TARTU ÜLIKOOL
Majandusteaduskond
Rahvamajanduse instituut
Majanduse modelleerimise õppetool

Kaarel Aus

KLASTERANALÜÜSI KASUTAMINE LOOMULIKE KLIENDISEGMENTIDE TUVASTAMISEKS EESTI LEIBKONDADE HULGAS

Magistritöö sotsiaalteaduse magistri kraadi taotlemiseks majandusteaduses

Juhendaja: dotsent Kaia Philips

Tartu 2014

Soovitan suunata kaitsmisele ……..

(juhendaja allkiri)

Kaitsmisele lubatud “ “ ..... 2014. a.

Majanduse modelleerimise õppetooli juhataja Jaan Masso .....……

(allkiri)

Olen koostanud töö iseseisvalt. Kõik töö koostamisel kasutatud teiste autorite tööd, põhimõttelised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

……..

(töö autori allkiri)

SISUKORD

SISSEJUHATUS

Informatsiooni ja kapitali üha vabam ning kiirem liikumine on loonud keskkonna, kus konkureerivate ettevõtete poolt klientidele pakutavad tooted ja teenused ei erine üksteisest kuigivõrd palju. Efektiivne tootmisprotsess on ettevõtte püsimise tarvilik nähtus. Edukas olemiseks ei piisa enam vaid paremast tootmise korraldamisest, sest efektiivselt suudetakse toota igal pool maailmas ning tihtipeale omavad turul tegutsevad konkurendid tugevat konkurentsieelist mastaabiefektist tulenevalt. Kuna toodete ja tootmisega konkurentidest eristuda on keeruline, tuleb pöörata vaade klientide poole. Konkurentidest eristumiseks tuleb tegeleda kliendisuhete loomise ja arendamisega. Selleks peab iga organisatsioon, kes soovib olla edukas, tundma oma kliente.

Klientide ootused erinevatele organisatsioonidele sõltuvad nii organisatsiooni poolt pakutavatest toodetest ja teenustest kui ka kliendi konkreetsetest eelistustest. Organisatsioonidel on võimalus vastata klientide ootustele, pakkudes äärmustena kas kõigile klientidele täpselt ühesugust või kõigile klientidele täiesti erinevat, kliendi personaalsest eelistusest lähtuvat lahendust. Erilahenduste pakkumine on kõige kulukam lahendus kliendi vajaduste rahuldamiseks ja ei ole seega üldjuhul majanduslikult mõistlik. Kõigile klientidele ühetaolise lahenduse pakkumine ei pruugi olla eelistatuim lahendus kõigile klientidele ning organisatsioonid võivad kaotada oma kliente konkurentidele, kes suudavad pakkuda kliendi eelistustega enam kokku sobivaid lahendusi. Seetõttu on mõistlik leida kompromiss, pakkudes sarnaste eelistustega kliendigruppidele nende eelistustest lähtuvaid lahendusi ning arendada pakutavat sihtgrupi eelistustest lähtuvalt. Organisatsiooni esmaseks ülesandeks on seega leida sarnase käitumisega klientide grupid ning töötada neile välja nende eelistustest lähtuvad lahendused. Nii luuakse üheaegselt täiendavat väärtust nii kliendile kui ka hüvist pakkuvale organisatsioonile.

Eestis ettevõtluse arendamisega tegelev riiklik sihtasutus Ettevõtluse Arendamise Sihtasutus jõuab oma turu segmenteerimise teemalisel kodulehel samuti järeldusele, et konkurentsis edukaks olemiseks on mõistlik kliente kõnetada sarnase käitumisega klientide kogumite ehk segmentide kaudu. Ettevõtluse Arendamise Sihtasutus toob samas materjalis välja ka turu segmenteerimiseks enamlevinud näitajad, milleks on geograafilised, demograafilised, sotsiaalmajanduslikud ning isiksuse- ja hoiakutepõhised tunnused. Eraldi tuuakse välja, et segmenteerimise juures on oluline, et saadud segmendid oleksid eristuvad, tuvastatavad, ligipääsetavad, olulised ja stabiilsed. (Turu segmenteerimine 2014) Kuidas selliseid segmente leida, jääb aga selgitamata.

Üheks võimaluseks on valida organisatsioonile subjektiivsetel alustel huvi pakkuvad näitajad ning valitud näitajate alusel kliente klassifitseerides moodustada sarnase käitumisega kliendigrupid. Alternatiivseks võimaluseks on andmekaeve kasutamine andmetes peituva loomuliku struktuuri uurimiseks ja seal loomulikult eksisteerivate kliendisegmentide leidmiseks; miski, milleni organisatsioonid tavapärase arutluskäigu raames ei pruugi jõuda. Sellised kliendigrupid kujutavad endast organisatsioonide seni ignoreeritud kliente, keda on turule siseneval organisatsioonil võimalik endale võita ning kelle olemasolust teadlik olemine aitab juba tegutsevatel organisatsioonidel neid paremini hoida ning seega oma kasumlikkust kasvatada.

Maailmas on andmekaevet sihtklientide tuvastamiseks kasutatud laialdaselt. Näiteid leiab nii ostukeskuste (Dennis *et al.* 2001), *online* mängutööstuse (Lee *et al.* 2004), reisibüroode (Liao *et al.* 2010) kui ka juuksurisalongi (Wei *et al.* 2013) klientide uurimisest. Eestis on andmekaeve kasutamine orgaisatsioonide poolt seni aga tagasihoidlik.

Käesoleva magistritöö eesmärk on tutvustada klasteranalüüsi võimalusi loomulike kliendisegmentide tuvastamisel tarbimiseelistuste alusel. Loomulike kliendisegmentide tuvastamine loob tugeva pinnase tulemuslikuks organisatsiooni kliendisuhete juhtimise korraldamiseks. Nii loob magistritöö aluse ja annab soovitusi tarbimiseelistustele tuginevate sihtkliendigruppide leidmiseks klasteranalüüsi abil juba tegutsevale või alles turule siseneda plaanivale organisatsioonile kliendisuhete juhtimise korraldamiseks.

Magistritöö eesmärgi saavutamiseks on püstitatud viis uurimisülesannet:

1) anda ülevaade kliendisuhete juhtimise raamistikust ja andmekaevest,
2) tutvustada klasteranalüüsi metoodikat,
3) selgitada välja sobivaim klasteranalüüsi tehnika kliendisegmentide moodustamiseks,
4) kirjeldada andmetes eksisteerivad loomulikke segmente ning tuvastada neile iseloomulikke tunnuseid,
5) anda soovitusi klasteranalüüsi abil kliendisegmentide moodustamiseks.

Magistritöö jaotub teoreetiliseks ja empiiriliseks osaks. Mõistmaks loomulike kliendisegmentide tuvastamise vajalikkust, annab teoreetiline osa esmalt ülevaate kliendisuhete juhtimise raamistikust. Seejärel tutvustatakse andmekaevet kui kliendisuhete juhtimist toetavat analüüsimeetodit. Lähemalt selgitatakse klasteranalüüsi kui ühe olulisima kliendisuhete juhtimist toetava ja loomulike kliendisegmentide tuvastamiseks sobilikuma tehnika metoodikat. Empiirilise osa ülesandeks on tuvastada Eesti leibkondade hulgas eksisteerivad loomulikud segmendid tarbimiseelistuste alusel ning leida tunnused, mis aitavad selgitada segmentide erinevaid eelistusi. Kliendisegmentide moodustamiseks kasutatakse indiviidide asemel leibkondi kui ühiselt tegutsevaid majandusüksuseid. Leibkonna tasandil tehakse tarbimisotsused ka indiviidide eest, kes küll tarbivad organisatsioonide poolt pakutavaid hüviseid, kuid ei suuda, ei saa või ei oska oma eelistusi väljendada (näiteks teovõimetud isikud ja lapsed). Samuti tehakse leibkonna tasandil ühised otsuseid suuremate kulutuste osas (elamispinna valik, kestvuskauba soetamine jne.). Eesti leibkondade segmenteerimine tarbimiseelistuste alusel viiakse läbi erinevaid klasteranalüüsi tehnikaid rakendades, et selgitada välja sobivaim klasteranalüüsi tehnika kliendisegmentide moodustamiseks. Analüüsile tuginedes püstitatakse uurimisküsimusi, mida Eesti leibkondade tarbimiseelistuste edasisel uurimisel täiendavalt silmas pidada ning antakse soovitusi kliendisegmentide moodustamiseks klasteranalüüsi abil.

Kliendisuhete juhtimise raamistiku ja andmekaeve olemuse tutvustamisel tugineb magistritöö autor põhiliselt andmekaeve konsultantide Berry ja Linoffi (2004) lähenemisele. Klasteranalüüsi metoodika tutvustamisel tugineb autor Byrne ja Uprichardi koostatud klasteranalüüsi-teemaliste võtmetekstide kogumikule (2012) ja

juhtivate statistikatarkvarade kasutusjuhenditele. Klasteranalüüsi arvukatest erinevatest tehnikatest võrreldakse kliendisegmentide tuvastamiseks sobivaima tehnika tuvastamiseks üksiku seose, keskmise seose, täieliku seose, Wardi ja K-keskmiste tehnikaid. Klasteranalüüs viiakse läbi Statistikaameti poolt teostatava Eesti leibkonna eelarve uuringu (edaspidi LEU) andmetele tuginedes. LEU viiakse läbi terve kalendriaasta kestel ning uuringu käigus kogutud andmed avaldatakse mitte vähem kui kolm kuud pärast kalendriaasta lõppu (Tikva ja Arnik 2012: 25). Kuna 2013. aasta andmed on uurijatele kättesaadavad parimal juhul alates 2014. aasta aprillist, viiakse käesolev magistritöö läbi 2012. aasta andmetele tuginedes, mis on värskeimad saadaval olevad andmed magistritöö teostamise hetkel.

1. KLIENDISUHETE JUHTIMINE JA ANDMEKAEVE

1.1. Kliendisuhete juhtimine

Kliendisuhete juhtimine (*customer relationship management* ehk CRM) on laialdaselt kasutuses olev mõiste ning tunnustatud ärijuhtimise meetod. Seda tõestab Rigby ja Ledinghami (2004: 118) poolt juba 2003. aastal läbiviidud uuring, kus küsitleti 708 juhtivat globaalset jaekaubandusettevõtte juhti, kellest 82% väitis, et kasutavad või plaanivad kasutama hakata kliendisuhete juhtimist oma ettevõtetes. Sellele vaatamata puudub kliendisuhete juhtimisel ühene definitsioon (Ngai *et al.* 2009: 2592).

Nimest tulenevalt viitab kliendisuhete juhtimine kliendikesksele äristrateegiale (Chalmeta 2006: 1016). Ling ja Yen (2001: 83) avavad kliendisuhete juhtimise mõiste kui äristrateegia, mis ühendab endas kogumit protsesse ja süsteeme, ehitamaks pikaajalisi ja kasumlikke suhteid valitud klientidega. Kliendisuhete juhtimise fookus on klientidel ning seetõttu erineb kliendisuhete juhtimine tavapärasest ettevõtte ressursside planeerimisest (e*nterprise resource planning* ehk ERP). Ettevõtte ressursside planeerimise eesmärk on ettevõttesisene ressursside kasutamise optimeerimine ning seetõttu on ettevõtte ressursside planeerimise tegevused suunatud pigem tagatoale (finantsjuhtimine, tootmine, inimressurss). Kliendisuhete juhtimise eesmärk on reageerida klientidele ning on seetõttu suunatud enam eesliinile (kõnekeskused, müük ja turundus) (King ja Burgess 2008: 422). Kliendisuhete juhtimine ei piirdu aga vaid klientide teenindamise ja müügi korraldamisega. Kliendisuhete juhtimine on rohkem kui erinevate turundusideede ümberpakendamine. Kliendisuhete juhtimine seostub kliendisuhtest lisandväärtuse otsimisega läbi tehnoloogia rakendamise kliendi kohta info omandamiseks, selle info jagamiseks, kliendisuhete pikaajaliseks arendamiseks ning olemasolevate protsesside integreerimiseks erinevatest valdkondadest (Boulding *et al.* 2005: 157). Kliendisuhete juhtimist käsitleva kirjanduse ülevaates leiavad Ngai, Wiu ja Chau (2009: 2592) kokkuvõtvalt, et kõigi erinevate kliendisuhete juhtimise mõiste

definitsioonide ühisosaks on arusaam, et kliendisuhete juhtimine on kompleksne protsess klientide hankimiseks ning säilitamiseks, kasutades ärianalüüsi ja maksimeerides seeläbi klientide väärtust organisatsioonile. Käesolevas magistritöös vaadatakse kliendisuhete juhtimist kui organisatsiooni juhtimisstrateegiat, mis kasvatab organisatsiooni väärtust läbi kliendisuhtlusest kogutud informatsioonile reageerimise.

Kliendisuhete juhtimise protsessi saab jagada operatiivseks ja analüütiliseks (Ngai *et al.* 2009: 2592). Operatiivse kliendisuhete juhtimise eesmärk on protsesside automatiseerimine. Sellel tasandil toimuvad tehingud, reklaampostitused, info kogumine ja talletamine ning palju muud. Analüütilise tasandi eesmärk on kliendi karakteristikute ja käitumise analüüsimine. Analüütiline kliendisuhete juhtimine võimaldab organisatsioonil paremini eraldada ja efektiivsemalt suunata ressursse kõige kasulikumale kliendigrupile. Rõhutades, et kliendisuhete juhtimine on äristrateegia, mitte pelgalt protsess, pakub Chalmeta (2006: 1021) täiendavalt välja strateegilise tasandi, kus tehakse kliendi käitumisest lähtuvalt ümberkorraldusi organisatsioonis. Laiemalt vaadates võib strateegilist tasandit vaadata osana analüütilisest kliendisuhete juhtimisest.

Organisatsioonide suhetes klientidega valitseb teatud seaduspära. Esmalt on organisatsioonil tarvis endale kliendid leida ja seejärel suhetarendada. Kracklauer, Mills ja Seifert (2004: 4) jagavad kliendisuhete juhtimise nelja etappi:

1) kliendi identifitseerimine (*client identification*),
2) kliendi meelitamine (*client attraction*),
3) kliendi hoidmine (*client retention*),
4) kliendi arendamine (*client development*).

Kliendisuhete juhtimine saab alguse klientide identifitseerimisest. Selle etapi eesmärk on sihtida populatsiooni osa, mis on kõige tõenäolisemalt valmis organisatsioonile kliendiks hakkama või on kõige kõrgema kasumlikkusega. Kliendi identifitseerimise etapi elementideks on klientide segmenteerimine ja sihtkliendi analüüs. Segmenteerimine tegeleb kogu turu jagamisega väiksemateks gruppideks, millel on sarnased tunnused. Sihtkliendi analüüsi eesmärk on leida kasumlik kliendisegment läbi klientide karakteristikute uurimise.

Klientide meelitamine järgneb klientide identifitseerimisele. Selle sammu eesmärk on varem defineeritud kliendigruppidele lähenemine. Põhimeetodiks on otseturundus. Näiteks võib klientidele saata kuponge, sooduspakkumisi, kliendilehti jne.

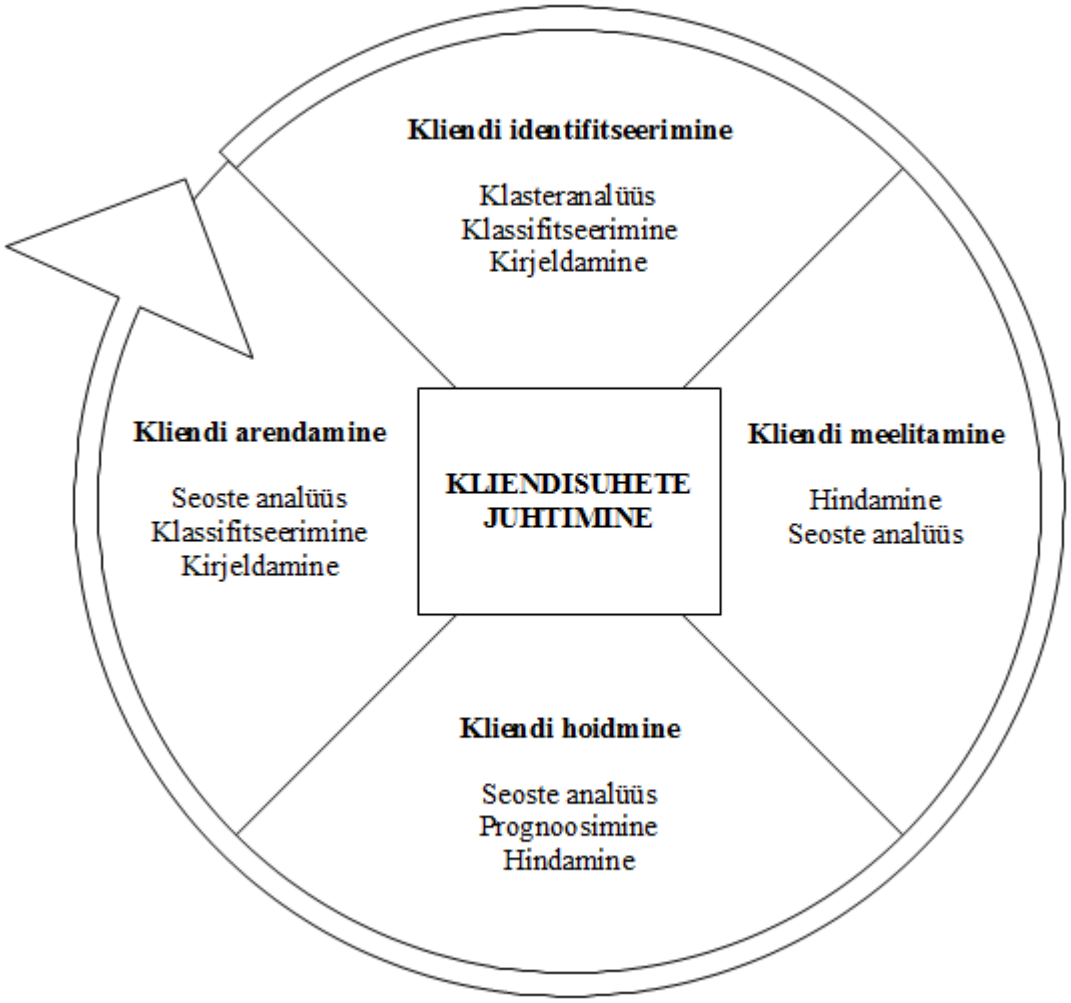
Kliendisuhete juhtimise keskseks mureks on klientide hoidmine organisatsiooni juures. Kliendi püsimine organisatsiooni juures sõltub kliendi rahulolust, mis kujutab endast kliendi taju selle osas, kuidas organisatsioon rahuldab tema ootusi (Kracklauer *et al.* 2004: 5). Organisatsioonile on oluline selles kliendijuhtimise etapis tunda kliendi profiili ja sellest lähtuvalt kasutada ressursse optimaalselt kliendi ootustele vastamiseks. Kliendi hoidmise meetoditeks on lojaalsusprogrammid, personaalne turundus ja kliendikaebuste haldamine.

Kliendi arendamise etapp tegeleb kliendi tehingute sageduse, mahu ja individuaalse kasumlikkuse suurendamisega. Seda eesmärki aitavad täita kliendi ostukorvi ja elukaare analüüs ning sellest lähtuv organisatsioonipoolne rist- ja pealemüügi korraldamine.

Üheskoos moodustavad kliendi identifitseerimine, kliendi meelitamine, kliendi hoidmine ja kliendi arendamine tervikliku kliendijuhtimise ringi (vt. joonis 1 lk. 12). Kui organisatsioon on tuvastanud oma sihtkliendid, neile lähenenud, neid hoidmas ja suurendamas kliendikasumlikkust, on eelnevalt omandatud teadmistele tuginedes võimalik paremini sooritada sihtkliendi analüüsi ning alustada tsükliga otsast peale.

Vaatamata sellele, et kliendisuhete juhtimine on üldtunnustatud äristrateegia, ei õnnestu kliendisuhete juhtimise edukas juurutamine organisatsioonides kuigi hästi. Anderson, Jolly ja Fairhurst (2007) uurisid viieaastasel perioodil (aastatel 2000 kuni 2005) maailma juhtivate tööstus- ja kaubandusteemaliste teadusajakirjade artikleid ning suutsid tuvastada 149 kliendisuhete juhtimise teemalist artiklit. Neile artiklitele tuginedes täheldati, et vaatamata asjaolule, et kliendisuhete juhtimine on laialt kasutuses, ei rakendata kliendisuhete juhtimist kui terviklikku strateegiat. Vaid vähesed jaekaupmehed leidsid, et kliendisuhete juhtimist võiks kasutada klientide segmenteerimiseks. Läbi töötatud artiklitest 7% identifitseerisid segmenteerimise kui eesmärgi või ülesande ja ainult 3% identifitseerisid segmenteerimise kui praeguse kliendisuhete juhtimise strateegia osa. Jaekaubanduses kasutuses olevad andmekaeve meetodid on pühendatud põhiliselt turundusele, mitte aga eri klientidele suunatud

lojaalsusprogrammidele, mis on nende samade ettevõtete poolt mainitud kui olulisimad valdkonnad. See annab tunnistust kõrgemal tasemel strateegilisest kliendisuhete juhtimise metoodika kasutamise puutumisest.



Joonis 1. Kliendisuhete juhtimise etapid ja andmekaeve meetodid (autori koostatud).

