doc. Ing. Iveta Stankovičová, PhD., Katedra informačných systémov, FM UK Bratislava

Posudok oponentky na dizertačnú prácu

Autorka – doktorandka: Ing. Lucie Zapletalová
Názov práce: Metody snížení dimenze pro modelování stavu zdraví

Téma práce

Zdravotný stav populácie krajiny je veľmi dôležité sledovať, lebo je to jeden z hlavných faktorov rozvoja ekonomiky. V stratégii EU je to dlhodobo sledovaný ukazovateľ a na zlepšenie zdravia sú prijímané rôzne opatrenia na úrovni krajín ale aj regiónov EU.

Ukazovatele stavu zdravia a jeho determinantov sú rozsiahle viacrozmerné databázy. Pri práci s takýmito údajmi je potrebné znižovať ich dimenziu, aby boli efektívnejšie využité pri tvorbe rôznych modelov a viacrozmerných analýz. Kvantitatívne metódy vyžadujú splnenie rôznych podmienok pri ich použití. Napríklad metódy zhlukovej analýzy vyžadujú nekorelovanosť vstupných premenných, ktorá je veľmi ťažko v reálnych údajoch dosiahnuteľná. Metódy zníženia dimenzie pomáhajú zabezpečiť takéto predpoklady vo viacrozmerných dátach a ich analýzach.

Téma práce je z tohto dôvodu veľmi dôležitá a zaujímavá. Aj keď v odbornej literatúre je venované veľa pozornosti tejto problematike, tak sú tu aj nezodpovedané otázky. Autorka sa rozhodla venovať hlavne problematike výberu metód na zníženie dimenzie v konkrétnej oblasti, v oblasti modelovania stavu zdravia a jeho determinantov. V tejto oblasti je definovaných príliš veľa rôznorodých ukazovateľov a na základe nich zhodnotiť zdravotný stav populácie v krajine alebo porovnať viaceré krajiny (napr. krajiny EU 27) je potom veľmi neprehľadné a komplikované.

Štruktúra a cieľ práce

Štruktúra práce je logická a vyvážená. Práca sa skladá z 8 častí (kapitol) a záveru. Prvá kapitola popisuje stav skúmania zdravia v krajinách Európy, ukazovatele a determinanty zdravia. V druhej kapitole autorka popísala súčasné prístupy znižovania dimenzie údajov. V tretej časti si vytýčila a popísala ciele práce. Štvrtá kapitola je venovaná použitej metodológii a údajom. V piatej kapitole sú popísané metódy použité v práci, konkrétne metódy pre znižovanie dimenzie, metódy zhlukovania objektov, hybridný prístup a vizualizácia geografických údajov v programe R. V šiestej kapitole sú tieto metódy aplikované a výsledky interpretované. Siedma časť obsahuje diskusiu a ôsma popisuje vedecké a praktické prínosy tejto predkladanej práce. V práci je celkovo 69 obrázkov a 26 tabuliek.

Práca je rozsiahla. Rozsah práce možno považovať až za prekročený. Spolu je to 165 strán, plus zoznam použitej literatúry na stranách 166 až 189, takmer 24 strán. Časť prílohy obsahuje 15 príloh, strany 194 až 237, t.j. spolu 43 strán. Prílohy obsahujú použité kódy programu R (40 strán kódov, aj na priloženom CD), výsledky použitých metód, zoznamy krajín EU podľa výsledných zhlukov, rôzne grafy a mapy.

Hlavný cieľ dizertačnej práce bol jasne stanovený a formulovaný už v abstrakte a úvode predkladanej práce takto:

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.

Na strane 62 v časti 3 Ciele dizertačnej práce takto autorka vytýčila 6 čiastkových cieľov, ktoré je potrebné splniť, aby bol dosiahnutý hlavný cieľ práce:

- *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.*

Metodika a metódy

Pre spracovanie tejto problematiky autorka použila štandardné postupy a vedecké metódy – predovšetkým deskripciu, analýzu, syntézu a komparáciu. Špecifické metódy, ktoré aplikovala na databázu údajov z Eurostatu. WHO a OECD o stave zdravia v krajinách EU-27 a údaje o determinantoch zdravia, boli z oblasti viacrozmerných metód. Celý proces získavania údajov z dostupných databáz, metód a vyhodnotenia výsledkov je názorne zobrazený na obrázku 4 (str. 66).

