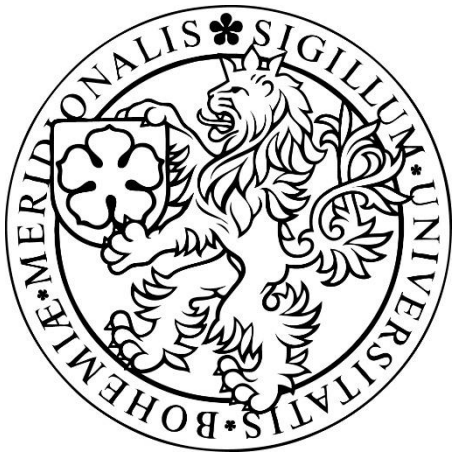
Jihočeská univerzita v Českých Budějovicích
Ekonomická fakulta



Segmentace a klasifikace jako nástroje pro zvýšení efektivity rozhodování v období velkých dat

Habilitační práce

Souhrn uveřejněných vědeckých a odborných prací doplněný komentářem

RNDr. Marta Žambochová

2020

Abstrakt

Habilitační práce se věnuje aplikacím pokročilých statistických metod, a to shlukové analýzy a rozhodovacích stromů v segmentačních a klasifikačních úlohách v oblasti společenských věd. Uvažované metody jsou nejprve stručně popsány, a poté se práce podrobněji soustřeďuje na jejich využití pro řešení reálných problémů v oblastech obchodu, řízení školství, ale i podpory podnikání. Habilitační práce má formu sjednocujícího komentáře k sedmi odborným pracím publikovaným ve vědeckých časopisech či sbornících z vědeckých konferencí zahrnutých do databáze Web of Science. Jednotlivé kapitoly představují přístup k aplikacím zmíněných metod při řešení v rámci konkrétních případových studií. Závěry práce pak poukazují na důležité fáze zpracování v segmentačních a klasifikačních úlohách, na možné problémy vznikající při jejich provádění a navrhují jejich řešení. Studie uvedené v této práci jsou reálnou ukázkou toho, jak je statistická analýza reálných dat užitečná pro praxi a jak mohou být její výsledky přínosné v rozhodovacích procesech.

Abstract

The habilitation thesis deals with the application of advanced statistical methods, namely cluster analysis and decision trees in segmentation and classification tasks in the field of social sciences. The considered methods are first briefly described, furthermore the thesis focuses in detail on their use to solve real problems in the fields of business, education, management as well as business support. The habilitation thesis has the form of a unifying commentary on seven professional papers published in scientific journals or scientific conferences proceedings included in the Web of Science database. The individual chapters represent the approach of the mentioned methods application in solving specific case studies. The conclusions of the work subsequently specify the important stages of processing in segmentation and classification tasks, point to the possible problems arising in their implementation and suggest their solutions. The studies presented in this work are a real example of how statistical analysis of existing data is useful for practice and how its results can be beneficial in decision-making processes.

Obsah

1 Předmluva

Předkládaná habilitační práce je sjednocujícím komentářem sedmi uveřejněných vědeckých a odborných prací, na nichž se autorka této habilitační práce podílela, přičemž ve všech případech byl podíl na přípravě, realizaci a interpretaci výsledků větší nebo stejný jako všech ostatních spoluautorů. Všechny tyto publikace jsou zahrnuty do databáze Web of Science, případně Scopus.

Publikované práce je možno rozdělit do několika okruhů:

1. Rozhodovací stromy a jejich použití pro:
    - klasifikaci objektů,
    - interpretaci výsledků shlukové analýzy,
    - data mining.
2. Shluková analýza a její využití pro:
    - segmentaci objektů a marketing,
    - segmentaci objektů a analýzu financování.
3. Hledání charakteristických rysů získaných tříd mnohorozměrných objektů.

Práce se věnuje problematice segmentace, klasifikace a jejich využití v oblasti společenských věd. Jmenovitě se zabývá shlukovou analýzou jako zástupcem skupiny metod úloh učení bez učitele vhodných pro segmentaci, a rozhodovacími stromy jako zástupcem skupiny metod úloh učení s učitelem pro klasifikaci v užším slova smyslu.

V jednotlivých kapitolách se práce soustřeďuje na rozhodovací stromy, shlukovou analýzu a hledání charakteristických rysů získaných skupin dat. Uvažované metody jsou nejprve stručně představeny, a poté se práce podrobněji soustřeďuje na jejich využití pro řešení reálných problémů jak ekonomické, tak marketingové povahy. Práce se též stručně zabývá výpočetní problematikou, která hraje při praktické realizaci klíčovou roli.

Tématem první části práce jsou rozhodovací stromy a jejich využití pro klasifikaci potenciálních zájemců o podnikání (článek Hlaváček, Žambochová a Siviček, 2015). Následuje popis postupu využívajícího rozhodovací stromy pro interpretaci výsledků shlukové analýzy, jejímž cílem byla identifikace vhodné platformy pro umístění informací pro studenty (článek Žambochová, 2012b).

Závěrem první části je uveden popis a vývoj problematiky využití rozhodovacích stromů v data mining (článek Žambochová, 2008).

Tématem druhé části je rozbor shlukové analýzy jako jednoho z nejčastěji využívaných prostředků pro segmentaci (článek Žambochová, 2014). Následuje popis využití segmentace objektů ve dvou případových studiích. První z nich se zabývala marketingovým průzkumem prostorového chování zákazníků maloobchodních řetězců v Bratislavě (článek Grossmanová, Kita a Žambochová, 2016). Cílem druhé studie byla identifikace skupin potencionálních zahraničních studentů ochotných podílet se na financování svého studia (článek Žambochová, 2012a).

Závěrečná část je věnována shlukové analýze specifického druhu mnohorozměrných dat, tak zvaných funkcionálních dat. Na analýze vývoje počtu odbavených pasažérů je ukázán způsob identifikace specifických rysů těchto dat (článek Žambochová, 2017).

Zahrnuté články:

1. Žambochová, M. (2008). Data mining methods with trees. *E+M Ekonomie a Management,* 11(1), 126-131, ISSN 1212-3609.
2. Žambochová, M. (2012a). Typology of foreign students interested in studying at Czech universities. *E+M Ekonomie a Management,* 15(2), 141-154, ISSN 1212-3609.
3. Žambochová, M. (2012b). Classification in terms of students' preferences for information sources. 9*th International Conference on Efficiency and Responsibility in Education*, Praha, 612-620, ISBN 978-80-213-2289-9.
4. Hlaváček, P., Žambochová, M. and Sivček, T. (2015). The Influence of the Institutions on Entrepreneurship Development: Public Support and Perception of Entrepreneurship Development in the Czech Republic. *Amfiteatru Economic*, 17(38), 408-421, ISSN: 1582 – 9146.
5. Žambochová, M. (2014). The modification of the k-means method for creating non-convex clusters. 8*th International Days of Statistics and Economics*, Praha, 1722-1730, ISBN 978-80-87990-02-5.
6. Grossmanová, M., Kita, P. and Žambochová, M. (2016). Segmentation of consumers in the context of their space behaviour: Case study of Bratislava. *Prague Economic Papers,* 25(2), 189-202, ISSN: 1210-0455, DOI: 10.18267/j.pep.554.
7. Žambochová, M. (2017). Cluster Analysis of World's Airports on the Basis of Number of Passengers Handled (Case Study Examining the Impact of Significant Events). *Statistika – Statistics and Economy Journal,* 97(1), 74-88, ISSN: 0322-788X.

2 Úvod

Rychlé tempo ekonomické globalizace v několika málo posledních dekádách, a tím vznikající ekonomické napětí spolu s rostoucím tlakem konkurence na trhu vedlo podnikové manažery k tomu, aby se zaměřili na výběr správné strategické rozhodovací politiky s cílem zvyšování zisků (You a kol., 2015).

Reakce na potřeby zákazníků správným způsobem ve správný čas mohou pomoci při budování dlouhodobých vztahů mezi společností a jejími zákazníky (Tsai a kol., 2015). Pochopení a rozlišení zákazníků podle jejich potřeb a reakce na marketingový mix pak hraje zásadní roli v řízení podniku. K analýze údajů o zákaznících se používají různé statistické metody nebo metody data miningu (McCarty a Hastak, 2007).

Připomeňme, že data mining je interdisciplinární obor s obecným cílem extrahování struktury, skrytých vlastností, vztahů a anomálií z velkého množství dat. Závěry získané pomocí metod data miningu, pomocí nichž je možno získat užitečné informace o zákaznících a objevit skryté chování zákazníků z velkých dat, má v dnešní době velký vliv na rozhodování managementu (Rud, 2001).

Segmentace trhu je základní marketingový koncept, který je sice poměrně snadno definovatelný a srozumitelný, ale těžko měřitelný a mnohdy výpočetně obtížný. Segmentace zákazníků je nejdůležitější aplikací v oblasti řízení vztahů se zákazníky, tzv. CRM (customer relationship management). Důvodem použití segmentace je nevhodnost nabízení stejného marketingového modelu vzhledem k heterogenním zákazníkům. Cílem je navrhnout marketingové strategie pro různé typy zákazníků (Bassi, 2007). Rozdělením zákazníků do segmentů je možno přizpůsobit marketingový mix cíleným skupinám zákazníků, a tím zlepšit jejich spokojenost a dosáhnout maximální efektivity. Neméně důležitá v rámci CRM je predikce a klasifikace, jejichž cílem je roztřídit zákazníky do předem určených skupin, a tím včas odhalit bonitní, respektive problémové zákazníky.

Segmentace trhu je obsahem mnoha studií, v nichž autoři navrhují různé přístupy k řešení. Klíčovým prvkem v CRM a segmentaci zákazníků jsou celkové informace o nich. V současnosti se dají data o zákaznících více či méně snadno získat prostřednictvím podnikových informačních systémů, podnikových datových skladů a internetu (Holmbom a kol. 2011).

Existuje velké množství statistických metod, které se využívají v oblasti segmentačních a klasifikačních úloh. Jako příklad lze vyjmenovat metody shlukové analýzy (Gilboa, 2009), rozhodovací stromy (You a kol., 2015), evoluční algoritmy (Liu a kol., 2012), či Kohonenovy samoorga-nizující se mapy SOM (Holmbom a kol., 2011).

Předkládaná práce se zabývá aplikací shlukové analýzy a rozhodovacích stromů pro různé úlohy podnikového managementu, a to převážně v oblasti řízení vztahů se zákazníky.

3 Rozhodovací stromy a jejich využití

Odborníci v oblasti marketingu si uvědomili, že segmentace trhu je problém s více kritérii. Jedním kritériem je požadavek, aby zákazníci v segmentu měli podobné profily (identifikovatelnost). Dle druhého kritéria by ale také měli podobně reagovat na marketingový mix (Smith, 1956). Problematika zabývající se řešením druhého kritéria využívá k analýze převážně různé zástupce ze třídy metod učení s učitelem, kde se rozhodovací pravidla pro zařazení objektů do tříd vytváří na základě učící (trénovací) množiny. A právě jedné z těchto metod, rozhodovacím stromům, se bude věnovat tato kapitola (Antoch, 1988; Savický a kol., 2000; Berikov a Litvinenko, 2009).

Velmi rozšířenou skupinou reprezentantů dat, kterých se využívá v datových modelech, jsou různé typy tak zvaných rozhodovacích stromů. Jedná se o struktury, které rekurzivně rozdělují zkoumaná data dle určitých rozhodovacích kritérií. Zatímco kořen stromu zpravidla reprezentuje celý výběrový soubor, vnitřní uzly stromu reprezentují podmnožiny výběrového souboru. V listech stromu můžeme vedle dalších informací vyčíst hodnoty vysvětlované proměnné.

Rozhodovací strom se většinou vytváří rekurzivně dělením prostoru hodnot prediktorů (vysvětlujících, nezávislých proměnných). Jinými slovy, rozhodovací strom se vytváří na základě výběru prediktorů s vhodným počtem kategorií, podle nichž se soubor objektů dělí na podsoubory s danými vlastnostmi. Tomuto postupu se říká „shora dolů“. Existuje však i postupy, které jej vytvářejí pomocí aglomerace „zdola nahoru“.

Máme-li strom s jedním listem, hledáme prediktor s vhodnými kategoriemi (podmínku větvení), který co nejlépe rozděluje prostor zkoumaných objektů do

podmnožin. Při hledání optimálního prediktoru se maximalizuje zvolené kritérium kvality dělení (tzv. splitting criterium).

Takto nám vznikne strom s více listy. V dalších krocích pro každý nový list hledáme prediktor s vhodnými kategoriemi, který množinu odpovídajících objektů náležících tomuto listu dále co nejlépe dělí do podmnožin. Jedná se o rekurzivní postup dělení dat na stále homogennější podmnožiny.

Proces dělení se zastaví, pokud je splněno zvolené kritérium pro zastavení (tzv. stopping rule). Omezení obsažená v kritériu pro zastavení mohou být např. „hloubka" stromu, počet listů stromu, stupeň homogenity množin dat v listech, atd.

Dalším krokem algoritmů je „prořezávání" takto vzniklého stromu (prunning). Během této fáze je nutno určit „správnou" velikost stromu, neboť příliš malé stromy dostatečně nevystihují všechny zákonitosti v datech, zatímco příliš veliké stromy zahrnují do popisu i nahodilé vlastnosti dat a mají malou zobecnitelnost. Za tím účelem se uvažují podstromy rozhodovacího stromu, který získáme použitým algoritmem dělení, a porovnává se kvalita jejich generalizace a vysvětlení dat, penalizovaná mírou složitosti výsledného řešení.

Prakticky vždy se postupuje tak, že se rozhodovací stromy nejprve vytváří na tzv. trénovacích datech, a poté se jejich kvalita ověří na tzv. testovacích datech.

Jiným způsobem je tzv. křížové ověřování (cross validation), kdy se k vytváření rozhodovacího stromu a jeho podstromů použijí téměř všechna data. Přitom dojde k rozdělení dat na několik disjunktních, přibližně stejně velkých částí, a postupně se vždy jedna část ze souboru vyjme. Zatímco nevyjmutá data slouží k vytvoření rozhodovacího stromu, vyjmutá část posléze poslouží k ověření „kvality jeho rozhodování". Výsledkem je *k* rozhodovacích stromů, jež je následně třeba vhodně zkombinovat. Například, ale ne nutně, tak, že se vybere takový podstrom, který má nejnižší odhad skutečné chyby. Pokud existuje více podstromů se srovnatelným odhadem skutečné chyby, vybírá se dle pravidla tzv. Occamovy břitvy ten nejmenší.

Jednotlivé postupy a k nim příslušné algoritmy vytváření rozhodovacích stromů se liší následnými charakteristikami:

- pravidlem dělení (splitting rule)
- kritérem pro zastavení (stopping rule)

- typem podmínek větvení
    - multivariantní (použije se několik prediktorů)
    - univariantní (v daném kroku se pro dělení vždy použije pouze jeden z prediktorů)
- způsob větvení
    - binární (každý z uzlů, kromě listů, se dělí na dva následníky)
    - *k*-ární (každý z uzlů se může rozdělit na více než dvě části, a přitom ne vždy na stejný počet)
- typ výsledného stromu, popis obsahu listů
    - klasifikační stromy (každému listu je přiřazena třída)
    - regresní stromy (každému listu je přiřazena konstanta – odhad hodnoty závisle proměnné)
- typ prediktorů

  lze použít libovolnou datovou strukturu, přičemž nejčastější jsou proměnné spojitého či diskrétního typu.

3.1 Algoritmy pro vytváření rozhodovacích stromů

Pro vytváření rozhodovacích stromů bylo navrženo velké množství algoritmů. Nejčastěji používané jsou CART, ID3, C4.5, AID, CHAID a QUEST, a jejich varianty. V prostředí statistického systému SPSS jsou implementovány algoritmy CART (pod označením CRT), CHAID a QUEST.

Algoritmus CART

Algoritmus poprvé popsali jeho autoři Breiman a kol. (1984). Základní algoritmus je použitelný v případě, že máme k dispozici jednu nebo více vysvětlujících proměnných. Tyto proměnné mohou být buď spojité nebo kategoriální (ordinální i nominální). Dále je uvažována jedna závisle proměnná, která také může být buď kategoriální (nominální i ordinální) nebo spojitá (Timofeev, 2004).

Výsledkem základního algoritmu jsou binární stromy, protože daná metodika používá pouze prediktory využívající pouze otázky s dichotomickou odpovědí, na které je možno odpovědět ano/ne (Je věk menší než 30 let? Je pohlaví mužské? …), nebo

prediktory vzniklé překódováním původně vícekategoriálních proměnných na dichotomické.

Algoritmus v každém kroku prochází všechna možná dělení pomocí všech přípustných hodnot všech vysvětlujících proměnných, a hledá „nejlepší“. Kritériem umožňujícím rozhodnout, které dělení je (nej)lepší, je zvýšení „čistoty dat“, přičemž „čistota dat“ je zpravidla popsána jejich homogenitou. To znamená, že jedno dělení je lepší než druhé, pokud jeho použitím obdržíme dva homogennější (z pohledu vysvětlované proměnné) soubory dat než pomocí jiného dělení. Algoritmus dělení se pro klasifikační stromy a pro regresní stromy poněkud liší.

Klasifikační stromy

Klasifikační stromy používáme v případě, že je vysvětlovaná proměnná kategoriální. To znamená, že se soubor původních dat snažíme v závislosti na vysvětlujících proměnných rozdělit do skupin (podmnožin), přičemž, v ideálním případě, každá skupina obsahuje prvky, jenž patří k téže kategorii.

Homogenita uzlů-potomků je měřena pomocí tzv. funkce znečištění (impurity function) $i(t)$. Funkce znečištění pro daný uzel by měla nabývat hodnoty 0, pokud pro všechny objekty obsažené v tomto uzlu nabývá vysvětlovaná proměnná stejné hodnoty, tedy všechny objekty náleží do jedné kategorie. Naopak maximální hodnoty nabývá funkce znečištění v případě, že jsou objekty obsažené v daném uzlu rozloženy do všech kategorií rovnoměrně. Cílem štěpení uzlů je dosáhnout maximální změny (snížení) znečištění $\Delta i(t)$, přičemž $\Delta i(t)$ můžeme definovat vztahem

$$\Delta i(t) = i(t) - P_l \cdot i(t_l) - P_p \cdot i(t_p) \tag{3.1}$$

kde t je štěpený uzel, t_l je jeho levý potomek, t_p je pravý potomek P_l, respektive P_p je podíl prvků, jenž padnou do levého, respektive pravého, uzlu-potomku. To znamená, že

$$P_l = \frac{N_l}{N}, \text{ respektive } P_p = \frac{N_p}{N}, \tag{3.2}$$

kde N značí počet objektů v uzlu t, N_l je počet objektů v uzlu t_l, N_p je počet objektů v uzlu t_p. Je nutno poznamenat, že P_l i P_p mohou být a často jsou interpretovány jako pravděpodobnosti.

Algoritmus CART se snaží pro každý uzel najít takové dělení, které maximalizuje pokles nečistoty přes všechna možná dělení uzlu, to znamená, že hledá to dělení, které

přináší maximální zlepšení homogenity dat v podmnožinách vzniklých dělením rodičovského uzlu. Funkci *Δi(t)* je možno definovat různými způsoby. Mezi dva nejpoužívanější patří tzv. Giniho index a Twoing pravidlo.