Kliendisuhete juhtimise edukaks juurutamiseks organisatsioonis ja seeläbi organisatsioonile maksimaalse täiendava väärtuse loomiseks peab kliendisuhete juhtimist vaatama kui tervikstrateegiat. Organisatsioonid rakendavad kliendisuhete juhtimise erinevaid etappe valikuliselt ja jätavad tihti kliendi identifitseerimata ehk strateegiale vundamendi loomata. Küllap on see tingitud asjaolust, et tegutsevatel organisatsioonidel on varasemalt kliendibaasid välja kujunenud ning kliendid varasemalt segmenteeritud, kuid valdkond väärib selgelt enam tähelepanu.

Kliendisuhete juhtimise osad, mis on organisatsioonide poolt alarakendatud, on kohad, kus on võimalik luua enim täiendavat väärtust. Seetõttu vaadeldakse käesolevas magistritöös lähemalt kliendi identifiseerimise aluseks olevat turu segmenteerimist, mis aitab organisatsioonidel oma kliente täpsemini kõnetada ning klientidelt kogutud informatsioonile reageeria.

1.2. Andmekaeve

Operatiivse kliendisuhete juhtimise protsessi käigus koguvad organisatsioonid tohutul hulgal andmeid, mida analüütilise kliendisuhete juhtimise käigus analüüsitakse. Selleks, et analüüsida suuri andmehulki, eesmärgiga leida andmete hulgast seniteadmata ja kasulikku informatsiooni, vajatakse selleks sobivaid analüüsitehnikaid. Kasulike andmetükkide otsimine suurtest andmemahtudest on saanud tuntuks kui andmekaeve (Jain *et al.* 2012: 366). Sarnaselt kliendisuhete juhtimise mõistele puudub ka andmekaeve mõistel ühene definitsioon. Hormazi ja Giles (2004: 63) vaatavad lähemalt kuute erinevat andmekaeve definitsiooni ning jõuavad järeldusele, et kõikidel vaatlusalustel definitsioonidel on läbivalt üks idee, milleks on olulise info eraldamine olemasolevatest andmetest, võimaldamaks paremate otsuste langetamist.

Andmekaeve käigus leitakse olemasolevaid andmeid lähemalt uurides uusi teadmisi. Nii suudab andmekaeve anda vastuseid ka küsimustele, mida otsustajad organisatsioonides isegi ei oska küsida (Moin ja Ahmed 2012: 738). Organisatsioonid ei pea tuginema enam vaid oma juhtide teadmistele ja kogemusele tähtsate otsuste langetamisel. Andmekaeve võimaldab analüüsida keskkonnast kogutud andmeid ning neile tuginedes leida uusi tegevussuundi ning uurimisvaldkondi.

Andmekaevet on pikalt kasutatud panganduses. Kuulsaim näide andmekaevsest tulenevast teadmisest pärineb krediidivõimekuse hindamisest. Panga klientide maksekäitumise uurimisel avastati, et kliendid, kes kasutavad pangakaarti kasiinos, jäävad kõrgema tõenäosusega oma maksetega panga ees hätta (Hormazi ja Giles 2004: 67). Tegu on aimatava järeldusega ning samale järeldusele võib jõuda teooriast tulenevate hüpoteeside testimisel, kuid hüpoteeside püstitamisel ollakse piiratud teooria ning parima teadmisega sel hetkel. Andmekaeve võimaldab leida ka selliseid

seaduspärasusi, mille seletamiseks teooria veel puudub. Leitud seaduspärasustele tuginedes on võimalik püstitada hüpoteeside paare nende leidude kontrollimiseks ning hüpoteesi vastuvõtmise korral luua uusi teadmisi ja teooriaid. Eelnevale tuginedes vaadatakse magistritöös andmekaevet kui meetodit, mis võimaldab olulise info eraldamist olemasolevatest andmetest ilma eelneva konkreetse probleemi püstituseta.

Andmekaeve on lai valdkond ning andmekaevet toetavaid statistilisi meetodeid ning lähenemisi on hulgaliselt. Hormazi ja Giles (2004: 65) jagavad andmekaeve operatsioonid analüüsitehnikate sarnasusele tuginedes seitsmesse gruppi:

1) klasterdamine,
2) visualiseerimine,
3) ennustav modelleerimine,
4) seoste analüüsimine,
5) kõrvalekallete avastamine,
6) sõltuvuse modelleerimine,
7) kokkuvõtlik kirjeldamine.

Ngai, Wiu ja Chau (2009: 2593) jagavad samuti andmekaeve kasutatavatele tehnikatele tuginedes seitsmeks grupiks, kuid eristavad klassifitseerimist ja klasteranalüüsi kui teineteisest erinevaid tehnikaid. Samas ei erista nende andmekaeve tehnikate klassifikatsioon kokkuvõtlikku kirjeldamist ülejäänud tehnikatest.

Andmekaeve juures ei ole aga oluliseks mitte erinevate tehnikate rakendamine, vaid uute teadmiste avastamine ja uurimisprobleemide lahendamine. Seega on mõistlik lähtuda andmekaeve meetodite eristamisel probleemidest, mida uuritakse. Lähtuvalt uurimisprobleemist võib andmekaeve tehnikad jagada kuueks (Berry ja Linoff 2004: 8):

1) klassifitseerimine,
2) hindamine,
3) prognoosimine,
4) sarnasuste ja seoste leidmine,
5) klastrite leidmine,
6) kirjeldamine ja profileerimine.

Klassifitseerimine, hindamine ja prognoosimine tegelevad kindla muutuja väärtuse leidmisega, eesmärgiks ennustada spetsiifilist muutujat läbi teiste muutujate andmetes. See on sarnane inimese õppimisele, kus harjumus tähendab, et sarnaseks võivad osutuda ka järgmised vaatlused (Hormazi ja Giles 2004: 64). Seoste leidmine ja klasterdamine tegelevad andmetes peituva struktuuri otsimisega ilma muutujatevahelistesse seostesse süvenemata. Kirjeldamise ja profileerimise käigus antakse ülevaade andmetest.

Klassifitseerimise korral ollakse veendunud, et uuritavad objektid kuuluvad etteantud arvu etteantud tunnustega klassidesse. Klassifitseerimise ülesandeks on anda igale objektile klassi tunnus ning jagada objektid klasside vahel. Enamlevinud klassifitseerimise moodusteks on otsustuspuu (*dicision tree*) ja lähima naabri (*nearest neighbor*) tehnikad (*Ibid*: 9). Klassifitseerimine on üks enamlevinud õppimise viise andmekaeves (Ngai *et al.* 2009: 2595). Objekti klassi tunnuse hindamine teadaolevatele tunnustele tuginedes võimaldab leida huvipakkuva objekti tunnuse seda eelnevalt teadmata. Klassi tunnus hinnatakse, tuginedes samasse klassi kuuluvate teiste objektide tunnustele. Nii on võimalik objekte klassifitseerida näiteks tunnuste alusel, millest ollakse huvitatud, nende tunnuste väärtusi teadmata. Klassifitseerimist võib kasutada näiteks kliendi soo, vanusegrupi, suhtumise, kliendiprofiili jne. leidmiseks, seda otseselt kliendilt küsimata.

Kui klassifitseerimine tegeleb diskreetsete tunnuste hindamisega, siis hindamine tegeleb pidevate tunnuste väärtuste leidmisega (Berry ja Linoff 2004: 9). Otsitakse muutujate vahelisi seoseid, et hinnata tundmatu muutuja väärtust. See võimaldab tundmatu muutujaga objekte reastada. Näiteks selgitada läbi muude tunnuste objekti vanust, kliendi sissetulekut või valmisolekut toodet tarbida. Nii saab näiteks klassifitseerimise käigus leitud potentsiaalsete klientide hulgast välja selgitada kõige kasumlikumad ning fokusseerida oma tegevus neile. Enamlevinud hindamistehnikateks on regressioon ja tehisnärvivõrgustikud (*neural networks*) (*Ibid*: 10).

Prognoosimise korral leitakse samuti väärtus tundmatule tunnusele, kuid võetakse arvesse varem hinnatud muid muutujaid ning antakse muutujale tulevikuväärtus. Ainus võimalus prognoosimise täpsuse kontrollimiseks on oodata tuleviku sündmuse toimumist. Prognoosimine on eristatud klassifitseerimisest ja hindamisest, kuna täiendavaid probleeme tekitab küsimus, kas tänane muutujatevaheline seos on ka

tulevikus sarnane. Võimalik on kasutada kõiki klassifitseerimise ja hindamise tehnikaid. (Berry ja Linoff 2004: 10). Prognoosimine võimaldab seega modelleerida gruppi kuuluva objekti tulevikukäitumist, tuginedes olemasolevatele ajaloolistele andmetele. Seda on võimalik teha eeldades, et grupi siseselt käituvad objektid sarnaselt. Tüüpiliseks näiteks on nõudluse prognoosimine (Ngai *et al.* 2009: 2595), kuid sarnaselt prognoositakse ka ilma, pankrotistumise tõenäosust ehk krediidiriski, kliendi lahkumist organisatsiooni juurest etteantud ajaperioodil, kliendipoolset tõenäosust tarbida konkreetset toodet tulevikus jne.

Sarnasuste ja seoste leidmise eesmärk on andmebaasis olevate objektide ja muutujate vahel leiduvate seoste identifitseerimine. Eksisteerib kolm põhilist tehnikat: seoste, ajaliselt järjestatud mustrite ja sarnase ajalise järjestuse uurimine (Hormazi ja Giles 2004: 64). Seoste uurimist kasutatakse selleks, et välja selgitada, millised objektid või tunnused eksisteerivad koos. Tüüpiline näide on ostukorvi analüüs. Ajaliselt järjestatud mustrite uurimist kasutatakse kliendi pikaajalise ostukäitumise uurimiseks. Sarnast ajalist järjestatust uuritakse toote elukaare analüüsiks. Sarnasuste ja seoste teadmine andmetes võimaldab rist- ja pealemüügi paremat planeerimist ning atraktiivsete toodete või teenuste väljatöötamist. Seoste uurimine on küllaltki lihtne meetod andmetest reeglite genereerimiseks. Kui kaks objekti leiduvad üheaegselt piisavalt tihti, võib genereerida reegli, et need objektid sobivad kokku. Lihtsale matemaatikale tuginedes on võimalik leida ka objektide koos esinemise tõenäosus (Berry ja Linoff 2004: 11). Seoste uurimise enamlevinud tehnikateks on lihtne statistika, eelnevalt paika pandud teooria ja ette antud algoritmid (Ngai *et al.* 2009: 2595).

Klastrite leidmine ehk klasteranalüüs kujutab endast heterogeense andmekogu segmenteerimist väiksematesse homogeensematesse gruppidesse ehk klastritesse. Klasterdamine erineb klassifitseerimisest selle poolest, et protsessi käigus määratakse ka klassid, millesse objektid liigitatakse. Klassifitseerimise korral tuleb klassid eelnevalt ise määratleda. (Moin ja Ahmed 2012: 740). Teisisõnu pole ette antud kriteeriume, mille alusel objektid klassifitseerida (Ngai *et al.* 2009: 2595). Klasteranalüüsi käigus grupeeritakse objektid omavahelise sarnasuse alusel. Uurija ülesandeks jääb määrata, kas tulemuseks saadud klastritel on tähendus või mitte. Klastrite leidmiseks kasutatakse peamiselt hierarhilisi ja mittehierarhilisi tehnikaid (Berry ja Linoff 2004: 11).

Klasteranalüüs võib anda lõpliku tulemuse uurijale huvipakkuva grupi leidmise näol. Näiteks klientide klasterdamine sarnasuse alusel annab tulemuseks erinevad kliendisegmendid. Kuid tihti on klastrite leidmine protsess, mis eelneb mõnele teisele andmekaeve meetodile. Näiteks võib sihtklientide omakorda klastriteks jagamine eelneda klientide kõnetamisele kliendi meelitamise etapis. Selle asemel, et kõnetada kõiki sihtkliente sarnaselt, võib kliendid jagada esmalt sarnaselt käituvateks gruppideks ja uurida, mis tegevused sobivad just nende gruppide kõige tulemuslikumaks kõnetamiseks.

Mõnikord piisab vaid keeruliste ja mahukate andmete kirjeldamisest lihtsamal kujul, et parandada arusaama andmete sisust. Kirjeldamine ja profileerimine on tehnika, mis võimaldab keerukaid andmebaase teha lihtsamini arusaadavaks, püstitamaks uurimisprobleeme ning mõistmaks andmete struktuuri. Piisavalt hea andmete kirjeldus võib tihti pakkuda välja ka vastuse uurimisprobleemile. Kui ei leita koheselt vastust, siis annab andmete parem mõistmine arusaama uurimissuunast, mida võtta uurimisprobleemi lahendamiseks. Hea näite leiab Ameerika Ühendriikide poliitikast, kus lihtne andmete kirjeldus, mille sisuks oli, et naised toetavad demokraate rohkem kui mehed, vallandas suure hulga põhjalikumate uurimuste läbiviimise ja ajakirjandusliku huvi. (Berry ja Linoff 2004: 12)

Kirjeldamise ja profileerimise osana võime vaadata andmete visualiseerimist ja üldistamist. Visualiseerimine on andmete kirjeldamine graafikutel ja joonistel. Visualiseerimist kasutatakse saamaks selgemat arusaama andmetest, püstitamaks täpsemaid uurimisülesandeid ja mõistmaks teiste andmekaeve meetoditega leitud mustreid (Ngai *et al.* 2009: 2595). Nii panustab visualiseerimine paremasse andmete mõistmisesse ja annab võimaluse peidetud mustrite avastamiseks, eriti kui tegu on keeruliste ja mittelineaarsete seostega.

Andmete üldistamine on tehnika, mille käigus antakse ülevaade alamhulga kohta läbi horisontaalse või vertikaalse summeerimise. Esimeses vaadatakse alamhulga kokkuvõtet, teises ennustatakse väljade vahelisi seosesid. Andmete üldistamine võib olla piisavaks tegevuseks näiteks ostukorvi analüüsil (Hormazi ja Giles 2004: 65).

Aina kasvavate andmemahtudega maailmas on andmekaeve roll juhtimisotsuste langetamise juures üha olulisem. Erinevad andmekaeve tehnikad võimaldavad otsustajatel hoomata suuremaid andmehulkasid ja fokuseerida oma tähelepanu olulisele infole. Leidmaks andmekaeve abil vastuseid küsimustele, mida otsustajad veel isegi ei oska küsida, tuleb erinevaid andmekaeve tehnikaid kasutada teadlikult ning süstemaatiliselt. Nii luuakse otsustajatele alternatiiv tavapärasele otsuste laengetamisele, mis tugineb vaid varasematele kogemustele ning teada olevale teooriale.

1.3. Andmekaeve kasutamine kliendisuhete juhtimisel

Andmekaevet võib pidada kliendisuhete juhtimise lahutamatuks osaks, kuna lisandväärtus kliendisuhete juhtimisest tekib klientidelt kogutud andmetest olulise eraldamisel ja sellele reageerimisest. Ilma andmekaeveta on aga tänapäeval kogutavatest andmetest keeruline leida veel seni teadvustamata informatsiooni. Vastamaks paremini klientide ootustele, tuleb tegeleda süstemaatiliselt uute avastuste tegemisega. Kuna kliendisuhete juhtimise analüütilises osas tegeletakse andmete analüüsiga erinevate nurkade alt, kasutades selleks erinevaid andmekaeve tehnikaid, võib vaadata kliendisuhete juhtimist kui andmekaeve strateegiat. Bahrami (2012: 61) väidab koguni, et kliendisuhete juhtimine ise on modernne ja arenenud tööriist süstemaatiliseks andmekaeveks.

Kliendisuhete juhtimine ei lõppe kliendiandmete analüüsil tugineva ühekordse uute automaatsete protsesside juurutamisega. Pidev andmete analüüs ning hüpoteeside jätkuv testimine on vajalikud, kuna kliendisuhete juhtimise tegevused ei ole staatilised. Inimesed otsivad pidevalt võimalusi oma heaolu parandamiseks. Seetõttu muutuvad ajas ka klientide eelistused ja käitumine. (Boulding *et al.* 2005: 163) Selleks, et klientidega kaasas käia, tuleb neist muudatustest teadlik olla ning pidevalt kliendisuhete juhtimise protsesse üle vaadata.

Erinevates kliendisuhete juhtimise etappides on andmekaeve eesmärgid erinevad ning seega on ka kasutatavad andmekaeve tehnikad erinevad. Ülevaate erinevates etappides kasutatavatest tehnikatest annab joonis 1 (vt. lk. 12). Alljärgnevalt kirjeldame andmekaeve tehnikate kasutamist kliendisuhete juhtimise etappides lähemalt.

Kliendi identifitseerimise faasis on eesmärgiks välja selgitada, kes on ettevõtte kliendid. Oma klientide tundmine on efektiivse kliendisuhete juhtimise süsteemi loomise eelduseks (Chalmeta 2006: 1019). Kui organisatsioonid mõistavad kliendi eelistusi, on neil võimalik välja töötada adekvaatsed turunduse strateegiad ja vastata klientide ootustele. Nii saavutatakse kõrgem kliendi rahulolu ja valmisolek tarbida organisatsiooni poolt pakutavaid tooteid ning teenuseid. Seega uuritakse selles etapis, millised on sarnase käitumisega klientide grupid, kes neisse kuuluvad, kuidas nad käituvad ning milline on nende kasumlikkus. Andmekaeve tehnikatest aitavad selles etapis klasteranalüüs, klassifitseerimine ja kirjeldamine ning profileerimine.

Klientide meelitamine on kliendisuhete juhtimise järgmiseks etapiks. Kliendi meelitamise etapis on oluline kõnetada õiget klienti, õigel ajal ja kliendi jaoks õigel viisil. Andmekaeve on selleks sobiv vahend võimaldades olla proaktiivne, mitte reaktiivne, pidevalt muutuvas maailmas (Hormazi ja Giles 2004: 65). Organisatsioonid ei pea ootama, et klient pöörduks oma muutunud või uute soovidega nende poole, vaid organisatsioon võib ise oma kliente kõnetada neid huvitavates valdkondades. Samuti on võimalik kokku hoida ressursse kõnetades vaid kliente, kelle ostuvalmidus on kõige kõrgem ja maksevõime parim. Kliendi meelitamise etapil on olulisteks küsimusteks, milliseid kliente tasub kõnetada, kuidas seda teha ja milliseid uusi tooteid turule tuua. Neid uurimisprobleeme aitavad kõige paremini lahendada hindamine ja seoste analüüs.

Kliendi hoidmise etapi eesmärk on hoida kliendi rahulolu piisavalt heal tasemel, et klient konkurendi poole ei pöörduks. Ühelt poolt tähendab see kulude kokkuhoidu liigse teenindustaseme pakkumise vältimisest ja üleinvesteerimisest hoidumise läbi. Teiselt poolt otsitakse täiendavaid tulusid pealemüügi edendamise kaudu. Põhilisteks küsimusteks selles etapis on kuidas kliendid käituvad, milliseid ressursse vajatakse klientide ootuste täitmiseks ning milline on klientide lahkumise tõenäosus. Seega on sobivateks uurimismeetoditeks seoste analüüs, prognoosimine ja hindamine.

Klientide arendamise etapis tegeletakse kliendi kasumlikkuse kasvatamisega läbi müügi suurendamise. Selleks uuritakse kliendi kogu eluea väärtust organisatsioonile, peale- ja ristmüügi võimalusi ning kliendi ostukorvi. Selles etapis leitud teadmised on sisendiks omakorda kliendi identifitseerimise protsessile, parandamaks nii klientide

segmenteerimist ja segmentide väärtuse hindamist. Kasutatavad tehnikad selles etapis on seoste analüüs, klassifitseerimine ja kirjeldamine.

Tihti vajatakse erinevate andmekaeve tehnikate kombinatsioone, et saavutada kliendisuhete juhtimise soovitavaid efekte. Näiteks soovitatakse otseturunduse tarvis kliendid kõigepealt klasteranalüüsi teel segmenteerida, leitud segmente lähemalt uurida ja saadud tulemuse alusel uued kliendid klassifitseerida (Ngai *et al.* 2009: 2595). Nii selgitatakse välja erinevate ootustega kliendid ning töötatakse välja just neile sobiv lähenemine – neile sobivad reklaamid, kaasa- ja pealemüügi taktikad jne. Kliendisuhete juhtimine moodustabki ühtse terviku (vt. joonis 1 lk. 12), kus erinevates etappides kasutatakse erinevaid andmekaevetehnikaid, mis teineteist ja kogu kliendijuhtimise süsteemi täiustavad. Kahjuks vaatavad enamik organisatsioone kliendisuhete juhtimist kui tehnoloogilist lahendust konkreetses kitsas valdkonnas, mis annab tulemuseks suure hulga omavahel kordineerimata tegevusi (Mendoza *et al.* 2007: 919). See on tingitud asjaolust, et paljud organisatsioonid, eriti finantsorganisatsioonid, omavad toote keskset kultuuri (*Ibid*: 913). Sellised organisatsioonid seavad üles oma protsesse toodete keskselt, selle asemel, et pühenduda klientidele (*Ibid*: 914). Nii otsitakse lokaalset optimumi globaalse asemel. Saavutamaks kogu potentsiaalselt lisandväärtust kliendisuhete juhtimise süsteemist, peab aga kliendisuhete juhtimist vaatama kui tervikprotsessi.

