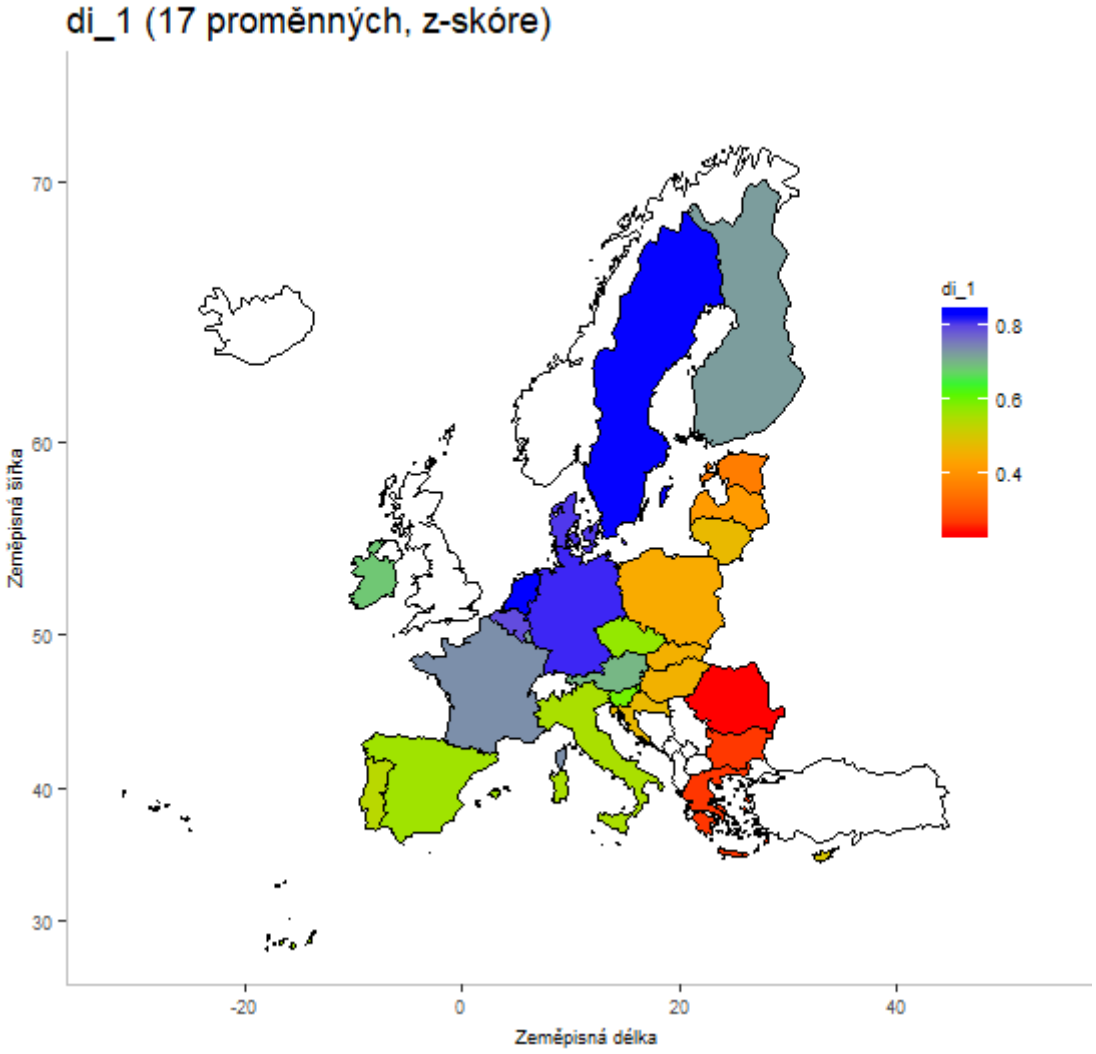
Zdroj: vlastní zpracování v programu R

Obrázek 67: Lineární uspořádání států, datový soubor 1B, determinanty stavu zdraví