Giniho index

Tak zvaný Giniho indexu je v praxi asi nejpoužívanější funkce znečištění. Giniho index je míra variability pro nominální proměnnou. Jde o analogii rozkladu rozptylu. Celková variabilita je vyjádřena jako vnitroskupinová variabilita a meziskupinová variabilita. Například v SPSS se meziskupinová variabilita označuje jako „improvement“. Funkce *i(t)* je definována následovně:

$$i(t) = \sum_{k \neq l} P(k \mid t) \cdot P(l \mid t) \tag{3.3}$$

kde *t* je uzel, *k, l* jsou indexy kategorie vysvětlované proměnné, *k, l = 1, ..., K*; *P(k|t)* a *P(l|t)* jsou určité váhy, které mohou být interpretovány jako podmíněné pravděpodobnosti kategorie *k* (resp. *l*) v uzlu *t*, tj. podíl počtu objektů uzlu *t* spadajícího do *k*-té (resp. *l*-té) kategorie vysvětlované proměnné a celkového počtu objektů v uzlu *t*.

Dosazením této funkce do (3.1) dostáváme:

$$\Delta i(t) = -\sum_{k=1}^{K} P^2(k|t) + \sum_{k=1}^{K} P^2(k|t_l) + \sum_{k=1}^{K} P^2\left(k\middle|t_p\right) \tag{3.4}$$

Giniho index hledá v trénovacích datech největší homogenní kategorii vysvětlované proměnné, a odděluje ji od ostatních dat. Praktické zkušenosti ukazují, že Giniho index dobře funguje pro „znečištěná“ data.

Twoing pravidlo

Na rozdíl od Giniho indexu Twoing pravidlo hledá takové dvě třídy, které dohromady obsáhnou více než 50 % dat. Twoing pravidlo maximalizuje následující změnu funkce znečištění.

$$\Delta i(t) = \frac{P_l \cdot P_p}{4} \left[\sum_{k=1}^{K} \left| P(k \mid t_l) - P(k \mid t_p) \right| \right]^2 \tag{3.5}$$

kde *t* je rodičovský uzel, *k, l* jsou indexy třídy závislé proměnné, *k, l = 1, ..., K*; P_p a P_l, mají stejný význam, jako je uveden u vztahu (3.3).

Vytváření stromů s pomocí Twoing pravidla je pomalejší než za použití Giniho indexu. Výhodou ovšem je, že vytváříme lépe vybalancované stromy.

Regresní stromy

Regresní stromy se používají v případě, že vysvětlovaná proměnná není kategoriální. Každá její hodnota může být (obecně) různá. V tomto případě algoritmus hledá nejlepší dělení zpravidla na základě minimalizace součtu rozptylů v rámci jednotlivých vzniklých uzlech-potomcích. Algoritmus pracuje na základě minimalizace součtu čtverců reziduí.

CHAID

Metodu CHAID (Chi-squared Automatic Interaction Detektor) navrhl a rozpracoval v roce 1980 G. V. Kass. Tato metoda je určena pro kategoriální vysvětlovanou proměnnou. Výsledkem jsou nebinární stromy. Metoda využívá k testování chí-kvadrát test. Používá se test buď na základě Pearsonovy statistiky, nebo na základě věrohodnostního poměru. Z důvodu vysoké časové náročnosti původního algoritmu autor hledá pouze suboptimální dělení místo prohledávání všech možných dělení (Wilkinson, 1992).

Algoritmus dělení probíhá následovně. V rámci jednoho listového uzlu se vytvoří kontingenční tabulka rozměrů $m \times k$ pro prediktor s m kategoriemi a vysvětlovanou proměnnou mající k kategorií. Dále se najde ta dvojice kategorií prediktoru, pro které má podtabulka rozměrů $2 \times k$ nejméně významnou hodnotu chí-kvadrát testu. Tyto dvě kategorie se sloučí a získáme novou kontingenční tabulku o rozměrech $(m–1)$ x k. Proces slučování opakujeme až do doby, kdy klesne významnost chí-kvadrát testu pod předem zadanou hodnotu, zpravidla 5 %. Tímto způsobem je ukončen proces dělení jednoho rodičovského uzlu na několik uzlů-potomků. Dále se pokračuje obdobně pro každý listový uzel až do doby, kdy výsledek chí-kvadrát testu je statisticky nevýznamný na zvolené hladině významnosti. Tato metoda byla dále rozpracována řadou dalších autorů a existuje více jejich implementací v různých SW systémech.

QUEST

Algoritmus QUEST, který je použitelný pouze pro nominální vysvětlovanou proměnnou, popsali poprvé Loh a Shih (1997). Obdobně jako v případě CART jsou vytvářeny pouze

binární stromy. Na rozdíl od metody CART, která výběr proměnné pro štěpení uzlu a výběr dělícího bodu provádí v průběhu budování stromu současně, provádí metoda QUEST toto odděleně. Metoda QUEST (Quick, Unbiased, Efficient, Statistical Tree) odstraňuje některé nevýhody algoritmů používajících vyčerpávající hledání (např. CART), jako je například výpočetní náročnost zpracování.

Tato metoda je vylepšením algoritmu FACT, který popsali Loh a Vanichsetakul (1988). V prvním kroku algoritmus převede všechny kategoriální vysvětlující proměnné na „ordinální“ pomocí transformace CRIMCOORD.

V každém listovém uzlu je pro každou spojitou vysvětlující proměnnou prováděn ANOVA F-test. Pokud největší ze vzniklých F-statistik je větší než předem daná hodnota F_0, pak příslušná proměnná je vybrána pro dělení uzlu. Pokud tomu tak není, je pro všechny proměnné proveden Levenův F-test. Pokud je největší Levenova F-statistika větší než F_0, pak je příslušná proměnná vybrána pro dělení uzlu. Pokud tomu tak není (žádná ANOVA F-statistika ani Levenova F-statistika není větší než hodnota F_0), je pro dělení vybrána proměnná s největší hodnotou ANOVA F-statistiky. Pro každý kategoriální prediktor se spočte Pearsonův chí-kvadrát test nezávislosti. Pro dělení uzlu je pak vybrána ta vysvětlující proměnná, která je se vysvětlovanou proměnnou nejvíce asociována.

Pro hledání dělícího bodu pro vybranou vysvětlovanou proměnnou je využívána kvadratická diskriminační analýza (QDA). To je podstatný rozdíl od algoritmu FACT, kde je využívána lineární diskriminační analýza (LDA).

Výše popsaný postup je rekurzivně opakován až do jeho zastavení vycházejícího ze zvoleného kritéria pro zastavení, jímž může být například splnění některé z následující podmínek:

1. Pokud se uzel stane čistým, to znamená, že všechny objekty v uzlu náleží do stejné třídy vysvětlované proměnné.
2. Pokud mají všechny objekty v uzlu stejné hodnoty pro každý prediktor.
3. Pokud aktuální hloubka stromu dosáhne uživatelem stanovené maximální hodnoty hloubky stromu.
4. Pokud je počet objektů v uzlu menší než uživatelem zadaná minimální hodnota velikosti uzlu.

5. Pokud rozdělení uzlu vede k podřízenému uzlu, v němž počet obsažených objektů je menší než zadaná minimální velikost velikosti podřízeného uzlu.

Rozpracování této metody se i dále věnovali další autoři a v různých variantách byla implementována to statistických SW systémů.

3.2 Využití rozhodovacích stromů pro klasifikaci

Rozhodovací stromy jsou vhodným a oblíbeným nástrojem pro klasifikaci objektů, tj. pro predikci hodnoty vysvětlované proměnné, protože jejich grafické znázornění je jednoduše čitelné, srozumitelné a interpretovatelné. Atraktivnost těchto metod je z velké míry ovlivněna faktem, že rozhodovací stromy představují pravidla, a tato pravidla se dají vyjádřit v běžném jazyce. Toto jsou nejdůležitější faktory, proč se algoritmy pro rozhodovací stromy úspěšně využívají právě pro klasifikaci objektů na základě dané vysvětlované proměnné. Rozhodovací stromy vedle toho často umožňují „pohodlně“ nalézt nové skryté zákonitosti, či odhalit a popsat strukturu zkoumaného souboru objektů. Jak bylo uvedeno výše, rozhodovací stromy rekurzivně dělí množinu objektů výběrového souboru dle algoritmem vybraných dělících kritérií, a to tak, že na poslední úrovni dělení, tedy v listech tvořeného stromu, jsou podmnožiny objektů co nejvíce homogenní z pohledu k vysvětlované proměnné (You, 2015).

Pokud má výsledný rozhodovací strom dostatečnou kvalitu, pak můžeme většinu jeho listů přiřadit jedné hodnotě vysvětlované proměnné, která je pro objekty obsažené v listu převažující. V ideálním případě, který však nastane zcela výjimečně, se může stát, že pro všechny objekty tohoto listu je hodnota vysvětlované proměnné stejná.

Pokud následně budeme procházet strom po cestě od kořene směrem k listu, pak můžeme u každé větve této cesty vyčíst rozhodovací pravidlo. Jednotlivá pravidla na dané cestě tvoří soubor vlastností objektů daného listu vzhledem k vysvětlujícím proměnným. Na základě toho pak můžeme určit, jaké vlastnosti určené vysvětlujícími proměnnými mají objekty s danou hodnotou vysvětlované proměnné.

3.2.1 Zkoumání vnímání podpor a bariér pro začínající podnikatele – případová studie

Velmi diskutovaným tématem v oblasti ekonomiky jsou dopady podpor a bariér pro podnikatele. Zvláštní kategorii pak tvoří dotace pro začínající podnikatele. Žambochová a Tišlerová (2011) se ve své studii zabývají pohledem na vstup do podnikání. Autorky srovnávají dvě formy podnikání, a to klasickou formu vlastního podnikání a tzv. franšízing. Speciálně se soustřeďují na otázku, zda záštita know-how získaného v rámci podnikání formou franšízingu usnadní odpověď na otázku, zda začít podnikat. Výzkum o ochotě začít podnikat jak formou franšízingu, tak klasickým způsobem podnikání, byl proveden formou dotazníků, a cílem výzkumu bylo nalézt ty skupiny obyvatel, u kterých je největší pravděpodobnost začít vlastní podnikání. Po definování znaků této skupiny mohou být přijata vhodná opatření pro investiční pobídky. K odhalení typů osob ochotných podnikat jednou z nabízených forem použily autorky metodu klasifikačních stromů. Mezi nejdůležitějšími rysy pro ochotu podnikat se ukázal jednak mladý věk, ale také nezaměstnanost.

Na zkoumání faktu zda, a za jakých podmínek, jsou nezaměstnaní schopni a ochotni řešit svou situaci vstupem do podnikání, se zabývá Žambochová (2013a). Studie se soustřeďuje na situaci v Ústeckém kraji jako jednom z regionů nejvíce postižených nezaměstnaností. Jak vyplynulo z předchozí studie, východiskem z nezaměstnanosti totiž nemusí být jen nalezení nového zaměstnání, ale i podnikání. Lidé ovšem často mají ze vstupu do podnikání strach, obávají s tím spojeného rizika, nemají finance, odrazuje je přílišná administrativa, nemají zkušenosti a hlavně, nemají informace a často ani nevědí, kde je získat. Cílem průzkumu bylo především zjistit míru informovanosti, všeobecný zájem o jednotlivé podpory a míru ochoty řešit nezaměstnanost vstupem do podnikatelské sféry. Nejenom začínající, ale i stávající podnikatelé mohou využívat různou formu podpory, mohou žádat o finanční podporu ve formě dotací ze strukturálních fondů EU, nebo využít jiná zvýhodnění jako třeba daňové úlevy či zvýhodněné úvěry. Další podpora je možná prostřednictvím poradenství a dalších služeb. Podpora podnikání neprobíhá jen na unijní úrovni, ale také v rámci republiky a regionů. Cílem šetření bylo zjistit, jak jsou na tom čeští občané, a zvlášť malí podnikatelé, se znalostí podpor a jejich využíváním, jak jsou spokojení s podnikáním, a zda si myslí, že stát a Evropa malé a střední podniky podporuje dostatečně. Jedním z hlavních cílů bylo identifikovat skupiny respondentů (dle věku, pohlaví, dosaženého vzdělání, pracovní situace, atd.), které jsou ochotny začít

podnikat, jaká odvětví považují dané skupiny za nejvhodnější pro podnikání, jaké překážky v podnikání považují za nejzávažnější, zda dotázaní jedinci znají a využívají některé podpory podnikání, a o jakou formu podpory by měli největší zájem. K této identifikaci opět bylo využito klasifikačních stromů. Studie ukázala, že nezaměstnaní v podnikání příliš nevidí řešení své nezaměstnanosti. Podobně podnikatelé příliš nejeví zájem o podpory ze strany státu a EU, a to převážně z důvodu, že o těchto podporách nemají přehled. Přesto jsou tito podnikatelé se svou situací víceméně spokojeni.

Na tuto studii navazovala práce popsaná v Hlaváček, Žambochová a Siviček (2015) uvedená v příloze 4. Šetření se zúčastnilo 836 respondentů z různých obcí napříč Českou republikou. Cílem tohoto průzkumu bylo především získat odpovědi na následující otázky:

- Jaké skupiny respondentů (dle věku, pohlaví, dosaženého vzdělání, pracovní situace, atd.) jsou ochotny začít podnikat?
- Které odvětví považují za nejvhodnější pro podnikání?
- Jaké překážky v podnikání považují za nejzávažnější?
- Zda dotázaní jedinci znají a využívají alespoň některé podpory podnikání? Případně které?
- O jakou formu podpory by měli největší zájem?
- Jaké jsou hlavní motivy pro rozhodnutí začít podnikat?
- Považují respondenti úroveň podpory podnikání za dostatečnou?

Konečná fáze hledání odpovědí se zaměřila na klasifikaci stanovisek respondentů s ohledem na jejich ochotu začít s podnikáním v podmínkách současného institucionálního prostředí a jeho přístupu k podpoře podnikání. K analýze získaných dat byla použita metodika klasifikačních stromů. Přesněji, ve statistickém systému SPSS byly vytvořeny pomocí metod CRT, CHAID a QUEST klasifikační stromy. Za vysvětlovanou proměnnou byla vybrána ochota k vstupu do podnikání, a za vysvětlující proměnné byly vybrány odpovědi na jednotlivé otázky. Kvalita modelů byla až překvapivě dobrá. Pomocí hodnoty odhadu rizika byl následně vybrán strom s nejlepší kvalitou, a na jeho základě byla vytvořena typologie osob (ne)ochotných podnikat, a to včetně upřesnění znalostí o dotacích. Ze získané stromové struktury bylo možno učinit následující závěry.

Nejvíce ochotni podnikat jsou:

- Lidé středního věku s nižším vzděláním, kteří pocházejí z menších obcí. Rádi by podnikali v oboru služeb a obchodu, s podporami nemají zkušenosti, nejraději by proto využili různých dotací. V jejich okolí je někdo, kdo podniká.
- Mladší muži s vyšším vzděláním. O možnostech podpor vědí, rádi by využili nejenom finanční podpory, ale i různé formy poradenství.

Naopak, nejméně jsou ochotni podnikat:

- Starší ženy, které pocházejí z obcí střední velikosti. O podporách neví nic a nevěří jim.
- Ženy se základním vzděláním, které nemají rády možné riziko podnikání. O podporách nic neví.
- Starší lidé se středoškolským vzděláním, kteří jsou zaměstnáni a v zaměstnání upřednostňují jistý výdělek. O podporách vědí maximálně z médií.

Výzkum ukázal, že z pohledu zaměstnavatelů je rozvoj podnikání nejvíce omezen institucionálními překážkami, legislativním prostředím a přístupem veřejné správy. Naopak, při zahájení podnikání se respondenti mnohem více obávají selhání a nedostatku finančních prostředků, což nepřímo poukazuje na nedostatečně rozvinutou síť podpory pro objekty vstupující do podnikání, omezené zdroje rizikového kapitálu, a obecně nízkou úroveň komunikace mezi podnikateli a institucemi v oblasti podpory a rozvoje podnikání. Efektivita a funkčnost právního prostředí, přístup instituce veřejné správy a konfigurace rozvojových programů pro začínající malé a střední podnikatele jsou institucionálními faktory, které mají silný dopad na rozvoj podnikání. Institucionální překážky, jako je nedostatek informací a zpětné vazby o potřebách podnikatelů, jsou často považovány za důležitější než sociální a ekonomické překážky.

Efektivností dotací pro podnikatele se zabývá např. práce Dvouletý a kol. (2018). Autoři ve své studii oslovili formou dotazníkového šetření mezi příjemci veřejné podpory pro začínající podnikatele v rámci programu START. Dvouletý a Orel (2020) pak využili vybrané informace, které publikovali Hlaváček, Žambochová a Siviček (2015), ke srovnání v rámci zemí Visegrádské čtyřky.

3.3 Využití rozhodovacích stromů pří interpretaci výsledků shlukové analýzy

Je nutno poznamenat, že rozhodovací stromy lze využít i v jiném kontextu, než pouze ke klasifikaci. Tak například Žambochová (2012b), viz příloha 3, použila rozhodovací stromy v návaznosti na shlukovou analýzu. Jak bude uvedeno ve 4. kapitole, důležitou součástí shlukové analýzy je podrobná interpretace jednotlivých shluků. To však někdy nebývá zcela jednoduché. A právě přehledná struktura rozhodovacích stromů spolu s popisem pravidel náležejících k jednotlivým dělením nabízí silný interpretační nástroj. K tomuto účelu za vysvětlovanou proměnnou, jež vstupuje do algoritmu na tvorbu rozhodovacího stromu, volíme proměnnou vzniklou na základě shlukové analýzy, jejíž hodnoty identifikují příslušnost jednotlivých objektů k daným shlukům. Metody pro shlukování implementované ve statistickém systému SPSS nabízejí takovouto proměnnou automaticky vytvořit v rámci výstupu ze shlukovací procedury. Pokud však tuto možnost metoda automaticky nenabízí, je nutné ji vytvořit. Za vysvětlující proměnné, tedy proměnné, pomocí nichž se provádí dělení, se pak zvolí všechny veličiny souboru, které slouží k identifikaci objektů, přičemž v rámci dotazníkových šetření se většinou jedná o socio-demografické otázky.

3.3.1 Průzkum názorů osob na preference jednotlivých zdrojů informací – případová studie

Vzdělání, především vysokoškolské, je velmi důležitým faktorem ve snaze o vyřešení mnohých sociálních a ekonomických problémů každé země. Vysoké školství silně přispívá k rozvoji celé společnosti i ekonomiky. Zlepšování kvalit vysokého školství, a to jak po kvantitativní, tak i po kvalitativní stránce je, nebo by alespoň mělo být, prioritou politiky každé země.

Od roku 1989 prošlo vysoké školství v ČR zásadními změnami. Jedním z důležitých důvodů je integrace do Evropské unie. Terciální školství již přestává být doménou gymnazistů, ale účastní se ho také absolventi odborně zaměřených středních škol a dokonce i učilišť s maturitou. Druhým důležitým aspektem je fakt, že získání vysokoškolského diplomu nebo jiného obdobného certifikátu již není závěrečnou fází vysokoškolského studia, ale nutností se stává i následné celoživotní vzdělání. Velmi expanzivní nárůst počtu studentů a struktura studujících ovlivňuje mnohé aspekty

terciálního vzdělávání, od financování, přes formy studia až po standardy kvality (Prudký a kol., 2010).