Je možné konštatovať, že použité metódy a algoritmy sú vhodné pre skúmanie, resp. analýzu, modelovanie, vizualizáciu a aplikáciu v danej oblasti.

Zhodnotenie výsledkov a prínos

Vybraná problematika dizertačnej práce je dostatočne analyzovaná na úrovni teoretickej aj empirickej. Pri spracovaní oboch častí práce autorka preukázala odbornú kompetentnosť, schopnosť samostatnej analytickej práce ako aj kreatívny vedecký prístup.

Možno konštatovať, že vytýčený hlavný cieľ práce a aj 6 čiastkových cieľov bolo splnených. Za prínosy práce možno považovať predovšetkým:

- Zostavenie prehľadu súboru 29 ukazovateľov stavu zdravia (str. 101, tab. 3) a následná ich analýza v krajinách EU-27 pomocou metód zníženia ich dimenzie (PCA, SPCA a kPCA).
- Interpretácia a vizualizácia troch až štyroch získaných rotovaných (Varimax rotácia) hlavných komponentov zo sady ukazovateľov o stave zdravia RC1 až RC4 (str. 106-109).
- Použitie, interpretácia a vizualizácia SPCA (riedka PCA), výsledky ktorej potvrdili výsledky PCA (str. 109-112).

- Použitie, interpretácia a vizualizácia KPCA (Kernel PCA), výsledky ktorej zlepšili analýzu, lebo medzi ukazovateľmi stavu zdravia sú nelineárne vzťahy (str. 117-120).
- Zhodnotenie stavu zdravia v krajinách EU-27. Vytvorenie 4 až 6 zhlukov krajín podľa stavu zdravia rôznymi zhlukovacími metódami. Geografická vizualizácia získaných zhlukov.
- Zostavenie prehľadu 27 ukazovateľov – determinantov zdravia (str. 136, tab. 11) a ich následná analýza v krajinách EU-27 pomocou metód zníženia dimenzie (PCA, SPCA a kPCA). Boli použité rovnaké metódy na zníženie dimenzie ako pri ukazovateľoch stavu zdravia (časť 6.6, od str. 137).
- Identifikácia a vizualizácia štátov EU-27 s podobnou úrovňou determinantov zdravia (časť 6.8, od str. 149).
- Identifikácia a vizualizácia štátov EU-27 s podobnou celkovou úrovňou stavu zdravia (vertikálna dimenzia) a determinantov zdravia (horizontálna dimenzia) na 2-rozmerných grafoch (časť 6.9, od str. 151). Krajiny sú zoskupené do štyroch kvadrantov (obr. 19 a 20) a úroveň dvoch dimenzií je interpretovaná. Čiže v práci bol vytvorený iným, novým spôsobom index celkového stavu zdravia. Ide o hybridný prístup.
- V časti 7 Diskusia (tab. 17, str. 156) sú prehľadne porovnané vlastnosti jednotlivých metód na zníženie rozmernosti ukazovateľov (premenných) a tiež porovnanie metód zhlukovej analýzy (tab. 18, str. 157).
- Práca podáva prehľad o použití rôznych balíkov programu R, ktoré sa môžu použiť na hodnotenie stavu zdravia v krajine a výsledky sa môžu použiť v riadení verejnej správy.