Rakendades organisatsioonis ainult üksikuid kliendisuhete juhtimise etappe, võivad tekkida ootamatud probleemid. Näiteks kuidas jagada kliendisuhete juhtimisest tekkinud lisandväärtus kliendi ja organisatsiooni vahel? Ideaalis soovib iga organisatsioon oma kasumit maksimeerida ning omandada kogu loodud lisandväärtuse. Näiteks rakendada võimalusel hinna diskrimineerimist. Nii maksimeeritakse oma kasumit müües identseid tooteid ja teenuseid erinevatele klientidele ja kliendigruppidele erinevate maksimaalsete hindadega, mida vastavad kliendid on valmis maksma. Selleks, et jaekaupmees oleks edukas hinna diskrimineerimisel, peab ta olema suuteline välja selgitama, mis hinnaga erinevad inimesed või grupid on valmis toodet soetama. (Phan ja Vogel 2010: 70) Et hinnadiskrimineerimine saaks toimida, on oluline, et kliendid ei saa teada, mis hinda maksavad teised kliendid. See eeldab klientide segmenteerimist nii, et moodustuvad üksteisest selgelt eristatud kliendigrupid. Vastasel korral riskib

organisatsioon väärtuse loomise asemel selle hävitamisega läbi negatiivse kuvandi tekkimise. Hoiatavaks näiteks on Amazon.com'i 2000. aasta katse (Ramassastry 2005). Amazon.com pakkus veebilehitseja poolt kliendi kohta salvestatava informatsiooni (*cookies*) abil soodsamaid hindu uutele klientidele. Kui regulaarsed kliendid said teada, et maksavad sama kauba eest rohkem kui uued kliendid, said nad vihaseks ja protestisid. Paljud hävitasid oma arvutist varasema veebikülastuste ajaloo, mis neid kui regulaarseid kliente identifitseeris. Tänu negatiivsele kajastusele ja suhete halvenemise klientidega, Amazon.com lõpuks vabandas ning pakkus pahastele klientidele hinnavahe tagasimakseid.

Teooria ei anna selgeid juhiseid, kuidas tekkiv lisanväärtus kliendi ja organisatsiooni vahel peaks jagunema. Saab ainult järeldada, et kumbki osapool ei saa tunda end olevat halvemas situatsioonis, kui enne kliendisuhete juhtimise kasutusele võttu. Kui ettevõtja tekitab olukorra, kus klient arvab, et on kehvemas olukorras, paneb ettevõte end sellega suure riski ette. Ideaalses olukorras luuakse väärtust samaaegselt nii ettevõttele kui ka kliendile (Boulding *et al.* 2005: 159) ning saavutatakse see läbi klientide ootustele vastamise efektiivse kliendibaasi segmenteerimise teel. Nii on võimalik müüa klientidele identseid tooteid ja teenuseid erinevatel viisidel erinevate hindadega ilma, et kliendid pahandaks. See saavutatakse läbi kliendi taju, et pakutavad tooted ja teenused on siiski erinevad.

Turu segmenteerimise olulisusele vaatamata leiavad Ngai, Wiu ja Chau (2009: 2597) oma uuringus, mis põhineb perioodil 2000 kuni 2006 avaldatud 87 teaduslikul artiklil, et kliendisuhete juhtimise juurutamisel on seni küllaltki vähe tähelepanu pööratud sihtkliendi analüüsile ja tagasihoidlikult kliendi identifitseerimisele ning segmenteerimisele. Kõige enam tähelepanu on läbi aegade pööratud aga klientide säilitamisele. Eksisteerib kummaline olukord, kus organisatsioonid üritavad säilitada kliente keda nad ei tunne. Nii võib klient tunda, et tema vajadustega ei arvestata. Organisatsioonid on pühendunud kitsalt konkreetsete probleemide lahendamisse läbi nende tekkepõhjuste otsimise, omamata seejuures objektiivset ettekujutust neid ümbritsevat keskkonnast.

Sotsioloog Andrew Abbott (2012) arutleb, et majandusteadus, ja sotsiaalteadused laiemalt, on liialt kiindunud kitsasse muutujate vahelise kausaalsuse uurimisele ning

pööranud liiga vähe tähelepanu andmete struktuuri uurimisele. Kausaalsuses peitub aga palju probleeme – kas seosed on juhuslikud või tegelikud, mispidine on muutujatevaheline mõju jne. Sotsiaalteadused ei tegele aga niivõrd muutujatevaheliste seoste kuivõrd gruppide ning nende käitumise uurimisega. Inimekäitumise osaks on grupeerumine enda sarnastega ning eristumine endast erinevatest (Bailey 2012: 90). Seetõttu on mõistlik uurida millised naturaalsed grupid eksisteerivad ning seejärel süveneda nende tekkepõhjuste uurimisesse. Klasteranalüüs sobib naturaalsete gruppide tuvastamiseks ideaalselt, kuna tegeleb just sarnaste objektidega gruppide leidmise, mitte muutujatevaheliste seoste uurimisega. Wang (2010: 6421) väidab koguni, et kliendisuhete juhtimise aluseks on klasteranalüüs, millest lähtuvalt tegeletakse järgmistes kliendisuhete juhtimise etappides erinevate kliendigruppidega.

Organisatsioonid, mis soovivad muutuda kliendikesksemateks ning saavutada lisandväärtuse loomet läbi kliendisuhete juhtimise, peavad esmalt tundma oma kliente ning seejärel sättima oma tegevusi klientide vajadustest lähtuvalt. Seetõttu on käesoleva magistritöö fookus kliendi identifiseerimise aluseks oleval turu segmenteerimisel. Turu segmenteerimiseks sobivad kõige paremini turul naturaalselt eksisteerivad kliendigrupid. Loomulike kliendigruppide tuvastamiseks sobivaim andmekaeve meetod on klasteranalüüs. Seetõttu on käesolevas magistritöös lähema vaatluse alla võetud klasteranalüüs kui tähtis esimene ja ettevalmistav samm paljudele sammudest nii andmekaeve kui kliendisuhete juhtimise protsessis.

1.4. Klasteranalüüsi olemus ja tehnikad

Klasteranalüüsi eesmärk on heterogeense andmekogu jagamine väiksematesse homogeensematesse gruppidesse ehk klastritesse. Vaatamata asjaolule, et klasteranalüüs on laialt levinud analüüsi meetod, on klasteranalüüsi keeruline üheselt defineerida. Termin ise ei viita ühelegi kindlale meetodile ega mudelile (Norušis 2011: 376). Statistika tarkvara Stata 11. versiooni, mida kasutatakse käesoleva magistritöö läbiviimiseks, kasutusjuhend väidab koguni, et klasteranalüüsi defineerimine võib olla võimatu ja vaatab klasteranalüüsi kui katsetusuurimislikku andmeanalüüsi tehnikat (StataCorp. 2009: 85).

Statistika tarkvara Stata kasutusjuhend rõhutab terminit katsetusuurimuslik, kuna klasteranalüüs üldjuhul ei kaasa leitud tulemuste tõenäosusliku korrektsuse hindamist (StataCorp. 2009: 85). Seda pole ka tarvis teha, kuna klasteranalüüsi meetodite eesmärk on suuresti hüpoteeside genereerimine, mitte nende paikapidavuse testimine. Automaatseid klastrite määramise tehnikaid kasutataksegi harva eraldiseisvatena, kuna klastrite leidmine pole tihtipeale uurimise lõppeesmärk. Kui klastrid on välja selgitatud, kasutatakse teisi meetodeid mõistmaks, mida need klastrid endast kujutavad (Berry ja Linoff 2004: 350). Käesolevas magistritöös vaadeldakse klasteranalüüsi selle eesmärgist lähtuvalt kui meetodit heterogeense andmekogu jagamiseks väiksematesse homogeensematesse gruppidesse ehk klastritesse.

Erinevaid klasteranalüüsi läbiviimise tehnikaid ja meetodeid on arvukalt. Kuna klasteranalüüs sõltub uurija valikutest, väidetakse koguni, et eksisteerib lõpmatult palju viise klasteranalüüsi sooritamiseks (StataCorp. 2009: 85). Siiski on klasteranalüüsile tüüpilised etapid andmete ettevalmistus, uuritavate objektide sarnasuse määramine, objektide grupeerimine ning leitud klastrite kirjeldamine ja hindamine (Jain *et al.* 2012: 307–308).

Klasteranalüüsi käigus grupeeritakse objektid klastritesse omavahelise sarnasuse alusel. Sarnasus on aga subjektiivne mõiste. Kasutades erinevaid sarnasuse määramise viise, on võimalik sama andmebaasi, sõltuvalt uurimisprobleemist ja analüüsi läbiviijast, erinevalt grupeerida. Uurija peab valima muutujad mille alusel objekte klasterdada, otsustama, kas muutujaid on tarvis transformeerida, määrama viisi objektide sarnasuse leidmiseks, valima meetodi klastrite moodustamiseks ning otsustama mitmesse klastrisse objektid jagatakse. Järgnevalt vaatame klasteranalüüsi etappe lähemalt.

1.4.1. Andmete ettevalmistus

Andmete kogumine ja ettevalmistus on iga statistilise analüüsi lahutamatu osa. Kuna klasteranalüüs tegeleb andmete struktuuri uurimisega, mitte kausaalsuse uurimisega, pole klasteranalüüsi tarvis vajalik teha andmete kohta eelnevaid eeldusi ning kontrollida nende paikapidavust (Norušis 2011: 376). See lihtsustab tunduvalt analüüsi läbiviimist ning võimaldab uurida ka väikeseid valimeid. Samas võib muutujate eelnev põhjalik uurimine ja lihtsad muutujate transformatsioonid oluliselt parandada klasteranalüüsi

tulemusi (Jain *et al.* 2012: 312). Seetõttu on oluline klasteranalüüsile eelnevalt süveneda andmetesse.

Klasteranalüüsi nõrkuseks on tulemuste sõltuvus erinditest andmete hulgas. Erinditega kaasneb oht, et nad kas venitavad saadud klastrid liialt suureks või luuakse liiaga palju klastreid (Norušis 2011: 398). Samas on ärilistest huvidest lähtuvate uuringute korral tihti huvipakkuvamad just erandlikud objektid – väga kasumlikud või kahjumlikud kliendid. Üheks võimaluseks vältida erindite soovimatut mõju klasteranalüüsile, kaotamata seejuures väärtuslikku informatsiooni, on erindite koondamine eraldi klastrisse ning nende eraldiseisev uurimine.

Tihti kirjeldavad objekte andmebaasis väga suur arv erinevaid muutujaid. Kõik muutujad ei pruugi aga endas kanda uurimisprobleemi lahendamise tarvis olulist informatsiooni. Nende muutujate kaasamine analüüsi, võib muuta tulemuseks saadud klastrid edasiseks uurimiseks ebasobivateks (*Ibid*.: 378) ning suurendab analüüsi arvutuslikku mahtu. Üheks võimaluseks on muutujate valik *ad hoc* printsiibil proovides erinevaid võimalusi, kuni jõutakse parimate muutujate väljaselgitamiseni (Jain *et al.* 2012: 313). Probleemile võib läheneda ka teisiti. Klasteranalüüsi eesmärk on objekte grupeerida sarnasuse alusel. Seega võib muutujad valida selle järgi, mille alusel soovitakse, et klastrid omavahel sarnaneksid (Norušis 2011: 376).

Kliendi identifitseerimise eesmärk on leida kliendisegmente, mis on lojaalsed või kasumlikud või suure tõenäosusega reageerimas kindlale pakkumisele. Seega on mõistlik kasutada muutujaid, mis kirjeldavad just huvipakkuvat käitumist. Kui aga eesmärgiks on kliendibaasi segmenteerida stimuleerimaks diskussiooni uue toote väljatöötamiseks erinevatele naturaalselt eksisteerivatele kliendigruppidele, on juhuslik muutujate valik sobilikum (Berry ja Linoff 2004: 372).

Objektide kohta teada olevate muutujate väärtused võivad olla nii kvantitatiivsed kui kvalitatiivsed. Kvalitatiivseid muutujaid pole võimalik aga kujutada matemaatiliste asukohavektoritena ning seega on nende objektidevahelise sarnasuse matemaatiline määramine problemaatiline. Kõiki muutujaid on küll võimalik transformeerida kvantitatiivseteks muutujateks, kuid sellised muudatused lisavad andmetesse võltsi

informatsiooni (Berry ja Linoff 2004: 360). Seetõttu kasutatakse käesolevas magistritöös klasteranalüüsi läbiviimiseks vaid kvantitatiivseid pidevaid tunnuseid.

Geomeetrias võib aga ka pidevate muutujate omavahelisel võrdlemisel tekkida probleeme. Nimelt võivad erinevad muutujad varieeruda erinevas ulatuses. See annab suuremas ulatuses kõikuvale muutujale objektide sarnasuse määramisel suurema kaalu, kuigi uurimisprobleemist lähtuvalt võib ühe muutuja väärtuse väike muutus olla tunduvalt olulisem suurest muutusest teises. Seetõttu mängivad klasteranalüüsis olulist rolli andmete skaleerimine ja kaalumine. Skaleerimise käigus seadistatakse andmeid võttes arvesse, et erinevaid muutujaid mõõdetakse erinevates mõõtühikutes või erinevates suurusjärkudes. Kaalumine pakub võimaluse kohandada andmeid võttes arvess, et mõned muutujad võivad olla olulisemad kui teised.

Muutujate esinemine erinevates mõõtskaalades on sotsiaalteaduslikes uurimustes küllaltki levinud. Inimese vanust mõõdetakse aastates, sissetulekut rahaühikutes ning elukoha kaugust pealinnast kilomeetrites. Skaleerimine annab võimaluse need muutujad omavahel võrreldavaks transformeerida, muutes muutujate väärtused kindlasse vahemikku. Näiteks -1 kuni 1 või 0 kuni 1. Peale sellist muutujate väärtuste transformatsiooni, võrreldakse omavahel muutujate suhtelist muutust, mitte enam muutust absoluutväärtustes. Objektidevaheline kahekordne sissetuleku erinevus rahaühikus muutub võrreldavaks kahekordse kauguse erinevusega pealinnast kilomeetrites.

Berry ja Linoff (2004: 364) toovad välja kolm kõige levinumat viisi andmete skaleerimiseks:

1) normaliseerimine,
2) indekseerimine,
3) standardiseerimine.

Normaliseerimine tugineb muutuja väärtuse asukoha määramisele kogu muutuja varieeruvuse ulatuses. Muutuja normaliseeritud väärtuse leidmiseks jagatakse iga muutuja väärtus, millest on lahutatud muutuja väikseim väärtus, muutuja kogu varieeruvusega. Nii saadakse igale objektile muutuja väärtus, mis jääb matemaatiliste arvutuste tarvis sobivasse vahemikku 0 kuni 1.

Indekseerimise korral jagatakse iga muutuja väärtus läbi muutuja kõikide väärtuste keskmisega. Nii saadakse tulemus, mis näitab, kui suure osa väärtuse keskmisest muutuja väärtus moodustab. Muutuja võib omada nii ülevalt kui alt piiramata väärtust, kuid suurte väärtuste mõju on pehmendatud ning erinevates skaalades mõõdetud muutujad muutuvad omavahel võrreldavaks, kuna esitatakse kui suure osa moodustab muutuja väärtus muutuja keskmisest väärtusest.

Standardiseerimise korral lahutatakse iga muutuja väärtusest muutuja keskväärtus ning jagatakse tulemus läbi muutuja väärtuste standardhälbega. Tulemuseks saadakse Z-skoor, mis näitab, mitme standardhälbe kaugusel on muutuja väärtus muutuja keskväärtusest. Sarnaselt indekseerimisele on leitud uued muutuja väärtused nii ülevalt kui alt piiramata, kuid suurte väärtuste mõju on pehmendatud ning erinevates skaalades mõõdetud muutujad on muudetud omavahel võrreldavaks.

Skaleerimise tulemusena saadakse lahti mõõtühikute ja suurusjärkude probleemist ning võrdsustatakse kõikide muutujate tähtsuse. Geomeetrias on saadud tulemus hea, kuid uurimisprobleemi lahendamiseks ei pruugi saadud tulemus parimaks osutuda. Mõni muutuja võib subjektiivsetel põhjustel olla uurijale olulisem kui teine. Näiteks laste arv peres mõjutab ilmselt leibkonna tarbimisharjumusi enam kui elamispinna suurus või telefoninumbrite arv peres. Kui uurija on sellisel seisukohal ning peab uurimisprobleemist lähtuvalt vajalikuks mõne muutuja tähtsust suurendada, on tal võimalik seda teha muutujate kaalumise teel. Selleks korrutab uurija huvipakkuva muutuja väärtuse läbi muutujale subjektiivselt määratud kaaluga. Kaaluga korrutamise tulemusel muutuja väärtused vastavalt kaalule, kas suurenevad, või vähenevad soovitud ulatuses, muutes nii muutuja tähtsust objektide sarnasuse määratlemisel.

Klasteranalüüsi läbiviimiseks pole otseselt tarvis teha andmete kohta eelnevaid eeldusi ning kontrollida nende paikapidavust. Küll aga lihtsustab klasteranalüüsile eelnev põhjalik töö andmetega saadud tulemuste sisulist tõlgendamist ja võimaldab andmeid uurimisprobleemile paremini sobilikuks muuta. Oluline mõju klasteranalüüsi tulemustele on nii kaasatud muutujate relevantsusel, andmetes esinevate erinditel ja nende võimaliku mõju teadvustamisel ning uurimisprobleemist lähtuvate transformatsioonide läbiviimisel.

1.4.2. Sarnasuse määramine

Klasteranalüüsi tuumaks on objektidevahelise sarnasuse leidmine ning sellele tuginedes sarnaseid objekte koondavate klastrite loomine. Inimesed on intuitiivselt väga head klastrite loojad, suutes iseeneselegi märkamatult sarnaseid asju omavahel grupeerida ning nii oma maailma seeläbi lihtsustada. Inimesed on väga head klasterdajad kahe ja kolmemõõtmelises maailmas, kuid enamik sotsiaalteaduslikke probleeme avaldub kõrgemamõõtmelistes ruumides (Jain *et al.* 2012: 310). Sellises keskkonnas objektide omavahelise sarnasuse määramiseks vajatakse matemaatilist lahendust.

Objektidevahelist sarnasust määrates mõõdetakse tegelikult objektidevahelist kaugust (*Ibid.*: 314). Mida väiksem on kahe objekti vaheline kaugus, seda sarnasemad objektid omavahel on. Kuna kõiki objektidevahelisi sarnasusi mõõta üritavaid lahendusi ei saa kutsuda otseselt kauguse mõõduks, kasutatakse sagedamini mõisteid sarnasuse või erinevuse mõõdik (StataCorp. 2009: 85). Käesolevas magistritöös kasutatakse mõistet sarnasuse mõõdik. Eksisteerib sadu akadeemilistes väljaannetes avaldatud tehnikaid andmeobjektide omavahelise sarnasuse mõõtmiseks (Berry ja Linoff 2004: 360). Mõõdikuid on kohandatud vastavalt konkreetsele uurimisprobleemile ja andmetele. Kuna paljud mõõdikud on teiste sarnasusmõõdikute erijuhud, võib matemaatiliselt väita, et eksisteerib lõputu hulk erinevaid sarnasuse määramise mõõdikuid (StataCorp. 2009: 85–86).

Berry ja Linoff (2004: 360–363) tutvustavad kliendisuhete juhtimiseks läbiviidava klasteranalüüsi tarvis kolme eri tüüpi sarnasusmõõdikuid – objektidevahelisele geomeetrilisele kaugusele, vektoritevahelisele nurgale ja jagatud tunnustele tuginevaid sarnasusmõõte.

Objektidevahelise geomeetrilise kauguse korral tuginetakse andmete esitamisele numbriliselt n-dimensionaalses ruumis. Iga muutuja kohta eksisteerib üks dimensioon ning iga objekti on võimalik kujutada kui punkti selles ruumis. Kahe punkti vaheline kaugust selles ruumis väljendab nende vahelist sarnasust. Mida lähemal punktid üksteisele asuvad, seda sarnasemad objektid omavahel on. Kauguste mõõtmiseks geograafiliselt on mitmeid võimalusi. Levinuimad neist on eukleideline- ning absoluutkaugus.

Teiseks võimaluseks on vaadata objekte mitte kui punkte, vaid kui vektoreid. Vektoril on lisaks pikkusele olemas ka suund, mis võimaldab mõõta vektoritevahelist nurka. Kui absoluutväärtus annab ülevaate objektidevahelisest skaala erinevusest, siis suund näitab muutujatevahelist suhet. Nii on võimalik sarnaseks tunnistada objektid, mis on küll erinevate suurustega, kuid sarnaste muutujatevahelise suhtega. Sisuliselt tehakse vektoritevahelise nurga leidmise korral läbi ka andmete skaleerimise. Nii suudetakse tuvastada, et kuldkala on sarnasem haikalale kui hiirele, kuigi oma suuruse poolest sarnaneb kuldkala enam hiirele.

Kategooriliste muutujate korral ei ole aga geomeetrilised ja vektoritevahelised mõõdud parim valik. Võimaluseks on vaadata objektide poolt jagatud tunnuste määra erinevates dimensioonides. Mida enam kattuvaid muutujaid erinevates dimensioonides, seda sarnasemad on objektid.

Eelnevalt on juba mainitud, et eksisteerib arvukalt erinevaid sarnasuse mõõdikuid, mis on mõeldud erinevatele konkreetsete juhtude tarvis. Käesolevas magistritöös vaadatakse kvalitatiivseid andmeid ning vajadusel skaleeritakse muutujad andmeid ettevalmistavas faasis. Seega vaadatakse edaspidi lähemalt ainult geograafilisi sarnasuse mõõdikuid.