Změna struktury studentů vedla k provedení studie týkající se preferencí zdrojů informací. Cílovou skupinou dotazníkového šetření byli potencionální studenti Univerzity J. E. Purkyně, a to v jakékoliv formě i stupni vzdělávání (prezenční, kombinované, celoživotní; bakalářské, navazující, doktorské). Průzkumu se zúčastnilo 1173 respondentů. Výsledky této studie shrnula Žambochová (2012b), viz příloha 3.

V dotazníku bylo respondentům mimo jiné nabídnuto osm různých typů zdrojů informací, a pro každý z nich měli dotázaní uvést míru oblíbenosti na škále 0–10, kde 0 znamenala, že respondent daný typ nevyužívá nikdy, a 10 znamenala nejvyšší míru obliby. Navíc mohli respondenti využít možnosti vypsat jakýkoliv jiný zdroj informací, který je pro ně důležitý.

S využitím Friedmanova testu a následné post hoc analýze bylo zjištěno, že míra oblíbenosti se u jednotlivých nabízených zdrojů informací významně liší. Největší oblibu v době studie měl veřejný internet, následován učebnicemi a přímou výukou. Naopak jako nejméně užitečná byla uváděna mimoškolní výuka.

V segmentačním procesu byl aplikován postup, kdy nejprve byla provedena shluková analýza, která respondenty rozdělila do skupin osob vzájemně si podobných z hlediska preference jednotlivých informačních zdrojů. Všechny proměnné vstupující do shlukování byly škálového typu s rozmezím 1–10, proto bylo možno použít jak TwoStep metodu, tak i metodu *k*-means. Výborné kvality shlukování bylo dosaženo při aplikaci TwoStep metody, která vytvořila dva shluky. Dobré kvality shlukování se ale také dosáhlo při aplikaci metody *k*- means, a to jak při vytváření dvou, tak při vytváření tří shluků. Na základě shlukové analýzy byla vytvořena nová proměnná reprezentující příslušnost k dané skupině vzniklé v předchozím kroku.

V druhé fázi segmentačního procesu je vždy nutné výsledné shluky vzniklé v první fázi řádně interpretovat. V případě velkých datových souborů je tato fáze velmi obtížná. Jak bylo uvedeno výše, dobrým a efektivním nástrojem mohou být rozhodovací stromy, a to díky jejich názorné a dobře čitelné struktuře. Při praktickém použití je užitečné tento nástroj aplikovat dvakrát. Jednou pro popis a interpretaci segmentů ze sociodemografického pohledu, podruhé z důvodu získání popisu charakteristického chování objektů z jednotlivých segmentů. K tomu účelu byly sestrojeny dvě skupiny

rozhodovacích stromů, přičemž v každé skupině byly vytvořeny stromy pomocí tří metod implementovaných v statistickém systému SPSS, a to CRT, QUEST a CHAID. Vysvětlovanou proměnnou vždy byla nově vzniklá proměnná vyjadřující příslušnost ke shluku. Vysvětlujícími proměnnými byly v případě první skupiny stromů proměnné obsahující identifikační údaje jako např. věk, pohlaví, vzdělání, velikost bydliště, studijní zaměření a podobně, zatímco v případě druhé skupiny stromů pak proměnné vystihující oblibu jednotlivých zdrojů informací. Ve všech případech byl vybrán strom s nejlepší kvalitou, tj. nejmenší hodnotou odhadu rizika, který vyjadřuje míru špatně klasifikovaných případů. Odhad rizika se u všech vytvořených stromů pohyboval v rozmezí od 0,067 do 0,315. To znamená, že míra úspěšnosti zařazení objektů se pohybovala v rozmezí od 93,3% do 68,5%. Kvalita modelů tedy byla velmi vysoká.

V závěrečné fázi zpracování byl ze struktury těchto dvou vybraných stromů identifikován popis skupin vytvořených shlukovou analýzou. Tím bylo zjištěno, jaký typ respondentů upřednostňuje jaký typ informačních zdrojů.

Z procesu segmentace vyplynulo, že:

- osoby středního věku mající nižší vzdělání technického či uměleckého zaměření vůbec nevyužívají Intranet[1] ani Wikipedii, využívají učebnice;
- mladí lidé s nižším vzděláním přírodovědného, humanitního či jazykovědného zaměření využívají Intranet, velmi využívají učebnice a znají Wikipedii;
- osoby mladšího a středního věku mající nižší vzdělání a byly podprůměrné v humanitních předmětech, nemají rády přímou výuku;
- osoby mladšího mladší a středního věku mající vyšší vzdělání upřednostňují přímou výuku a učebnice a nevyužívají Intranet;
- mladší ženy uměleckého a humanitního zaměření upřednostňují přímou výuku, ale internet využívají jen průměrně;
- osoby středního věku mající vyšší vzdělání upřednostňují přímou výuku, hodně využívají internet a znají Wikipedii;
- muži mladší a středního věku mající nižší vzdělání technického či uměleckého neupřednostňují přímou výuku.

[1] Pod pojmem Intranet se rozumí privátní neboli „soukromý" internet, který používá stejné technologie jako veřejný internet, avšak využívat jej může pouze omezený okruh uživatelů (v rámci firmy, školy), informace z Intranetu nejsou veřejně dostupné.

Průzkum ukázal, že v době, kdy byl průzkum prováděn, byl nejpopulárnějším informačním zdrojem veřejný internet, a jeho popularita stále rostla, což potvrdil i vývoj v následujících letech. Respondenti napříč spektrem však stále preferují učebnice, a to jak v papírové, tak i elektronické formě.

Mladší muži a muži středního věku s nižším technickým nebo uměleckým vzděláním se vyhýbají přímému vyučování. To je v protikladu k mladým ženám se zaměřením na humanitní a umění, které přímou výuku upřednostňují.

3.4 Využití rozhodovacích stromů v oblasti data mining

Množství dat uložených v různých databázích exponenciálně roste. Pochopení a uchopení takového množství dat je přitom čím dál tím těžší a náročnější. Získání přehledu nad těmito daty je jedním z hlavních cílů data miningu. Data mining je proces analýzy dat z různých úhlů pohledu a jejich shrnutí do užitečných informací (Žambochová, 2008), viz příloha 1. Tento proces probíhá v rámci tak zvané business intelligence analysis (BIA), která zahrnuje, mimo jiné, „chytrou“ analýzu nejenom firemních dat. Úlohou data miningu je získávání skrytých informací z velkých souborů dat. Často se zde zkoumají data obsažená v podnikových databázích (Olszak, 2016).

Data mining slouží k dvěma základním cílům, a to k jejich porozumění a predikci (Kohavi, 2001). To znamená, že úkolem data miningu je jednak odhalit a pojmenovat pravidla a vzory, na jejichž základě je možno identifikovat výhodné, nebo naopak problémové objekty zájmu, například neplatící klienty či zákazníky, u nichž je velká pravděpodobnost, že produkt sice koupí, ale nezaplatí. Na základě těchto vzorů a pravidel je posléze vybudován model, pomocí něhož je možno predikovat na základě vstupních údajů, zda se objekt zařadí do výhodné, či naopak problémové kategorie. Data mining se tak stává účinným nástrojem při efektivním rozhodování (Rud, 2001; Berry a Linoff, 2004).

Dle vytyčeného cíle pak můžeme úlohy z oblasti data mining rozdělit do několika kategorií:

- klasifikace,
- odhady hodnot vysvětlované proměnné,
- segmentace,
- analýza vztahů,

- predikce v časových řadách,
- detekce odchylek.

Nástroje pro data mining využívají celé řady relativně různorodých statistických i nestatistických metod, jako jsou například:

- rozhodovací stromy,
- shluková analýza,
- neuronové sítě,
- genetické algoritmy,
- induktivní pravidla,
- Bayesovská klasifikace.

Jak píše Maheswari a kol. (2014), v dnešní době se produktově zaměřený pohled na podnikání mění na pohled orientovaný především na zákazníka. To vytváří silnější tlak na dobré řízení vztahů se zákazníky (Customer relationship management, neboli CRM). Mezi hlavní nákladové cíle CRM patří zvyšování tržeb prostřednictvím spokojenosti zákazníků a snižování nákladů. Aby se těchto cílů mohlo dosáhnout, je potřeba co nejlepší identifikace zákazníků a jejich diferenciace. Průběžně je nutno sledovat změny chování a potřeb zákazníků, a těmto změnám se průběžně přizpůsobovat. Kvalitní CRM pomáhá firmě nalézt a udržet „správné“ zákazníky z celkového velkého množství potenciálních zákazníků. Data mining napomáhá organizaci identifikovat vhodné potencionální zákazníky, ale i stávající zákazníky, u nichž hrozí jejich ztráta. Žambochová (2008) uvádí, že rozhodovací stromy jako nástroj pro klasifikaci jsou jedním z nevyužívanějších nástrojů data miningu v oblasti CRM, a na názorném příkladu ukazuje funkci takovýchto stromů.

4 Segmentace objektů a její využití v ekonomii a managementu

Klíčovou roli v publikovaných pracích, které jsou podkladem pro tuto habilitační práci, ale i v disertační práci autorky (Žambochová, 2010b), hrají metody shlukování, především BIRCH a *k*-means. Jejich myšlenka je stručně popsána v následujících podkapitolách.

4.1 Shluková analýza

Shluková analýza se zabývá podobností datových objektů. Řeší dělení množiny objektů do několika předem nespecifikovaných skupin (shluků) tak, aby si objekty uvnitř jednotlivých shluků byly co nejvíce podobny, zatímco objekty z různých shluků si byly podobny co nejméně (Everit a kol., 2001; Řezanková a kol., 2007).

Shlukovou analýzu lze provádět mnoha různými metodami. Jednotlivé metody se od sebe liší jednak různým způsoby určování podobnosti objektů (měr podobnosti), a jednak způsoby shlukování (např. hierarchické a nehierarchické).

Při výběru vhodné metody shlukové analýzy záleží jednak na tom, zda máme k dispozici přímo zdrojová data či agregované údaje, např. tabulku četností či matici podobností. Pokud máme k dispozici zdrojová data, je výběr metody závislý na typu proměnných (nominální, ordinální či kvantitativní).

Statistické programové systémy obvykle nabízejí jednak hierarchický algoritmus, jehož výsledek bývá zobrazován ve formě tzv. dendrogramu, a také nehierarchický iterativní algoritmus *k*-means. V statistickém systému SPSS je od verze 11.5 implementována metoda TwoStep Cluster Analysis.

Hierarchické shlukování

Jednou z nejznámějších a nejpopulárnějších skupin algoritmů shlukové analýzy jsou metody hierarchického shlukování (Žambochová, 2010). Jejich popularita vychází z jednoduchosti, přehlednosti a dobré interpretovatelnosti výsledků. V hierarchických metodách je vytvářena hierarchie rozdělení do shluků, a to buď aglomerativním nebo divizivním způsobem. Grafickým výstupem hierarchických metod je většinou speciální typ stromového grafu, který se nazývá dendrogram. Jedná se o binární strom, jehož každý list představuje jeden z objektů a vnitřní vrcholy reprezentují shluk obsahující všechny objekty náležející listům podřízeným tomuto vrcholu. Horizontální řezy dendrogramem jsou jednotlivá rozdělení do shluků. Prezentace výsledku shlukování formou dendrogramu je velkou výhodou hierarchických metod. Umožňuje uživateli na základě grafického znázornění výsledků rozhodnout o počtu výsledných shluků. Aglomerativní způsob začíná v situaci, kdy je každý objekt samostatným shlukem. Postupně se spojují vždy dva shluky, které jsou si v daném okamžiku nejvíce podobny. Spojování se provádí až do stavu, kdy jsou všechny objekty v jednom společném shluku. Za kritérium zastavení

může být využito i dosažení určitého počtu shluků. Divizivní způsob postupuje opačným směrem, začíná v situaci, kdy jsou všechny objekty v jednom společném shluku, který je postupně rozdělován až do stavu, kdy je každý objekt v samostatném shluku.

Metoda k-means

Metoda *k*-means je speciálním zástupcem skupiny algoritmů využívajících tak zvané centroidy, jenž tvoří významnou podskupinu metod rozkladu. Jedná se o velmi oblíbený a v praxi patrně nejvíce používaný iterativní shlukovací postup, který je vhodný pro shlukovou analýzu v případě kvantitativních dat. Základním cílem je nalezení takového rozkladu objektů do předem daného počtu shluků, pro který je součet čtverců vzdáleností jednotlivých objektů od center shluků (tak zvaných centroidů), do kterých náleží, nejmenší. Přesněji, nechť **X** označuje množinu všech sledovaných objektů; **x** je libovolný objekt, $c(\mathbf{x})$ značí centroid nejbližší objektu $\mathbf{x} \in \mathbf{X}$ a cílem je nalézt optimální rozklad do *k* shluků, potom minimalizujeme

$$Q = \sum_{\mathbf{x} \in \mathbf{X}} \left\| \mathbf{x} - c(\mathbf{x}) \right\|_2^2 = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{il} \sum_{j=1}^{d} \left(x_{ij} - c_{lj} \right)^2 , \qquad (4.1)$$

kde w_{il} je rovno 1, pokud *i*-tý objekt leží v *l*-tém shluku a w_{il} je rovno 0, pokud tam neleží.

V literatuře je pod různými názvy uvedeno mnoho variant základního postupu *k*-means, ve kterých je centroid tvořen průměrnými hodnotami proměnných. Mimo jiné jsou to Forgyova (Forgy, 1965), Janceyova (Frank a Todeschini, 1994), Llyodova (Frank a Todeschini, 1994; Kanungo a kol., 2002), MacQueenova (MacQueen, 1967), Wishartova (Frank a Todeschini, 1994) metoda, ale i mnoho dalších.

Základní algoritmus *k*-means je tvořen následujícími kroky (Hartigan a Wong, 1979; Faber, 1994):

0. Vstup: datová matice **X**, požadovaný počet shluků *k*.
1. Prvotní „náhodné“ rozdělení objektů do *k* shluků, tj. $\mathbf{C}^0 = \{C_1^0, \ldots, C_k^0\}$.
2. Výpočet centroidů všech shluků, tj. pro $l = 1, \ldots, k$ vypočítat centroid $\mathbf{c}_l$.
3. Přiřazení všech objektů *k* centroidům, tj. pro množinu centroidů $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ vzniklých v kroku 2 vytvořit rozdělení $\mathbf{C}^{i+1} = \{C_1^{i+1}, \ldots, C_k^{i+1}\}$ množiny **X** do shluků.
4. Pokud došlo ke změně, tj. pokud se změnila množina shluků oproti předchozí

iteraci, neboli pokud existuje objekt $x \in \mathbf{X}$, který byl přiřazen do jiného shluku než v předchozí iteraci, návrat na krok 2.

5. Výstup: *k* centroidů a přiřazení objektů matice **X** do *k* výsledných shluků.

Výhody metody *k*-means:

- jednoduchý princip,
- přijatelná rychlost - použitelnost pro relativně velké soubory dat,
- relativně dobré výsledky (díky minimalizaci intra-variability).

Nevýhody metody *k*-means:

- hledá pouze lokální minimum,
- použitelná pouze pro kardinální[2] data,
- je nutné zadat požadovaný počet shluků,
- hledá pouze konvexní shluky,
- silný vliv odlehlých hodnot,
- je časově náročný pro obzvlášť velké soubory,
- vliv počátečního rozdělení dat do shluků je poměrně výrazný.

Odstraňováním, či alespoň snížením nevýhod se zabývalo mnoho autorů. Problém nalezení pouze lokálního optima je dán principem algoritmu, a proto jej nelze odstranit. Existují však varianty metody *k*-means pro jiná než kardinální data. Například Řezanková a Löster (2013) se zabývají hodnocením shluků v případě kategoriálních dat. Způsob umožňující na základě předběžné analýzy stanovit požadovaný počet shluků, navrhli Řezanková a kol. (2008). Velký dopad odlehlých hodnot diskutuje (Žambochová, 2009b, 2010a). Řezanková a kol. (2007) zdůrazňují omezující časové nároky, zejména u velkých souborů. Návrh na snížení vlivu počátečních rozdělení do shluků představuje Žambochová (2009a).

Základním principem fungování algoritmu *k*-means je minimalizace součtu čtverců vzdáleností jednotlivých objektů od jistých center. Z tohoto faktu vyplývá sférický tvar vytvořených shluků. Tento fakt lze považovat za jednu z negativních

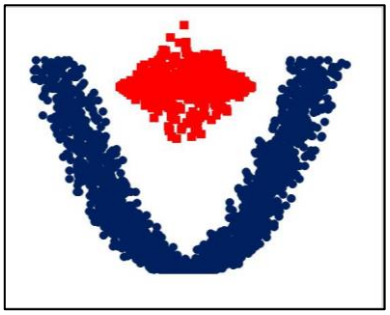
[2] Existuje více typologií veličin, v této práci je používána terminologie dle Stevens (1946). V rámci této typologie jsou veličiny děleny do tří základních skupin, a to nominální, ordinální a kardinální. Poslední ze jmenovaných typů se ještě dělí na dvě podskupiny, a to spojité a diskrétní. V literatuře jsou někdy kardinální veličiny nazývány též kvantitativní, měřitelné či metrické.

vlastností algoritmu *k*-means. Žambochová (2014) proto navrhla možné řešení, viz příloha 5.

Nevhodné dělení datových souborů obsahujících nekonvexní shluky je dobře viditelné z dvou sad jednoduchých dat zobrazených na obr. 1 až obr. 4. Jsou zde vždy graficky znázorněny jednak přirozené shluky, tak jak se opravdu jeví v souborech, a jednak skupiny daného počtu shluků tak, jak je vytvořil základní algoritmus *k*-means při různě zadaném parametru, který představuje zadání konkrétního počtu požadovaných shluků. Je vidět, že pokud pomocí metody *k*-means vytvoříme přesně požadovaný počet shluků, je rozdělení naprosto nevyhovující. Pokud však postupně rozdělujeme zpracovávaný soubor na větší počet shluků než je požadovaný, dostaneme se do situace, že se každý z nekonvexních shluků rozdělí do několika konvexních částí.

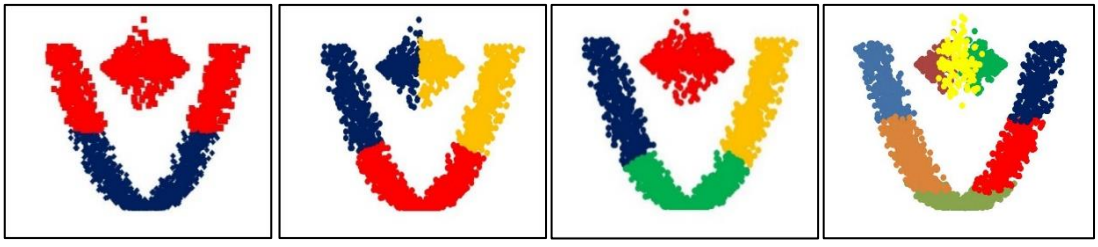
Odtud vznikl nápad provádět shlukování ve dvou fázích. V první fázi se použije algoritmus *k*-means a provede rozdělení do většího počtu shluků, než je požadováno. V druhé fázi zpracování pak některé shluky „vhodně pospojujeme“ a tím dostaneme požadované množství shluků. Tyto shluky budou tvarově blíže přirozeným shlukům než shluky vytvořené přímo metodou *k*-means.