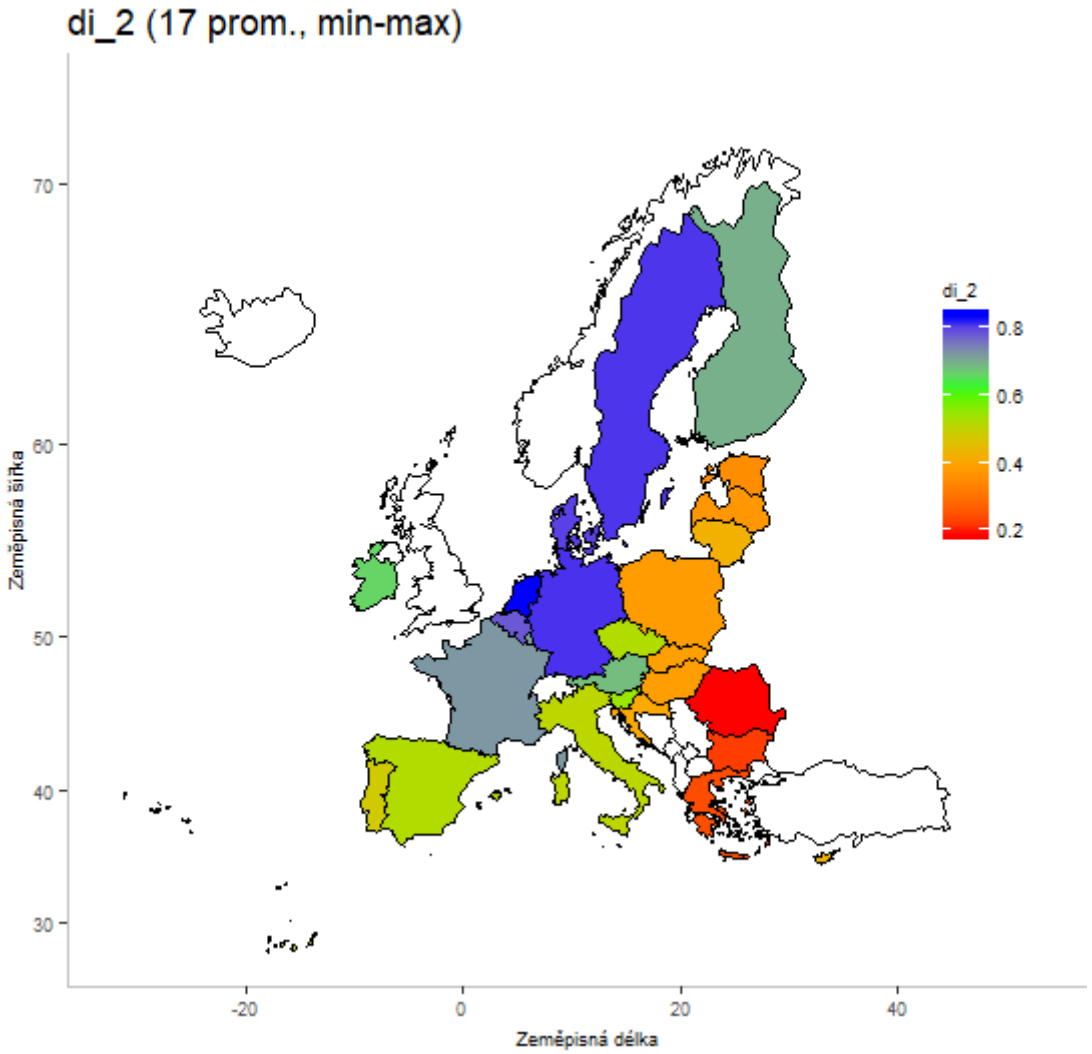
Zdroj: vlastní zpracování v programu R

Obrázek 68: Vizualizace agregované míry (1A), determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R

Obrázek 69: Vizualizace agregované míry (1B), determinanty stavu zdraví



Zdroj: vlastní zpracování v programu R