Enamik geograafilise sarnasuse mõõdikud on Minkowski kauguse erijuhud (StataCorp. 2009: 85–86). Minowski kaugus summeerib kahe objekti kõikide tunnuste erinevused ning lubab summat omakorda kohendada argumendi *p* võrra. Minkowski kaugus avaldub valemina järgmiselt:

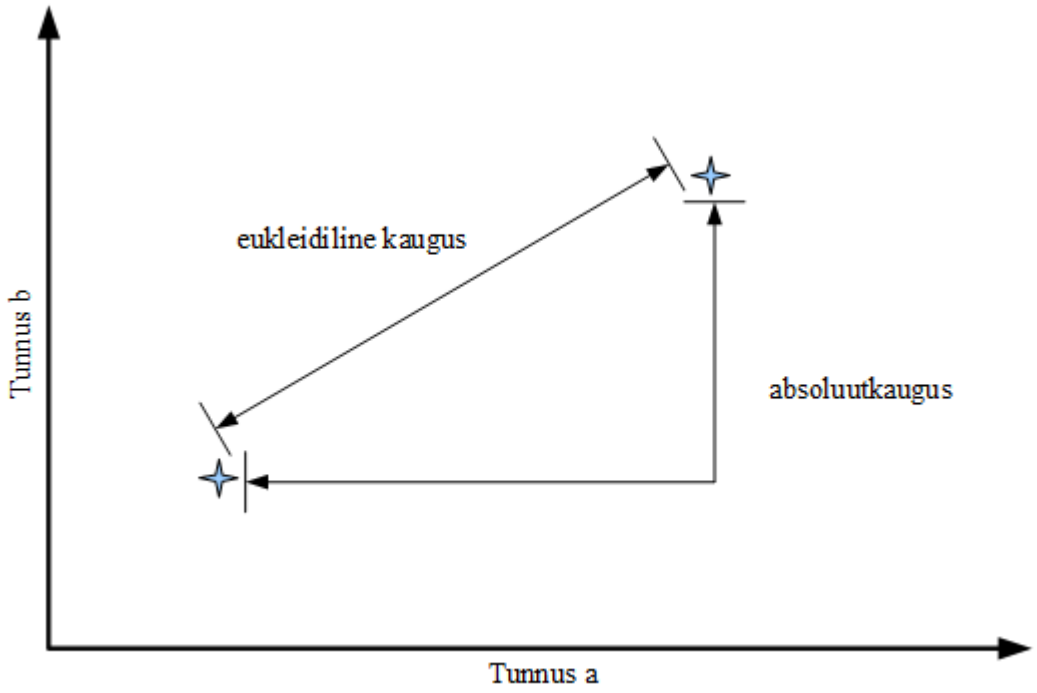
(1) $$\left(\sum_{i=1}^{n} | x_i - y_i |^p \right)^{1/p}$$

kus *p* – muudetav argument,
n – tunnuste arv,
x_i – objekti *X* tunnuse *i* väärtus,
y_i – objekti *Y* tunnuse *i* väärtus.

Asendades Minowski kauguses argumendi $p = 1$ saadakse tulemuseks absoluutkaugus. Absoluutkauguse valem avaldub matemaatiliselt järgmiselt:

(2) $$\sum_{i=1}^{n} | x_i - y_i |$$

Kuna klasteranalüüsi on arendatud erinevates teaduse valdkondades eraldiseisvalt on kõikidel kasteranalüüsi meetoditel hulganiselt sünonüüme (StataCorp. 2009: 502). Nii nimetatakse absoluutkaugust tihti ka *Cityblock* või *Manhattan* kauguseks. Meetod on oma sünonüümide nimed saanud seetõttu, et tehnika sarnaneb linnatänavate pikkuse mõõtmisega kahe punkti vahel liikudes. Meetod ei otsi mitte kahe punkti vahelist kaugust sirgjoones vaid pikki erinevaid muutujaid (vt. joonis 2).



Joonis 2. Vaatlustevahelise sarnasuse määramine geomeetriliselt (autori koostatud)

Asendades Minowski kauguses argumendi *p=2*, saadakse tulemuseks eukleideline kaugus. Eukleideline kaugus on intuitiivselt tajutav kahe ja kolmemõõtmelises maailmas, kus eukleideline kaugus kujutab endast sirget kahe uuritavat objekti kujutava punkti vahel (vt. joonis 2). Eukleideline kaugus avaldub matemaatiliselt järgmiselt:

(3) $$\left\{\sum_{i=1}^{n}(x_i - y_i)^2\right\}^{1/2}$$

Kasteranalüüsis toimib eukleideline kaugus hästi kui andmed on kompaktsed või asuvad isoleeritud klastrites. Statistika tarkvara Stata 11. versioon kasutab eukleideline kaugust vaikeseades sarnasuse mõõdikuna enamikele kvantitatiivsetele klasterdamistehnikatele

tuginedes mõõdiku lihtsusele ja intuitiivsusele, mis võimaldab ka saadud tulemusi lihtsalt tõlgendada (StataCorp. 2009: 504).

Mitemete klasterdamise tehnikate korral ei soovitata kasutada muud sarnasuse mõõdikut peale eukleidelise kauguse ruudus (StataCorp. 2009: 91). Eukleideline kaugus ruudus avaldub valemina järgmiselt:

(4) $$\sum_{i=1}^{n}(x_i - y_i)^2$$

Ruutu tõstetud eukleideline kaugus kannatab aga puuduse käes, et see sõltub tugevasti muutujate mõõtühikutest (Norušis 2011: 378), kuna ruutu tõstmise läbi suurendatakse suuremas skaalas mõõdetud muutuja rolli sarnaasuse määramisel veelgi.

Sarnasuse määramiseks on nii kirjanduses kui ka statistika tarkvara Stata 11. versioonis väljapakutud veel mitmeid erinevaid Minowski kauguse erijuhte. Tihti annavad need sama tulemuse, mis eukleideline kaugus või eukleideline kaugus ruudus, erinedes vaid tulemuseks saadud dendrogrammi kõrguste erinevuses (StataCorp. 2009: 90). Käesolevas magistritöös ei vaadata rohkemaid erijuhte lähemalt, kuna klasteranalüüsi tulemuste tõlgendamisel dendrogrammi kõrguste erinevuses üldjuhul informatsiooni ei lisa (StataCorp. 2009: 91).

Käesolevas magistritöös viiakse klasteranalüüs läbi kasutades pideval skaalal kvantitatiivseid andmeid. Seetõttu vaadeldi lähemalt enamlevinud Minkowski kauguse erijuhte. Kõikide Minkowski kauguse variatsioonide puuduseks on suurima skaalaga muutuja domineerimine klastrite moodustamisel. Absoluutkauguse korral on domineerimine küll väiksem, kuid siiski olemas. Seetõttu on andmete ettevalmistamise faasis oluline suure variatsiooniga muutujate domineerimise vähendamine läbi andmete teadliku transformeerimise.

1.4.3. Klasterdamise tehnikad

Objektide liitmiseks sarnasuse alusel, kasutades eelnevalt valitud sarnasuse mõõdikuid, on mitmeid erinevaid tehnikaid. Klasteranalüüsi tehnikad jagunevad hierarhilisteks ning mittehierarhilisteks. Mittehierarhilised klasterdamistehnikad on tänu väiksemale

ressursitarbele kiiremad ning lubavad analüüsida suuremahulisi andmebaase (StataCorp. 2009: 87). Hierarhilised algoritmid on aga palju mitmekülgsemad kui mittehierarhilised algoritmid. Hierarhilised tehnikad saavad paremini hakkama keerulise kujuga klastritega (Jain *et al.* 2012: 322) ning võimaldavad uurijale paremat ülevaadet analüüsi käigust (StataCorp. 2009: 87).

Hierarhiliste meetodite korral määratakse igas sammus objekt konkreetsesse klastrisse, kust objektil enam välja, teise klastrisse, liikuda pole võimalik. See omadus võimaldab kogu hierarhilise klasteranalüüsi protsessi visualiseerida puuna (Berry ja Linoff 2004: 370), kus iga objekt on justkui suure puu üks muutumatu haru. Ühelt poolt on klastrisse määramise lõplikkus hea, kuna võimaldab analüüsi järgselt uurida klastri tekkimist ning valida andmetele sobivaim klastrite arv, kuid teiselt poolt võib tekkida täiendavaid probleeme mürarikkas ja kõrgemõõtmelises ruumis (Tan *et al.* 2006: 526). Klastrid võivad hakata kujunema analüüsi tarvis ebasobivalt ning analüüsi käigus ei saa olukord enam paraneda.

Hierarhilised tehnikad omakorda jagunevad liitvateks ning jagavateks. Jagavad tehnikad alustavad ühest suurest klastrist, kuhu kuuluvad kõik objektid, mida hakatakse samm-sammult jagama väiksemateks klastriteks. Jagamine viiakse läbi eesmärgiga leida osad, mis on *„puhtamad“* kui esialgne klaster (Berry ja Linoff 2004: 371). *„Puhtus“* defineeritakse puhtuse funktsiooniga, mis kas, minimeerib keskmise klastritesisese erinevuse uutes klastrites või maksimeerib uute klastrite vahelise erinevuse (*Ibid:* 372). Kuna jagavad tehnikad on väga arvutusressursi mahukad, ei ole tegu levinud tehnikatega. Kirjanduses on neid mainitud vähe ning ka statistika tarkvara Stata 11. versioonil puuduvad jagavad hierarhilised klasterdamise tehnikad (StataCorp. 2009: 88). Seetõttu jäetakse käesolevas magistritöös jagavad tehnikad edaspidise vaatluse alt välja.

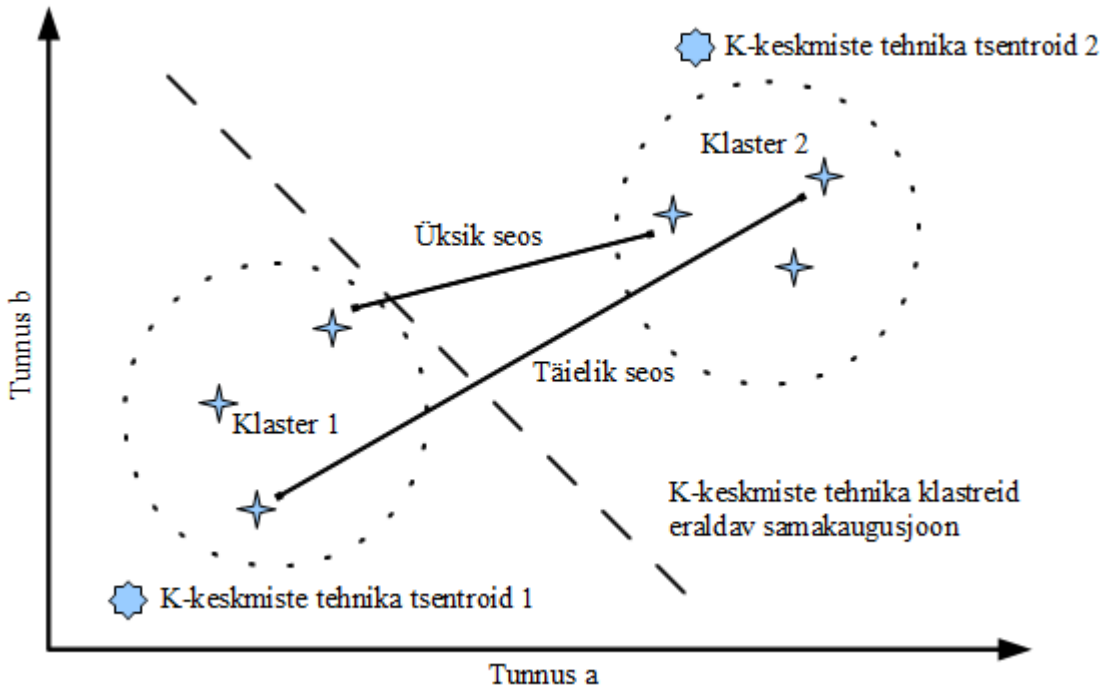
Liitvate tehnikate puhul on tegu sisuliselt ümberpööratud jagava klasteranalüüsiga. Protsessi alustatakse suure hulga klastritega, kus iga objekt moodustab omaette klastri. Samm-sammult liidetakse objektid omavahel aina suurenevatesse klastritesse. Protsess lõppeb ühe suure klastri, kuhu kuuluvad kõik objektid, moodustamisega. Protsessi alguses on klastrid väga väikesed ja omanäolised. Protsessi lõpu poole muutuvad klastrid suuremaks ning heterogeensemateks. Hierarhilistele tehnikatele omaselt, jääb kogu protsessi ajalugu uurijale hiljem järgitavaks. See võimaldab peale protsessi lõppu,

kui on moodustunud üks suur klaster, leida uurijal klasterdamise tase, mis sobib kõige paremini uurimisprobleemiga. (Berry ja Linoff 2004: 368)

Tehnika esimeseks sammuks sarnasuse maatriksi loomine. Sarnasuse maatriksisse leitakse kõikide objektide vahelised sarnasuse mõõdiku väärtused vastavalt eelnevalt valitud sarnasuse mõõdikule. Sarnasuse mõõdiku valib uurija lähtuvalt uurimisprobleemist.

Teise sammuna liidetakse omavahel uueks klastriks kaks omavahel kõige sarnasemat objekti ehk objektid, millede vaheline väärtus sarnasuse maatriksis on väikseim. Sarnasuse maatriksis asendatakse objektid uue tekkinud klastriga ning arvutatakse klastrile välja sarnasus kõigi ülejäänud objektidega. Objektide ja klastrite liitmist korratakse senikaua, kuni kõik objektid kuuluvad ühte klastrisse.

Kui objektidevaheline kaugus leitakse vastavalt valitud sarnasuse mõõdikule, siis selleks, kuidas leida kahest üksikust objektist tekkinud klastri sarnasust ülejäänud objektidega, on mitmeid võimalusi (vt. joonis 3). Selles kuidas leitakse liitmise käigus objektidest tekkinud klastrile sarnasus teiste objektiga, peitubki erinevate liitvate tehnikate erinevus (Tan *et al.* 2006: 516).



Joonis 3. Klastrite moodustamine erinevate tehnikate korral (autori koostatud).

Enim tuntud liitvad hierarhilised tehnikad on: üksik seos, täielik seos, keskmine seos, Wardi meetod, tsentroidi seos, mediaani seos ja kaalutud keskmine seos (StataCorp. 2009: 88). Enamik tehnikaid on erinevad üksiku seose, täieliku seose, keskmise seose ja Wardi meetodi edasiarendused (Jain *et al.* 2012: 320). Seetõttu vaadatkse magistritöös neid nelja tehnikat lähemalt.

Üksiku seose tehnika korral määravad klastri sarnasuse teiste klastritega objektid võrreldavatest klastritest, mis on omavahel kõige lähedamal (vt. joonis 3, lk 32). Klastritevaheliseks sarnasuseks loetakse nende objektide omavaheline sarnasus (StataCorp. 2009: 88). Oluliseks üksiku seose tehnikaga saadud tulemuste omaduseks on asjaolu, et iga objekt on klastri siseselt vähemalt ühe objektiga lähemalt seotud kui ükskõik millise objektiga klastrist väljas pool klastrit (Berry ja Linoff 2004: 369). Meetodi puuduseks on ketistumise efekti esinemine (StataCorp. 2009: 88). Ketistumise efekt kujutab endast pika välja venitatud klastri loomist, kus igal sammul liidetavad objektid on küll omavahel väga sarnased, kuid iga uus liidetud objekt võib oluliselt erineda esimestest klastrisse liidetud objektidest. Teatud uurimisprobleemide korral võib see aga olla just uurija poolt soovitud omadus. Nii sobib tehnika hästi väljavenitatud klastrite tuvastamiseks, millel on ebavõrdsed variatsioonid ja ebavõrdsed suurused (Norušis 2011: 388). Samuti on võimalik tehnikaga leida klastreid, mis on peidetud teiste klastrite sisse ja klastreid, mis neid ümbritsevad (Jain *et al.* 2012: 321). Samas on üksiku seose tehnika tundlik müra ja erindite suhtes, mis võivad anda tulemuseks omavahel vähe seotud objektidega klastrid (Tan *et al.* 2006: 519).

Täieliku seose tehnika kasutab klastritevahelise sarnasuse määramiseks, vastupidiselt üksiku seose tehnikale, objekte võrreldavatest klastritest, mis asuvad üksteisest kõige kaugemal (vt. joonis 3 lk. 32). Üheks klastriks liidetakse, sarnaselt üksiku seose tehnikaga, klastrid, millel võrreldavate objektide vaheline kaugus on kõige väiksem (StataCorp. 2009: 88). Sellise meetodiga saadakse tulemuseks klastrid, kus kõik klastrisse kuuluvad objektid asuvad üksteisest ühel kindlal maksimumkaugusel (Berry ja Linoff 2004: 369). Tegu on üksiku seose tehnikale vastanduva tehnikaga, mis saab tulemuseks väga kompaktsed ja mitteväljavenitatud klastrid (StataCorp. 2009: 88). Täieliku seose tehnika on vähem tundlik müra ja erindite suhtes, kuid võib lahutada väljavenitatud kujuga klastrid erinevateks klastriteks (Tan *et al.* 2006: 520). Üksiku

seose tehnika leiab tulemuseks küll mitmekülgsemaid klastreid kui täieliku seose tehnika, kuid pragmaatilisest vaatest lähtuvalt saadakse täieliku seose meetodiga tihti paremaid tulemusi reaalelus kasutamiseks (Jain *et al.* 2012: 321).

Keskmise seose tehnika on täieliku ja üksiku seose tehnika vahepealne lahendus (Tan *et al.* 2006: 521). Keskmise seose tehnika kasutab klastritevahelise sarnasuse määramiseks kõikide võrreldavatesse klastritesse kuuluvate objektide omavaheliste sarnasusmõõtude keskmist. Tegu on küllaltki robustse meetodiga, mis vähendab ketistumise efekti ohtu, kuid ei anna tulemuseks liialt kompaktseid klastreid (StataCorp. 2009: 88).

Wardi tehnika eesmärk on hoida klasterdamise käigus klastrid võimalikult homogeensed ja minimeerida klastrisisest variatsiooni. Esmalt leitakse kõikide muutujate keskmised väärtused. Seejärel leitakse igale objektile ruutu tõstetud eukleidiline kaugus klastri keskmisest. Omavahel liidetakse kõikide klastrisse kuuluvate objektide kaugused keskmisest ning omavahel liidetakse klastrid, mille liitmise tulemusena tekib väikseim võimalik suurenemine klastri siseste kauguste ruutude summas. (Norušis 2011: 387) Wardi meetod on väga sarnane keskmise seose tehnikale ning seda on võimalik näidata ka matemaatiliselt, kui keskmise seose tehnika objektidevahelise sarnasuse mõõtmise aluseks on eukleidiline kaugus ruudus (Tan *et al.* 2006: 523). Wardi meetod saab hästi hakkama sfäärjate klastrite tuvastamisega, kuid ei anna väga häid tulemusi eri suurustega gruppide tuvastamisel (StataCorp. 2009: 88). Eelnevalt vaadatud hierarhilise klasterdamise meetoditest erineb tehnika selle poolest, et klastrite liitmisel võib tekkida olukord, kus omavahel liidetavad klastrid on sarnasemad, kui eelnevalt liidetud klastrid (Tan *et al.* 2006: 523). Teiste käesolevas magistritöös tutvustatud hierarhiliste tehnikate korral on omavahel liidetavad klastrid üksteisest igal sammul aina enam erinevad.

Mittehierarhiliste klasterdamise meetodite korral ei jagata objekte samm-sammult lõplikesse klastritesse. Käesolevas magistritöös vaadatakse lähemalt kõige sagedamini kasutatavat klasterdamise tehnikat ja seega ka kõige sagedasemini kasutatavat mittehierarhilise klasterdamise tehnikat, milleks on K-keskmiste meetod. *„K“* K-keskmiste tehnika nimes viitab sellele, et otsitakse eelnevalt fikseeritud arvu *K* klastrit tuginedes objektide sarnasusele. (Berry ja Linoff 2004: 350)

K-keskmiste tehnikast on mitmeid versioone, mis peamiselt erinevad üksteisest esialgsete klastrite keskkoha valimise osas. Mõned üritavad leida koheselt võimalikult häid üksteisest eristuvaid stardipunkte, teised kasutavad juhuslikke objekte. (Jain *et al.* 2012: 325) Kõik K-keskmiste tehnikad koosnevad kolmest sammust, mida korratakse kuni lõpliku tulemuseni jõudmiseni.

Esimeses sammus valitakse andmete hulgast tehnikale K objekti stardipunktideks ehk seemneteks. Stardipunktideks on võimalik valida näiteks esimesed objektid andmebaasis või lasta tarkvaral määrata need juhuslikult. Kui andmetes on mingi tähenduslik järjekord, on mõistlik valida üksteisest võimalikult eristuvad andmepunktid. Igast seemest saab eraldiseisva klastri esimene element. (Berry ja Linoff 2004: 354) Kui uurijal on hea aimdus, millisteks klastrid kujunevad või millised on stereotüüpsed objektid, võib kasutada neid kui seemneid (Norušis 2011: 399).

Teises sammus määratakse iga ülejäänud objekt, vastavalt valitud sarnasuse mõõdikule, endale lähima seemne (tsentroidi) juurde ja seega sellega samasse klastrisse. Klastreid võib kujutada üksteisest eraldatuna piiriga, kuhu erinevate klastrite tsentroidist on sama pikk distants (vt. joonis 3 lk. 32).

Kui kõik objektid on määratud lähima seemne juurde ja saavutatud esialgne klasterdamine, arvutatakse kolmanda sammuna uutele klastritele tsentroidid. Leitud tsentroidid võetakse kasutusele seemnetena algoritmi järgmises tsüklis. Taaskord tehakse läbi eelnimetatud teine ja kolmas samm. Võib juhtuda, et mõni objekt liigub uue seemne tõttu teise klastrisse. Tsüklit korratakse ja tsentroide arvutatakse ümber kuni klastrite piirid enam ei muutu ning objektid ei vaheta enam klastreid. Praktikas leiab algoritm stabiilse lahendi peale tosinat tsüklit. (Berry ja Linoff 2004: 355).