Obr. 1: Soubor I. – přirozené shluky



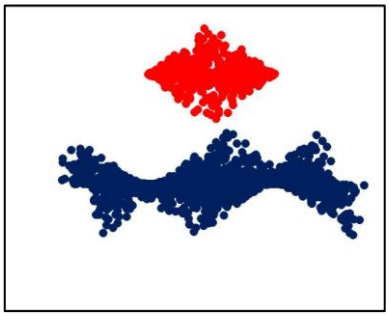
Obr. 2: Soubor I. – shluky vytvořené pomocí metody *k*-means pro různé počty shluků

2 shluky 3 shluky 4 shluky 8 shluků



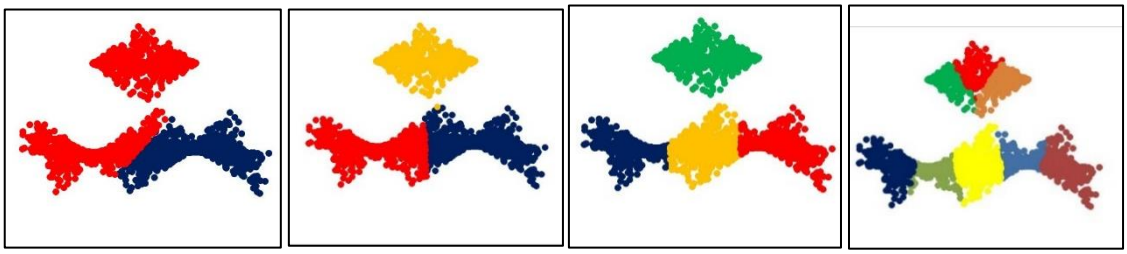
Zdroj: Vlastní zpracování (Žambochová, 2014)

Obr. 3: Soubor II. – přirozené shluky



Obr. 4: Soubor II. – shluky vytvořené pomocí metody *k*-means pro různé počty shluků

2 shluky 3 shluky 4 shluky 8 shluků



Zdroj: Vlastní zpracování (Žambochová, 2014)

Přitom je nutno vybrat vhodnou metodu spojování shluků. Lze využít například některou ze základních metod aglomerativního hierarchického shlukování (Řezanková a kol., 2007; Everit, a kol., 2001). Bohužel mnohé z nich nevytváří vhodné výsledné shluky. Některé jsou příliš pomalé a tím nepoužitelné pro velké soubory. Dále se nabízejí další možnosti v rámci jiných algoritmů. Varianty nabízejí například (Karypis a kol., 1999; Guha a kol., 2001; Kogan a kol., 2006). Výhody a nevýhody jednotlivých možností blíže diskutuje Žambochová (2014).

Dvoukroková (TwoStep) metoda

Princip TwoStep metody uvádí například Žambochová (2010). Tento postup využívá algoritmus BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), který blíže popsují Zhang a kol. (1996, 1997). Algoritmus nejprve vytváří tzv. CF-strom, do kterého zařazuje postupně přicházející data. Výhodou tohoto postupu je, že prochází datový soubor pouze jedenkrát. Nevýhodou je poměrně velká citlivost na pořadí vstupujících datových bodů.

Vlastní algoritmus shlukování probíhá ve třech hlavních fázích. V první fázi vytváří CF-strom, do kterého zařazuje postupně přicházející objekty. Ve druhé fázi kondenzuje vytvořený CF-strom a optimalizuje jeho velikost upravením prahové hodnoty (jeden z parametrů CF-stromu), a pomocí vhodného „přestavění“ stromu zároveň umožní

odstranění odlehlých objektů. Ve třetí fázi je minimalizován vliv na pořadí vstupních dat. Algoritmus shlukuje listové vrcholy pomocí aglomerativního hierarchického algoritmu shlukování. Metoda je implementována do statistického systému SPSS pod názvem „TwoStep Cluster Analysis“.

Kritéria pro určení optimálního počtu shluků

Pro určení optimálního počtu shluků bylo navrženo mnoho informačních kritérií (Řezanková a kol. 2008; Žambochová, 2017). V systému SPSS jsou implementována tři informační kritéria.

Prvním je Schwarzovo bayesovské kritérium, neboli Bayesian Information Criterion, které je označováno zkratkou BIC. Toto kritérium slouží ke stanovení optimálního počtu shluků při dvoukrokové shlukové analýze. Počítá se podle vztahu

$$BIC(C_k) = -2\sum_{i=1}^{k} \lambda_i + w_k \ln(N), \tag{4.2}$$

kde N značí počet objektů, λ_i je charakteristika pro i-tý shluk spočtená podle vzorce

$$\lambda_i = -n_i \sum_{j=1}^{m_1} \frac{1}{2} \ln\left(s_j^2 + s_{ij}^2\right) + \sum_{j=1}^{m_2} H_{ij}\ , \tag{4.3}$$

kde n_i je počet objektů v i-tém shluku, m_1 je počet kvantitativních spojitých proměnných, m_2 je počet kategoriálních proměnných, s_j^2 je výběrový rozptyl j-té spojité proměnné, s_{ij}^2 je výběrový rozptyl j-té spojité proměnné v i-tém shluku a H_{ij} je entropie daná vztahem

$$H_{ij} = -\sum_{l=1}^{r_j} \frac{n_{ijl}}{n_i} \ln\left(\frac{n_{ijl}}{n_i}\right), \tag{4.4}$$

kde r_j je počet kategorií j-té kategoriální proměnné a n_{ijl} představuje četnost l-té kategorie j-té kategoriální proměnné v i-tém shluku.

Hodnoty vah w_k ve vzorci (4.2) se spočtou podle vzorce

$$w_k = k\left(2m_1 + \sum_{j=1}^{m_2} r_j - 1\right), \tag{4.5}$$

Pro stanovení optimálního počtu shluků se nejprve vypočítají hodnoty Schwarzova bayesovského kriteria pro počty shluků v rámci předem stanoveného intervalu. Na základě těchto hodnot se podle minimální hodnoty Schwarzova bayesovského kritéria stanoví počáteční odhad počtu shluků.

Druhé kritérium je zváno Akaikeho, označuje se *AIC* (*Akaike Information Criterion*) a počítá se ze vztahu

$$AIC(C_k) = -2\sum_{i=1}^{k} \lambda_i + 2w_k \ , \tag{4.6}$$

kde w_k je opět dáno vztahem (4.5).

Při stanovení optimálního počtu shluků se postupuje stejně jako v případě BIC.

Pro hodnocení výsledných shluků získaných pomocí metod rozkladu se používá tzv. obrysový koeficient (silhouette coefficient), který vyjadřuje míru koheze a separace. Pro *i*-tý objekt z *h*-tého shluku se počítá podle vzorce

$$SC_i = \frac{\mu_i - \eta_i}{\max\{\mu_i - \eta_i\}} \ , \tag{4.7}$$

kde

$$\eta_i = \frac{\sum_{j \in C_h} D_{ij}}{n_h - 1} \qquad \text{a} \qquad \mu_i = min_{g \neq h}\left(\frac{\sum_{j \in C_g} D_{ij}}{n_g}\right),$$

kde D_{ij} představuje vzdálenost *i*-tého a *j*-tého objektu, tedy η_i je průměrná vzdálenost *i*-tého objektu od ostatních objektů společného shluku a μ_i je minimální vzdálenost *i*-tého objektu ze všech vzdáleností tohoto objektu od všech objektů z jiných shluků.

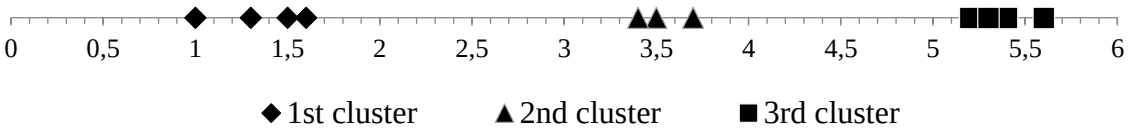
Obrysový koeficient se vypočítá jako průměr hodnot SC_i, vypočítaných dle (4.7), přes všechna data, tedy

$$SC = \frac{\sum_{i=1}^{n} SC_i}{n} \ . \tag{4.8}$$

Je zřejmé, že pro každé *i* platí $-1 \leq SC_i \leq 1$, tedy i hodnoty *SC* se pohybují v rozmezí od -1 do 1. Dále platí, že čím je hodnota *SC* vyšší, tím jsou shluky kompaktnější a separovanější. Obrysový koeficient lze použít nejenom pro porovnání metod, ale také pro stanovení optimálního počtu shluků.

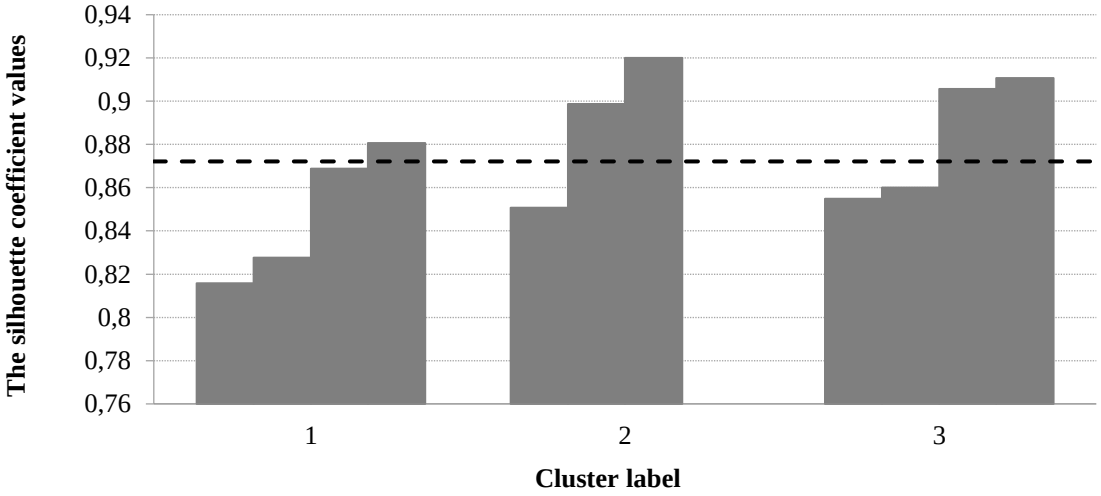
Obrázky 5. a 6. ukazují jednoduchý a ilustrativní příklad konstrukce koeficientu siluety. Obrázek 5 představuje situaci jedenácti objektů rozdělených do tří shluků. Na obrázku 6 je graficky zobrazena reprezentace jednotlivých hodnot obrysových koeficientů (šedé sloupce) a výsledného průměrného obrysového koeficientu (černá čárkovaná čára).

Obr. 5: Datový soubor obsahující 11 objektů rozdělených do tří shluků



Zdroj: Vlastní zpracování (Žambochová, 2017)

Obr. 6: Obrysový koeficient pro tři shluky dat zobrazené na obr. 5.



Zdroj: Vlastní zpracování (Žambochová, 2017)

4.2 Využití shlukové analýzy pro marketing

S rozvojem informačních technologií přichází i rozvoj databázových technologií v maloobchodě. Data týkající se transakcí se zákazníky jsou zaznamenávána a ukládána do velkých databází, což se stává výzvou pro různé výzkumné pracovníky (Peker a kol., 2017). Segmentace zákazníků je velmi důležitou úlohou v řízení vztahů se zákazníky, která primárně využívá shlukovou analýzu. Jedná se o seskupování zákazníků do skupin na základě jejich podobnosti, a to především podobnosti nákupních vzorů zákazníků či jejich preferencí. Jedním z hlavních cílů je identifikace problémových zákazníků. Dalším cílem segmentace zákazníků je i pomoc manažerům optimalizovat marketingové zdroje. Vytvoření segmentů může manažerům pomoci při vytváření strategie cílení na zákazníky na základě jejich atraktivity (Liao a kol., 2011; Carmichael a kol., 2018).

4.2.1 Segmentace spotřebitelů v kontextu jejich prostorového chování – případová studie

Studie (Grossmanová, Kita a Žambochová, 2016), která je součástí přílohy 6, analyzuje vývoj maloobchodní sítě hlavního města Slovenska Bratislavy ovlivňující nákupní chování a životní styl spotřebitelů. Cílem tohoto příspěvku bylo analyzovat tržní segmenty spotřebitelů ve městě Bratislava v kontextu jejich prostorového chování za účelem zvýšení obchodní atraktivity jednotlivých městských částí. Studium prostorového chování spotřebitelů nabízí příležitost, a to jak na strategické úrovni, tak i po taktické stránce. V této souvislosti jsou prostorové interakce výsledkem vlivu mnoha faktorů, a je potřeba odhalit vzájemnou provázanost. V případě maloobchodu se prostorové interakce vytvářejí hlavně mezi místem bydlení zákazníka a místem, které navštěvuje nejčastěji (práce, volný čas), a je silně ovlivněna mobilitou zákazníků. V rámci analýzy lokality jako předpokladu rozhodování o umístění maloobchodu je důležité si uvědomit, že toto rozhodnutí určuje vzdálenost k zákazníkovi, který musí vyvinout určité úsilí k překonání vzdálenosti, strávit nějaký čas a vynaložit náklady na možnou dopravu (Šveda a Križan, 2012). Za předpokladu určitých preferencí zákazníků jsou firmy, jejichž přirozeným cílem je maximalizovat podíl na trhu nebo zisk, nuceny přijímat rozhodnutí o jejich umístění a ovlivňovat svými akcemi výsledky a strategie svého konkurenta (Suárez-Vega a kol., 2014).

Jak argumentují Tsai a kol. (2004), pro dobrou segmentaci nestačí použít pouze základní obecné demografické proměnné. Proto byly k těmto základním proměnným přidány proměnné informující o konkrétních nákupech, zejména o druhu zboží, délce a datu nákupu, způsobu dopravy a délce cesty na nákup, o spokojenosti či o nedostatcích tohoto nákupu.

Cílem výzkumu provedeného v roce 2011 na území hlavního města Slovenska Bratislavy bylo vytvořit prostorové rozvržení maloobchodní sítě v území města s využitím znalostí geomarketingu a tvorby databázových dat o maloobchodní síti jako součásti geografického informačního systému (GIS). Pomocí standardizovaného dotazníku bylo osloveno 11 389 respondentů nakupujících v maloobchodních zařízeních umístěných v jednotlivých částech města Bratislava. Respondenti museli mít trvalé bydliště nebo přechodný pobyt v Bratislavě, a být plnoletí.

Jako nástroj pro segmentaci byla vybrána shluková analýza. Metoda *k*-means je použitelná pouze v případě kardinálních veličin, což v případě této studie splněno nebylo. Hierarchický postup shlukování byl zavržen převážně z důvodu špatné čitelnosti výsledného přiřazení objektů k jednotlivým shluků, a to z důvodu velkého počtu shlukovaných objektů. Nejvhodnější pro účely studie se ukázala TwoStep metoda.

Určení počtu shluků bylo ponecháno na automatickém vyhodnocení dle kritéria AIC, výsledným optimálním počtem byly dva shluky. Kvalita výsledného shlukování byla dobrá, hodnota průměrného obrysového koeficientu byla 0,6. Výstupem této fáze zpracování byla nová proměnná znázorňující příslušnost k danému shluku.

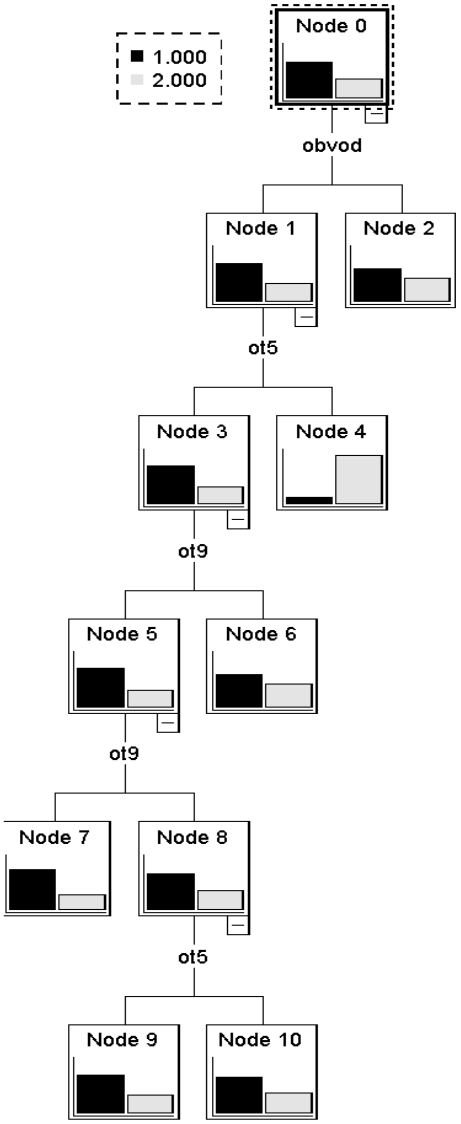
Následná fáze zpracování se týkala interpretace získaných shluků. Interpretace se řadí mezi nejpodstatnější fáze každé analýzy dat, v případě použití vícerozměrných statistických metod, mezi něž se řadí i shluková analýza, je interpretace obzvlášť důležitá. V případě velkých datových souborů je však situace velmi nepřehledná a interpretační vztahy jsou často těžko odhalitelné. Proto je užitečné využít vhodný nástroj, například rozhodovací stromy, které naleznou pravidla popisující vlastnosti objektů v jednotlivých shlucích, a to z pohledu zadaných vysvětlujících proměnných.

Byly sestrojeny dvě skupiny rozhodovacích stromů vytvořených na základě sociálně ekonomických hledisek, respektive typu nákupního chování. V obou případech byly použity metody CRT, QUEST a CHAID. Vysvětlovanou proměnnou byla vždy proměnná vyjadřující příslušnost ke shluku. Vysvětlujícími proměnnými byly při zpracování dle prvního hlediska proměnné obsahující identifikační údaje (např. věk, pohlaví, příjem, počet členů domácnosti, obvod). Cílem vytvoření těchto stromů bylo získat sociodemografický popis spotřebitelů v jednotlivých shlucích. V případě druhého hlediska byly za vysvětlující proměnné vybrány proměnné vystihující nákupní chování a preference. Cílem vytvoření stromů této skupiny bylo získat nákupní a názorové vlastnosti charakteristických zástupců jednotlivých shluků. V obou případech byl vybrán strom s nejlepší kvalitou, tj. nejmenší hodnotou odhadu rizika, která vyjadřuje míru špatně klasifikovaných případů. Jeden z každé skupiny těchto stromů je zobrazen na obr. 7 a 8. Z důvodu rozsáhlosti nebyly pro zobrazení vybrány stromy s nejlepší kvalitou, na jejichž základě bylo provedeno vyhodnocení.

Odhad rizika se u všech vytvořených stromů z první skupiny pohyboval v rozmezí od 0,349 do 0,352. To znamená, že míra úspěšnosti zařazení objektů se pohybovala

v rozmezí od 65,1 % do 64, 8%. Kvalita první skupiny modelů sice nebyla příliš vysoká, ale přesto jsou vytvořené modely dobře použitelné. Jednotlivé modely jsou opět kvalitativně velmi srovnatelné, a jsou tedy dobře použitelné k interpretaci výsledků.