Poznámky a pripomienky k práci:

1. Na str. 63 spomínate, že je vhodné použiť pri tejto problematike niektorú z metodológií data miningu, napr. SEMMA metodológiu (SAS). Ja si myslím, že túto metodológiu v práci nepoužívate a ani nerobíte „data mining". Nemáte tzv. big data. Používate len údaje za 27 krajín EU v sledovanom roku a tak nerobíte ani výberový súbor (sample). Máte síce veľa premenných a to býva bežné aj v data miningu, ale tie big data nemáte a nerobili ste ani ďalšie kroky podľa SEMMA metodológie. V tejto práci sa skôr rieši zabezpečenie podmienok pre korektné vykonanie výpočtov rôznych viacrozmerných metód. Lebo počet štatistických jednotiek n = 27 krajín EU je veľmi malý rozsah súboru v porovnaní s 29 premennými stavu zdravia, resp. 27 determinantami zdravia. Preto je potrebné znížiť dimenziu.
2. V prílohe 3, v tabuľkách 19 a 20 sú v posledných troch stĺpcoch použité symboly (skratky) h2, u2 a com (com aj v tabuľkách 21 a 22), ktoré nie sú v práci nikde vysvetlené. Prosím, vysvetlite ich.
3. Práca je veľmi rozsiahla. V časti 6 je veľa textu (interpretácií), ale výsledky výpočtov sú buď v prílohách alebo len slovne popísané. Často chýbajú výsledky priamo v tabuľkách a grafoch ako výsledky z použitého softvéru R.

Otázky do diskusie:

4. Ktorú z troch použitých metód na zníženie dimenzie (PCA, SPCA a kPCA) by ste odporučili pre prax v danej oblasti skúmania úrovne zdravia a jeho determinantov na regionálnej a národnej úrovni?

5. Ako sú dostupné údaje na úrovni NUTS2 za Českú republiku o stave zdravia a jeho determinantoch?

Záver

Predložená dizertačná práca napĺňa kritériá stanovené pre tento druh prác. Autorka preukázala potrebnú odbornú erudovanosť a schopnosť aplikovať teoretické poznatky na riešenie reálnych problémov teórie aj praxe.

Prácu odporúčam k obhajobe pred príslušnou komisiou. Po úspešnej obhajobe odporúčam udeliť autorke práce vedeckú hodnosť Ph.D.

V Bratislave dňa 3.3.2023

POSUDOK ZÁVEREČNEJ PRÁCE

Téma: Metody snížení dimenze pro modelovaní stavu zdraví

Typ záverečnej práce: Dizertačná záverečná práca

Autor: Ing. Lucie Zapletalová

Oponent: doc. Ing. Mária Vojtková, PhD.

Kritériá hodnotenia záverečnej práce:

1. *Stanovenie cieľa a miera jeho splnenia*

Téma práce je z praktického hľadiska vysoko aktuálna, vyhodnotenie stavu zdravia v krajinách EÚ poskytuje relevantný prehľad o zdravotnom stave a systémoch zdravotnej starostlivosti v EÚ so zameraním na osobitné charakteristiky a problémy jednotlivých krajín na základe porovnania medzi krajinami. Hlavný cieľ práce je sformulovaný veľmi široko, zahŕňa nielen porovnanie a využitie výsledkov lineárnych a nelineárnych techník metód zníženia dimenzie, následne zhlukovanie a lineárne usporiadanie, ale aj praktický cieľ zameraný na modelovanie celkovej úrovne stavu zdravia a determinantov stavu zdravia v krajinách EÚ-27. Stanovený cieľ práce doplnený praktickými čiastkovými cieľmi doktorandka v plnej miere splnila.

Na druhej strane názov práce nie úplne korešponduje so stanoveným cieľom. Vzhľadom k veľkému počtu ukazovateľov charakterizujúcich stav zdravia a determinantov tohto stavu je zrejmé, že vybrané ukazovatele sú vo veľkej miere prepojené a využitie metód zníženia dimenzie je nesporné. Samotná práca však zahŕňa aj využitie ďalších viacrozmerných metód.

2. *Jazyková úroveň a používanie správnej odbornej terminológie*

Práca má požadovanú štylistickú a jazykovú úroveň. V dizertačnej práci je vo všeobecnosti použitá správna odborná terminológia. Otázne je však použitie metodológie data miningu SEMMA, ktorú autorka v práci spomína. Z formálneho hľadiska jej možno vytknúť zopár typografických chýb, pričom jednotlivé pojmy, prípadne štatistiky sa využívajú v texte skôr (kapitola 2) ako sú zadefinované.