K-keskmiste meetodi olulisemaks puuduseks asjaolu, et saadud tulemused ei anna informatsiooni selle kohta, mitu klastrit andmetes naturaalselt esineb. Klastreid moodustatakse täpselt nii palju, kui uurija on eelnevalt defineerinud. Erinevad *K* väärtused võivad viia väga erinevate tulemusteni, mis on kõik on iseenesest matemaatiliselt õiged. Kõik sõltub sellest, mille tarvis klasterdamist kasutatakse. Üheks võimaluseks seda probleemi ületada on K-keskmiste meetodit rakendada erinevate

eelnevalt valitud klastrite arvuga ning võrrelda seeläbi saadud tulemuste vastavust uurimisprobleemiga ning omavahel (*Ibid*.: 357).

K-keskmiste meetod on sarnaselt muudele klasterdamise tehnikatele tundlik erindite suhtes. Kui tehnika stardipunktideks valitakse üksteisest võimalikult erinevad objektid, eelistatakse stardipunktidena just erindeid (Tan *et al.* 2006: 503–504). See võib viia kompaktsete andmete jõulisele üksteisest eraldamiseni või leitakse tulemuseks üks suur keskne klaster, mida ümbritsevad mitmed väga väikesed klastrid. Selline tulemus võib olla kasulik, kui otsitakse pettusi või tootmisvigasid (Berry ja Linoff 2004: 358). Muudel juhtudel on aga mõistlik need erandjuhud andmetest eemaldada või andmeid skaleerida, kui seda varasemalt andmete ettevalmistuse faasis juba tehtud ei ole.

Erindid esindavad äärmuslikke objekte, mis kommertssuunitlusel läbiviidavates uuringutes võivad olla tihti enim huvipakkuvad. Kui sellised objektid pakuvad eraldi uurijale huvi, on võimalik erindid eemaldada ka analüüsi käigus. Leitud väikesed klastrid kujutavad endast tihti erindeid ning seetõttu on mõistlik need andmetest eraldada ning sooritada analüüs uuesti erinditele ja ülejäänud andmetele eraldi (Tan *et al.* 2006: 506). Sarnaselt erindite poolt loodud klastrite eraldi uurimisele, on võimalik eraldi uurida ka alles jäänud andmehulka. Erindite eemaldamine ülejäänud andmetest, võimaldab eemaldada erindite poolt loodud klastrite tõmbe ning seeläbi leida uusi klastreid andmetest, mis jäid alles. Nii muutuvad tulemused sisukamateks. (Berry ja Linoff 2004: 373)

K-keskmiste meetod on arvutuslikult kiire, kuid omab mitmeid puudusi. Tehnika leiab sfäärjaid klastreid ning ei saa hästi hakkama naturaalsete klastrite leidmisega kui neil ei ole sfäärjas kuju või on neil väga suuresti erinevad suurused (Tan *et al.* 2006: 510). Meetod ei saa hakkama ka omavahel kattunud klastrite tuvastamisega (Berry ja Linoff 2004: 365). Lisaks kipuvad tulemused takerduma lokaalsetesse optimumidesse (Jain *et al.* 2012: 339).

Neist puudustest saab mõnel määral üle. Lokaalsetesse optimumidesse takerdumist on võimalik vältida esialgsete stardipunktide põhjalikul valikul, valides stardipuntideks üksteisest piisavalt erinevad objektid ning vältides erindite kaasamist (Jain *et al.* 2012: 339). Kui uurija soovib tuvastada keerulisema kujuga klastreid, mis on erinevate

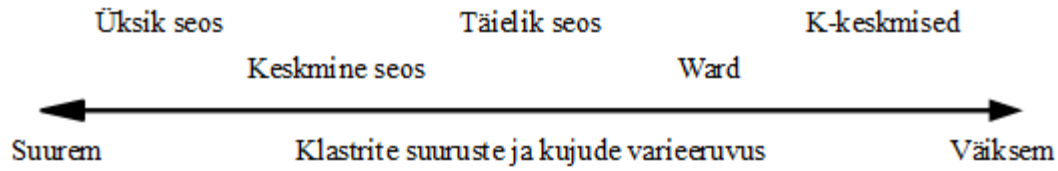
tiheduste ja kujudega, võib seda teha läbi suurema hulga klastrite leidmise. Selleks leiab uurija esmalt suura hulga klastreid ning liidab klastrid omavahel arvestades uurimisprobleemi eripärasid (Tan *et al.* 2006: 510). Näiteks on praktikas juuksurisalongi klientide käitumise analüüsil otsitud esialgu 11 üksteisest eristuvat klastrit ning seejärel grupeeritud need klastrid neljaks huvipakkuvaks kliendigrupiks (Wei *et al.* 2013: 7516).

K-keskmiste algoritmi korral võib juhtuda, et tulemuseks saadakse teiste hulgas ka tühi klaster. See juhtub kui peale uute tsentroidide arvutmist asuvad kõik objektid ülejäänud tsentroididele lähemal kui tühjaks jäänud klastri tsentroidile (Tan *et al.* 2006: 506). See võib meile inditseerida, et andmetes ei leidu nii palju sisukaid klastreid, kui oleme otsimas ning piirab liialt suure arvu klastrite otsimist. Samas võib tegu olla erindite negatiivse mõjuga klasterdamise protsessile.

K-keskmiste tehnikat on kasutatud väga laialdaselt paljude erinevate probleemide lahendamisel. Näiteks on leitud optimaalsed naiste rõivasuurused Ameerika Ühendriikide sõjaväele kasutades 3 000 naise 100 erinevat mõõtu ja saades tulemuseks mõõdud, mis ei ole tavapärased moest tuntud mõõtude järk-järgulised suurenemised, vaid erinevatele kehatüüpidele sobivad suurused (Ashdown 1998). Juuksurisalongide tarvis on K-keskmiste meetodil leitud klientide ostukäitumisele tuginedes kliendisegmendid (Wei *et al.* 2013) ning turismiettevõtted on kasutanud K-keskmiste tehnikat sisendi saamiseks uute reisipakettide väljatöötamiseks (Liao *et al.* 2010).

Erinevalt hierarhilistest klasterdamise tehnikatest, ei anna mittehierarhilised klasterdamise tehnikad, seega ka K-keskmiste tehnika, uurijale samm-sammulist ülevaadet klastrite kujunemisest. Kui uurijal on soov omada ülevaadet klastrite kujunemisest on sobivaks hierarhiline analüüs. K-keskmiste tehnika kohta on ülal välja toodud ka teisi puudusi, mille osas hierarhilised tehnikad on paremad. Hierarhiliste tehnikate puuduseks omakorda on aga limiteeritud andmehulk, mida saab töödelda, kuna tegu on väga ressursimahuka tehnikaga (Tan *et al.* 2006: 518). Probleemi lahendab erinevate tehnikate sümbioos.

Erinevad klasterdamise tehnikad käituvad klastrite moodustamisel erinevalt ning annavad erinevad tulemused. Magistritöös lähemalt vaadeldud tehnikate poolt leitud klastrite suuruste ja kujude varieeruvusest annab ülevaate joonis 4 (lk. 38).



Joonis 4. Erinevate klasterdamise tehnikate leitavate klastrite suuruste ja kujude varieeruvus (autori koostatud).

Kuna K-keskmiste tehnikal esineb probleeme klastrite arvu ja stardipunktide valimisel, on võimalik kasutada teisi tehnikaid nende määramiseks. Üks võimalus on kasutada hierarhilisi klasterdamise meetodeid andmetest võetud valimi baasil esialgsete klastrite leidmiseks ning kasutada nende esialgsete klastrite tsentroide K-keskmiste tehnika seemnetena. Hierarhiliste tehnikate korral puudub probleem algpunktide valimisega (Tan et al. 2006: 525) ja klastrite arvu on võimalik määrata tuginedes analüüsi käigule. Küll aga piirab hierarhilist analüüsi kasutatavate andmete maht. Seetõttu kasutatakse hierarhilises klasteranalüüsis väiksemamahulisi valimeid andmetest ning hiljem kaastatakse analüüsi K-keskmiste tehnikat kasutades suurem hulk andmed. See lahendus toimib hästi, kui hierarhilise klasteranalüüsi tarvis kasutatakse väikeseid valimeid, soovitatavalt paar sada kuni paar tuhat objekti, ja K-keskmiste tehnikaga leitavate klastrite arv on suhteliselt väike võrrelduna hierarhilises klasteranalüüsis kasutatud valimi suurusesse (Tan et al. 2006: 503).

Kõikidel klasterdamise tehnikatel on nii omad tugevused kui ka puudused. Seetõttu sobivad erinevad klasterdamise tehnikad erinevate klasterdamise probleemide lahendamiseks. Kliendisegmentidena kasutamiseks sobivad küllaltki võrdsete suurustega klastrid, kuna väga väikesed klastrid kujutavad endast kulukaid erilahendusi ootavaid kliendigruppe ning väga suurte klastrite korral ei saavutata piisavat personaalsust. Seetõttu osutub teooriale tuginedes sobivaimaks tehnikaks kliendisegmentide moodustamiseks K-keskmiste tehnika, mis annab tulemuseks kõige väiksema klastrite suuruste ja kujude varieeruvusega klastrid.

1.4.4. Klastrite hindamine ja kirjeldamine

Klasteranalüüsi eesmärk on andmehulga jagamine sarnaste objektidega gruppideks, mida on seejärel võimalik uurimisprobleemist lähtuvalt kasutada. Tulemuseks saadud klastrite juures huvitab uurijaid, mille poolest klastritesse kuuluvad objektid omavahel sarnanevad ning teistesse segmentidesse kuuluvatest objektidest erinevad. Paraku leiab peaaegu iga klasterdamise tehnika klastreid ka andmehulgast, kus naturaalseid klastreid tegelikult ei eksisteerigi (Tan *et al.* 2006: 532). Seega huvitab uurijat, kas leitud klastrite arv annab üldse parima võimaliku klasterdamise tulemuse. Seda saab otsustada lähtuvalt tulemuste sobivusest uurimisprobleemile, kuid eksisteerivad ka matemaatilised testid sobiva klastrite arvu määramiseks andmehulgas.

Parima klastrite arvu määramiseks kasutatakse klasteranalüüsi peatamise reegleid. Klasteranalüüsi peatamise reeglid leiavad igale võimalikule klastrite hulgale väärtuse, mille järgi otsustatakse, milline klastrite arv on andmehulgale sobivaim (StataCorp. 2009: 159). Klasteranalüüsi tehnikatele iseloomulikult eksisteerib kirjanduses rohkearvuliselt erinevaid klasteranalüüsi peatamise reegleid (*Ibid*.: 92). Milligan ja Cooper (2012: 246) võrdlevad omavahel 30 enimlevinud tehnikat ning leiavad, et ainult mõned tehnikatest saavad sobiva klastrite arvu määramisega hästi hakkama. Kaheks parimaks tunnistatakse Calinski-Harabaszi ja Duda-Hart indeksid. Calinski-Harabaszi indeks sobib kasutamiseks nii mittehierarhiliste kui ka hierarhiliste klasterdamise tehnikate korral, Duda-Hart töötab aga ainult hierarhilise klasteranalüüsi korral (StataCorp. 2009: 160). Calinski-Harabaszi ja Duda-Hart indeksid on ka ainsad klasteranalüüsi peatamise reeglid, mida pakub statistika tarkvara Stata 11. versioon (StataCorp. 2009: 160). Käesolevas magistritöös tutvustatakse neid tehnikaid lähemalt.

Calinski-Harabaszi indeks avaldub järgmisel kujul (Calinski ja Harabasz 1974: 10–11):

(5) $$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

kus *CH(k)* – Calinski-Harabaszi indeks väärtus,

B(k) – klastritevaheline varieeruvus,

W(k) – klastritesisene varieeruvus,

k – klastrite arv,

n – objektide arv.

Hästi eristuvatel klastritel on suur klastrite vaheline varieeruvus ja väike klastrite sisene varieeruvus. Seega on optimaalne klastrite arv väärtus *k*, mis maksimeerib funktsiooni *CH(k)*. Mida suurem on indeksi väärtus, seda selgemalt on klastrid üksteisest eristatud, väiksemad väärtused iseloomutavad omavahel vähem eristuvaid klastreid (StataCorp. 2009: 165). Indeksi väärtust nimetatakse vahel ka Calinski–Harabaszi pseudo-F väärtuseks (*Ibid*.: 163).

Calinski–Harabaszi indeksi eeliseks on tema sobivus nii hierarhilisele kui ka mittehierarhilisele klasteranalüüsile. Samuti leiab indeks globaalse optimumi, kuna indeksi arvutusse on kaasatud kõik objektid (StataCorp. 2009: 165).

Üldjuhul Calinski-Harabaszi indeksi väärtus klastrite arvu suurenemisel väheneb ning seetõttu saavutatakse parim tulemus kõige väiksema arvu klastrite korral. Turu segmenteerimiseks vajame aga rohkem kui ühest klastrist koosnevat lahendit. Seetõttu on mõistlik uurida Calinski-Harabaszi indeksi väärtuse tippe – väärtusi, mis on samaaegselt suuremad, kui väiksema ja suurema klastrite arvuga lahendi korral. Need inditseerivad andmetes peituvat struktuuri ja viitavad loomulikele klastritele. Kui sellist tippu ei eksisteeri, ei ole andmete struktuurist tulenevalt alust ühe lahendi eelistamiseks teisele.

Hierarhilise klasteranalüüsi tarvis on võimalik kasutada ka Duda-Hart indeksit. Duda-Hart indeks võrdleb omavahel klastrit ning kahte selle alamklastrit ning annab hinnangu, kas parem on ühine või jagatud klaster. Täpsemalt jagatakse kahest klastrist koosneva lahendi klastrite keskmisest arvestatud ruutvigade summa, ruutvigade summaga olukorras, kus tegu on ühe klastriga, millel on üks keskmine. Protseduuri viiakse läbi ainult selle andmeosa ulatuses, kus toimub klastrite liitmine. Suurem Duda-Hart indeksi väärtus inditseerib paremat lahendi. (Milligan ja Cooper 2012: 240) Matemaatiliselt avaldub indeks järgmiselt:

(6) $$DH = \frac{Je(2)}{Je(1)}$$

kus DH – Duda-Hart indeks väärtus,

$Je(1)$ – ruutvigade summa jagamisele minevas grupis,

$Je(2)$ – ruutvigade summa kahes tulemuseks saadavas alamklastris.

Duda-Hart indeksist saab tuletada ka pseudo-T^2 väärtuse, mis avaldub järgmiselt (StataCorp. 2009: 166):

(7) $$\frac{1}{Je(2)/Je(1)} = 1 + \frac{T^2}{N_1 + N_2 - 2}$$

kus T^2 – Duda-Hart pseudo-T^2 väärtus,

N_1 – objektide arv esimeses alamgrupis,

N_2 – objektide arv teises alamgrupis.

Väike Duda-Hart pseudo-T^2 väärtus inditseerib üksteisest eristuvaid klastreid. Mida suuremaks muutub väärtus, seda vähem klastrid üksteisest eristuvad. Kuna Duda-Hart indeks tegeleb lahendi, kas parem on ühendklaster või kaks alamklastrit, hindamisega ja ei kasuta selleks informatsiooni ülejäänud objektide kohta, on tegu lokaalse hinnanguga, mis ei pruugi kokku langeda globaalse optimumiga. (StataCorp. 2009: 165) Seetõttu sobib indeks pigem klasterlahenduse sobivuse kinnitamiseks kui parima lahendi leidmiseks. Näiteks võib leida huvipakkuva lahendi Calinski–Harabaszi indeksit kasutades ja lahendi paikapidavuse uurimiseks kasutada Duda-Hart indeksit ja pseudo-T^2 väärtust.

Tarkvarapaketi SAS/STAT kasutusjuhend (SAS Institute Inc. 2013: 227) soovitabki leida konsensus erinevate lahendite statistikute vahel uurides Calinski–Harabaszi indeksi tippe täiendavalt Duda-Hart indeksi ja pseudo-T^2 väärtuse abil. Duda-Hart indeksi suur väärtus ja väike pseudo-T^2 väärtus, millele järgneb suurem väärtus klastrite liitmisel, sobib hästi kompaktsete naturaalsete klastrite esinemise määramiseks.

Klasteranalüüsi läbiviimisel on tulemuseks saadud sarnaste objektidega klastrid, on järgmiseks ülesandeks välja uurida, mis teeb need klastrid eriliseks. Kas on võimalik leida reegleid, mis kirjeldavad neid klastreid ning mille poolest erinevad klastrid teistest klastritest?

Lihtsaim viis on uurida klastri objektide iga muutuja keskmist väärtust ning võrrelda seda kogu andmekogumi muutuja keskmisega ning seejärel järjestada muutujad erinevuse suuruse või Z-skoori alusel. Nende muutujate uurimine, mis näitavad suurimat erinevust ülejäänud andmegrupis, aitab juba suuresti lahti seletada klastri

olemust. Lisaks muutujate väärtuste keskmisele erinevusele, võib vaadata ka suhteid teistesse muutujatesse. (Berry ja Linoff 2004: 373).

Klastrite keskmiste ja tsentroidi kirjeldamine on levinuim viis klastri kirjeldamiseks. See sobib hästi sfäärjatele klastritele, kuid ei ütle palju väljavenitatud klastrite kohta. Sellisel juhul annab parema tulemuse klastri äärmiste punktide või piiride kirjeldamine. (Jain *et al.* 2012: 330)

Kui klasterdamise käigus ei kasutatud kõiki objektide kohta teada olevaid tunnuseid objektide grupeerimiseks, võime uurida neid tunnuseid ka peale klastrite moodustamist. Nii näiteks on võimalik leida klastrid uurijat huvitavate omaduste alusel ning seejärel otsida muid tunnuseid, mis selliste objekte klastreid kirjeldavad. Selline lähenemine sobib hästi, kui puudub võimalus koguda objektide kohta andmeid uurijat konkreetselt huvitavate omaduste osas, kuid teatakse teisi tunnuseid, mis sarnaste objektide klastritele on omased. Samuti võimaldab selline lähenemine sooritada klasteranalüüsi ainult pidevaid andmeid kasutades ning hiljem kategoorilisi tunnuseid uurides neid klastreid iseloomustada. Näiteks on ühe Taiwani juuksurisalongi kliendibaasi segmenteerimisel kasutatud klientide kohta teada olevaid pidevaid muutujaid klasterdamise läbiviimiseks ning seejärel on kaasatud analüüsi klientide kohta teada olevad kategoorilised tunnused klastrite paremaks iseloomustamiseks (Wei *et al.* 2013: 7515–7516). Klasterdamine viiakse läbi kasutades kliendi viimasest külastusest möödunud aega, külastussagedust ning ostusummat. Seejärel on uuritud klastrisse kuuluvate klientide kohta teada olevaid muid muutujaid, nende poolt soetatud tooteid, ning leitud tooted, mis esinevad teatud klastrites sagedamini kui teistes. Selline teadmine võimaldab paremini kirjeldada leitud klastreid ning annab võimaluse uusi kliente hiljem nende ostuhuvi alusel klassifitseerida sobivasse kliendisegment.

Klasteranalüüsi ülesandeks käesolevas magistritöös on välja selgitada loomulikult eksisteerivad kliendigrupid ilma nende kohta eelnevalt eeldusi tegemata. Seetõttu on mõistlik kliendisegmentide arvu määramiseks kasutada matemaatilisi teste. Kõikide magistritöös lähemalt tutvustatud klasterdamise tehnikate korral on naturaalsete klastrite arvu tuvastamiseks andmetes võimalik kasutada Calinski–Harabaszi indeksit ning hierarhiliste tehnikate korral ka Duda-Hart indeksit. Kuna kliendisegmentide tarvis otsitakse küllaltki sfäärijaid klastreid, sobib klastrite kirjeldamiseks neisse kuuluvate

objekte muutujate keskmiste väärtuste võrdlemine andmekogumi keskmisega. Täiendava võimaluse klastrite iseloomustamiseks annab klastrite kirjeldamine ka klasterdamisse kaasamata jäetud kategooriliste tunnuste abil.

2. LOOMULIKE KLIENDISEGMENTIDE LEIDMINE

2.1. Tarbimiseelistusi iseloomustavad andmed

Kasumit taotleva organisatsiooni eesmärk on rahaliselt mõõdetava väärtuse loomine. Rahalist väärtust loovad organisatsioonid oma klientidele tasu eest hüviste pakkumisega. Organisatsioonide klientideks on nende poolt pakutavate hüviste tarbijad ning klientide eelistusi kajastavad kõige paremini nende poolt tehtud tarbimisotsused.