Obr. 7: Ukázka rozhodovacího stromu z první skupiny (s využitím algoritmu QUEST)



Vysvětlivky:

ot5: počet členů domácnosti

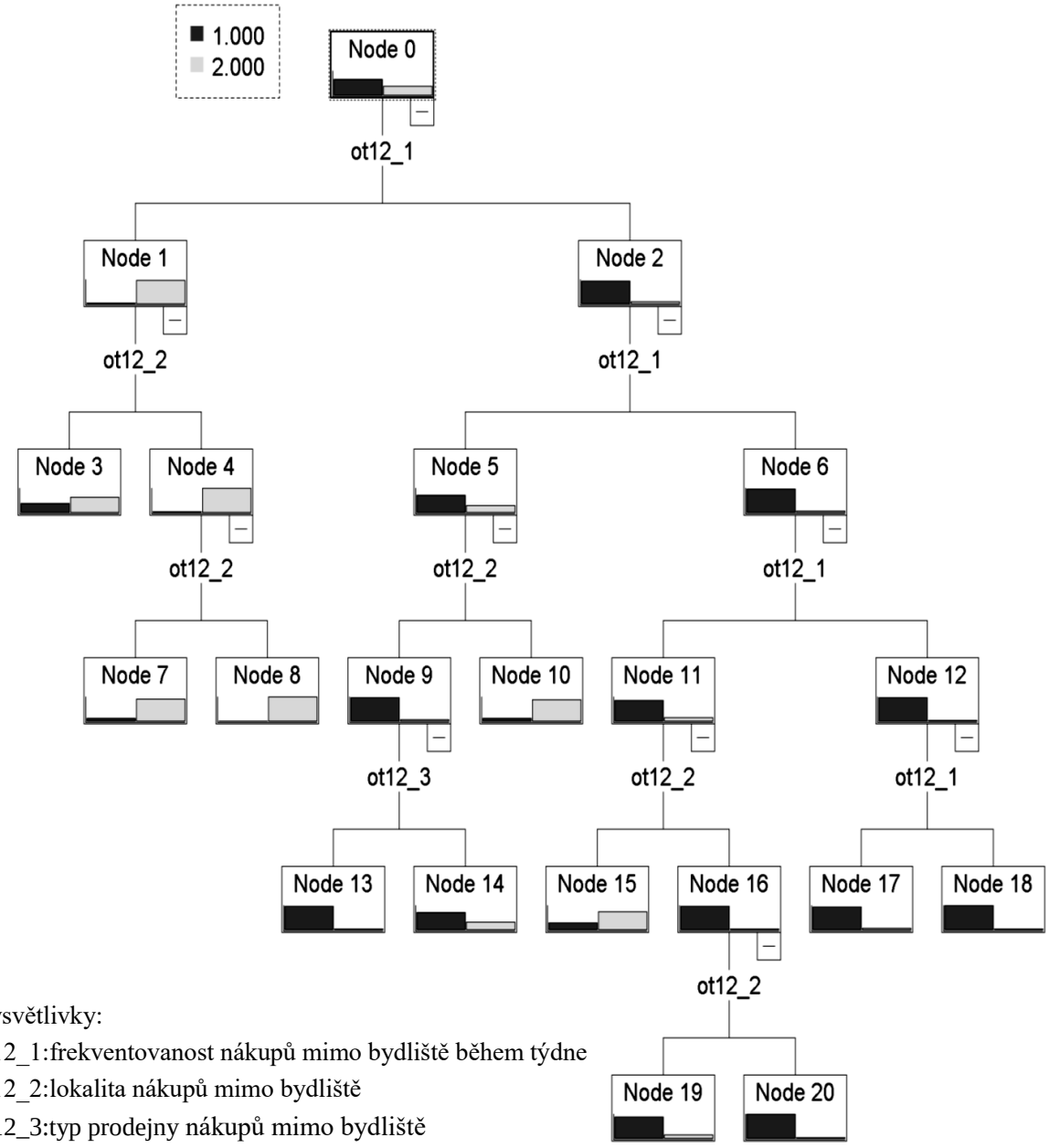
ot9: frekventovanost nákupů během týdne

Zdroj: Vlastní zpracování (Grossmanová, Kita a Žambochová, 2016)

Všechny použité algoritmy pro vytváření rozhodovacích stromů na základě prvního hlediska ukázaly PSČ jako položku, která nejvíce ovlivňuje přiřazení do shluku. Z důvodu nepřehlednosti výstupů byla tato položka vyjmuta ze zpracování a jako identifikace místa bydliště byla ponechána pouze položka obvod. Kvalita modelů se tímto krokem téměř nezměnila (pouze o setiny procenta), což znamená, že tento krok nevedl ke zhoršení výsledných výstupů a je dobře použitelný. Úlohu nejvlivnější položky v těchto

nových modelech převzala položka obvod. Druhou nejvlivnější sledovanou položkou byla informace o počtu členů domácnosti, třetí v pořadí pak informace o frekvenci nákupů a času stráveném v obchodě. Některé modely ještě poukázaly na vliv věku a příjmu respondentů.

Obr. 8: Ukázka rozhodovacího stromu z druhé skupiny (s využitím algoritmu CRT)



Vysvětlivky:
ot12_1:frekventovanost nákupů mimo bydliště během týdne
ot12_2:lokalita nákupů mimo bydliště
ot12_3:typ prodejny nákupů mimo bydliště

Zdroj: Vlastní zpracování (Grossmanová, Kita a Žambochová, 2016)

Odhad rizika se u všech vytvořených stromů z druhé skupiny pohyboval v rozmezí od 0,029 do 0,034. To znamená, že míra úspěšnosti zařazení objektů se pohybovala v rozmezí od 97,1 % do 96,6 %. Kvalita této skupiny modelů byla tedy velmi vysoká,

jednotlivé modely jsou kvalitativně velmi srovnatelné, všechny jsou tedy výborně použitelné k interpretaci výsledků.

Ze struktury těchto vybraných stromů jsme vyčetli popis skupin vytvořených shlukovou analýzou. Tím jsme zjistili, jaký typ respondentů upřednostňuje jaký typ nakupování.

Z hlediska typu nákupního chování se jako položka s největším významem pro přiřazení do shluku ukázala informace o volbě obchodního řetězce. Z tohoto pohledu se těsně sdružili respondenti upřednostňující známé řetězce oproti respondentům upřednostňujícím méně známé nebo malé obchody. Na druhém místě v síle vlivu se společně umístily položky obsahující informace o četnosti a délce nákupu v jiných lokalitách. Z tohoto pohledu se těsně sdružili respondenti vyjíždějící za nákupem někam mimo pravidelně, a to minimálně několikrát za týden oproti lidem vyjíždějícím za nákupem mimo svou lokalitu pouze příležitostně, a to v delším než týdenním intervalu. Z interpretačního hlediska tyto položky přinesly pouze málo informací, a svým vlivem z modelů vyřadily položky ostatní. Proto byl proveden pokus je z dalšího zpracování vyřadit. Kvalita modelů se téměř nesnížila, takže výsledné modely byly dobře akceptovatelné. V nově vytvořených modelech se poté jako důležité projevily položky s informacemi o typu nejčastěji navštěvovaných prodejen. Dalšími z položek majícími silný vliv na přiřazení do shluků jsou i položky obsahující informaci o důvodu volby navštíveného řetězce, o nedostatcích, které respondent té či oné prodejně pociťuje, či informace o úrovni spokojenosti s prodejnou.

Z vytvořených modelů vyplynulo, že názory a chování nakupujících nejvíce ovlivňuje lokalita jeho bydliště, velikost domácnosti a četnost a délka nákupů, upřednostňovaný sortiment, v menší míře pak i věk a příjem respondenta.

Ukázalo se, že ženy-hospodyňky s průměrným příjmem, bydlící v 1. a 3. obvodu, které nakupují převážně ve všední dny a nestráví nákupem průměrně více než 1,5 hodiny, si vybírají svůj obchod (převážně potraviny, masny, pekařství, dětské oblečení, hračky) na základě sortimentu, výše cen či jen ze zvyku, přičemž si nemyslí, že by v poslední době došlo v jejich prodejně k nějaké výrazné změně v kvalitě nákupu. Upřednostňují nákup ve známých obchodních řetězcích. Při svých nákupech by přivítaly zlepšení v oblasti dostupnosti prodejny, bankovních služeb a byly by vděčné za zvýšení množství zdravých a bio potravin.

Respondenti z 2. a 5. obvodu nakupující méně často, spíše náhodně, a to převážně v prodejnách s bytovými doplňky, domácími potřebami, sportovními potřebami či ve zlatnictví, stráví průměrně svými nákupy i několik hodin. Tito respondenti si nejsou v poslední době vědomi žádné změny k horšímu v jejich prodejně, nejsou však spokojeni s dostupností prodejen, chybí jim častější prodejní akce a dětský koutek v prodejně.

Další ze skupin tvoří respondenti 2. a 5. obvodu, kteří nakupují velmi málo, a to převážně v neděli nebo pondělí, ale také mladí lidé z 1. a 3. obvodu, kteří nákupu věnují velmi dlouhou dobu (průměrně nad 175 min). Tato skupina je tvořena z velké části osobami nejčastěji nakupujícími v prodejnách elektro, obchodech s počítači, trafikách, sex shopech či bufetech, jednak respondenty, kteří jsou se svými prodejnami spokojeni z hlediska přístupu prodavačů, organizace prodejny i bankovními službami. Přivítali by delší trvání prodejních akcí, zlepšení vybavení prodejen a bezbariérový přístup do prodejny.

Obyvatelé 2. a 5. obvodu nakupující vícekrát týdně, kteří mají nadprůměrný příjem, jsou starší 29 let a nejčastěji nakupují v masně, trafice, prodejnách elektro, květinářství či hračkářství, si nejsou vědomi výraznějších zlepšení v kvalitě nákupu a nejsou se svými obchody ani výrazně spokojeni ani nespokojení.

Obyvatelé 2. a 5. obvodu, kteří bydlí společně s maximálně dvěma dalšími osobami a chodí v místě bydliště nakupovat pravidelně, cestující do svého obchodu pouze krátce, nejčastěji nakupující v pekařství, drogerii, trafice, papírnictví či obuvi, nejsou spokojeni s nabízeným sortimentem, s dlouhým časem stráveným ve frontách, chtěli by zlepšit kvalitu služeb a chybí jim dostatečné informace o nabízených výrobcích.

Studie (Grossmanová, Kita a Žambochová, 2016) si kladla za cíl dva hlavní cíle. Prvním z nich bylo zjistit, jak zákazníci bratislavských nákupních středisek vnímají nákupní centra, a to jejich segmentací podle jejich aktivit a nákupních zvyklostí. Druhým cílem bylo zkoumat význam sociálnědemografických charakteristik pro každý segment. Zjištění této studie ukazují, jak segmentace může pomoci porozumět preferencím a vzorcům chování zákazníků během jejich cest do obchodního centra. Kromě toho další studie mohou těžit z tohoto výzkumu, který je považován za součást časové osy průzkumu vývoje maloobchodní sítě a chování spotřebitelů při budování časoprostorového modelu spotřebitelského chování občanů Bratislavy. Z metodologického hlediska studie ukazuje využití nástroje ve formě rozhodovacích stromů k podrobné interpretaci výsledků

shlukové analýzy a nalezení popisu vlastností charakteristických spotřebitelů jednotlivých segmentů vzniklých při první fáze procesu segmentace.

4.3 Využití shlukové analýzy pro analýzu financování školství

Financování terciálního školství je velmi diskutovaným tématem jak na úrovni teoretické ekonomie, tak ještě silněji v rámci ekonomické praxe. Česká republika patří v rámci OECD dlouhodobě k zemím s podprůměrnými výdaji na vysokoškolské vzdělávání. Přitom navyšování veřejných prostředků do tohoto sektoru brání v České republice fiskální omezení EU (Matějů a kol., 2009). Kvantitativní rozvoj českého vysokého školství na přelomu tisíciletí vedl k přechodu od elitního k univerzálnímu vzdělávání. Nedostatek veřejných zdrojů na tuto masifikaci a polemika o tom, kdo má prospěch ze vzdělání, vedou nutně k novým podnětům do diskuze o reformě českých vysokých škol (Vomáčková, Žambochová a Tišlerová, 2011). Proto jsou nutné nové koncepce financování veřejných vysokých škol. Zavedení kvalitativních kritérií do financování mělo zásadní dopad na fungování jednotlivých škol (Taušer a Žamberský, 2012). Průchodnost často diskutované participace studentů na financování jejich studia je podmíněna její návratností (Finardi a kol., 2012).

4.3.1 Zkoumání faktorů ovlivňujících ochotu zahraničních studentů platit školné – případová studie

Hlavní motivací studie (Žambochová, 2012a), která je součástí přílohy 2, byla analýza alternativních možností financování terciálního školství. Primárně se studie zaměřila na školné placené zahraničními studenty. Na základě získaných informací byla vytvořena segmentace zahraničních zájemců o studium na českých vysokých školách. Byly popsány hlavní charakteristiky jednotlivých kategorií, jejich finanční možnosti a motivace, ale také jejich vzájemné odlišnosti. Výzkumný vzorek tvořilo 1093 studentů ze 6 zemí, a to Slovenska, Řecka, Ukrajiny, Ruska, Běloruska a Číny. Nejvíce respondentů bylo z Číny, čínští studenti tvořili téměř polovinu celkového sledovaného souboru. Naopak nejmenší vzorek dat byl sebrán v Ruské federaci.

V rámci zpracování byla nejprve provedena shluková analýza. Přesněji, byla vybrána procedura TwoStep ze statistického systému SPSS, a to z důvodu zpracování převážně nominálních veličin. Druhým důvodem této volby byla právě možnost vytvořit

novou veličinu představující příslušnost ke shluku. Určení počtu shluků bylo ponecháno na automatickém vyhodnocení dle kritéria AIC. Metoda nabídla jako optimální řešení čtyři shluky. Tři z nich byly podobné velikosti, jeden byl co do počtu zařazených objektů velmi malý. Bližším průzkumem se zjistilo, že jsou v něm zahrnuti studenti, kteří nemají zájem o studium v zahraničí, a otázky týkající se případného studia v zahraničí převážně nevyplnili.

Velikost datového souboru, a to jak z pohledu dimenze, neboli počtu sledovaných veličin, tak i z pohledu počtu objektů, činila interpretaci výsledků shlukové analýzy velmi obtížnou. Proto byl opět využit nástroj v podobě rozhodovacích stromů.

Z důvodu velkého množství veličin získaných z dotazníku bylo před použitím algoritmů na tvorbu rozhodovacích stromů použito předzpracování dat, sestávající se z výběru pouze těch veličin, které v rámci chí-kvadrát testu nezávislosti vykázaly asociaci s novou veličinou určující příslušnost ke shluku. To znamená, že do algoritmů na tvorbu stromů vstupovaly jako vysvětlující proměnné pouze tyto signifikantní veličiny. K tvorbě byly použity tři algoritmy, které jsou implementovány ve statistickém systému SPSS, tj. CRT, CHAID a QUEST. Hodnota odhadu rizika vyšla pro všechny vytvořené stromy podobná, a to v rozsahu od 0,282 do 0,325, což sice není ideální, ale dostačující úroveň. Z hlediska interpretovatelnosti poskytl nejzajímavější řešení algoritmus CRT (obr. 9), zatímco algoritmus QUEST dopadl nejhůře.

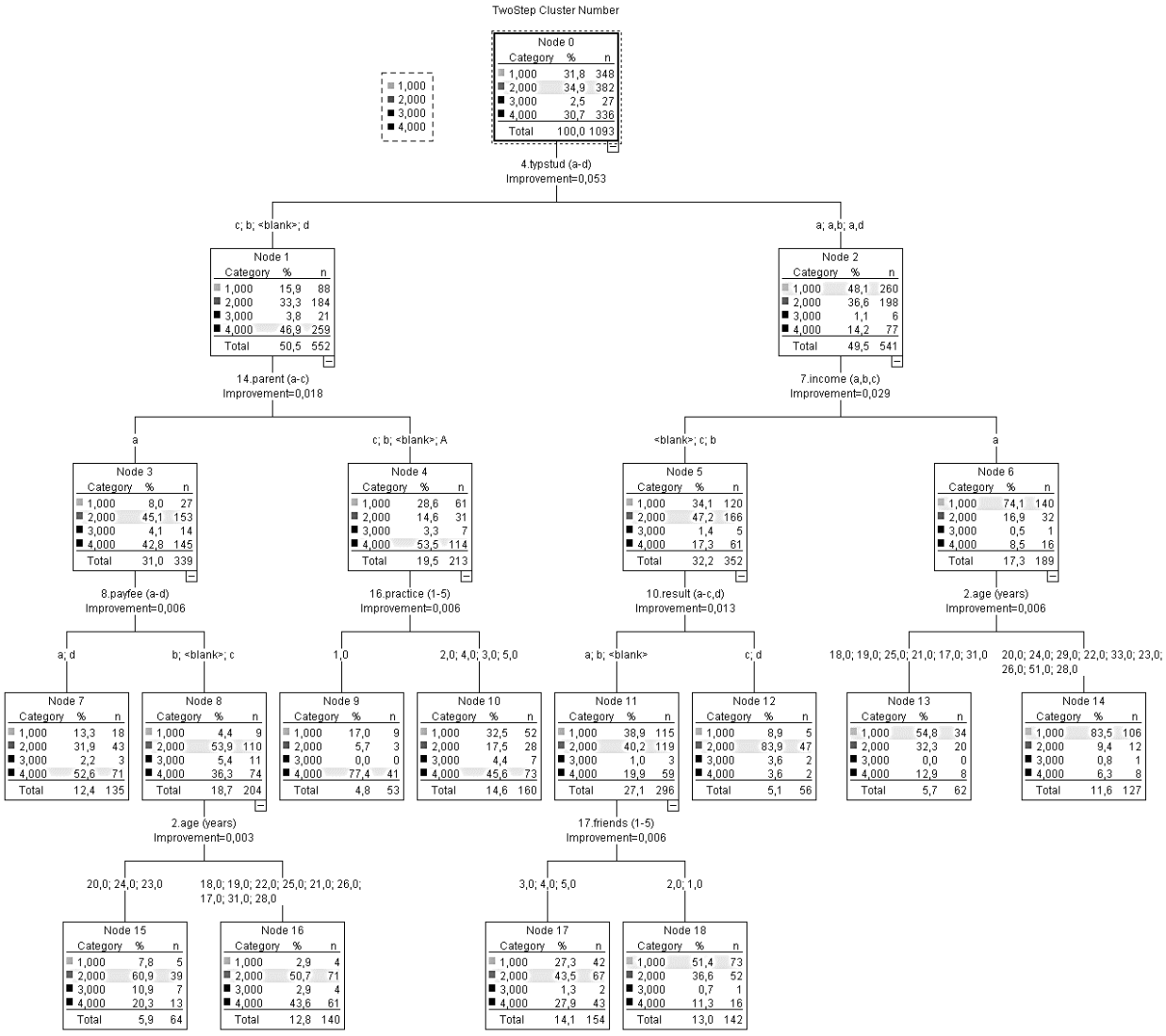
Dle výstupů z jednotlivých algoritmů je možno charakterizovat příslušníky daných shluků z hlediska odpovědí na vybrané otázky. Výše zmíněný čtvrtý shluk, co do počtu objektů zanedbatelný, se v ani jednom typu stromu nijak nevyčleňoval, což bylo z interpretačního hlediska příznivé, neboť to znamenalo, že v něm jsou zařazeny objekty, jež bylo možné označit jako blíže nespecifikovaný šum.

V běžných případech se pro výslednou interpretaci vybírá vesměs strom s nejmenší hodnotou odhadu rizika. V tomto případě byly všechny tři vzniklé stromy podobné kvality, a proto se mohlo k popisu jednotlivých shluků využít kombinace těchto tří stromů.