3. *Vhodnosť použitých metód, metodológia*

Problematiku metodológie a použitých metód skúmania autorka podrobne opísala v kapitole 4 a 5, pričom samotný proces názorne prezentuje na obrázku 4 na str. 66. Úvodná časť je venovaná opisu databázy s rozdelením dát do dvoch skupín a to ukazovatele stavu zdravia a determinanty stavu zdravia v krajinách EÚ. Nasleduje popis troch základných metód a to metód zníženia dimenzie, metód zhlukovania objektov a hybridný prístup kombinujúci viacrozmerné škálovanie s lineárnym usporiadaním. V závere tejto časti práce autorka uvádza možnosti vizualizácie geografických dát v programe R prostredníctvom konkrétnych balíkov programov. Na všetkých spomenutých metódach je následne postavená analýza dát v kapitole 6.

Z hľadiska použitých metód a metodológie je možné konštatovať, že autorka dizertačnej práce preukázala schopnosť samostatne vedecky pracovať, použité metódy považujem za vhodné a aplikáciu za relevantnú s ohľadom na predpoklady, ktoré si autorka stanovila.

4. *Zhodnotenie poznatkovej bázy*

Nevyhnutným predpokladom spracovania dizertačnej práce bolo naštudovanie poznatkov z oblasti použitých metód ako aj stavu skúmania zdravia v krajinách Európy. Doktorandka pri spracovaní dizertačnej práce využila aktuálnu zahraničnú a domácu literatúru z predmetných oblastí, pričom v

predloženej práci sa odvoláva a cituje veľký počet zdrojov uvedených na str. 166 až 189. Z toho následne pramení práca napísaná v rozsahu 165 strán textu (237 strán vrátane použitej literatúry a príloh), pričom možno polemizovať o prekročení odporúčanej hranici rozsahu. Kapitola 6 obsahuje rozsiahle interpretácie s odvolávkami na prílohy, čo prácu nesprehľadňuje avšak nemá vplyv na kvalitu práce.

Doktorandka preukázala schopnosť pracovať s odbornou a vedeckou literatúrou, prepájať poznatky a vhodne ich aplikovať na reálnych dátach prostredníctvom „open source" softvéru v programe R.

5. Vedecký prínos a originalita práce

Doktorandka pri spracovaní dizertačnej práce realizovala vedecký výskum na databáze 27 krajín EÚ, ktoré charakterizovala pomocou ukazovateľov rozdelených do dvoch skupín: ukazovatele stavu zdravia (29) a determinantov stavu zdravia (27). Databáza pochádza z údajov Eurostatu, WHO a OECD. Aj keď závery z vlastných analýz sa nedajú zovšeobecniť, práca poskytuje všeobecné postupy:

- lineárnych a nelineárnych metód zníženia dimenzie (ich koncepčné porovnanie je prehľadne zhrnuté v tabuľke 17 na str. 156),
- porovnania a nastavenia optimálnych hyperparametrov pri metóde riedkej analýzy hlavných komponentov (SPCA), Kernel analýzy hlavných komponentov (KPCA) ako aj optimálneho počtu zhlukov v metóde k-priemerov,
- identifikácie skupiny štátov s podobnou situáciou v stave zdravia pomocou vybraných metód zhlukovej analýzy (ich koncepčné porovnanie je prehľadne zhrnuté v tabuľke 18 na str. 157),
- výberu vhodných vizualizačných techník v programu R.

6. Aplikačné prínosy práce pre prax

Dizertačná práca poskytuje poznatky a postupy aplikácie viacrozmerných štatistických metód, ktoré sú prínosné hlavne v oblasti verejného zdravotníctva na:

- nájdenie skupín krajín EÚ s podobnou úrovňou stavu zdravia a ako aj determinantov stavu zdravia,
- navrhnutie agregovaných mier celkovej úrovne stavu zdravia a celkovej úrovne determinantov stavu zdravia v krajinách EÚ,
- prehľad použitých balíkov programu R, ktoré môžu byť ďalej využité pre praktické účely nielen v oblasti zdravotníctva, ale aj v iných oblastiach verejnej správy.

