SHLUKOVÁ ANALÝZA V PROBLEMATICE DANÍ

Jarmila Rybová

Abstract

Cluster analysis is a method of multivariate data processing. It can be easily applied, for example, by means of statistical software. The article focuses on an application of cluster analysis in the field of taxation. The possibility of applying the method is compared for each tax or groups of taxes at two levels: by an individual state and between selected countries. It is also possible to determine, on the basis of cluster analysis, the impact of taxes on economic operators, which can be sorted according to various criteria (income, types of households represented by their members, etc.). An overview of studies in which the selected method of cluster analysis applied to selected datasets shows that cluster analysis has been supplemented by another statistical method or that the clustering process is repeated on a selected part of a set of objects or characters. The reason is usually better interpretation of results and complementary broader context.

Keywords: cluster analysis, taxation, comparison of taxes, multivariate data analysis
JEL Classification: C38, H21, H23, H24

Úvod

Cílem tohoto článku je zmapovat možnosti použití shlukové analýzy na soubory dat z oblasti daní k jejich mezinárodnímu porovnávání. Článek by se mohl stát návodem aplikace shlukové analýzy při zkoumání odlišností daňových systémů, skupin či jednotlivých daní mezi vybranými státy.

Shluková analýza je využívána v široké škále oborů lidské činnosti od přírodních věd až po oblast ekonomických analýz. Metodické postupy v rámci shlukové analýzy nabízejí možnost výběru nejvhodnější použitelné metodiky zpracování dat, zobrazení výstupů a interpretaci výsledků daného problému. Výhodou dané metody je její jednoduchost, názorný grafický výstup zjištěných výsledků a možnost použití statistických softwarů, které tuto metodu zpravidla nabízejí v rámci analýz vícerozměrných dat.

Nevýhody shlukové analýzy lze spatřovat v její náchylnosti ke zkreslování výsledků, jsou-li v souboru dat přítomny objekty výrazně odchýlené od ostatních. Soubory použitých dat je třeba kontrolovat, otestovat použití zvolené metody shlukové analýzy, zvážit standardizaci dat, případně odlehlé objekty zcela odstranit. Další nevýhodou může být nutnost opakování shlukovacích procesů nebo doplnění shlukové analýzy další statistickou metodou případně další prací s daty. Důvodem je zpravidla potřeba podrobnější interpretace výsledků.

1. Popis analýzy shluků

Analýza shluků, shluková analýza (*cluster analysis*) nebo také clusterová analýza (CLU) je řazena mezi metody, které se zabývají zjišťováním podobnosti vícerozměrných objektů (tj. objektů, u nichž je měřeno více proměnných) a jejich klasifikací do tříd či shluků. Tato metoda je vhodná u objektů, které mají přirozenou tendenci se seskupovat [Meloun a kol. 2005, s. 268].

Autoři studijního textu *Shluková analýza* – Jan Kelbel a David Šilhán – uvádějí, že cílem shlukové analýzy je „v dané množině objektů nalézt její podmnožiny – shluky objektů – tak, aby si členové shluku byli navzájem podobní, ale nebyli si příliš podobní s objekty mimo tento shluk " [2015, s. 1]. Shluk neboli cluster definuje Meloun a Militký jako „skupinu objektů, jejichž vzdálenost (nepodobnost) je menší než vzdálenost (respektive nepodobnost), které mají objekty do shluku nepatřící" [2004, s. 461]. Dále autoři formulují tři hlavní cíle analýzy shluků:

- popis systematiky, jenž je tradičním využitím shlukové analýzy pro průzkumové cíle a taxonomii, což je empirická klasifikace objektů;
- zjednodušení dat, kdy analýza shluků poskytuje při hledání taxonomie zjednodušený pohled na objekty;
- identifikaci vztahu, kdy po nalezení shluků objektů, a tím i struktury mezi objekty, je snadnější odhalit vztahy mezi objekty [ibid., s. 454].