Segmentace dala zajímavé výsledky, neboť vytvořené shluky byly velmi dobře interpretovatelné. Dále se ukázalo, že v každém ze vzniklých shluků je dominantní příslušnost k jednomu (resp. několika) státům původu studenta. Odtud vzniklo doporučení pro marketing, že bude nejjednodušší se cíleně zaměřit pro daný typ studia na

studenty daných zemí, protože se vyplatí investovat do segmentů s dostatečným počtem studentů. Jednotlivé algoritmy pro tvorbu rozhodovacích stromů poskytly také řadu dalších a rozumně interpretovatelných výsledků.

Obr. 9: Rozhodovací strom vzniklý metodou CRT



Zdroj: Vlastní zpracování v SPSS (Žambochová, 2012a)

Záměrem provedeného šetření v šesti zemích bylo mimo jiné zjistit a nasimulovat model tzv. ideálního zákazníka, tedy specifikovat, jaký student vykazuje největší ochotu studovat v České republice (na bázi samofinancování).

Studenti byli mimo jiné dotazováni na svá zázemí (příjmová skupina, vzdělání rodičů, stávající obor a stupeň studia, atd.), dále byly zkoumány jejich motivy, obavy a mnoho dalších charakteristik.

Na základě tohoto šetření (jehož součástí byla mimo jiné i výše popsaná segmentace) byl vyprofilován „ideální“ typ studenta z hlediska ochoty a výše finančních prostředků, které by za své studium v ČR investoval.

Ideální zákazník:

- země původu: Čína (resp. Řecko),
- jeho rodiče jsou oba (nebo alespoň jeden z nich) vysokoškolsky vzdělaní,
- plánuje v zahraničí dlouhodobější doktorské studium, případně střednědobý pobyt v zahraničí v rámci magisterského studia,
- bez rozdílu věku (nebyla prokázána závislost),
- bez rozdílu pohlaví (nebyla prokázána závislost).

Studie poskytla nové pohledy a podněty do stále aktuální diskuze o reformě českých vysokých škol a umožnila rozšířit prostor kritického vnímání a posuzování faktorů ovlivňujících jejich kvalitu. Provedla hloubkovou analýzu zájmu zahraničních studentů vybraných zemí o případné studium v ČR, a jejich ochotu participovat na financování jejich studia. Po metodologické stránce studie ukazuje možnosti segmentace při hledání alternativních možností financování vysokých škol.

5 Hledání charakteristických rysů získaných tříd mnohorozměrných objektů

Ve všech studiích popsaných v předchozích kapitolách byla data získána pomocí dotazníkového šetření. Tím byla získána mnohorozměrná data, kde každý objekt představoval jednoho respondenta. Každý objekt pak byl reprezentován pomocí několika proměnných vypovídajících o jednotlivých vlastnostech daného objektu.

Ve studii (Žambochová, 2017), která posloužila jako základ a inspirace této kapitoly, však jsou zpracovávána mnohorozměrná data zcela jiného charakteru. Data byla získána opakovaným zjišťováním jedné vlastnosti pro každý ze sledovaných objektů.

5.1 Funkcionální data

V mnoha vědních oborech se setkáváme s tak zvanými funkcionálními daty, např. v meteorologii, medicíně, strojírenství, ale také ekonomii. Jedná se o data získaná opakovaným sledováním daných veličin na množině objektů. Tato měření mohou být prováděna buď nepřetržitě během časového intervalu, nebo diskrétně v několika samostatných po sobě následujících časových okamžicích. Obecně není nutno, aby měření pro jednotlivé objekty probíhalo po stejně dlouhých časových úsecích, ani po stejně dlouhé období. Jak název napovídá, předpokládá se existence nějaké funkce, jež danou veličinu popisuje a kterou data, která máme k dispozici, dobře reprezentují. Tato funkce většinou není známa, ale je charakterizována naměřenými hodnotami (Ramsay a Silverman, 2005) Analýzou funkcionálních dat (FDA) se zabývá mnoho publikací. Jednou z často používaných analýz je shluková analýza (James a Sugar, 2003; Tarpey a Kinateder, 2003).

Aplikací shlukové analýzy funkcionálních dat se zabývala Žambochová (2017), viz příloha 7. Tento článek je případová studie zkoumající dopad významných událostí na počet odbavených pasažérů. Ve studii bylo analyzováno 838 letišť, u kterých byl v období od ledna 2000 do konce roku 2013 měsíčně sledován počet odbavených pasažérů. Důležitým rysem těchto dat bylo, že informace byly sbírány ve stejných okamžicích a po stejně dlouhé období, tedy pro všechna sledovaná letiště bylo sebráno stejné množství údajů. Data jsou tak úplná a poměrně rozsáhlá.

5.2 Zkoumání dopadu významných událostí – případová studie

Letecká doprava patřila do roku 2020 mezi nejrychleji rostoucí odvětví dopravy, stala se nejrychlejší, nejkomfortnější a nejbezpečnější dopravou na světě, a tím i jedním z nenahraditelných způsobů dopravy. Mezinárodní asociace letecké dopravy IATA předpovídá, že mezinárodní letecká doprava poroste v následujícím desetiletí v průměru o 6,6 % ročně, blíže viz Kasturi a kol. (2016).

Do současné doby se velká část údajů o leteckých společnostech nevyužívá pro analytické účely. Hlavním důvodem je fakt, že data jsou v nestrukturované, respektive polostrukturované podobě. Studie FAA udává, že objem produkovaných dat se pohybuje v desítkách terabytů za rok. Kritická analýza 63 empirických studií odhaluje, že na rozdíl od medicíny je využití analýzy dat v sektoru leteckých společností stále v počáteční fázi,

a schopnost těchto studií generovat znalosti není dostatečné (Akpinar a Karabacak, 2017). Zlepšení digitalizace leteckých údajů a zlepšení jejich analýzy si dává za úkol letectví v kontextu strategie Průmysl 4.0, detailně viz Kasturi a kol. (2016). Důležitým nástrojem se v těchto případech stává analýzy velkých dat (BIG DATA analysis).

Je dobře známo, že existuje mnoho faktorů ovlivňujících oblibu letecké dopravy a intenzitu jejího využívání. K tomu, aby letecká doprava úspěšně fungovala, je zapotřebí nejen dopravních prostředků, tedy letadel, ale i jejího zázemí, tj. letiště a letištní plochy. Letiště je běžným hospodářským subjektem, u kterého se jeho úspěšnost hodnotí dle provozních a ekonomických ukazatelů. Mezi základní ukazatele patří výkonové ukazatele, jako je počet pohybů letadel, počet tun přistání, počet odbavených tun nákladu, a v neposlední řadě počet odbavených cestujících. Hledání a nalezení skupin letišť s podobným trendem vývoje počtu odbavených pasažérů může dát lepší porozumění o intenzitě a šíři dopadů různých faktorů ovlivňujících leteckou dopravu, jako je sezónnost, výkyvy počasí, změny klimatu, přírodní či antropogenní katastrofy a podobně. Faktorem, jakým je počet odbavených cestujících, se bude zabývat tato podkapitola. Je založena především na článku Žambochová (2017), který uvádí nový přístup k analýze těchto vlivů a popisuje možnost klasifikace světových letišť z pohledu trendu vývoje počtu odbavených pasažérů. Pro tuto klasifikaci byly zvoleny metody shlukové analýzy.

Důležitostí počtu pasažérů na chod letišť se například zabývá Akamavi a kol. (2015). Změnami cestovatelského chování lidí v době extrémních povětrnostních podmínek se zabývají mimo jiné Lu a kol. (2014) či Hassan a kol. (1999).

Vstupní data

Data byla shromážděna v rámci diplomové práce (Darda, 2014), a to částečně z Civilního ústavu letectví (Service technika de l'aviation civile) se sídlem v Paříži, a od francouzského ministerstva pro ekologii, udržitelný rozvoj a energii (Ministère de l'écologie, du développement trvalý et de l'énergie) se sídlem v Paříži.

Od začátku roku 2000 do konce roku 2013 byly shromážděny údaje týkající se 838 letišť z celého světa. Ve studii nešlo o srovnávání velikosti jednotlivých letišť, ale především o srovnání trendů v počtu odbavených cestujících, proto byly údaje za každé letiště normalizovány. Proměnnými vstupujícími do shlukové analýzy proto byly

přepočítané údaje o počtu pasažérů v jednotlivých měsících. Matice vstupních údajů obsahovala 838 řádků a 168 sloupců, všechny tyto proměnné byly kardinální a spojité. Jednotlivé řádky odpovídají jednotlivým letištím, zatím co sloupce datu. To znamená, že první sloupec odpovídá lednu 2000, zatímco poslední prosinci 2013.

Metodologie

Světová letiště se vzájemně velmi silně odlišují co do velikosti dle různých ukazatelů. Počet odbavených pasažérů v měsíci je jedním z nich. Jak již bylo výše řečeno, ve studii nebyly důležité absolutní hodnoty tohoto ukazatele, ale především chování a vývoj v průběhu celého sledovaného období.

Byly diskutovány dvě varianty standardizace dat, a to standardizace průměrem a směrodatnou odchylkou (normalizace), kdy se pro každý řádek datové matice od dané hodnoty odečte aritmetický průměr všech hodnot v daném řádku, a tento rozdíl se vydělí směrodatnou odchylkou spočtenou ze všech hodnot v daném řádku, tj.

$$z_{ij} = \frac{x_{ij} - \overline{x_i}}{s_i} \quad , \tag{4.8}$$

kde x_{ij} jsou původní hodnoty *i*-tého řádku datové matice, $\overline{x_i}$ je aritmetický průměr hodnot *i*-tého řádku a s_i značí směrodatnou odchylku hodnot *i*-tého řádku. Je zřejmé, že nově vzniklá matice $\mathbf{Z} = (z_{ij})_{i=1,\dots,838}^{j=1,\dots,168}$ má nulové řádkové součty a jednotkové řádkové rozptyly.

Druhou diskutovanou variantou standardizace byla standardizace průměrem, kdy se pro každý řádek datové matice vydělí hodnota každého jeho prvku aritmetickým průměrem všech hodnot daného řádku.

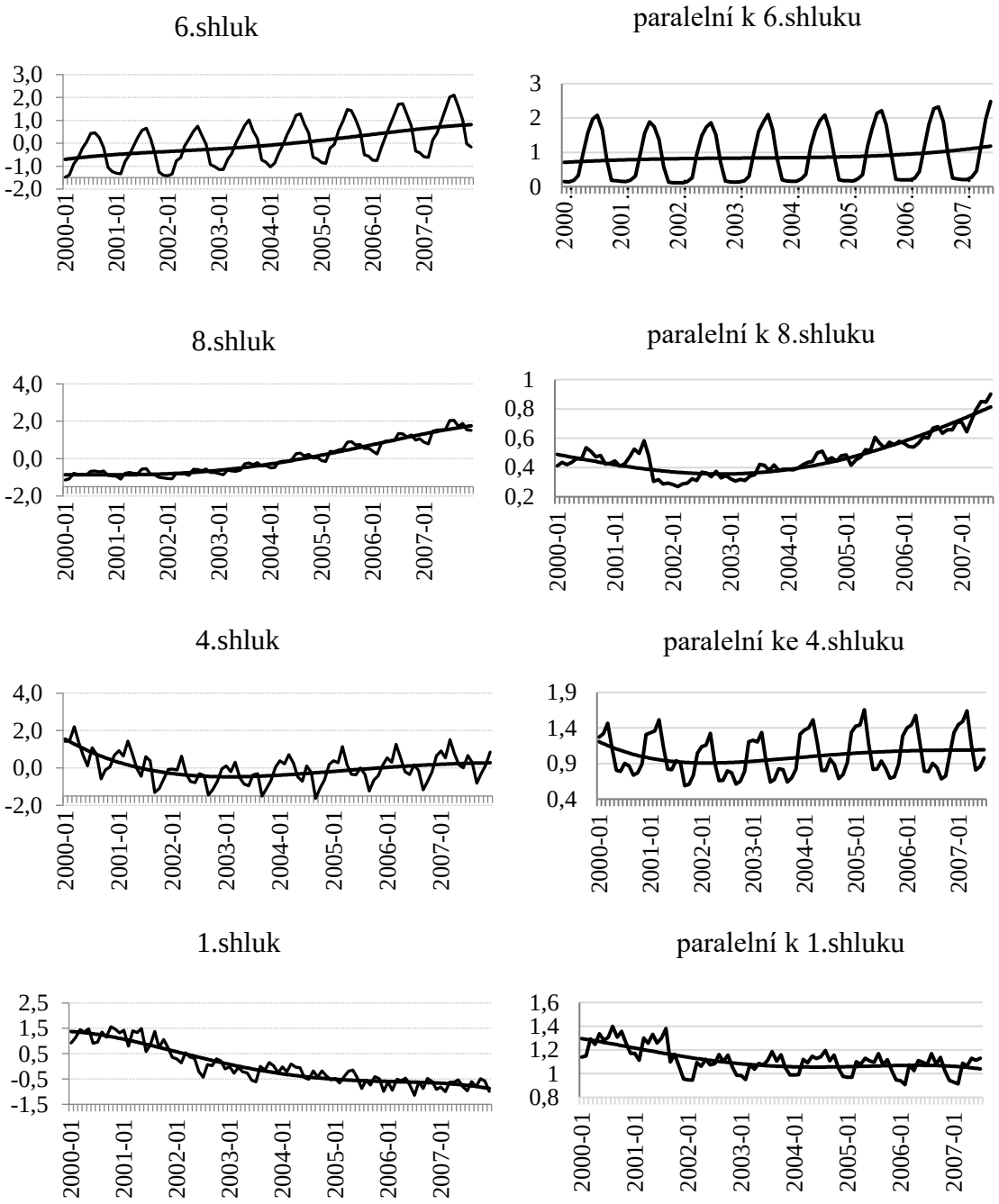
$$z_{ij}^{*} = \frac{x_{ij}}{\overline{x_i}} \quad , \tag{4.9}$$

kde x_{ij} jsou původní hodnoty *i*-tého řádku datové matice, $\overline{x_i}$ je aritmetický průměr hodnot *i*-tého řádku. Je zřejmé, že nově vzniklá matice $\mathbf{Z}^{*} = (z_{ij}^{*})_{i=1,\dots,838}^{j=1,\dots,168}$ má řádkové součty rovny počtu sloupců, tedy 168.

Ve studii bylo pracováno s oběma druhy standardizace. Výsledná shlukování přitom byla dosti podobná, například v případě rozdělení všech objektů do osmi shluků se podařilo spárovat vzniklé shluky z obou případů standardizace. Vzniklé centroidy projevovaly obdobný charakter vývoje v průběhu sledovaného období (viz obr. 10),

a takto spárované shluky se lišily pouze v několika málo zařazených objektech. Je nutné podotknout, že v případě funkcionálních dat není centroid „běžným“ vícerozměrným objektem, který si lze představit jako „těžiště“ v mnohorozměrném prostoru, ale jedná se o funkcionální aproximaci.

Obr. 10: Srovnání chování vybraných centroidů vzniklých metodou *k*-means v případě osmi shluků. Vlevo průběh a trend centroidu v případě standardizace dle (4.8), vpravo průběh a trend centroidu v případě standardizace dle (4.9)



Zdroj: Vlastní zpracování

Pro vyhodnocení však byl nakonec vybrán první typ standardizace, a to z důvodu názornějšího chování centroidu v průběhu sledovaného období a tím pádem snazší interpretace.

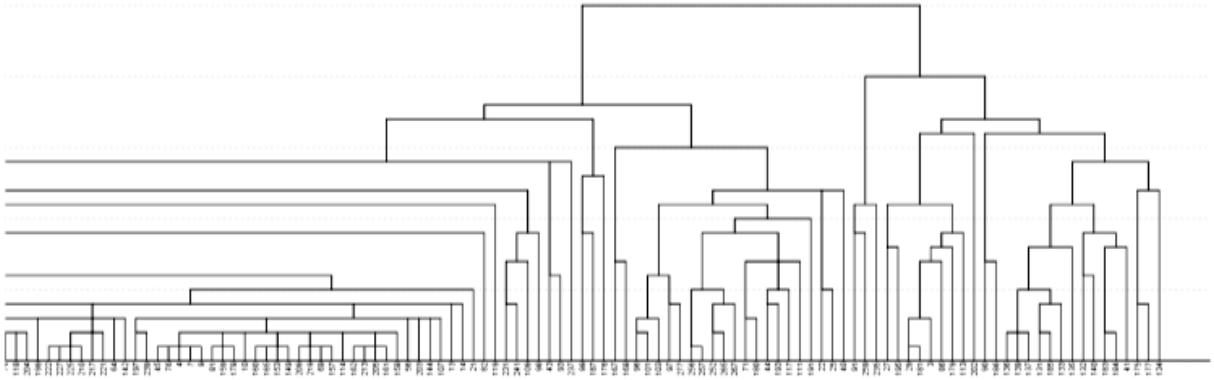
Jak již bylo zmíněno v kapitole 4.1.2, jednou z nevýhod metody *k*-means je nutnost předem zadat požadovaný počet shluků. Naproti tomu u hierarchických metod se určení počtu shluků může provádět až po samotném průběhu shlukovacího algoritmu, například na základě vzniklého dendrogramu. Toto určení může být výsledkem diskuze, počet výsledných shluků není striktně ani algoritmem ani jinými pravidly předepsán. Metoda TwoStep sama optimální počet shluků navrhuje, a to na základě pravidel, která uživatel zadává jako vstupní parametr. Tato metoda ale umožňuje provést shlukovací algoritmus i s předem stanoveným počtem shluků. Uživatel pak má možnost sám posoudit kvalitu výsledného shlukování pomocí SC koeficientu (koeficientu siluety).

Shluková analýza, stejně jako většina vícerozměrných statistických metod, je zatížena subjektivitou zpracovatele. Jednou z nejdůležitějších částí této analýzy je interpretace výsledku. Někdy se může stát, že z matematického hlediska nejlepší výsledek není dobře interpretovatelný. Proto je v takovém případě vhodné přistoupit k dalšímu shlukování, které má sice poněkud horší kvalitu, ale lepší interpretabilitu. Někdy se může stát, že ač z matematického hlediska dobrý výsledek existuje, neexistence jeho rozumné interpretace jej činí nepoužitelným.

V případě dat ve studii Žambochová (2017) byly diskutovány různé metody a různá kritéria kvality výsledného shlukování. Pro sledování kvality shlukování byly použity všechny tři indexy implementované v SPSS, a to BIC, AIC a SK. Za použití jednotlivých ukazatelů se ukázala jako nejlepší volba dvou, nebo tří shluků. Na dva shluky jako optimální počet poukazoval i dendrogram vytvořený hierarchickou metodou (viz obr. 11).

Bohužel se ukázalo, že „optimální“ rozdělení do dvou, nebo tří shluků však rozumnou interpretaci nemělo, proto se muselo přistoupit k variantě přijmutí shlukování suboptimálního, které nabízelo přijatelně interpretovatelné řešení. Z hlediska interpretace se dokonce podařilo nalézt dvě takováto řešení, z nichž každé přineslo něčím zajímavé výsledky. Vyhodnocení a interpretace těchto řešení budou stručně popsány v následujících odstavcích.