Käesolevas magistritöös vaatame kliendina leibkondi kui organisatsioonide poolt pakutavate hüviste lõpptarbijaid. Leibkondi kasutatakse indiviidi asemel, kuna leibkond esindab ühiseid tarbimisotsuseid tegevate indiviidide üksust. Leibkonna tasandil tehakse tarbimisotsused ka indiviidide eest, kes küll tarbivad organisatsioonide poolt pakutavaid hüviseid, kuid ei suuda, ei saa või ei oska oma eelistusi väljendada (näiteks teovõimetud isikud ja lapsed). Samuti tehakse leibkonna tasandil ühised otsuseid suuremate kulutuste osas (elamispinna valik, kestvuskauba soetamine jne.). Magistritöös kasutatakse LEU 2012. aasta (edaspidi LEU 2012) andmeid Eesti leibkondade tarbimiskulutuste kohta (Leibkonna eelarve ... 2013). LEU on Statistikaameti poolt riikliku statistika seaduse alusel läbiviidav läbilõikeline ebaregulaarne uuring, mis annab ülevaate Eesti leibkondade tarbimiskulutustest leibkonnaliikme kohta kalendriaastas (Tikva ja Arnik 2012: 5). LEU eesmärgiks on anda usaldusväärne ülevaade Eesti leibkondade kulutuste ja tarbimisharjumuste kohta. Täiendavalt pakub uuring infot leibkondade põhiliste sotsiaalsete näitajate, elamistingimuste ja demograafilise koosluse kohta. Seetõttu kasutatakse LEU tulemusi eestimaalaste tarbimiskulude analüüsimisel ja kulutuste trendide jälgimisel. (*Ibid*: 5) Toodete ja teenuste pakkujatel võimaldab uuring saada infot selle kohta, kuidas jagunevad Eesti tarbija tarbimiskulutused ning mis iseloomustab erinevate tarbimiseelistustega leibkondi. Käesoleva magistritöö empiirilise osa ülesandeks ongi uurida Eesti tarbijate loomulikku struktuuri, moodustades leibkonna tarbimiskulutustele tuginedes võimalikult homogeensete tarbimiseelistustega

segmendid. Nende segmentide iseloomustamine annab aimu segmentide võimalikest tekkepõhjustest ning aitab tarbijale suunatud organisatsioonidel korraldada tulemuslikumat kliendisuhete juhtimist.

LEU 2012 sobib Eesti turu üldistavaks segmenteerimiseks hästi, kuna uuringu üldkogumisse kuuluvad kõik tavaleibkondades elavad uuringu aasta 1. jaanuari seisuga vähemalt 15-aastased Eesti alalised elanikud. Üldkogumit esindava loendina kasutab LEU siseministeeriumi hallatavat Eesti rahvastikuregistrit. Isikute valikuks valimisse kasutatakse mitteproportsionaalset süstemaatilist kihtvalikut. Selle valiku puhul jagatakse üldkogum kattumatuteks osadeks ehk kihtideks. Igas osas tehakse teistest osadest sõltumatu süstemaatiline valik, rakendades kihtides erinevaid kaasamistõenäosusi. (Tikva ja Arnik 2012: 13)

Leibkonna kulutuste liigitamisel kasutatakse Euroopa Liidu statistikaameti väljatöötatud tarbimiskulutuste klassifikaatorit *COICOP-HBS*, mis jagab tarbimiskulutused kaheteistkümnesse alagruppi (*Ibid*: 24):

1) toit ja alkoholita joogid,
2) alkoholsed joogid ja tubakatooted,
3) rõivad ja jalatsid,
4) eluase,
5) majapidamiskulud,
6) tervishoid,
7) transport,
8) side,
9) vaba aeg,
10) haridus,
11) restoranid ja hotellid,
12) mitmesugused kaubad ja teenused.

LEU materjalideks on ankeet ja tarbimispäevik. Ankeedis uuritakse leibkonna liikmete erinevaid tausttunnuseid, sissetulekut, kulutusi üle 100 eurot maksnud kaupadele ja teenustele viimase 12 kuu jooksul ning sise- ja välisreisidel käimisi ning neil tehtud kulutusi. Tarbimispäevikut täidetakse kahenädalase perioodi jooksul ning päevikute täitmise asemel on lubatud uuringusse valitud leibkonnal jätta kahenädalase perioodi

ostutšekid. (Tikva ja Arnik 2012: 11) Puudu olevad tarbimispäevikud imputeeritakse ankeetidele tuginedes lähima naabri meetodiga (*Ibid*: 24). Aasta jooksul kogutud andmed üldistatakse terve aasta tarbimiskulutusteks.

LEU uuringu tulemuste tõlgendamisel tuleb tähele panna, et uuritakse leibkonna liikmete keskmist tarbimist, mitte individuaalset tarbimist. Seetõttu uuritakse magistritöös leibkonna, mitte indiviidi tarbimiskulutusi.

LEU 2012 andmefail koosneb 9080 vaatlusest 75 muutuja kohta. Iga vaatlus kujutab endast ülevaadet ühe leibkonna demograafilistest, geograafilistest ja sotsiaalmajanduslikest tunnustest ning ankeeti täitnud isiku isiklikest tunnustest. Muutujad on jagatud nelja gruppi. Esimene annab ülevaate leibkonna liikmetest, teine uuringule vastanud isikust, kolmas leibkonnast tervikuna ning neljas tarbimiskulutustest leibkonna liikme kohta aastas.

Magistritöös kasutatakse klientide eelistuste kirjeldamiseks nende tarbimiskulusid Euroopa Liidu statistikaameti väljatöötatud tarbimiskulutuste klassifikaatori COICOP-HBS alusel 12 kulugrupis. Kuna Euroopa Komisjon kohustab riiklikel statistikaametitel imputeerida kõik sissetulekutega seotud tunnuste puuduvad väärtused (*Ibid*: 24), siis tarbimiskulutuste kohta andmefailis puuduvaid väärtusi ei eksisteeri.

Kõik segmenteerimise aluseks võetavad tunnused on pidevad ning mõõdetud samas mõõtühikus, milleks on *euro*. Kuna kõik muutujad on mõõdetud samas mõõtühikus, pole põhjendatud andmete skaleerimine. Läbi skaleerimise võrdsustataks muutujate kaalud ning kaotataks andmetes juba esialgu eksisteerivad loomulikud klientide eelistusi esindavad kaalud.

Kuna tegemist on tarbimiskulutustega, ei saa tunnustel olla negatiivseid väärtusi. Tunnuste väärtused on ülalt piiramata, seega hajuvad vaatlused suuremate väärtuste suunas ning esineb erindlikke vaatlusi. Kuna need erindid kujutavad endast väga suurte tarbimiskulutustega leibkondi, on nad huvipakkuvad. Seetõttu erindeid andmetest ei eraldata.

Tarbimiskulud korreleeruvad liigiti väga tugevalt ning olulisuse nivool 0,05 statistiliselt oluliselt leibkonna tarbimise kogukuluga, kuna suurem eelarve võimaldab leibkondadel

kõigele rohkem kulutada (vt. lisa 1). Suuremad tarbimiskulutused ei pruugi näidata erinevaid tarbimiseelistusi, vaid kajastavad pigem leibkonna suuremat tarbimisvabadust. Paremini selgitab kliendi eelistusi kulutuste osakaal tema kogukulutustest. Seetõttu transformeeritakse erinevad tarbimiskulud erinevate tarbimiskulutuste osakaaluks leibkonna tarbimise kogukulust. Nii saadakse tunnustele väärtused vahemikus 0 kuni 1. See transformatsioon muudab tunnused ülevaatlikumaks ning pehmendab samaaegselt erindite mõju.

Tarbimiskulutuste transformeerimine tarbimiskulutuste osakaaluks tarbimiskulude kogusummast ei eemalda terves ulatuses korrelatsiooni tarbimise kogukulu ja erinevate tarbimiskulutuste osakaalude vahel (vt. lisa 2). Tarbimise kogukulu on positiivselt korreleeritud transpordikulutuste ja negatiivselt toidu ning eluasemekulutuste osakaaluga tarbimise kogukulust. Toidu ja eluasemekulutusi võib nende vältimatuse tõttu nimetada sundkulutusteks. Erinevatesse sissetulekute gruppidesse kuuluvatel leibkondadel on erinevad võimalused tarbida ning tarbimiskulutuste eelarve kasvades väheneb sundkulutuste osakaal tarbija eelarves. Sellest lähtuvalt on mõistlik uurida tarbimiseelistusi erinevatel tarbimiskulutuste tasemetel. Andmebaasi jagamist väiksemateks gruppideks toetab ka asjaolu, et hierarhilise kasteranalüüsi läbiviimiseks sobilik vaatluste hulk on arvutusressursi mahukusest tulenevalt soovitatavalt kuni paar tuhat vaatlust (Tan et al. 2006: 503).

Käesoleva magistritöö autor on otsustanud leibkonnad jagada tarbimiskulutuste summa alusel nelja kvartiili, igaühes 2270 vaatlust. Tarbimiskulud leibkonna liikme kohta aastas vahemikus 112 kuni 1704 eurot moodustavad esimese kvartiili, vahemikus 1704 kuni 2641 eurot teise kvartiili, vahemikus 2641 kuni 3970 eurot kolmanda kvartiili ning vahemikus 3970 kuni 26460 eurot neljanda kvartiili. Keskmised tarbimiskulude osakaalud leibkonna tarbimise kogukulust valimis tervikuna ning kvartiilides on toodud tabelis 1 (lk. 48), kolm kõige suurema osakaaluga kululiiki on tähistatud halli taustaga. Tabelist selgub, et tarbimise kogukulu suurenedes vähenevad toidule ja mittealkohoolsetele jookidele ning eluasemele tehtavate kulutuste osatähtsused tarbimise kogukulust ning kasvavad kõikide teiste tarbimiskulutuste liikide osatähtsused, välja arvatud alkoholsetele jookidele ja tubakale tehtavad kulutused, mille

osatähtsus on kõrgeim II kvartiilis. Samasuunalisele seosele viitavad ka muutujatevahelised korrelatsioonid (vt. lisa 2).

Tabel 1. Tarbimiskulutuste osakaalud leibkonna tarbimise kogukulust.

Tarbimiskulutuste liik	**Kogu valim**	**Kogu tarbimiskulu kvartiil**			
		I	**II**	**III**	**IV**
Toit ja mittealkohoolsed joogid	0,327	0,384	0,356	0,320	0,246
Alkohoolsed joogid ja tubakas	0,041	0,038	0,045	0,042	0,038
Riided ja jalanõud	0,039	0,019	0,035	0,047	0,058
Eluase	0,203	0,277	0,218	0,180	0,135
Majapidamiskulud	0,049	0,032	0,043	0,053	0,067
Tervishoid	0,033	0,022	0,029	0,040	0,041
Transport	0,098	0,035	0,075	0,108	0,172
Side	0,067	0,097	0,072	0,058	0,044
Vaba aeg	0,079	0,058	0,072	0,081	0,104
Haridus	0,005	0,002	0,003	0,005	0,008
Hotellid, kohvikud, restoranid	0,017	0,005	0,010	0,018	0,034
Mitmesugused kaubad ja teenused	0,044	0,033	0,042	0,049	0,053

Allikas: (Leibkonna eelarve ... 2013); autori arvutused.

Kuna leibkonna tarbimiskulud kokku moodustavad terviku, peab ühe kululiigi osakaalu suurenemine kaasa tooma mõne teise kululiigi osakaalu vähenemise. Seetõttu on mõistlik uurida klasteranalüüsi kaasatud muutujate vahelisi korrelatsioone. Tugevamaid seoseid kui $r = 0,2$ ($p < 0,05$) esineb I kvartiilis rohkem kui ühe muu kululiigiga järgmistel kululiikidel: toit ja mittealkohoolsed joogid; eluase; transport (vt. lisa 3). Kulutused toidule ja mittealkohoolsetele jookidele; eluasemele ning transpordile omavad suurimaid osakaale leibkondade tarbimise kogukulust (vt. tabel 1). Sellele tuginedes on oodata klastrite kujunemist I kvartiilis eelmainitud tunnuste alusel. II kvartiilis on tugevamaid seoseid kui $r = 0,2$ ($p < 0,05$) rohkem kui ühe muu muutujaga lisaks I kvartiilis välja toodud tunnustele ka tunnusel vaba aeg (vt. lisa 4). III kvartiili muutujate vahelised korrelatsioonid viitavad klastrite kujunemisele sarnaselt I kvartiilile tunnuste toit ja mittealkohoolsed joogid, eluase ning transport alusel (vt. lisa 5). IV kvartiilis on tugevamaid seoseid kui $r = 0,2$ ($p < 0,05$) rohkem kui ühe muu muutujaga vaid tunnustel toit ja mittealkohoolsed joogid ning transport (vt. lisa 6). Küll aga esineb muid statistiliselt olulisi seoseid, mille korrelatsioonikordaja absoluutväärtus jääb pisut alla *0,2*. Seetõttu on oodata IV kvartiilis kõige mitmekülgsemat klastrite kujunemist.

Eelolevale tuginedes on kogu magistritöö ulatuses oodata klastrite kujunemist põhiliselt toidu ja mittealkohoolsete jookide; eluaseme ja transpordikulutuste osakaalude alusel.

2.2. Eesti leibkondade segmenteerimine

Leibkondade segmenteerimine viiakse läbi tarbimiseelistuste alusel klasteranalüüsi kasutades neljas eraldiseisvas andmegrupis. Andmegrupid on moodustatud LEU 2012 andmebaasi vaatluste jagamisel nelja kvartiili tarbimise kogukulu alusel. Kuna madalamates kvartiilides on sundkulutuste osakaal kõrgem, on oodata erinevates kvartiilides erinevat klastrite kujunemist.

Igas kvartiilis viiakse klasteranalüüs läbi kasutades nelja enimlevinud hierarhilise klasterdamise tehnikat: üksiku seose, täieliku seose, keskmise seose ja Wardi tehnikat. Täiendavalt kasutatakse mittehierarhilise tehnikana K-keskmiste tehnikat. K-keskmiste tehnika korral valitakse tehnika alguspunktideks juhuslikult andmetest valitud *K* valimi keskmised. Kui magistritöö autoril on alust arvata, et hierarhiliste tehnikatega leitud klastrite keskpunktid võivad parandada K-keskmiste tehnika tulemusi, seda ka proovitakse.

Klastrite moodustamise eesmärgiks on üksteisest eristuvate turusegmentide leidmine, seega otsitakse klastreid, mis on võimalikult kompaktsed ehk sfäärjad ning üksteisest eristuvad. Selliste klastrite tuvastamiseks andmetes sobib hästi Calinski-Harabaszi indeks. Erinevate tehnikatega leitud igale klasterdamise tulemustele arvutatakse Calinski-Harabaszi indeksi väärtus. Magistritöös võrreldakse erinevate tehnikatega leitud lahendeid, mis koosnevad kuni kümnest klastrist, kuna suurema klastrite arvu korral, ei pruugi saadud lahendid olla piisavalt ülevaatlikud. Indeksi kõrgeim väärtus viitab parimale klasterdamise tulemusele ning sobivamale tehnikale ja lahendile. Kuna üldjuhul Calinski-Harabaszi indeksi väärtus klastrite arvu suurenemisel väheneb, saavutatakse sageli indeksi kõrgeim väärtus kõige väiksema arvu klastrite korral. Turu segmenteerimiseks vajame aga rohkem kui ühest klastrist koosnevat lahendit. Seetõttu uuritakse parimate lahenditena loomulikke klasterlahendusi näitavaid Calinski-Harabaszi indeksi väärtuste tippe – väärtusi, mis on samaaegselt suuremad kui väiksema ja suurema klastrite arvuga lahendite korral. Kui sellist tippu ei suudeta tuvastada, ei ole

alust konkreetse tehnikaga leitud ühe lahendi eelistamiseks teisele Calinski-Harabaszi indeksi alusel. Parim klastrite arv määratakse sellisel juhul parima alternatiivse tehnikaga tuvastatud klasterdamise tasemel ning võrreldakse sel tasemel Calinski-Harabaszi indeksi väärtuseid tehnikate omavahelise võrdluse tarvis. Hierarhiliste tehnikate korral kasutatakse täiendavalt Duda-Hart indeksit ja pseudo-T^2 väärtust kinnitamaks Calinski-Harabaszi indeksile tuginedes leitud parimaid lahendeid.

Kuna Wardi tehnika korral ei soovita kirjandus kasutada muid sarnasuse mõõdikuid, kui eukleidiline kaugus ruudus (StataCorp. 2009: 91) ning magistritöös soovitakse erinevate klasterdamistehnikatega saadud tulemusi omavahel võrrelda, kasutatakse käesolevas magistritöös sarnasuse mõõduna eukleidilist kaugust ruudus. Ruutu tõstetud eukleidiline kaugus kannatab puuduse käes, et see sõltub tugevasti muutujate mõõtühikutest (Norušis 2011: 378) ja annab läbi ruutu tõstmise suuremal skaalal mõõdetud muutujatele veelgi olulisema rolli objektidevahelise sarnasuse kujundamisel. Kõnealune puudus ei ole käesolevas magistritöös probleemiks, kuna tarbimiskulutused on mõõdetud kõik samas mõõtühikus ning viidud läbi osakaaludeks transformeerimise samale skaalale vahemikus 0 kuni 1.

Parimaks tunnistatud klasterdamise lahendit kirjeldatakse kui kvartiilis esinevaid loomulikke turusegmente. Segmente kirjeldatakse läbi segmendi tarbimiskulutuste võrdlemise kvartiili keskmiste tarbimiskulutustega. Keskmiste alusel tulemuste võrdlemine on piisav, kuna tulemuseks oodatakse küllaltki sfäärjaid klastreid. Täiendavalt uuritakse segmentide kirjeldamiseks ning nende olemuse selgitamiseks erinevaid LEU 2012 andmetes leiduvaid demograafilisi, geograafilisi, sotsiaalmajanduslikke ja isiksuse tunnusteid, mida klasterdamise läbiviimiseks ei kasutatud.

2.2.1. Tabrmise kogukulu I kvartiil

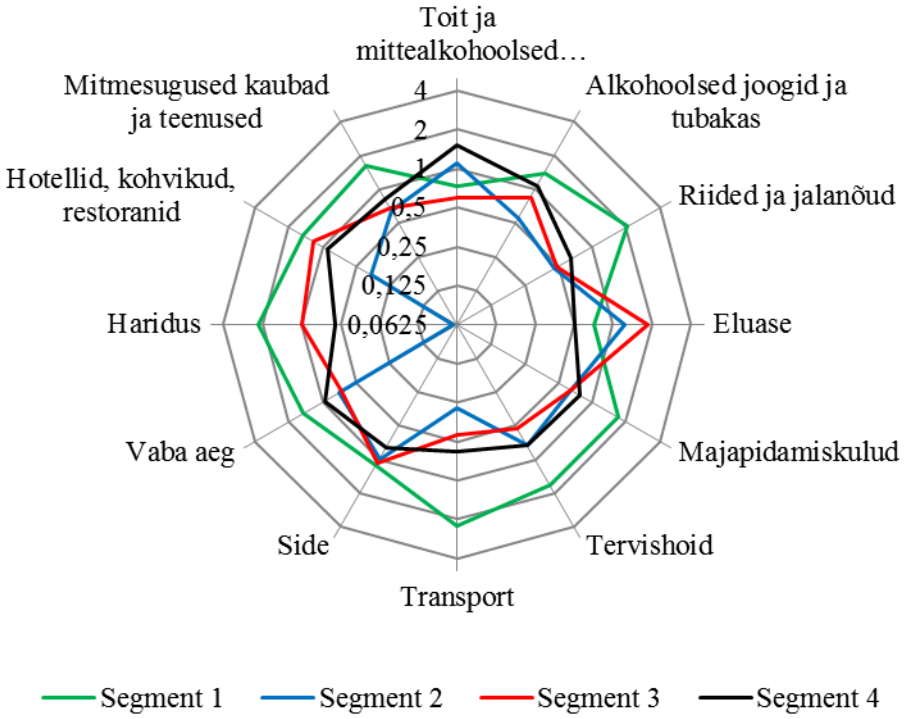
Klasteranalüüs viiakse läbi erinevaid tehnikaid rakendades. Erinevatele lahenditele leitud Calinski-Harabaszi indeksi väärtused on koondatud alljärgnevasse tabelisse 2.

Tabel 2. Calinski-Harabaszi indeksi väärtused erinevatele lahenditele I kvartiilis.

Klastrite arv	**Calinski-Harabasz'i indeksi väärtus**				
	Üksik seos	**Keskmine seos**	**Täielik seos**	**Ward**	**K-keskmine**
2	14,56	40,95	-	785,83	907,14
3	9,63	41,64	335,85	635,74	759,9
4	7,69	46,48	**389,95**	520,75	604,43
5	9,05	54,89	341,58	468,87	566,58
6	13,28	48,92	345,96	428,23	500,1
7	12,15	41,39	296,76	405,33	472,77
8	13,15	77,99	256,86	390,74	446,74
9	15,01	87,72	237,74	380,76	425,16
10	13,8	86,89	228,02	370,63	417,27

Allikas: (Leibkonna eelarve ... 2013); autori arvutused.

Tabelist 2 selgub, et parim lahend üksiku ja keskmise seose tehnika korral on üheksast klastrist koosnev ning täieliku seose tehnika korral neljast klastrist koosnev. Wardi ja K-keskmiste tehnikat kasutades ei leidu ühtegi lahendit, mille korral Calinski-Harabaszi indeksi väärtus oleks üheaegselt suurem nii väiksema kui ka suurema arvuga klastritega lahendi indeksi väärtusest. See ei anna alust üht lahendit teisele eelistada. Leitud parimatest lahenditest on kõrgeim Calinski-Harabaszi indeksi väärtus täieliku seose tehnikaga leitud nelja klastri korral. Lokaalse maksimumi asumist nelja klastrilise lahendi juures kinnitavad ka Duda-Harti indeks ja pseudo-T^2 väärtus. Kuna Wardi ja K-keskmiste tehnika Calinski-Harabaszi indeksi väärtused on nelja klastri tasemel kõrgemad, kui täieliku seose tehnika korral, viiakse läbi K-keskmiste analüüs kasutades algpunktidena täieliku seose ja Wardi tehnikaga leitud klastrite keskpunkte nelja klastrilise lahendi korral. Calinski-Harabaszi indeksi väärtused uutele lahenditele on vastavalt 569,71 ja 634,82. Seega osutub parimaks lahendiks nelja klastriline lahend, mis on leitud K-keskmiste tehnikaga, rakendades klastrite algpunktidena Wardi seosega leitud klastrite keskpunkte. Leitud nelja klastrilist lahendit vaadatakse kui nelja üksteisest eristuvat turusegmenti I kvartiilis. Segmentide tarbimiseelistustest suhtena kvartiili keskmisesse logaritmskaalal, annab ülevaate joonis 5 (lk. 52).