Základní myšlenkou analýzy je podobnost zkoumaných objektů. Objektem může být předmět nebo jev charakterizovaný určitými stavy vybraných kvalitativních či kvantitativních znaků vyjádřených číselnými hodnotami. Do shluku jsou seskupeny objekty, které vykazují určité známky podobnosti. Nalezené shluky vystihují strukturu dat pouze s ohledem na vybrané znaky. Všechny znaky jsou v této metodě považovány za stejně významné. Nesprávné zařazení či výběr znaků vede ke zkreslení výsledků analýzy.

Shluky vytvořené analýzou graficky znázorňují vztahy mezi objekty, jedná se o tzv. dendrogram či stromový diagram s vertikálním či horizontálním zobrazením spojů mezi objekty. Podobnost objektů mezi sebou může být zjišťována různými způsoby, tj. použitím míry korelace, míry vzdálenosti nebo míry asociace. Každá z uvedených měr představuje specifický pohled na podobnost objektů. „Korelační a vzdálenostní míry jsou míry metrických dat, zatímco asociační míry jsou určeny spíše pro nemetrická data" [Meloun, Militký, 2004, s. 455].

Některé míry pro metrická data lze alespoň stručně uvést, jedná se o korelační a vzdálenostní míry. Korelační míry tvoří jednu ze tří skupin podobnostních měr. Základní mírou podobnosti dvou objektů či znaků Xi a Xj vyjádřených v kardinální škále (vzájemný poměr užitnosti jednotlivých variant podle 1 kritéria, intervalové a poměrové proměnné) může být Pearsonův párový korelační koeficient **r**. Podobnost objektů vyjádřených v ordinální škále (pořadová čísla) lze vyjádřit Spearmanovým korelačním koeficientem.

Vzdálenostní míry představují zobrazení objektů v prostoru. Lze uvést alespoň některé typy měr či metrik. Souřadnice objektu tvoří jednotlivé znaky. Zde je možné použít eukleidovskou vzdálenost, zvanou také geometrická metrika, která představuje délku přepony pravoúhlého trojúhelníka a její výpočet je založen na Pythagorově větě. Základem např. Wardovy metody je čtverec eukleidovské vzdálenosti, který je další

vzdálenostní měrou. Dále se lze setkat s Manhattanskou vzdáleností, zvanou též Hammingova metrika. Zde je třeba připomenout, že znaky by měly být vzájemně nezávislé, což ovlivňuje tvorbu shluků.

Metody shlukové analýzy

V procesu shlukové analýzy lze uplatnit metody, které dělíme na hierarchické a nehierarchické. Hierarchické shlukovací metody se podle směru shlukování člení na aglomerační, resp. aglomertativní metody a divizní metody shlukování.

Kelbel a Šilhán popisují hierarchické shlukování jakožto „sekvenci vnořených rozkladů, která na jedné straně začíná triviálním rozkladem, kdy každý objekt dané množiny objektů tvoří jednoprvkový shluk, a na druhé straně končí triviálním rozkladem s jedním shlukem obsahujícím všechny objekty“ [2015, s. 5]. Meloun a Militký [2004] k hodnocení aglomeračních hierarchických metod dodávají, že při použití různých metod hierarchického shlukování často dochází k velmi odlišným výsledkům. Důvodem je zpravidla to, že jednotlivé metody ovlivňují prostor mezi objekty. Je obtížné říci, která metoda je pro daný problém jednoznačně nejlepší. V mnoha případech nejlépe vyhovuje Wardova metoda, která prostor mezi objekty rozšiřuje tvorbou kompaktních shluků s velkým počtem objektů.