Obr. 11: Střed dendrogramu

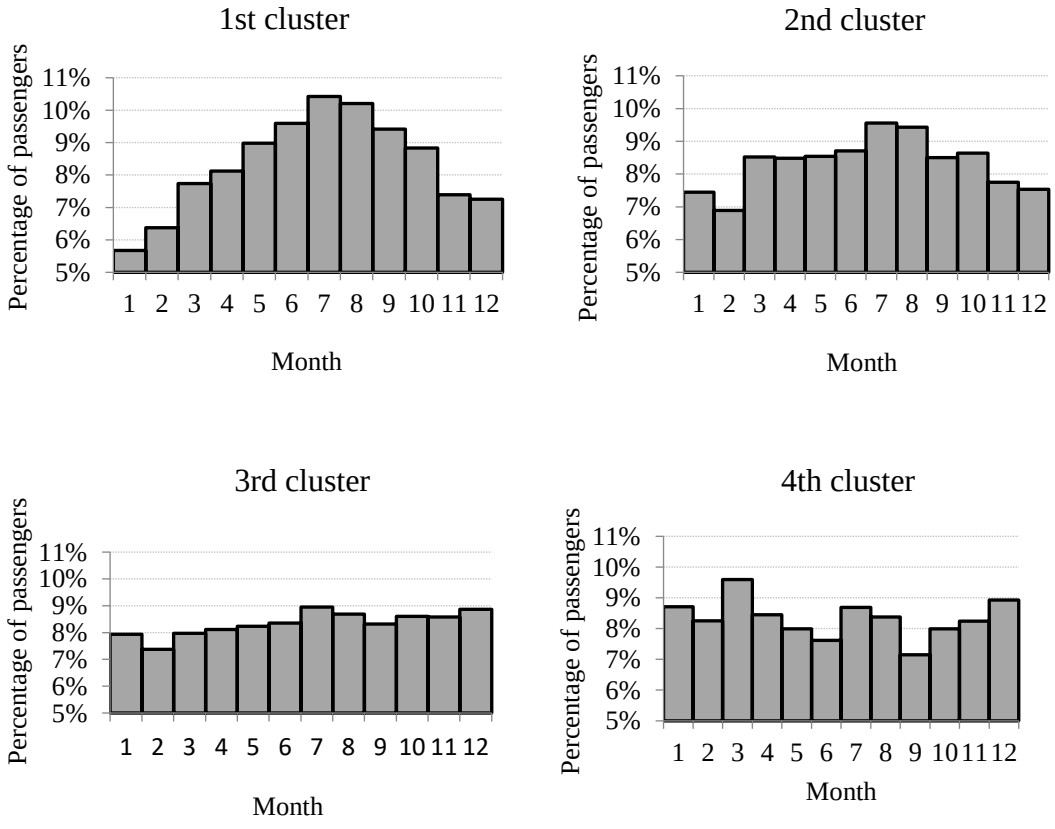


Zdroj: Vlastní zpracování (Žambochová, 2017)

Rozdělení letišť do čtyř shluků – vyhodnocení a interpretace výsledků

V případě volby hodnoty počtu shluků rovné čtyřem se od sebe centroidy jednotlivých shluků, jinými slovy fiktivní letiště jako představitelé jednotlivých shluků, od sebe vzájemně lišily především typem sezónního chování. Toto je znázorněno na obr. 12.

Obr. 12: Proporcionální rozdělení odbavených cestujících v průběhu roku, které je charakteristické pro letiště jednotlivých shluků



Zdroj: Vlastní zpracování (Žambochová, 2017)

Počet odbavených pasažérů na letištích v prvním shluku má typický periodický charakter. Na začátku roku dosahuje počet pasažérů minima, pak roste. V letních měsících červenci a srpnu počet pasažérů dosahuje největších hodnot, poté počet pasažérů klesá, na konci roku klesá k podobným hodnotám jako na začátku roku. Největší zastoupení v této kategorii mají evropská letiště a dále letiště ze Severní Ameriky. Mnohé letecké společnosti operují z těchto letišť tak zvané charterové a sezónní linky. Některá letiště z tohoto shluku odbaví v letních měsících od června do září přes 80 % pasažérů.

Pro letiště v druhém shluku má vývoj počtu odbavených pasažérů v průběhu roku opět periodický charakter, ale s menší amplitudou než u letišť z prvního shluku. Představitelé tohoto shluku se projevují vyrovnanějším provozem. Opět platí, že nejvyšší obrat odbavených pasažérů představují měsíce červenec a srpen. V tomto shluku jsou více méně zastoupeny tři kontinenty rovnoměrně, a to Severní Amerika, Evropa a Asie.

Typické letiště třetího shluku odbavuje v průběhu roku cestující bez velkých výkyvů. Zástupci tohoto shluku jsou rozmístěni po celém světě. Mezi letiště tohoto shluku patří mimo jiné většina tranzitních letišť, ale také mnoho letišť z ostrovních států.

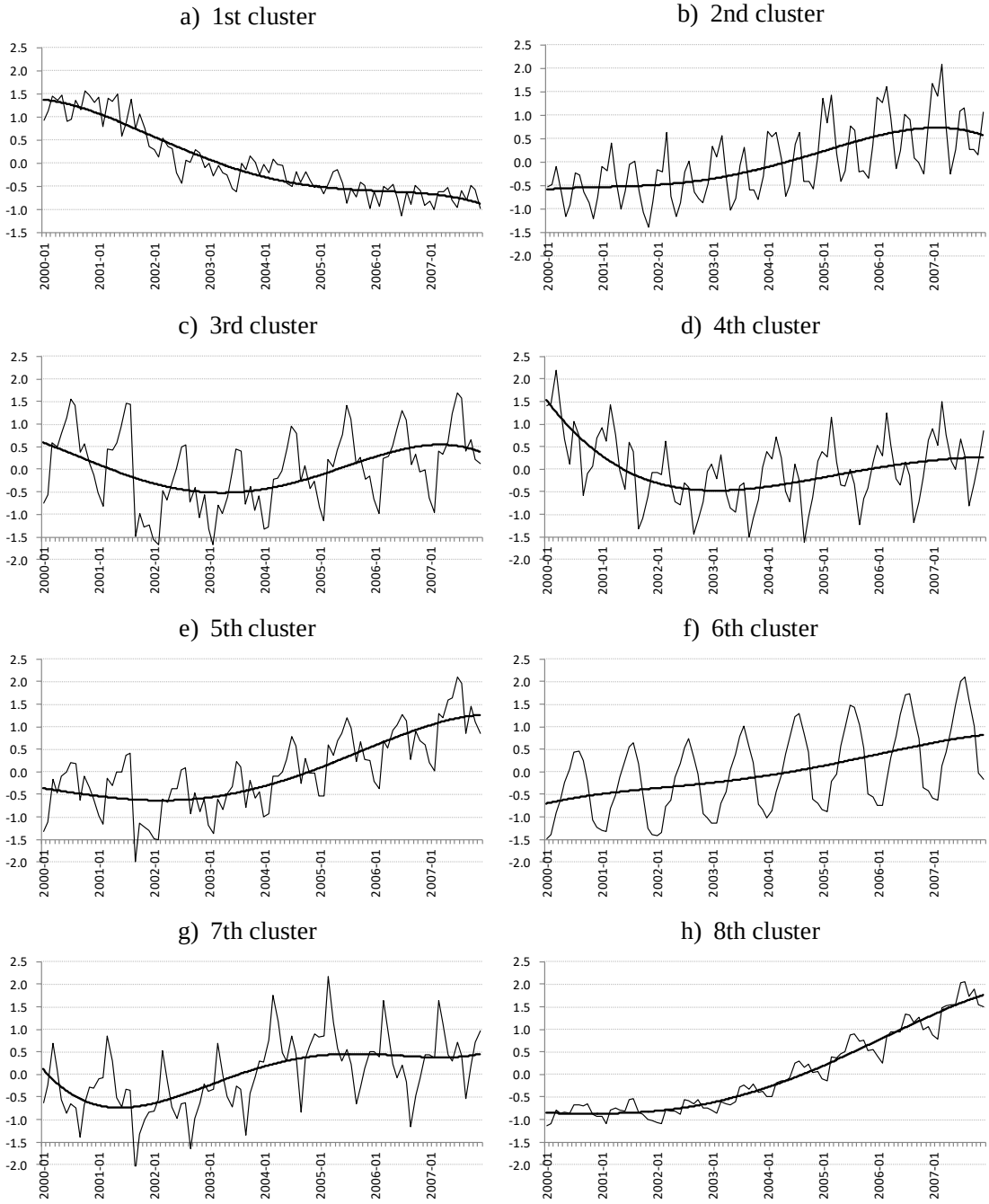
Roční průběh počtu odbavených pasažérů v případě charakteristického letiště čtvrtého shluku není tak jednoznačný, jak tomu bylo v předchozích třech případech. Zdá se, že tento shluk je tvořen dvěma skupinami. Po prozkoumání geografické polohy letišť tohoto shluku je zřejmé, že se jedná o letiště na jižní polokouli. Dá se předpokládat, že tato letiště mají největší obraty v době léta na jižní polokouli, tedy v měsících na přelomu roku. Velkých letišť je ale na jižní polokouli málo, proto pokles uprostřed roku je vyrovnán velkým výkonem letišť z druhé skupiny letišť tohoto shluku, a to letišť v teplých oblastech, která jsou cílem turistů ze severní polokoule.

Rozdělení letišť do osmi shluků – vyhodnocení a interpretace výsledků

Interpretačně velmi odlišné výsledky jsme obdrželi v případě rozdělení do osmi shluků. V tomto případě se průběh počtu odbavených pasažérů ve sledovaném období u charakteristických zástupců jednotlivých shluků odlišoval nejen z hlediska cykličnosti, ale též trendem a ojedinělými výkyvy z dlouhodobého průběhu (viz obr. 13). Sezónnost je nejméně podstatná pro 1. shluk, pro ostatní shluky je velmi výrazná. Po bližší analýze

těchto výkyvů byla odhalena spojitost s různými zásadními událostmi, a to jak přírodními, tak i antropogenními katastrofami.

Obr. 13: Časové řady normalizovaných měsíčních počtů odbavených pasažérů na charakteristických letištích jednotlivých shluků



Zdroj: Vlastní zpracování (Žambochová, 2017)

Pro letiště umístěná v prvním shluku je charakteristický pokles po září 2001. Dalším znakem je silný pokles na konci roku 2003. Tento pokles se začal zmírňovat až

koncem roku 2005. Ze znalosti světových událostí lze usoudit, že tato letiště byla jednak ovlivněna teroristickým útokem 1. září 2001, ale ještě silněji obdobím SARS, které propuklo koncem roku 2002. Důsledky epidemie SARS v letecké dopravě jsou popsány například v Loh a Elaine (2006).

Letiště v druhém shluku jsou charakteristická jednak výraznou sezónností, ale navíc se u nich projevuje silně rostoucí trend odbavených pasažérů ve sledovaném období. Události 11. 9. 2001 jsou mírně znatelné.

U letišť ve třetím shluku je patrný velmi silný propad od září 2001. Stav z počátku roku 2000 byl po tomto propadu dosažen až kolem roku 2007.

U letišť čtvrtého shluku se projevil menší pokles v počtu odbavených pasažérů v září 2001, ale výraznější je pokles koncem roku 2003. Oproti letištím z druhého shluku se těmto letištím nepodařilo vyrovnat stav z počátku roku 2000 ani do konce sledované doby.

Pro letiště pátého shluku je charakteristická výrazná sezónnost, ale také velmi silný pokles počtu odbavených pasažérů po září 2001. Od té doby je zřejmý rostoucí trend.

U letišť šestého shluku je patrný mírný rostoucí trend v průběhu celého období. Navíc se zde projevuje silné sezónní chování, nadprůměrné množství odbavených pasažérů v době prázdnin a naopak silně podprůměrné množství odbavených pasažérů kolem přelomu roku. Není tu zřejmý vliv žádné mimořádné události.

U zástupců sedmého shluku je charakteristický znatelný propad v období kolem září 2001 a pak pomalý návrat do původního stavu.

Pro letiště zařazená do osmého shluku je charakteristický výrazně rostoucí trend a minimální sezónnost. Popis tohoto shluku je v souladu například s popisem vývoje letecké dopravy v Číně, jak je popsán například ve Wang a kol. (2014).

Závěry ze studie

V posledních letech se řada publikací zabývá praktickým využitím shlukové analýzy dat sebraných na letištích. Například Mangortey a kol. (2020) polemizují s Žambochovou (2017) v ohledu praktického využití klasifikace letišť v celosvětovém měřítku. Tito autoři navrhují metodu shlukování denního provozu letišť použít v regionálním kontextu jako

nástroj pro provozovatele letišť, analytiky a výzkumné pracovníky FAA ke zlepšení provozu na letištích, a to díky identifikaci klíčových charakteristik a trendů a možnosti předpovědi následného vývoje a přijmutí vhodných opatření. Autoři dalších článků, např. Malighetti a kol. (2009) či Ayyildiz a Yalcin (2018), uvažují a následně interpretují pouze jednu hodnotu počtu shluků, která jim podle jimi zvoleného kritéria vyšla jako optimální. Žambochová, (2017) však diskutuje několik různých hodnot počtu shluků, na jejichž základě vznikne rozdělení do shluků dobré a prakticky přijatelné kvality. Navrhnutá řešení mají podstatně lepší interpretovatelnost a tím i praktické využití.

Žambochová (2017) ve studii navíc poukázala na nevhodnost striktního určení počtu shluků dle různých informačních kritérií optimality. Na rozdíl od toho je v literatuře poukázáno na mnoho takovýchto kritérií, jejichž autoři si staví za cíl co možná nejvyšší kvalitu výsledného shlukování na úkor dobré interpretovatelnosti.

Jsem přesvědčena, že z hlediska praktického využití shlukové analýzy je nutno brát v úvahu nejen matematickou stránku problému, ale především stránku interpretační. Jistým ústupkem od formální matematické „optimality“ je možno odhalit nové skryté závislosti v datech. Samozřejmostí by však mělo být dodržení přípustné kvality. Sklouznutí pod ni by nemělo být v žádném případě akceptováno.

6 Závěr

Předložená habilitační práce si klade za cíl rozšířit současný stav poznání v oblasti segmentace a klasifikace při analýze dat v oblasti společenských věd, o zhodnocení současných přístupů a o návrh vlastních modifikací. Obsahem práce je sjednocený náhled do problematiky, od matematického popisu vybraných metod používaných jak pro segmentaci, tak i pro klasifikaci, až po ukázky konkrétních aplikací v různých oblastech ekonomiky. Důraz je přitom kladen na interpretovatelnost výsledků provedených analýz s vědomím, že kvalitní interpretace získaných výsledků je základním nástrojem podpory rozhodování a zvýšení jeho efektivity.

Práce vychází z dlouhodobého zaměření autorky na tématiku segmentace a klasifikace, jmenovitě na rozhodovací stromy a shlukovou analýzu, a to jak po stránce statistické a algoritmické teorie, tak, a to především, po stránce využití a aplikace těchto metod při řešení různých praktických problémů, převážně z oblasti marketingu a řízení vztahů se zákazníky, nejenom v oblasti školství, ale i v oblasti obchodu a podnikání.

První skupina aplikací zahrnutých v této habilitační práci je zaměřena na témata z oblasti školství, jako například marketingové aktivity na vysokých školách, kdy zákazníkem je student, či zjišťování potřeb znalostí a schopností, které by měli získat studenti na vysokých školách.

Ačkoliv veřejné vysoké školy nejsou komerčními organizacemi, přesto je třeba používat různé, pokud možno optimální, marketingové aktivity. Marketingový koncept je orientován na zákazníka, jímž je v tomto případě student. Marketingový management pomáhá vytvářet konkurenční vzdělávací program a prostřednictvím tohoto programu může komunikovat se svými potenciálními zákazníky, tj. studenty. V zahraničí jsou marketingové průzkumy a koncepty běžnější než v České republice. Žambochová (2013b) popisuje jeden z průzkumů provedený mezi potencionálními studenty, který je zaměřen především na spádovou oblast Fakulty sociálně ekonomické z Ústeckého regionu, který patří mezi regiony v České republice s nejvyšší mírou nezaměstnanosti a současně nízkou úrovní vzdělání. Pomocí klasifikace dat bylo navrženo doporučení, jak oslovit studenty a povzbudit je v ochotě studovat na fakultě.

V rámci dvouletého grantového projektu interní grantové agentury UJEP v Ústí nad Labem (IGA) byla provedena rozsáhlá studie, jejímž cílem bylo poskytnout získané poznatky a podněty do stále aktuální diskuze o reformě českých vysokých škol. Výsledky studie byly diskutovány v několika publikovaných výstupech. Například publikace Žambochová a Tišlerová (2011), Vomáčková, Žambochová a Tišlerová (2011) či Žambochová (2012a) se snažily uplatnit ekonomický a marketingový přístup k interpretaci dat získaných tímto primárním výzkumem. V této souvislosti provedly hloubkovou analýzu zájmu oslovených zahraničních studentů o případné studium v České republice. Na základě analýz získaných dat, a to z velké části analýz využívajících metod a postupů z oblasti segmentace a klasifikace, zde byla s poměrně vysokou mírou obecnosti formulována marketingová doporučení.

Moderní doba vytváří stále větší tlak na individuální vzdělávání člověka, a tím klade stále větší důraz na co nejefektivnější získávání vědomostí. Pro poskytovatele těchto informací je důležité znát potřebu skladby požadovaných znalostí a způsobů předávání znalostí.

Žambochová (2012b) předkládá na základě segmentace a klasifikace sebraných dat vzniklá doporučení týkající se preferovaných způsobů a formy výuky, ale i zdrojů

informací na její podporu. Na základě dotazníkového šetření, kterého se zúčastnilo 1173 respondentů, byla provedena segmentace pomocí shlukové analýzy. K vyhodnocení a interpretaci výsledných segmentů byly využity rozhodovací stromy. Vývoj v následujících letech potvrdil některé ze závěrů studie, a to především výraznou a neustále vzrůstající populárnost internetu, oblibu učebnic jako zdroje informací, preferenci přímé výuky u mladých žen humanitního zaměření a naopak vyhýbání se přímé výuce u mužů technického a uměleckého zaměření.

Tvorbou nových osnov vyučované látky i novými přístupy k výuce se zabývá mnoho týmů pedagogů, a to na různých úrovních – počínaje vedením škol, přes státní úřady až po orgány Evropské unie. Žambochová a Kulhanová (2017) provedly dotazníkové šetření mezi absolventy různých úrovní škol v Ústeckém kraji. Na základě segmentace a klasifikace dat získaných z tohoto šetření sestavily doporučení týkající se potřeb získaných kompetencí absolventů všech úrovní škol, i když prioritou byla především vysokoškolská úroveň studia. Navržená doporučení pomáhají při tvorbě osnov a formy výuky. Autorky srovnáním s jinými studiemi zjistily, že situace v České republice není příliš odlišná od jiných zemí, byla však odhalena i mnohá specifika pro Českou republiku. Aesaert a kol. (2015) na základě rozsáhlého šetření v rámci základních škol v Belgii poukázali na rozdílný názor žáků základních škol a jejich rodičů na získávání potřebných kompetencí v oblasti informačních technologií. Obdobně i Žambochová a Kulhanová (2017) odhalily rozdíly v pohledu na důležitost těchto kompetencí mezi absolventy různých stupňů škol. Mezi českými absolventy se také potvrdil kritický přístup k výuce cizích jazyků, při níž se klade důraz především na gramatiku na úkor komunikačních schopností, tak jak popisují situaci ve Španělsku Martínez a kol. (2015). Poněkud smutným specifikem českého průzkumu bylo zjištění, že ne všichni, kteří cítí nedostatečné schopnosti v těchto oblastech, cítí potřebu dalšího vzdělávání mimo vlastní systém školství. Toto se týká především absolventů středního všeobecného vzdělávání a bakalářských humanitních, právnických oborů.