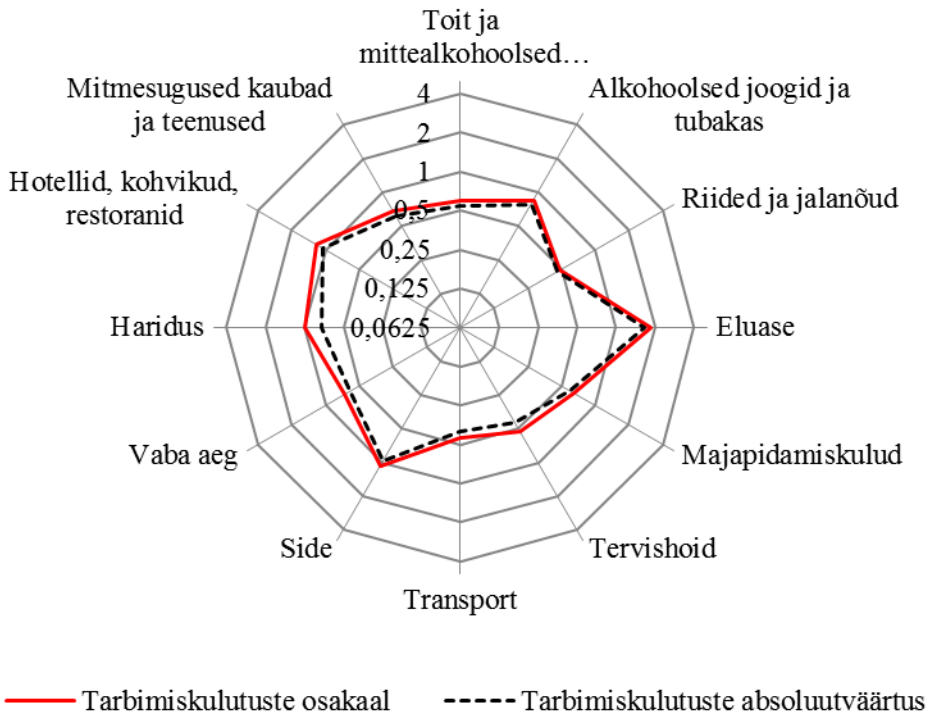
Joonis 5. I kvartiili segmentide tarbimiseelistused suhtena kvartiili keskmisesse (autori koostatud).

Esimesse segmenti kuulub 30,6% I kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse logaritmskaalal, annab ülevaate roheline joon joonisel 5. Segmenti iseloomustavad kvartiili keskmisest madalamad kulutused eluasemele ning toidule ja mittealkohoolsetele jookidele. Üle kahe korra on segmendi keskmised kulutused nii transpordile, riietele ja jalanõudele ning haridusele suuremad, kui I kvartiilis tervikuna. Sarnase tulemuse saame ka võrreldes segmendi tarbimiskulusid kvartiili keskmisesse absoluutväärtustena. Seega ei mõjuta kogu tarbimisieelarve suurus segmendi tarbimiskulutuste jagunemist osakaaludena tarbimiseelarvest. Kõrgete kulude tõttu transpordile; mitmesugustele kaupadele ja teenustele; hotellidele, kohvikutele, restoranidele; riietele ja jalanõudele; haridusele ja vabale ajale, annab käesoleva töö autor segmendile nimeks *„Kulutame transpordile ja kogemustele"*.

Sinine joon joonisel 5 annab ülevaate teise segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse logaritmskaalal. Segmenti kuulub 25,9% I kvartiili leibkondadest

ning segmenti iseloomustavad keskmisest kõrgem kulutuste osakaal eluasemele ning toidule ja mittealkohoolsetele jookidele. Segmendi kulutused sidele moodustavad kvartiili keskmisega võrreldava taseme. Kulutused kõigele muule jäävad segmendis tunduvalt alla kvartiili keskmise. Peaaegu täielikult puuduvad segmendil kulutused haridusele. Sarnase tulemuse saame ka võrreldes segmendi tarbimiskulusid kvartiili keskmisesse absoluutväärtustena. Seega ei mõjuta kogu tarbimisieelarve suurus tarbimiskulutuste jagunemist osakaaludena tarbimiseelarvest. Tulenevalt kõrgetest kulutustest eluasemele ning toidule ja mittealkohoolsetele jookidele, annab käesoleva magistritöö autor segmendile nimeks *„Kulutame toidule ja eluasemele"*.

Kolmandasse segmenti kuulub 18,1% kvartiili leibkondadest (punane joon joonisel 5, lk. 52). Tulenevalt segmendi kvartiili keskmisest madalamast tarbimiskulutuste eelarvest, annab segmendi tarbimiskulutuste võrdlemine kvartiili keskmisega absoluutväärtustena osakaaludest pisut erinevad pildi (vt. joonis 6).



Joonis 6. Segmendi 3 tarbimiskulutuste suhe kvartiili keskmisesse osakaaluna tarbimise eelarvest ja absoluutväärtusena (autori koostatud).

Jooniselt 6 (lk. 53) selgub, et kolmanda segmendi kulutused haridusele moodustavad osakaaluna kogukuludest kvartiili keskmise taseme, kuid absoluutnumbrina jäävad kvartiili keskmisele alla. Samuti ületavad osakaaluna segmendi kulutused hotellidele, restoranidele ja kohvikutele kvartiili keskmise, kuid jäävad absoluutnumbritena kvartiili keskmisele tasemele. Ülejäänud kululiikide osas annavad osakaalud ja absoluutnumbrid sarnase võrdluse. Segmendile on iseloomulik väga madal kulutuste tase transpordile, tervishoiule, vabale ajale ning riietele ja jalanõudele. Seevastu kulutused eluasemele on segmendis nii osakaaluna tarbimiseelarvest kui ka absoluutväärtusena pea kaks korda kvartiili keskmisest kõrgemad. Segmendi eluasemekulude kõrge taseme tõttu annab autor segmendile nimeks *„Kulutame eluasemele"*.

Neljandasse segmenti kuulub 25,4% I kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate roheline joon joonisel 5 (lk. 52). Segmenti iseloomustab kvartiili keskmisest tunduvalt madalam eluaseme, transpordi ja hariduse kulutuste tase. Kvartiili keskmisele tasemele jäävad alla ka teiste kulutuste osakaalud välja arvatud kaks erandit. Segmendi kulutused toidule ja mittealkohoolsetele jookidele on kvartiili keskmisest pea poole kõrgemad ning segmendi keskmist ületavad ka kulutused alkohoolsetele jookidele ja tubakale. Sarnase tulemuse saame ka võrreldes segmendi tarbimiskulusid kvartiili keskmisesse absoluutväärtustena. Seega ei mõjuta kogu tarbimisieelarve suurus tarbimiskulutuste jagunemist osakaaludena tarbimiseelarvest. Segmendi keskmisest kõrgemate kulutuste tõttu toidule, alkohoolsetele jookidele ja tubakale annab käesoleva magistritöö autor segmendile nimeks *„Sööme, joome ja suitsetame“*.

Täiendavat infot leitud segmentide kohta on võimalik leida uurides tunnuseid, mis on segmendi leibkondade kohta teada, kuid, mida pole kaasatud klasteranalüüsi klastrite moodustamiseks. Esimese kvartiili segmentide LEU 2012 andmetes olemasolevatest, kuid klasteranalüüsi mitte kaasatud leibkondi iseloomustavatest geograafilistest, demograafilistest, sotsiaalmajanduslikest ning isiksuse tunnuste esinemissageduste suhtest kvartiili keskmisesse samade tunnuste esinamise sagedusse, annab ülevaate lisa 7. Esinemissageduse alusel kvartiili keskmisest enim erinevad segmentide kategoorilised tunnused on koondatud tabelisse 3 (lk. 55).

Tabel 3. Segmente enim iseloomustavad kategoorilised tunnused I kvartiilis.

Segment	Keskmisest kõrgema esinemissagedusega tunnused	Keskmisest madalama esinemissagedusega tunnused
Kulutame transpordile ja kogemustele	Lastega paar, Kesk-Eesti, auto, mitteaktiivne.	Üksik, üksikvanem.
Kulutame toidule ja eluasemele	III taseme haridus, pensionär, üksikvanem, Kirde-Eesti.	Eestlane.
Kulutame eluasemele	Põhja-Eesti, korter, üüripind.	
Sööme, joome ja suitsetame	Pensionär, Lääne-Eesti, üksik, lasteta paar, I taseme haridus.	III taseme haridus.

Allikas: (Leibkonna eelarve ... 2013); autori arvutused.

Teooriaga kokku sobivalt andis parima tulemuse kliendisegmentide tuvastamisel tarbimise kogukulu I kvartiilis K-keskmiste tehnika. Tuvastati neli küllaltki sarnase suurusega ja üksteisest eristuvat kliendisegmenti. Leitud kliendisegmendid eristuvad üksteisest märgatavalt ka klasteranalüüsi kaasamata jäetud kategooriliste tunnuste alusel.

2.2.2. Tabrmise kogukulu II kvartiil

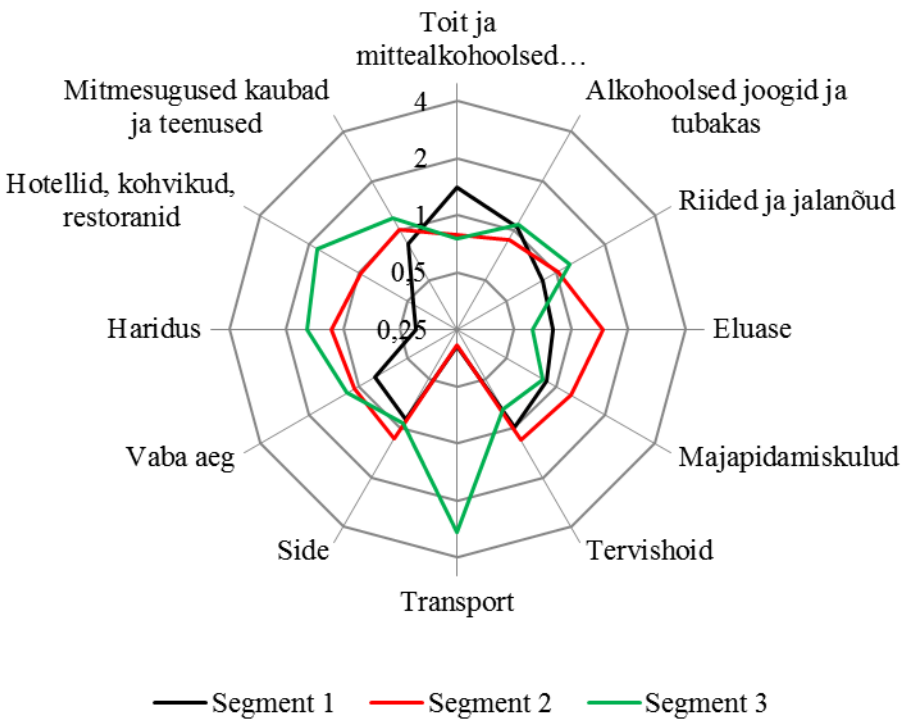
Sarnaselt I kvartiilile, viiakse ka II kvartiilis läbi klasteranalüüs erinevaid tehnikaid rakendades. Erinevate tehnikatega kuni kümne klastrilistele lahenditele leitud Calinski-Harabaszi indeksi väärtused on koondatud alljärgnevasse tabelisse 4.

Tabel 4. Calinski-Harabaszi indeksi väärtused erinevatele lahenditele II kvartiilis.

Klastrite arv	Calinski-Harabasz'i indeksi väärtus				
	Üksik seos	Keskmine seos	Täielik seos	Ward	K-keskmine
2	9,55	23,47	453,49	447,53	529,25
3	7,57	18,35	443	433,41	**590,52**
4	12,97	20,74	425,95	434,75	561,9
5	14,54	19,05	360,66	386,07	490,32
6	13,8	15,69	310,37	358,35	431,78
7	12,63	13,95	282,09	336,89	408,46
8	12,16	14,23	248,97	321,36	398,27
9	10,96	19,31	242,29	308,8	364,21
10	10,44	17,58	220,27	295,28	349,72

Allikas: (Leibkonna eelarve ... 2013); autori arvutused.

Tabelist 4 (lk. 55) selgub, et parim lahend on K-keskmiste tehnika korral kolmest, keskmise seose ja Wardi tehnika korral neljast ning üksiku seose korral viiest klastrist koosnev. Täieliku seose tehnikat kasutades ei leidu ühtegi lahendit, mille korral Calinski-Harabaszi indeksi väärtus oleks üheaegselt suurem nii väiksema kui ka suurema arvuga klastritega lahendi indeksi väärtusest. See ei anna alust üht lahendit teisele eelistada. Leitud parimatest lahenditest on kõrgeim Calinski-Harabaszi indeksi väärtus K-keskmiste tehnikaga leitud kolme klastriga lahendil. Hierarhiliste tehnikate korral jäävad kolme klastrilise lahendi korral Calinski-Harabaszi indeksi väärtused madalamaks. Kuna indeksi väärtused kolme klastri tasemel on küllaltki kõrged ka täieliku seose ja Wardi tehnika korral, kasutatakse K-keskmiste tehnika alguspunktidena alternatiivina ka täieliku seose ja Wardi tehnikaga leitud kolmest klastrist koosneva lahendi klastrite keskpunkte. Leitud lahendite Calinski-Harabaszi indeksi väärtused ei osutu paremaks. Seega kirjeldatakse parima lahendina K-keskmiste tehnikaga leitud kolmest klastrist koosnevat lahendit kui kolme üksteisest eristuvat turusegmenti. Segmentide tarbimiseelistustest suhtena kvartiili keskmisesse logaritmskaalal, annab ülevaate joonis 7.



Joonis 7. II kvartiili segmentide tarbimiseelistused suhtena kvartiili keskmisesse (autori koostatud).

Võrreldes segmentide tarbimiskulusid kvartiili keskmisesse absoluutväärtustena saame ülaltoodud joonisega pea kattuva pildi. Seega ei mõjuta teises kvartiilis segmentide kogu tarbimisieelarve suurus tarbimiskulutuste jagunemist osakaaludena tarbimiseelarvest.

Esimesse segmenti kuulub 36,1% teise kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate must joon joonisel 7 (lk. 56). Segmenti iseloomustab kvartiili keskmisest tunduvalt madalamad kulutused eluasemele, transpordile ja haridusele. Segmendi kulutused tervishoiule jäävad kvartiili keskmisele tasemele. Segmendi kulutused toidule ja mittealkohoolsetele jookidele on aga kvartiili keskmisest pea poole kõrgemad ning kulutused alkohoolsetele jookidele ja tubakale ületavad segmendi keskmist. Viimatinimetatule tuginedes annab autor segmendile nimeks *„Sööme, joome ja suitsetame“*.

Teise segmenti kuulub 37,6% II kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate roheline joon joonisel 7 (lk. 56). Segmenti iseloomustab keskmisest pea poole kõrgem kulutuste tase eluasemele ning keskmisest madalam kulutuste tase transpordile, toidule ja mittealkohoolsetele jookidele ning alkohoolsetele jookidele ja tubakale. Kõrgete eluasemekulude tõttu annab magistritöö autor segmendile nimeks *„Kulutame eluasemele“*.

Kolmandasse segmenti kuulub 26,3% II kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate punane joon joonisel 7 (lk. 56). Jooniselt selgub, et kvartiili keskmisest madalamad kulutused on segmendis vaid eluasemele, majapidamisele, tervishoiule ning toidule ja mittealkohoolsetele jookidele. Segmendile on aga iseloomulik pea kolm korda kvartiili keskmist ületav kulu transpordile. Samuti on kvartiili keskmisest tunduvalt kõrgemad segmendi kulutused hotellidele, restoranidele, kohvikutele ja haridusele. Seetõttu annab autor segmendile nimeks „*Kulutame transpordile ja kogemustele“*.

Täiendavat infot leitud segmentide kohta on võimalik leida uurides tunnuseid, mis on segmendi leibkondade kohta teada, kuid, mida pole kaasatud klasteranalüüsi klastrite

moodustamiseks. Esimese kvartiili segmentide LEU 2012 kategooriliste tunnuste esinemissageduste suhtest kvartiili samade keskmisesse tunnuste esinamise sagedusse, annab ülevaate lisa 8. Esinemissageduse alusel kvartiili keskmisest enim erinevad segmentide kategoorilised tunnused on koondatud tabelisse 5.

Tabel 5. Segmente enim iseloomustavad kategoorilised tunnused II kvartiilis.

Segment	Keskmisest kõrgema esinemissagedusega tunnused	Keskmisest madalama esinemissagedusega tunnused
Sööme, joome ja suitsetame	Pensionär, Kirde-Eesti, üksik, lasteta paar, I taseme haridus.	III taseme haridus.
Kulutame eluasemele	Üüripind, korter, Põhja-Eesti, üksik, III taseme haridus.	Auto.
Kulutame transpordile ja kogemustele	Lastega paar, mitteaktiivne, õpilane, Lääne-Eesti, Kesk-Eesti, auto.	Üksik, üksikvanem.

Allikas: (Leibkonna eelarve ... 2013); autori arvutused.

Parima tulemuse kliendisegmentide tuvastamisel tarbimise kogukulu II kvartiilis andis teooriaga kokku sobivalt K-keskmiste tehnika. Tuvastati kolm küllaltki sarnase suurusega ja üksteisest eristuvat kliendisegmenti. Leitud kliendisegmendid eristuvad üksteisest märgatavalt ka klasteranalüüsi kaasamata jäetud kategooriliste tunnuste alusel.

2.2.3. Tabrmise kogukulu III kvartiil

Sarnaselt I ja II kvartiilile, viiakse ka III kvartiilis läbi klasteranalüüs erinevaid tehnikaid rakendades. Erinevate tehnikatega kuni kümne klastrilistele lahenditele leitud Calinski-Harabaszi indeksi väärtused on koondatud tabelisse 6 (lk. 59). Tabelist selgub, et parim lahend K-keskmiste ja Wardi tehnika korral on neljast, üksiku seose tehnika korral kaheksast, täieliku seose tehnika korral üheksast ning keskmise seose tehnika korral kümnest klastrist koosnev lahend. Leitud parimatest lahenditest on kõrgeim Calinski-Harabaszi indeksi väärtus K-keskmiste tehnikaga leitud nelja klastriga lahendil. Hierarhiliste tehnikate korral jäävad nelja klastrilise lahendi korral Calinski-Harabaszi indeksi väärtused madalamaks. Wardi tehnikaga leitud neljast klastrist koosneva lahendi Calinski-Harabaszi indeksi väärtus on küllaltki kõrge ning selle tunnistavad parimaks ka Duda-Harti indeks ja pseudo-T^2 väärtus. Seetõttu kasutatakse

Wardi tehnikaga leitud klastrite keskpunkte K-keskmiste tehnika alguspunktidena ning saadakse uus lahend, mille Calinski-Harabaszi indeksi väärtus on 501,07, mis osutub esialgsest K-keskmiste tehnikaga leitud lahendi tulemusest kõrgemaks. Saadud parimat nelja klastrilist lahendit kirjeldatakse kui nelja üksteisest eristuvat turusegmenti III tarbimise kogukulu kvartiilis.

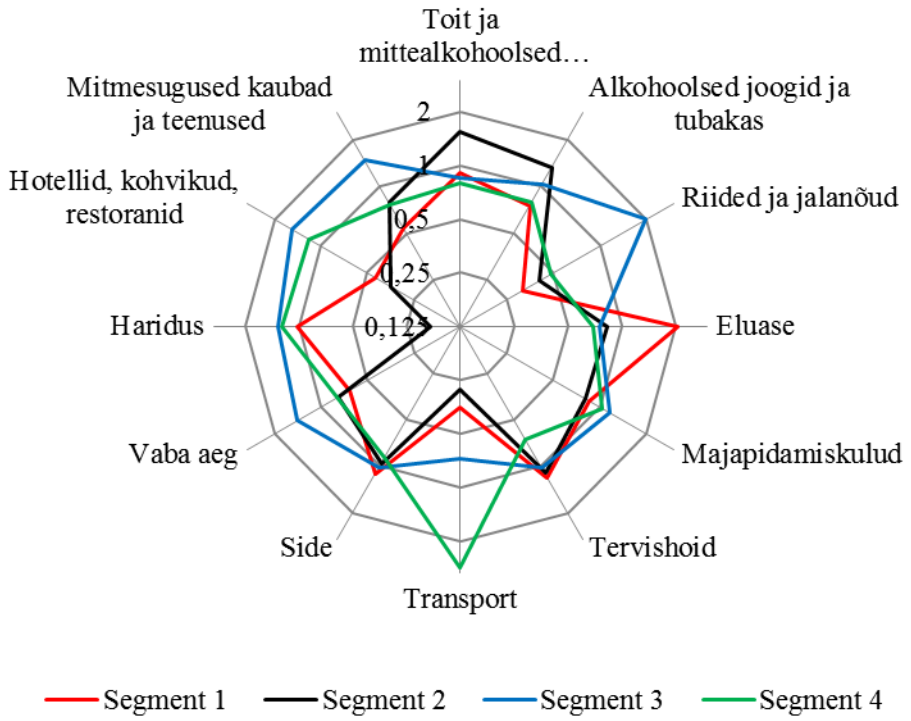
Tabel 6. Calinski-Harabaszi indeksi väärtused erinevatele lahenditele III kvartiilis.