Výstupem hierarchického shlukovacího procesu jsou dendrogramy, které lze vytvářet na podkladě podobnosti objektů či znaků. Dendrogram podobnosti objektů analyzuje vzdálenosti mezi objekty, je zde patrná struktura objektů uvnitř shluků. Dendrogram podobnosti znaků odhaluje velmi podobné znaky (*n*-tice znaků), které mohou být navzájem nahraditelné. Při sestrojování dendrogramu je třeba posoudit těsnost proložení, čili míru věrohodnosti. Pro hodnocení nejvhodnějšího dendrogramu lze použít dvě kritéria [Meloun, Militký, 2004, s. 469]. Prvním kritériem je kofenetický korelační koeficient (CC), měří Pearsonův korelační koeficient mezi skutečnou vzdáleností a vzdáleností získanou z dendrogramu. Druhým je kritérium delta Δ, které měří stupeň přetvoření struktury dat. Je žádoucí, aby kritérium delta Δ dosahovalo hodnot co nejbližších nule.

Hierarchické shlukovací postupy jsou rychlé, nejsou vhodné pro analýzy velkých výběrů a zároveň jsou velmi citlivé na existenci odlehlých objektů. Výsledky hierarchického shlukování mohou být zavádějící, neboť nežádoucí předešlé kombinace v průběhu analýzy přetrvávají a mohou vést k umělým výsledkům.

Nehierarchické shlukování může mít oproti hierarchickému několik výhod. Odlehlé body nemají tak silný vliv na výsledky shlukovacího procesu a lze k analýze použít i méně vhodné znaky. Kvalita dosažených výsledků se odvíjí od schopnosti uživatele vybrat tzv. zárodkové body. Při nehierarchickém shlukování je třeba nalézt optimální počet shluků, provést počáteční rozklad na výsledné shluky a zjistit tzv. středové body jednotlivých shluků. Optimální počet shluků lze stanovit prostřednictvím indexů, např. Kelbel a Šilhán [2015, s. 7] uvádějí Calinski–Harabascův index, C index, Goodman–Kruskal index. V dalších krocích je třeba stanovit shlukové body prostřednictvím vybrané metody či algoritmu a určit konečný počet shluků. Např. metoda K-means (MacQeenův algoritmus publikovaný v roce 1967) je poměrně jednoduchou metodou. Základní algoritmus této metody používá eukleidovskou vzdálenost. Autoři k metodě K-means doplňují: „Experimentální zkušenosti naznačují, že je tento algoritmus pro účely vektorové kvantizace

velmi vhodný jak z hlediska kvality rozkladu trénovací množiny do shluků, tak i poměrně rozumnými výpočetními nároky“ [ibid., s. 9].

2. Metodický postup při aplikaci shlukové analýzy

Před vlastní analýzou dat se stanoví jedna nebo více pracovních hypotéz. Důraz je třeba klást na přesnost stanovení pracovní hypotézy, která bude analýzou následně potvrzena či vyvrácena, tj. prostřednictvím výběru objektů a znaků. Výchozí hypotézy určují, co přesně chceme na základě dat zjistit, to znamená ujasnit si zkoumaný problém a předpokládané výsledky.

Posléze jsou vybrány objekty a jejich znaky vhodné pro analýzu. Při zkoumání podobností a odlišností např. vybraných daní v různých státech je třeba, aby ekonomické ukazatele, které byly vybrány jakožto znaky pro analýzu, byly zjistitelné ve všech vybraných státech ve stejných jednotkách nebo alespoň převoditelné na shodné jednotky. Pokud tomu tak není a některý stát neposkytuje potřebné údaje, je možné doplnit vlastní odhad. Nepřesnost odhadu se však může projevit na zkreslení výsledků shlukové analýzy. Totéž platí i pro zkoumání odlišností mezi jednotlivými daněmi nebo skupinami daní v rámci jednoho státu. Znaky, které charakterizují jednotlivé objekty, by měly být vzájemně nezávislé. Výběr znaků pro analýzu zajišťuje správnost výsledků a jejich interpretace.