Okrem uvedených prínosov môžeme vnímať aj prínos v rovine vzdelávacej činnosti.

7. Otázky pre autora pri obhajobe práce

a) V práci pri identifikácii štátov s podobnou celkovou úrovňou stavu zdravia ako aj determinantov stavu zdravia a ich lineárneho usporiadania na str. 133 a str. 150 ste k pôvodným dátam pripojili dva objekty vzor (P) a anti-vzor (AP). Môžete bližšie vysvetliť, ako boli tieto objekty definované?
b) Na ktoré ukazovatele stavu zdravia a determinanty stavu zdravia v Českej republike resp. v iných krajinách EÚ-27 mala najväčší dopad pandémia Covid-19? K akému posunu, čo sa týka celkovej úrovne stavu zdravia podľa Vášho názoru došlo?
c) Predpokladáte využitie výsledkov Vašej práce v reálnom prostredí?

8. Záverečné odporúčanie

Konštatujem, že predložená práca spĺňa požiadavky kladené na dizertačné práce z hľadiska obsahu, formy, prínosu a stanoveného cieľa. Prácu odporúčam na obhajobu a po jej úspešnom obhájení navrhujem udeliť Ing. Lucie Zapletalovej titul „philosophiae doctor", v skratke „Ph.D."

V Bratislave, dňa 6.3.2023

Podpis 

Posudek na disertační práci
Název práce: **Metody snížení dimenze pro modelování stavu zdraví**
Doktorandka: **Ing. Lucie Zapletalová**
Oponentka: prof. Ing. Hana Řezanková, CSc.

Disertační práce se zabývá zejména vyhodnocením a porovnáním lineárních a nelineárních metod pro snížení rozměrnosti množiny ukazatelů stavu zdraví a jejich determinantů. Získané poznatky jsou využity při předzpracování reálných dat, která jsou následně analyzována za účelem posouzení nerovností ve stavu zdraví v zemích EU-27. Práce se tedy zaměřuje na metody analýzy dat a má jednak komparační, jednak aplikační charakter.

Práce je poměrně rozsáhlá a vlastní text je doplněn o přílohy v rozsahu více než 40 stran. Text je přehledný bez gramatických chyb (jen výjimečně někde chybí čárka v souvětí) a dobře se čte. Grafy v obrázcích i tabulky jsou připraveny pečlivě. Práce obsahuje seznam obrázků, tabulek, použitých obecných zkratek a zkratek států.

I přes zřejmou pečlivost autorky lze nalézt drobné nepřesnosti a v některých částech mohlo být postupováno trochu jinak, což však nesnižuje celkovou kvalitu disertační práce. Některé mé níže uvedené poznámky k formální úpravě a řazení kapitol mohou být ovlivněny neznalostí zvyklostí při přípravě disertační práce na Fakultě ekonomicko-správní Univerzity Pardubice.

Kapitola obsahující cíle disertační práci je neobvykle zařazena až jako třetí a na rozdíl od ostatních kapitol je samozřejmě rozsahem podstatně úspornější. Cíle lze jistě formulovat dříve, než je popsána zkoumaná problematika. Ostatní text je uspořádán vhodně, a to jak co do obsahu a rozsahu kapitol, tak do jejich pořadí.

V publikacích většího rozsahu (knihy, kvalifikační práce) se obvykle používá dvouúrovňový způsob číslování tabulek a obrázků, stejně tak u vzorců (rovnic). I když předložená disertační práce obsahuje seznam obrázků a tabulek, přece jen orientace, do kterých kapitol jsou obrázky a tabulky zařazeny, by byla vhodnější.