Druhá skupina aplikací zahrnutých do této habilitační práce je úzce zaměřena na obchod. Segmentace a klasifikace se stává nezbytnou součástí při podpoře rozhodování v marketingu. Jednou ze základních součástí marketingového mixu je cena a její určení je jedním z nejtěžších úkolů marketingu. Cena by na jedné straně měla odrážet kvalitu nabízeného produktu, na druhé straně spotřebitelé na cenu reagují svým nákupním

chováním. Návrh na tvorbu ceny ojetých vozidel na základě klasifikace prodaných vozů nabízí například Žambochová (2007).

Dalšími součástmi marketingového mixu je distribuce a propagace. Při plánování podnikatelských aktivit je dobré provést segmentaci trhu. Na jejím základě je pak jednodušší plánování reklamy a rozhodnutí o distribuci (kamenné prodejny, online prodej, atd.) Pomocí dobře provedené segmentace je možno vybrat atraktivní segmenty zákazníků, na které je následně dobré se efektivně zaměřit pro dosažení dostatečného zisku. Jednu segmentaci trhu zaměřenou na časoprostorové chování spotřebitelů provedli ve své studii Grossmanová, Kita a Žambochová (2016). Autoři vycházejí z myšlenky, že pro úspěšnou lokalizaci prodejního místa musí distributor umět posoudit nákupní a prostorové chování dané zájmové skupiny. Kritéria pro umístění prodejního místa souvisí s prostorovou dimenzí geomarketingu. Provedená analýza dat, a to především segmentace a klasifikace, byla součástí průzkumu vývoje maloobchodní sítě s cílem vytvořit časoprostorový model spotřebitelského chování občanů Bratislavy. Segmentace spotřebitelů ze Slovenské republiky, kterou popsali Kita a kol. (2020a), měla za cíl specifikaci hlavních znaků tak zvaného odpovědného spotřebitele, tedy člověka, který využívá produkty a služby způsobem, jenž je nejméně škodlivý pro životní prostředí a zároveň člověka, který dodržuje zdravý životní styl. Závěry studie jsou využitelné výrobci a obchodníky k tvorbě vhodných modelů marketingové komunikace, přičemž jedním ze zásadních závěrů studie se ukázala potřeba výrazného zlepšení dostupnosti bio potravin tak, aby se mohlo zlepšit proekologické chování spotřebitelů na Slovensku. Podrobněji jsou výsledky této studie popsány odděleně v pohledu na zdravé stravování a na udržitelné chování v publikacích Kita a kol. (2020b, 2020c).

Třetí skupina aplikací zahrnutých do této habilitační práce se týká provozu letišť, jako jedné ze zásadních oblastí silně se vyvíjejícího leteckého odvětví. Stále zrychlující se tempo vývoje v celé společnosti vede svět k digitální transformaci. Průmysl 4.0, jak se tato digitální transformace nazývá, přináší řadu výhod, musí však čelit nově vznikajícím rizikům a výzvám spojeným s organizačními a lidskými faktory. Se čtvrtou průmyslovou revolucí musí digitalizující se podniky vytvořit inovační systémy, které jim umožní spolupracovat se všemi zúčastněnými stranami. Společnosti působící v odvětví letectví, cestování a cestovního ruchu jsou nuceny optimalizovat své zkušenosti se zákazníky shromažďováním údajů a neustálým získáváním znalostí (Sahin a kol., 2019). Využití analýzy dat v sektoru leteckých společností je však stále v počáteční fázi,

a schopnost těchto studií generovat znalosti není dostatečné (Akpinar a Karabacak, 2017). Počet odbavených cestujících je jedním z nejdůležitějších faktorů, který ovlivňuje chod letišť, jak uvádí Akamavi a kol. (2015). Segmentací letišť na základě vývoje měsíčních údajů o počtu odbavených cestujících se zabývala pátá podkapitola založená na článku Žambochová (2017). Studie mimo jiné přispěla k odborné diskuzi o možnostech využití segmentace a správné interpretace jejích výsledků v leteckém odvětví, viz Mangortey a kol. (2020).

Výše uvedené studie jsou reálnou ukázkou toho, jak je statistická analýza reálných dat a speciálně segmentace a klasifikace užitečná pro praxi, a jak data a jejich analýza mohou být přínosná v rozhodovacích procesech.

7 Prohlášení

Ráda bych na tomto místě poděkovala spoluautorům prací, které jsem prezentovala jako součást habilitační práce. Čestně prohlašuji, že ve všech případech byl můj podíl na přípravě, realizaci a interpretaci výsledků větší nebo stejný jako všech ostatních spoluautorů.

8 Použitá literatura

Aesaert, K., Van Braak, J., Van Nijlen, D., Vanderlinde, R. (2015). Primary school pupils' ICT competences: Extensive model and scale development. *Computers and Education*, 81, 326-344.

Akamavi, R. K., Elsayed M., Pellmann, K. a kol. (2015). Key determinants of passenger loyalty in the low-cost airline business. *Tourism management*, 46, 528-545.

Akpinar, M. T., Karabacak, M. E. (2017). Data mining applications in civil aviation sector : State-of-art review. *CEUR Workshop Proc*. 1852, 18-25.

Antoch, J., (1988) Klasifikace a regresní stromy. *Robust 1988*, [cit. 10. 9. 2020]. Dostupné z: https://www.statspol.cz/robust/1988_antoch88.pdf

Ayyildiz, E., ve Yalcin, S. (2018). Analysis of airports using clustering methods: case study in Turkey. *Pressacademia*, *5*(3), 194‑205.

Bassi, F. (2007). Latent class factor models for market segmentation: an application to pharmaceuticals. *Statistical Methods & Applications*, 16 (2), 279–287.

Berikov, V., Litvinenko, A. *Methods for statistical data analysis with decision trees*. [cit. 16. 5. 2009]. Dostupné z: http://www.math.nsc.ru/AP/datamine/eng/decisiontree.htm.

Berry, M., Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management*. John Wiley & Sons, New York, 643 s.

Breiman, L., Friedman, J., Stone, Ch. J., . Olshen R.A (1984). *Classification and Regression Trees*, Taylor & Francis, 368 s.

Carmichael, G., Chen, Y.W., Luo, C. (2018). Data-driven segmentation of consumers' purchase behaviour in the retail industry. In: 2018 *4th International conference on information management, ICIM* 2018, 215-219.

Darda, P. (2014). *Ekonomický význam cestujících pro letiště*. Diplomová práce, Univerzita J. E. Purkyně, FSE, Ústí nad Labem.

Dvouletý, O., Longo, M. C., Blažková, I., Lukeš, M., Andera, M. (2018). Are publicly funded Czech incubators effective? The comparison of performance of supported and non-supported firms. European Journal of Innovation Management, 21(4), 543-563.

Dvouletý, O., Orel, M. (2020). Determinants of solo and employer entrepreneurship in Visegrad countries: findings from the Czech Republic, Hungary, Poland and Slovakia. *Journal of Enterprising Communities-People and Places in the Global Economy*, 14(3), 447-464.

Everit, B. S., Landau, S., Leese, M. (2001). *Cluster analysi*. 4. vydání, Hodder Arnold, London, 237 s.

Faber, V. (1994). Clustering and the continuous *k*-means algorithm. *Los Alamos Science*, 22, 138–144.

Finardi, S., Mazouch, P., Fisher, J. (2012). Odhad míry návratnosti investic do vysokoškolského vzdělání podle oborů, pohlaví a regionů. *Politická ekonomie*, 60(5), 563-589.

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*, 21(3), 768 – 769.

Frank, I. E., Todeschini, R. (1994). *The data analysis handbook*. Elsevier Science Ltd, Amsterdam, 365 s.

Gilboa, S. (2009). A segmentation study of Israeli mall customers. *Journal of retailing and consumer services*, 16(2), 135-144.

Guha, S., Rastogi, R., Shim, K. (2001). CURE: An efficient clustering algorithm for large databases. *Information Systems*, 26 (1), 35–58.

Hartigan, J. A., Wong, M. A. (1979). A *k*-means clustering algorithm. *Applied statistics*, 28(1), 100–108.

Holmbom, A. H., Eklund, T., Back, B. (2011). Customer portfolio analysis using the SOM. *International Journal of business information systems*, 8 (4), 396-412.

IATA, *The Impact of September* 11 2001 *on Aviation*. [cit. 8. 10. 2016]. Dostupné z: http://www.iata.org/pressroom/documents/impact-9-11-aviation.pdf.

James, G. M. a Sugar, C. A. (2003). Clustering for sparsely sampled functional data. Journal of the American Statistical Association, 98(462), 397-408.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, CH. D., Silverman, R., WU, A. Y. (2002). An efficient *k*-means clustering algorithm. *Analysis and Implementation. IEEE Transactions on pattern analysis and machina inteligence*, 24(7).

Kasturi, E., Devi S., Kiran, S., Manivannan, S. (2016). Airline route profitability analysis and optimization using BIG DATA analyticson aviation data sets under heuristic techniques. *Procedia computer science,* 87, 86-92.

Karypis, G., Han, E-H., Kumar, V. (1999). CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, *IEEE Computer*, 32 (8), 68-75, ISSN: 0018-9162

Kita, P., Žambochová, M., Mazalán, P., Strelinger, J., Kitová M. V. (2020a) Consumer behaviour of Slovak households in the sphere of organic food in the context of sustainable consumption. *Central European Business Review* (v tisku)

Kita P., Žambochová M., Kita J. (2020b). *Spotrebiteľské správanie slovenských domácností v oblasti vybraných druhov potravín v kontexte spoločensky zodpovednej spotreby*. Univerzita Komenského v Bratislave, 196 s.

Kita P. a kol. (2020c). *Model marketingovej komunikácie na zdravie orientované nákupné správanie spotrebiteľov so zreteľom na postoje k spotrebe zdravých potravín*. GUPRESS s.r.o., Bratislava, 232 s.

Kogan, J., Nicholas, Ch., K.; Teboulle, M. (2006). *Grouping multidimensional data: recent advances in clustering*. Springer - Verlag Berlin Heidelberg, 268 s.

Kohavi, R. (2001). Data mining and visualization. In: *Sixth annual symposium on frontiers of engineering, National Academy Press, D. C*., 30-40.

Liao, S., Chen, Y., Hsieh, H. (2011). Mining customer knowledge for direct selling and marketing. *Expert systems with applications*, 38 (5), 6059‑6069.

Liu, Y., Kiang, M., Brusco, M. (2012). A unified framework for market segmentation and its applications. *Expert systems with applications*, 39 (11), 10292-10302.

Loh, W. Y., Shih, Y.-S. (1997). Split selection methods for classification trees, *Statistica Sinica*, 7, 815-840.

Loh W. Y, Vanichsetakul N. (1988) Tree-Structured Classification via Generalized Discriminant Analysis. *Journal of the American Statistical Association*, 83(403), 715-728.

Loh, E. (2006). The impact of SARS on the performance and risk profile of airline stoce. *International journal of transport economics*, 33(3), 401-422.

Lu, Q.-Ch., Zhang, J., Peng, Z.-R. et al. (2014). Inter-city travel behaviour adaptation to extreme weather events. *Journal of transport geography*, 41, 148-153.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of* 5*th Berkeley symposium on mathematical statistics and probability*, Berkeley, University of California Press, 281-297.

Maheswari, R. U., Mahesan, S. S., Tamilarasan, A., Subramani K. (2014). Role of data mining in CRM, *International journal of engineering research*, 3(2), 75-78.

Malighetti, P., Paleari, S., Redondi, R. (2009). Airport classification and functionality within the European network. *Problems & Perspectives in Management*, 7(1), 183-196.

Mangortey, E., Puranik, T. G., Pinon, O., Mavris, D. (2020). Classification, analysis, and prediction of the daily operations of airports using machine learning. *AIAA Science and technology forum* (*AIAA Scitech*).

Matějů, P. a kol. (2009). *Bílá kniha terciárního vzdělávání*. MŠMT ČR.

McCarty, J. A., Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, 60(6), 656‑662.

Olszak, C. M. (2016). Toward better understanding and use of business intelligence in organizations. *Information systems management* 33(2), 105-123.

Peker, S., Kocyigit, A., Eren, P.E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing intelligence & planning*, 35(4), 544-559.

Prudký, L., Pabian, P., Šima, K. (2010). *České vysoké školství: Na cestě od elitního k univerzálnímu vzdělávání* 1989-2009. Grada publishing. a.s., 168 s.

Ramsay J., Silverman B. W. (2005). *Functional Data Analysis*. Springer-Verlag New York, 2. vydání, 428 s.

Rud, O. P. (2001). *Data mining – Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM)*, Praha: Computer Press, 329.

Řezanková, H., Húsek, D., Snášel, V. (2007). *Shluková analýza dat*. Professional Publishing, Praha, 218 s.

Řezanková, H., Húsek , D., Snášel, V. (2008). Clusters number determination and statistical software packages. *DEXA 2008: 19th International conference on database and expert systems applications*, 549-553.

Řezanková, H., Löster, T. (2013). Shluková analýza domácností charakterizovaných kategoriálními ukazateli. *E+M. Ekonomie a Management*, 16(3), 139-147.

Savický, P., Klaschka, J., Antoch J. (2000). Optimální klasifikační stromy. *Sborník ROBUST 2000*, 267-283.

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing*, 21, 3-8.

Stevens. S. S. (1946). On the Theory of Scales of Measurement. *Science* 103(2684), 677-680.

Suárez-Vega, R., Santos-Peñate, D. R., Dorta-González, P., Rodríguez-Díaz, M. (2011), A multi-criteria GIS based procedure to solve a network competitive location problem. Applied Geography, 31(1), 282–291.

Šveda, M., Križan, F. (2012). The Manifestation of Commercial Suburbanization in the Selected Sectors of Economy in the Hinterland of Bratislava. *Journal of Economics*, 60(5), 460–481.

Tarpey, T. a Kinateder, K. K. J. (2003). Clustering functional data. *Journal of classification*, 20(1), 93–114.

Taušer, J., Žamberský, P. (2012). Kvalitativní kritéria ve financování veřejných vysokých škol a jejich dopad na Vysokou školu ekonomickou v Praze. *Acta Oeconomica Pragensia*, 4(20), 74–88.

Timofeev R. (2004). *Classification and regression trees (CART) Theory and applications,* CASE - Center of Applied Statistics and Economics, Humboldt University, Berlin, 39 s.

Tsai, C., Chiu, C. (2004). A purchase-based market segmentation methodology. *Expert systems with applications*, 27(2), 265-276.

Tsai, Ch.-F., Hu, Y.-H., Lu, Y.-H. (2015). Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. *Expert Systems*, 32 (1), 65-76.

Vomáčková H., Žambochová M., Tišlerová K. (2011). *Současné ekonomické křižovatky českých vysokých škol*. Univerzita Jana Evangelisty Purkyně v Ústí nad Labem, 158 s.

Wang, J., Mo, H., Wang, F. (2014). Evolution of air transport network of China 1930-2012. *Journal of transport geography*, 40(SI), 145-158.

Wilkinson, L. (1992). Tree structured data analysis: AID, CHAID and CART. *Sawtooth/ SYSTAT Joint software konference*, Sun Valley, ID.

Wilkinson, S. M., Dunn, S., Ma, S. (2012). The vulnerability of the European air traffic network to spatial hazards. *Natural hazards*, 60(3), 1027-1036.

You, Z., Si, Y.-W., Zhang, D., Zeng, X. X., Leung, S. C. H., Li T. (2015). A decision-making framework for precision marketing, *Expert systems with applications*, 42 (7), 3357-3367.

Zhang, T., Ramakrishnan, R., Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record*, 25(2), 103–114.

Zhang, T., Ramakrishnan, R., Livny, M (1997). BIRCH: A new data clustering algorithms and its applications. *Journal of data mining and knowledge discovery*, 1(2),141–182.

Žambochová, M. (2007). Odhad cen ojetých vozů pomocí rozhodovacích stromů, *Mezinárodní statisticko-ekonomické dny na VŠE v Praze* [CD-ROM], Praha, 1-6.

Žambochová, M. (2009a). Inicializační rozdělení do shluků a jeho vliv na konečné shlukování v metodách k-průměrů. *Sborník prací účastníků vědeckého semináře doktorského studia FIS VŠE Praha*, 243-250.

Žambochová, M. (2009b). Odlehlé objekty a shlukovací algoritmy. *Mezinárodní statisticko-ekonomické dny na VŠE v Praze* [CD-ROM], Praha, s. 1-6.

Žambochová, M. (2010a). Shlukování v souborech s odlehlými objekty pomocí metod *k*-průměrů. *Informační bulletin České statistické společnosti*, 22(3), 123-130.

Žambochová, M. (2010b). *Shluková analýza rozsáhlých souborů dat: nové postupy založené na metodě k-průměrů*. Disertační práce, Vysoká škola ekonomická Praha, Fakulta statistiky a informatiky.

Žambochová, M., Tišlerová, K. (2011). Classification of individuals: willigness to start their own business based on franchise systém. *Proceedings – Aplimat* 2011, [CD-ROM], Bratislava: Slovak University of Technology, 1647-1655.

Žambochová, M. (2013a). A statistical analysis of the support for small business as a solution for unemployment in the region Usti nad Labem. *Proceedings of the* 11*th International Conference LEF* 2013, Liberec: TU Liberec, 642-651.

Žambochová, M. (2013b). Statistical Analysis of the marketing activities of the university. 10*th International Conference on Efficiency and Responsibility in Education*, Praha, 664-670.

Žambochová, M., Kulhanová, A. (2017). Classification of individuals according to their opinions on acquiring necessary competencies within their studies. *ACC Journal*, 22(3), 31-43.

9 Seznam příloh

Příloha 1

Žambochová, M. (2008). Data mining methods with trees. *E+M Ekonomie a Management*, 11(1), 126-131, ISSN 1212-3609.

Příloha 2

Žambochová, M. (2012a). Typology of foreign students interested in studying at Czech universities. *E+M Ekonomie a Management*, 15(2), 141-154, ISSN 1212-3609.

Příloha 3

Žambochová, M. (2012b). Classification in terms of students' preferences for information sources. 9*th International Conference on Efficiency and Responsibility in Education*, Praha, 612-620, ISBN 978-80-213-2289-9.

Příloha 4

Hlaváček, P., Žambochová, M. and Sivček, T. (2015). The Influence of the Institutions on Entrepreneurship Development: Public Support and Perception of Entrepreneurship Development in the Czech Republic. *Amfiteatru Economic*, 17(38), 408-421, ISSN: 1582 – 9146.

Příloha 5

Žambochová, M. (2014). The modification of the k-means method for creating non-convex clusters. 8*th International Days of Statistics and Economics*, Praha, 1722-1730, ISBN 978-80-87990-02-5.

Příloha 6

Grossmanová, M., Kita, P. and Žambochová, M. (2016). Segmentation of consumers in the context of their space behaviour: Case study of Bratislava. *Prague Economic Papers*, 25(2), 189-202, ISSN: 1210-0455, DOI: 10.18267/j.pep.554.

Příloha 7

Žambochová, M. (2017). Cluster Analysis of World's Airports on the Basis of Number of Passengers Handled (Case Study Examining the Impact of Significant Events). *Statistika – Statistics and Economy Journal*, 97(1), 74-88, ISSN: 0322-788X.