Volba metody shlukové analýzy a algoritmů shlukovacích procesů je klíčovým prvkem úspěšnosti analytické činnosti, významně ovlivňuje zobrazení výsledků v dendrogramu, což také následně slouží interpretaci výsledků. Jednou z možností je podrobit data předběžně několika přijatelným metodám shlukové analýzy a podle zobrazení dendrogramu vybrat metodu nejvhodnější. Zobrazení shluků, patrnost odlišností a podobností mezi objekty a shluky je rozhodující. Předběžné zobrazení více metod shlukové analýzy může upozornit na odlehlé objekty, či potřebu standardizace dat.

Problémem této metody může být existence odlehlých objektů, které se výrazně odlišují od ostatních. Na výrazně odchýlené objekty v souboru je tato metoda velice citlivá. Zkreslení výsledků analýzy vlivem odlehlých objektů lze zabránit předběžným prověřením objektů pomocí profilového diagramu znaků či jiným způsobem znázornění dat a následně jejich uvážlivým odstraněním.

Další neméně důležitou otázkou je standardizace dat, která se může týkat znaků i objektů. Standardizaci dat je možné realizovat podle objektů nebo podle znaků. Znaky vyjádřené numerickými hodnotami mohou napříč objekty vytvářet odlišné stupnice číselných hodnot. Každý numericky vyjádřený znak může vykazovat jinou stupnici číselných hodnot a tyto rozdíly stupnic mohou ovlivňovat míru podobnosti objektů. Většina měr vzdálenosti je velmi citlivá na různé stupnice numerické velikosti znaků. Jsou-li numericky vyjádřené znaky uvedeny v odlišných měrných jednotkách, znaky je vhodné standardizovat. Dojde tím k eliminaci vlivu rozdílů stupnic mezi různými znaky. Meloun a Militký uvádějí výhody standardizace znaků: „Znaky v jednotném měřítku, kde je průměrná hodnota 0 a směrodatná odchylka 1, lze porovnávat snadněji. Se změnou měřítka nedojde k rozdílu mezi standardizovanými znaky“ [2004, s. 461]. Standardizace objektů, tzv. řádková, může být efektivní ve speciálních případech, např. pro odstranění nekonstantnosti rozptylu.

Kelbel a Šilhán ke standardizaci znaků zmiňují: „Hodnoty jednotlivých znaků objektů jsou často v různých jednotkách. To může způsobovat, že se určité znaky jeví jako

dominující a jiné znaky jen málo ovlivňují průběh shlukování. Někdy je proto výhodné data upravit tak, aby byly všechny znaky souměřitelné. Jedním ze způsobů, jak toho docílit, je standardizace dat“ [2015, s. 3].

Potřebu standardizace dat lze testovat zkušebním shlukovacím procesem dat (vybranou metodou nebo několika vybranými metodami), a to před standardizací i po uskutečnění standardizace a následně porovnat výsledky zobrazení dendrogramů. Standardizovat lze všechna data souboru objektů a znaků nebo jen vybrané znaky či objekty. Podle mého názoru je ve většině případů vhodné standardizovat všechny znaky celého souboru objektů, čímž by měla být zajištěna souměřitelnost všech výchozích dat.

Následně lze přistoupit k vlastní shlukové analýze. Nastavení podmínek shlukovacího procesu závisí na nabídce a možnostech použitého softwaru. Pro shlukovací proces mohou být vybrány všechny sledované znaky a objekty nebo jen některé z původního souboru. Grafické zobrazení výsledků, tj. dendrogram, může mít zpravidla horizontální či vertikální polohu.