Na str. 47 autorka uvádí pojem „Silhouetův koeficient". Jde skutečně o koeficient, jehož autorem je Silhouet? V textu je použit podobný termín, a to „koeficient siluet" (angl. Silhouette index); není zřejmé, zda jde o související pojmy. Z jakého důvodu je název metody „fuzzy k-průměrů" v názvech kapitol 5.2.2, 6.3.3 a 6.7.3 uváděn s velkým písmenem „F" (když v ostatním textu je správně používáno malé písmeno)?

V textu lze nalézt i v češtině neobvyklá spojení, jako „obyčejné nejmenší čtverce", což je sice doslovný překlad z angličtiny, ale v češtině se spíše používá pouze termín „metoda nejmenších čtverců". V R prostředí jsou spíše „balíčky" než „balíky". Stejný termín se vyskytuje ve dvou podobách: je používán jak zápis COVID-19, tak zápis Covid-19.

Z dalších drobností lze uvést nesprávné použití „viz.“. Slovo „viz“ není zkratkou, jedná se o rozkazovací způsob slovesa vidět, proto se za ním nepíše tečka. Jako symbol násobení by v textu neměla být uváděna hvězdička, ale některý ze speciálních symbolů (křížek nebo tečka). Pokud je název metody (indexu, koeficientu apod.) odvozen od jména (jmen) autorů, pak by měla být zohledněna pravidla českého jazyka. Stejně jako autorka používá termíny Pearsonův či Spearmanův koeficient, tak by také měl být uváděn Kaiserův-Meyerův-Olkinův index.

Z některých nepřesností lze zmínit vyjádření „pokud se nevyskytují průměrná pořadí hodnot náhodných veličin“. Náhodná veličina může obsahovat stejné hodnoty. Jednou z možností při stanovení pořadí je přiřadit těmto stejným hodnotám průměrná pořadí. Takže aby se průměrná pořadí vyskytovala, nejdříve se musí stanovit (podstatné jsou ty některé stejné hodnoty). Ale to je samozřejmě drobnost z hlediska formulace.

Vybrané metody analýzy dat byly aplikovány na datový soubor s údaji zaměřenými na problematiku stavu zdraví v zemích EU. Detailně jsou popsány jednotlivé ukazatele, je zřejmé, o kterých zemích byly údaje do analýz zahrnuty, a jsou řádně uvedeny zdroje. Co však lze v textu obtížně dohledat, je období, za které byly ukazatele zjištěny. Období lze zjistit teprve na str. 99: „Data jsou získána pro období před pandemií Covid-19, ve většině případů pro rok 2019 nebo nejbližší dostupný rok.“ Konkrétně jsou roky uvedeny na str. 101 v tabulce 3. Takže časové údaje v disertační práci uvedeny jsou, ale mohly být zdůrazněny poněkud dříve v textu.

Celkově lze k obsahu disertační práce uvést, že stanovené cíle autorka splnila, jak je detailně komentováno v kapitole 7. V ní kromě popisu porovnávaných metod pro snižování rozměrnosti ukazatelů a charakterizování vybraných metod shlukové analýzy doktorandka přehledně uvádí především výhody a nevýhody jednotlivých metod z těchto dvou skupin.

Z hlediska vědeckého přínosu pro vědní obor „systémové inženýrství a informatika“ lze vyzdvihnout fakt, že byly vytvořeny algoritmy sloužící pro výběr optimálních hyperparametrů některých metod pro snižování rozměrnosti ukazatelů a metod shlukové analýzy.

Dotaz k obhajobě mám následující:

Které z nově získaných poznatků o metodách analýzy dat by bylo možné zobecnit pro využití v jiných oblastech, než jsou data o stavu zdraví v zemích EU kolem roku 2019?

Závěrem lze konstatovat, že předložená disertační práce splňuje obvyklé požadavky kladené na kvalitu, rozsah a hloubku zpracování disertačních prací. Nevyskytují se v ní žádné závažné nedostatky. Proto disertační práci Ing. Lucie Zapletalové doporučuji k obhajobě před komisí pro obhajoby disertačních prací.

V Praze 27. 2. 2023

prof. Ing. Hana Řezanková, CSc.