K vyhodnocení výsledků dendrogramu můžeme použít více postupů. Prvním z nich je postup od vzájemně nejbližších objektů k nejvzdálenějším nebo naopak. Počet shluků není v tomto případě nastaven pevně. Při dalším postupu lze sledovat předem stanovený požadavek týkající se počtu shluků nebo vzdálenosti mezi objekty. Vždy je třeba zohlednit vazby mezi objekty či shluky a jejich těsnost. Vztahy mezi podobnými objekty jsou zobrazeny spojovacími čarami. Vzdálenost objektů je vyčíslena na ose grafu. Stanovení požadované vzdálenosti mezi objekty nebo počet shluků podstatně ovlivňuje interpretaci zjištěných výsledků. Např. od vzdálenostního bodu 5 jsou objekty považovány za velmi vzdálené a počet shluků se bude odvíjet od vzdálenosti mezi objekty, která převyšuje 5 bodů.

Pro podrobnější a přesnější vyjádření vztahů mezi objekty ve shluku je žádoucí mít k dispozici výchozí nestandardizovaná data zúčastněných objektů. Jsou-li za znaky objektů vybrány různé ekonomické veličiny, je možné zvážit, zda by mohlo být přínosné tyto veličiny vzájemně porovnávat a uvádět do vzájemných vztahů, např. použít pro vyhodnocení poměrové ukazatele vypočtené z výchozích hodnot znaků za jednotlivé objekty. Možností, jak rozšířit či doplnit interpretaci výsledků, je aplikace obou shlukovacích postupů hierarchických a nehierarchických, čímž lze získat výhody obou shlukovacích metod.

3. Přehled studií

Studie, které se zabývají shlukovou analýzou v oblasti daní, představují poměrně široké spektrum. Vzhledem k omezenému rozsahu článku není výčet studií úplný. Stručně jsou zde uvedeny studie, které se týkají daňové problematiky v letech 1994–2012, v nichž je použita metoda shlukové analýzy, zejména studie z on-line databáze Web of Science, konkrétně ProQuest.

První studie je součástí disertační práce. Její autorka, Michaela Čimborová [2007], prostřednictvím hierarchické shlukovací metody měřící eukleidovskou vzdálenost mezi objekty, identifikovala rozdíly daňových mixů ve vybraných 21 státech OECD v letech 1965–2003 s cílem analyzovat vliv globalizace na jejich vývoj. Shluková analýza byla doplněna další analytickou metodou, konkrétně diskriminační analýzou. Shlukovacímu procesu bylo tedy podrobeno 21 států OECD z různých kontinentů v období 1965–2003

s cílem vytvořit historický průřez vývojem jejich daňových mixů. Celkově bylo uskutečněno devět shlukovacích procesů.

Výsledky analýzy v daňových strukturách prokazují pokles podílu příjmových daní ve spojení s růstem podílu příspěvků na sociální zabezpečení a daní ze spotřeby. Daňové mixy vyspělých států se jeví relativně stabilními. Koncem vymezeného období je možné u těchto států pozorovat vlivy liberalizace a globalizace, zároveň nadále přetrvávají určité národní odlišnosti a zvyklosti dané hlubokými historickými, politickými, sociálními a ekonomickými znaky státu.

Italský autor A. M. Serrano [1994][1] zkoumá jeden typ daní. Konkrétně se jedná o osobní důchodové daně v členských zemích EU-12, které jsou charakterizovány několika znaky[2]. Výsledků je dosaženo pomocí dvou shlukovacích procesů aplikovaných opakovaně na vybrané státy EU. Při prvním shlukovacím procesu byly vytvořeny tři shluky, z toho jeden shluk obsahuje jeden stát, proto zde již další analýza neproběhla. Další dva shluky tvoří více států, oba shluky jsou následně podrobeny dalšímu shlukovacímu procesu s cílem podrobnější a přesnější interpretace výsledků.

Studie autorů Allana a Iglarshe [1996] se zabývá horizontální nespravedlností ve federálním systému daní z příjmů v roce 1988, která je způsobena rozdílným daňovým zacházením s různými zdroji příjmů. Horizontální daňovou spravedlnost definuje např. Musgrave [1959, s. 160] jakožto rovné zacházení s daňovými poplatníky, kteří dosahují stejných úrovní příjmů. Shluky zde oddělují poplatníky s různými zdroji příjmů do skupin. V různých shlucích se vyskytují různé úrovně efektivní daňové sazby a rovnoměrnosti explicitního zdanění. To vede k hypotéze, že zdroj příjmů je faktorem horizontální nespravedlnosti. Ve většině příjmových kategorií platí, že shluky obsahující mzdové příjmy ze zaměstnání, mají nejvyšší efektivní daňovou sazbu a zdanění je relativně rovnoměrnější než u poplatníků s jinými hlavními zdroji příjmů. Analýza rozptylu umožňuje zjistit, které zdroje příjmů mají vyšší jednotnost explicitního zdanění.

Autorky Gnip a Tomic [2010] analyzují prostřednictvím shlukové analýzy daňové zatížení příjmů na trhu práce daní z příjmů, platbami sociálního pojištění a DPH v Chorvatsku a dalších státech Evropské unie v kontextu zaměstnanosti a nezaměstnanosti. V rámci shlukové analýzy je použit algoritmus shlukování K-means (tato metoda rozděluje objekty do shluků, které mají srovnatelný prostorový rozsah) a hierarchický shlukovací proces. Studie ukazuje, že důvodem pro vysokou míru nezaměstnanosti je vysoké daňové zatížení práce uvedenými daněmi. Z výsledků shlukové analýzy lze odvodit, že v zemích EU s vysokým daňovým zatížením práce konkrétními daněmi se vyskytuje také poměrně vysoká nezaměstnanost. Podobné výsledky autorky zjistily při použití efektivní daňové sazby, ale zde podotýkají, že se nejedná o pravidlo. Existují státy, kde výše uvedené tvrzení neplatí.

Nejvyšší daňové zatížení na trhu práce nalezly autorky ve „starých" členských státech. Ve většině nových států EU je zatížení nižší (mimo Maďarska a Slovinska). Výše daňové zátěže na trhu práce (zejména u zaměstnanců) je subjektivně vnímána citlivěji v nových členských státech, neboť výše hrubé mzdy je v nových členských státech EU

1 Tohoto autora cituje také Čimborová [2007, str. 119].

2 Znaky osobní důchodové daně pro shlukovou analýzu: předmět daně, základ daně, minimální a maximální daňové sazby, počet daňových pásem, standardní odpočty, efektivní průměrné daňové sazby.

výrazně nižší než v těch starých. V případě Chorvatska je na podkladě výsledků této studie doporučeno snížit daňové zatížení, čímž by se zlepšily životní podmínky a zvýšila nabídka pracovní síly. S ohledem na situaci v Chorvatsku se jeví vhodným i snižování veřejných výdajů. Na straně zaměstnanců by vyšší disponibilní příjmy mohly generovat vyšší výdaje na spotřebu, a tím i zvýšení příjmů z jiných daní, např. DPH. Obdobnou problematiku zdanění příjmů z práce jako tato studie řeší také Trpeski a Tashevska [2012] tentokrát se zaměřením na Makedonii.

Problematika převodních cen, která je součástí účetnictví a přímo dopadá na určení výše daňové povinnosti, je řešena ve studii *The role of cluster analysis in assessing comparability under the U. S. transfer pricing regulations* [Johnson, 2001], a to pomocí několika konvenčních metod včetně hierarchické shlukovací metody. Shluková analýza může urychlit proces zjištění a stanovení podobnosti mezi kontrolovaným subjektem a nekontrolovaným srovnatelným souborem. V této studii je porovnáván v praxi používaný *screening* se shlukovou analýzou. Výsledkem obou postupů je určení, které podniky jsou vzájemně srovnatelné a které nikoliv. Výsledky dosažené oběma postupy jsou odlišné. Postupy v rámci *screeningu* jsou podle Johnsona více zatíženy zaujatostí, předsudky apod. než je tomu v případě shlukové analýzy. Dále autor uvádí, že v kontextu převodních cen lze ve shlukové analýze používat proměnné, např. hrubý příjem, růst tržeb, majetku apod. Shluková analýza může být využita k porovnávání převodních cen a hodnocení jejich srovnatelnosti v praxi, výstupem může být rozhodovací strom nebo screeningový přístup. Pomocí shlukové analýzy mohou být nalezeny různé atributy, které jednotlivé podniky vzájemně odlišují, mohou být porovnávány různé metodiky převodních cen např. s jejich následným doporučením pro použití v konkrétním podniku.

Ze zmiňovaných prací lze čerpat informace o možnostech aplikace shlukové analýzy, o citlivosti analýzy, o potřebě přesnosti určování hypotéz, výběru objektů a znaků. Je vhodné dodat, že shluková analýza je zpravidla doplněna podrobnějším zkoumáním vztahů objektů či znaků ve skupině, prostřednictvím jiné analytické metody nebo opakováním shlukovacího procesu u objektů seskupených v rámci jednoho shluku vytvořeného při prvním shlukovacím procesu, přestože tyto skutečnosti nejsou u některých studií jednoznačně uvedeny.

4. Možnosti aplikace shlukové analýzy v oblasti daní

Shluková analýza může být použita v oblasti daní z různých hledisek. Na podkladě uvedených zdrojů lze shrnout použití shlukové analýzy v oblasti daní v následujících bodech:

- zkoumání podobností a odlišností ucelených daňových systémů mezi vybranými státy;
- zkoumání skupiny daní nebo vybraného typu daně vzájemně mezi sebou nebo mezi vybranými státy;
- zkoumání daní ve vztahu k dalším ekonomických aktivitám v konkrétním státě či v mezinárodním porovnání;
- zkoumání daní ve vztahu k daňovým subjektům, poplatníkům daně či plátcům v konkrétním státě či v mezinárodním porovnání;
- zkoumání daní ve vztahu k veřejným rozpočtům v konkrétním státě či v mezinárodním porovnání.

Objektem analýzy mohou být jak jednotlivé daně, jejich skupiny, tak státy nebo ekonomické činnosti nebo daňové subjekty. U ekonomických a daňových subjektů lze aplikovat členění, která poskytují statistická data, např. podle zastoupení členů domácnosti nebo podle příjmových skupin. Ekonomickým subjektem se přitom rozumí subjekty v ekonomickém prostředí daného státu, tj. občan jednotlivec, domácnost, skupina jednotlivců. Daňovým subjektem se pak rozumí fyzická či právnická osoba, na kterou je daň legislativně uvalena. Pojem daňový subjekt se shoduje s pojetím daňového subjektu v jednotlivých daňových zákonech.

Tabulka 1 | Znaky a objekty ve shlukové analýze při zkoumání daní ve vymezeném období

Lokalizace	Objekty	Předmět zkoumání	Znaky
Mezinárodní porovnání	Státy	Jedna konkrétní daň (např. dopad daně v různých státech)	Znaky vybrané daně: základ daně, sazby daně, efektivní sazba daně, průměrná sazba daně, odpočty od základu daně, slevy na dani.
	Státy	Skupina daní	Znaky vybrané skupiny daní: efektivní sazby daní, průměrné sazby daní, podíly daní na HDP apod.
	Státy	Daňové soustavy, daňový mix	Znaky daňových systémů vybraných států: ekonomické ukazatele, např. daňová kvóta, podíl daní na HDP.
Porovnání v rámci státu	Skupina daní	Daně ve vztahu: • k veřejným rozpočtům, • ekonomickým či daňovým subjektům, • ekonomickým aktivitám apod.	Efektivní sazba daně, daňový výnos, daně ze spotřeby – spotřeba zdaněných produktů, majetek, příjmy, podíl daně na disponibilním příjmu poplatníka, typy domácností, při určité úrovni důchodu.
	Soustava všech daní vybíraných v konkrétním státě	Daně ve vztahu k veřejným rozpočtům, k ekonomickým či daňovým subjektům, ekonomickým aktivitám apod.	Znaky definující vybrané daně nebo skupiny daní (např. spotřeba zdaněných produktů, základ daně, sazba daně apod.).
Aplikace možná v mezinárodním měřítku i v rámci státu	Ekonomické a daňové subjekty	Dopad daní na: • ekonomické či daňové subjekty, • skupiny ekonomických subjektů, • ekonomické aktivity.	Daňové příjmy na osobu, průměrná spotřeba zdaněných statků na osobu, průměrné příjmy na osobu, průměrná daňová povinnost poplatníka, apod. Zkoumání lze omezit na jedinou daň, skupinu daní nebo lze porovnávat daně celkem vybírané na konkrétním území

Zdroj: Vlastní zpracování

Závěr

Použitelnost shlukové analýzy v oblasti zkoumání daní může být poměrně široká – od daňové spravedlnosti po dopad daně na vybrané poplatníky. Důležitými prvky při aplikaci dané metody jsou: stanovení výchozích hypotéz, výběr objektů a znaků, volba metody shlukové analýzy a nastavení parametrů shlukovacího procesu (počet shluků, standardizace dat apod.), přesnost interpretace dosažených výsledků, přítomnost odlehlých objektů, případně absence některých dat a jejich náhrada odhadem. Podrobnější zkoumání vztahů uvnitř shluku může být využito k dalším pokročilejším analýzám.

Literatura

ALLAN, R. G., IGLARSH, H. J., 1996. A cluster analysis of horizontal tax equity. *The Journal of the American Taxation Association*. Issue 1, s. 31–44. ISSN 0198-9073.

ČIMBOROVÁ, M., 2007. *Zmena daňových mixov v OECD krajinách vplyvom meniacich sa podmienok globalizujúceho sa světa* [Disertační práce]. Praha. Vysoká škola ekonomická, Fakulta financí a účetnictví.

GNIP, A. G., TOMIC, I., 2010. How hard does the tax bite hurt? Croatian vs. European worker. *Financial Theory and Practice.* Issue 2, s. 109–142. ISSN 1332-3970.

JOHNSON, R. E., 2001. The role of cluster analysis in assessing comparability under the U. S. transfer pricing regulations. *Business Economics*. Issue 2, s. 30–38. ISSN 0007-666X.

KELBEL, J.; ŠILHÁN, D., 2009. Shluková analýza [online]. Praha: ČVUT [cit. 23. 1. 2015]. Dostupné z: cmp.felk.cvut.cz/cmp/courses/recognition/zapis_prednasky/zapis_02/13/shlukovani.pdf.

LUKASOVÁ, A.; ŠARMANOVÁ, J., 1985. *Metody shlukové analýzy.* Praha: SNTL.

MELOUN, M.; MILITKÝ, J.; HILL, M., 2005. *Počítačová analýza vícerozměrných dat v příkladech*. Praha: Academia. ISBN 80-200-1335-0.

MUSGRAVE, R. A., 1959. *The Theory of Public Finance.* New York: McGraw-Hill. ISBN 978-0070441156.

----------, 1990. Horizontal equity, once more. *National Tax Journal*. Issue 2, s. 113–122. ISSN 0028-0283.

----------, 1984. *Public Finance in Theory and Practice* (4. vyd.). New York: McGrawHill. ISBN 978-0070441262.

SERRANO, A. M., 1994. *Divergences in Personal Income Taxes at a State and Local Level in the European Union: An Applied Multivariate Analysis*. Rome: Regional Science Association.

TRPESKI, P.; TASHEVSKA, B., 2012. Labour Tax Wedge in the Republic of Macedonia – trends and international comparison. *Annales Universitatis Apulensis: Series Oeconomica.* Issue 2, s. 571–585. ISSN 14549409.