Klastrite arv	**Calinski-Harabasz'i indeksi väärtus**				
	Üksik seos	**Keskmine seos**	**Täielik seos**	**Ward**	**K-keskmine**
2	4,29	37,56	64,28	370,53	532,08
3	7,7	36,43	124,51	373,52	482,62
4	7,44	26,72	142,52	376,21	**501,04**
5	6,14	21,61	168,83	353,58	454,69
6	10,15	23,63	198,47	341,11	410,53
7	9,99	26,89	176,97	323,97	375,47
8	10,38	24,48	188,64	311,57	373,16
9	9,99	22,72	207,25	301,77	351,27
10	9,52	41,46	197,69	289,23	343,57

Allikas: (Leibkonna eelarve ... 2013); autori arvutused.

Sarnaselt II kvartiili lahendile, annab III kvartiili segmentide tarbimiskulude keskmise võrdlemine kogu kvartiili keskmisega nii osakaaludena kogu tarbimise eelarvest kui ka absoluutväärtustena sarnase tulemuse. Seega ei mõjuta III kvartiilis segmentide kogu tarbimisieelarve suurus tarbimiskulutuste jagunemist osakaaludena tarbimiseelarvest. Segmentide tarbimiseelistustest suhtena kvartiili keskmisesse logaritmskaalal, annab ülevaate joonis 8. (lk. 60).

Esimesse segmenti kuulub 18,4% III kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate punane joon joonisel 8 (lk. 60). Segmenti iseloomustab kvartiili keskmisest enam kui kaks korda kõrgemad eluaseme kulud ning keskmisest oluliselt madalamad kulutused riietele ja jalanõudele; hotellidele, kohvikutele, restoranidele; transpordile ning vabale ajale. Kõrgete eluasemekulude tõttu annab käesoleva magistritöö autor segmendile nimeks *„Kulutame eluasemele“*.



Joonis 8. III kvartiili segmentide tarbimiseelistused suhtena kvartiili keskmisesse (autori koostatud).

Teise segmenti kuulub 22% III kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate must joon joonisel 8. Segmendile on iseloomulik kvartiili keskmisest märgatavalt kõrgem kulutuste osakaal toidule ja mittealkohoolsetele jookidele ning alkohoolsetele jookidele ja tubakale. Kulutused transpordile, riietele ja jalanõudele ning haridusele on segmendis keskmisest oluliselt madalamad. Eelmainitule tuginedes annab autor segmendile nimeks *„Sööme, joome ja suitsetame“*.

Kolmandasse segmenti kuulub 37,7 % III kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate sinine joon joonisel 8. Jooniselt selgub, et segmenti iseloomustavad keskmisest madalamad kulutused eluasemele, transpordile ning toidule ja mittealkohoolsetele jookidele. Keskmisest pea kaks korda kõrgemad on segmendi kulutused riietele ja jalanõudele. Segmendi kulutused on kvartiili keskmisest märgatavalt kõrgemad ka vabale ajale; restoranidele, hotellidele ja kohvikutele;

mitmesugustele kaupadele ja teenustele ning haridusele. Sellest lähtuvalt nimetab autor segmendi kui *„Meeldivad riided ja väljas käimine“*.

Neljandasse segmenti kuulub 21,9% III kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate roheline joon joonisel 8 (lk. 60). Segmenti iseloomustab keskmisest pea kolm korda kõrgemad kulutused transpordile ning keskmisest kõrgemad kulutused hotellidele, kohvikutele, restoranidele ja haridusele. Mõnevõrra ületavad kvartiili keskmist ka segmendi majapidamiskulud. Muude kulude osatähtsused jäävad kvartiili keskmisele alla. Lähtuvalt segmendi kulude struktuurist annab autor segmendile nimeks *„Kulutame transpordile ja kogemustele“*.

Täiendavat infot leitud segmentide kohta on võimalik leida uurides tunnuseid, mis on segmendi leibkondade kohta teada, kuid, mida pole kaasatud klasteranalüüsi klastrite moodustamiseks. Kolmanda kvartiili segmentide LEU 2012 andmetes olemasolevatest, kuid klasteranalüüsi mitte kaasatud leibkondi iseloomustavatest geograafilistest, demograafilistest, sotsiaalmajanduslikest ning isiksuse tunnuste esinemissageduste suhtest kvartiili keskmisesse samade tunnuste esinamise sagedusse, annab ülevaate lisa 9. Esinemissageduse alusel kvartiili keskmisest enim erinevad segmentide kategoorilised tunnused on koondatud tabelisse 7.

Tabel 7. Segmente enim iseloomustavad kategoorilised tunnused III kvartiilis.

Segment	**Keskmisest kõrgema esinemissagedusega tunnused**	**Keskmisest madalama esinemissagedusega tunnused**
Kulutame eluasemele	Üüripind, Kirde-Eesti, Põhja-Eesti, üksik, lasteta paar.	Auto.
Sööme, joome ja suitsetame	Pensionär, Kirde-Eesti, üksik, lasteta paar.	III taseme haridus, auto, üüripind.
Meeldivad riided ja väljas käimine	Mitteaktiivne, õpilane, lastega paar, üksikvanem.	
Kulutame transpordile ja kogemustele	Õpilane, Lääne-Eesti, Kesk-Eesti, auto, lastega paar, mitteaktiivne.	Üksikvanem.

Allikas: (Leibkonna eelarve ... 2013); autori arvutused.

Sarnaselt I ja II kvartiiliga andis ka III kvartiilis parima tulemuse kliendisegmentide tuvastamisel K-keskmiste tehnika. Tuvastati neli küllaltki sarnase suurusega ja

üksteisest eristuvat kliendisegmenti. Leitud kliendisegmendid eristuvad üksteisest märgatavalt ka klasteranalüüsi kaasamata jäetud kategooriliste tunnuste alusel.

2.2.4. Tabrmise kogukulu IV kvartiil

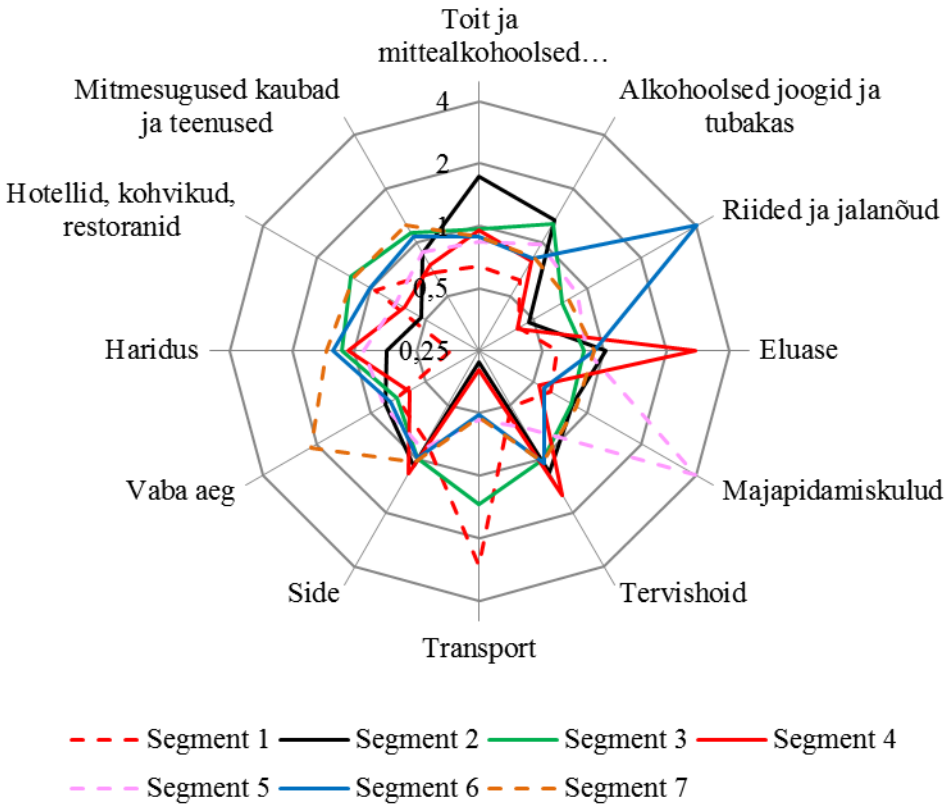
IV kvartiilis viiakse klasteranalüüs läbi sarnaselt ülejäänud kvartiilidega. Erinevate tehnikatega kuni kümne klastrilistele lahenditele leitud Calinski-Harabaszi indeksi väärtused on koondatud alljärgnevasse tabelisse 8.

Tabel 8. Calinski-Harabaszi indeksi väärtused erinevatele lahenditele IV kvartiilis.

Klastrite arv	**Calinski-Harabasz'i indeksi väärtus**				
	Üksik seos	**Keskmine seos**	**Täielik seos**	**Ward**	**K-keskmine**
2	7,03	7,03	643,45	629,83	722,75
3	5,09	5,09	376,31	453,7	560,32
4	4,61	20,71	305,32	387,96	492,4
5	4,33	19,56	274,64	356,72	443,47
6	3,96	115,02	241,4	342,03	400,3
7	4,09	106,05	229,83	334,95	**418,35**
8	3,82	92,28	225,31	321,93	392,75
9	3,64	98,51	201,36	310,98	373,01
10	3,44	88,41	197,26	303,79	358,68

Allikas: (Leibkonna eelarve ... 2013); autori arvutused.

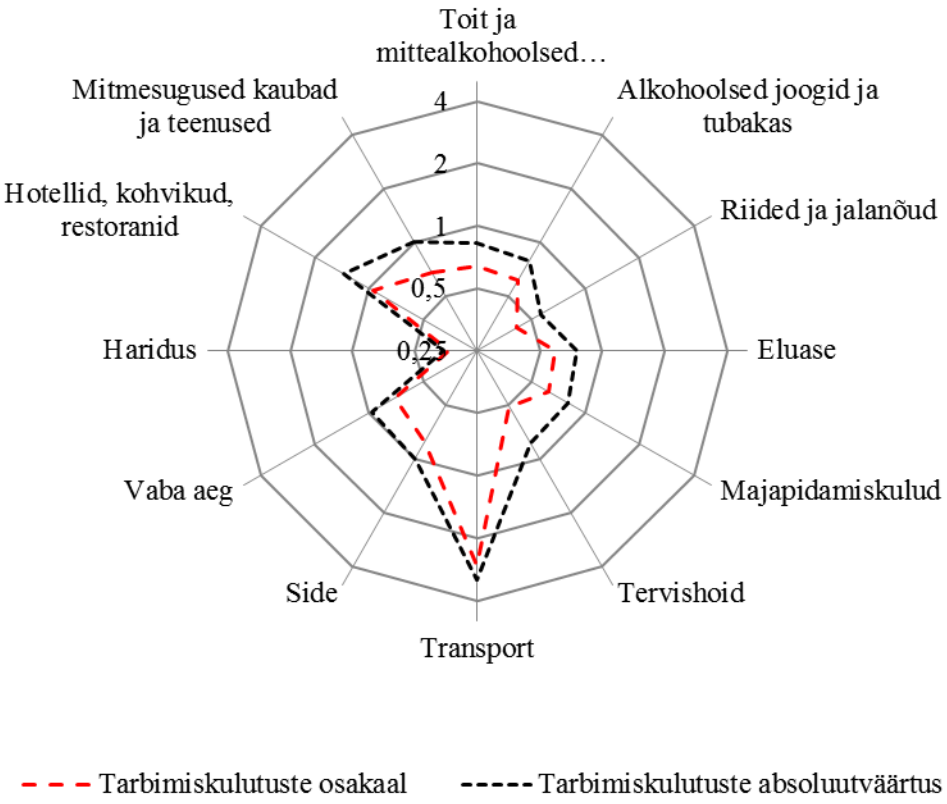
Tabelist 8 selgub, et parim lahend üksiku seose ja K-keskmiste tehnika korral on seitsmest ning keskmise seose tehnika korral kuuest klastrist koosnev. Täieliku seose ja Wardi tehnikat kasutades ei leidu ühtegi lahendit, mille korral Calinski-Harabaszi indeksi väärtus oleks üheaegselt suurem nii väiksema kui ka suurema arvuga klastritega lahendi indeksi väärtusest. See ei anna mingit alust üht lahendit eelistada teisele. Leitud parimatest lahenditest on kõrgeim Calinski-Harabaszi indeksi väärtus K-keskmiste tehnikaga leitud seitsme klastriga lahendil. Kasutades K-keskmiste tehnika alguspunktidena täieliku seose ja Wardi tehnikaga leitud neljast klastrist koosneva lahendi klastrite keskpunkte, ei osutu leitud tulemuste Calinski-Harabaszi indeksi väärtused esialgu leitud lahendi omast kõrgemaks. Seega kirjeldatakse parima lahendina K-keskmiste tehnikaga leitud seitsme klastrilist lahendit kui seitset üksteisest eristuvat turusegmenti. Segmentide tarbimiseelistustest suhtena kvartiili keskmisesse logaritmskaalal, annab ülevaate joonis 9. (lk. 63).



Joonis 9. IV kvartiili segmentide tarbimiseelistused suhtena kvartiili keskmisesse (autori koostatud).

Esimesse segmenti kuulub 13,7% IV kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate punane katkendlik joon joonisel 9. Tulenevalt segmendi kvartiili keskmisest erinevast tarbimiskulutuste eelarvest, annab segmendi tarbimiskulutuste võrdlemine kvartiili keskmisega absoluutväärtustena osakaaludest pisut erinevad pildi (vt. joonis 10, lk. 64). Jooniselt 10 selgub, et katkendlik must joon, mis tähistab segmendi keskmist tarbimiskulutuste suhet kvartiili keskmisesse absoluutväärtuses, ümbritseb terves ulatuses pidevat joont. Asjaolu on tingitud segmendi kvartiili keskmisest kõrgemast tarbimiskulutuste eelarvest. Tarbimiskulude osakaalu alusel kulutab segment kõigele, välja arvatud transpordile, kvartiili keskmisest vähem. Vaadeldes aga segmendi kulutusi absoluutnumbrites, selgub, et segment kulutab keskmisest enam ka hotellidele, kohvikutele, restoranidele ning kvartiili keskmisel tasemel vabale ajale. Segmenti iseloomustab enim kvartiili keskmisest umbkaudu kolm

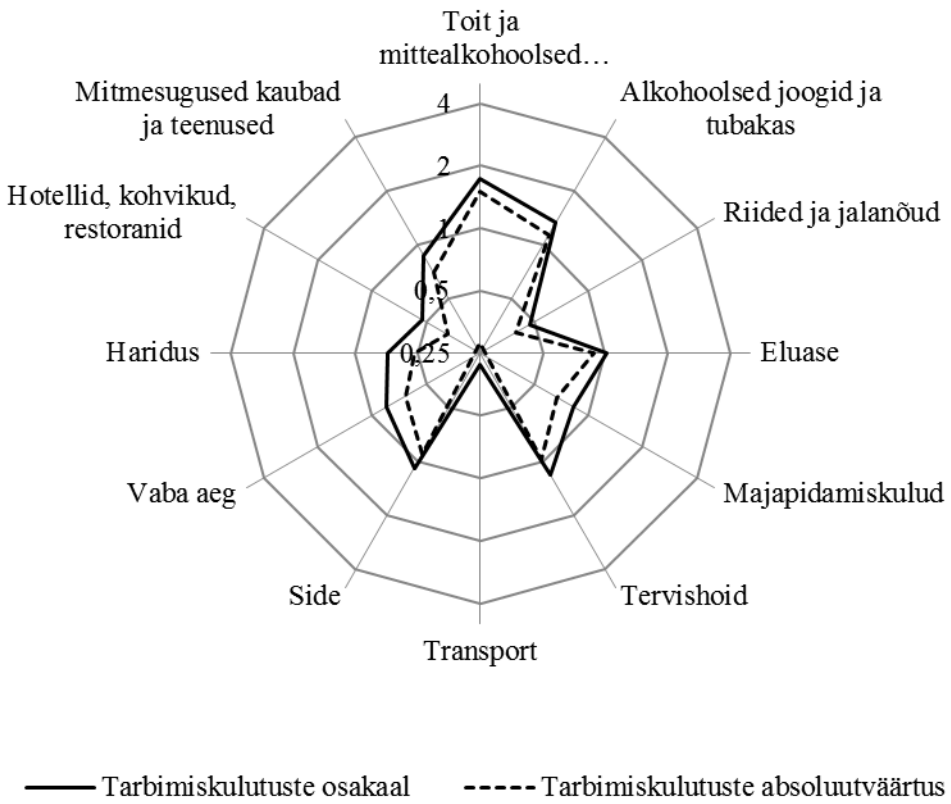
korda kõrgemad kulutused transpordile ning keskmisest oluliselt madalamad kulutused haridusele. Sellest lähtuvalt annab autor segmendile nimeks *„Kulutame transpordile“*.



Joonis 10. Segmendi 1 tarbimiskulutuste suhe kvartiili keskmisesse osakaaluna tarbimise eelarvest ja absoluutväärtusena (autori koostatud).

Teise segmenti kuulub 14,5% IV kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate must joon joonisel 9 (lk. 63). Ka teise segmendi korral annavad segmendi tarbimiskulutuste võrdlemine kvartiili keskmisega absoluutväärtustena ja osakaaludena pisut erinevad pildi (vt. joonis 11, lk. 65). Jooniselt selgub, et katkendlik must joon, mis tähistab segmendi keskmist tarbimiskulutuste suhet kvartiili keskmisesse absoluutväärtuses, on ümbritsetud terves ulatuses pideva musta joonega, mis tähistab segmendi keskmist tarbimiskulutuste suhet kvastiili keskmisesse osakaaluna kogu tarbimiseelarvest. See on tingitud segmendi kvartiili keskmisest madalamast tarbimise kogukulust. Seetõttu moodustavad segmendi kulutused toidule ja mittealkohoolsetele jookidele, alkohoolsetele jookidele ja tubakale, eluasemele, tervishoiule ja sidele osakaaluna kogu

tarbimise kulutustest kvartiili keskmise või keskmisest pisut kõrgema taseme. Absoluutväärtustes jäävad aga segmendi kõik kululiigid, välja arvatud toit ja mittealkohoolsed joogid ning alkohoolsed joogid ja tubakas, kvartiili keskmisest madalamateks. Seetõttu annab autor segmendile nimeks *„Sööme, joome ja suitsetame“*.



Joonis 11. Segmendi 2 tarbimiskulutuste suhe kvartiili keskmisesse osakaaluna tarbimise eelarvest ja absoluutväärtusena (autori koostatud).

Kolmandasse segmenti kuulub 26,4% IV kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate roheline joon joonisel 9 (lk. 63). Segmenti eristub keskmisest kõrgemate kulutuste poolest transpordile; hotellidele, kohvikutele, restoranidele ning alkohoolsetele jookidele ja tubakale. Kvartiili keskmisele jäävad alla segmendi kulutused riietele ja jalanõudele, vabale ajale ning majapidamisele. Sarnane tulemus saadakse ka võrreldes segmendi tarbimiskulusid kvartiili keskmisega absoluutväärtustena. Seega ei mõjuta kogu tarbimisieelarve suurus tarbimiskulutuste jagunemist osakaaludena tarbimiseelarvest. Lähtuvalt keskmisest kõrgematest kulutustest hotellidele, kohvikutele, restoranidele ning transpordile osakaaludena ning

keskmisest kõrgematele absoluutkulutustele haridusele annab autor segmendile nimeks *„Kulutame transpordile ja kogemustele“*.

Neljandasse segmenti kuulub 8,7% IV kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate punane joon joonisel 9 (lk. 63). Segmenti iseloomustab kvartiili keskmisest pea kolm korda kõrgemad eluaseme kulud ning keskmisest oluliselt madalamad kulutused riietele ja jalanõudele; hotellidele, kohvikutele, restoranidele; transpordile ja vabale ajale. Sarnase tulemuse annab ka segmendi tarbimiskulude võrdlemine kvartiili keskmisega absoluutväärtustes. Seega ei mõjuta segmendi kogu tarbimisieelarve suurus tarbimiskulutuste jagunemist osakaaludena tarbimiseelarvest. Kõrgete eluasemekulude tõttu annab autor segmendile nimeks *„Kulutame eluasemele“*.

Viiendasse segmenti kuulub 8% IV kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate lilla katkendlik joon joonisel 9 (lk. 63). Segmenti iseloomustab kvartiili keskmisest umbkaudu neli korda kõrgem majapidamiskulutuste tase. Muude kululiikide osatähtsus jääb alla segmendi keskmise. Sarnane tulemus saadakse ka võrreldes segmendi tarbimiskulusid kvartiili keskmisega absoluutväärtustena. Eelnevale tuginedes annab autor segmendile nimetuseks *„Kulutame majapidamisele“*.

Kuuendasse segmenti kuulub 10,8 % IV kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate sinine joon joonisel 9 (lk. 63). Segmendile on iseloomulikud keskmisest mõnevõrra madalamad kulutused majapidamisele, transpordile, alkohoolsetele jookidele ja tubakale ning vabale ajale. Kvartiili keskmisest enam kui neli korda kõrgemad on segmendi kulutused riietele ja jalanõudele. Segmendi kulutused on kvartiili keskmisega samal tasemel või kõrgemad restoranidele, hotellidele, kohvikutele; mitmesugustele kaupadele ja teenustele ning haridusele. Sarnase tulemuse annab segmendi tarbimiskulude võrdlemine kvartiili keskmisega absoluutväärtustes. Segmendi tarbimiskulude struktuurist lähtuvalt annab autor segmendile nimeks *„Meeldivad riided ja väljas käimine“*.

Seitsmendasse segmenti kuulub 17,9 % IV kvartiili leibkondadest. Segmendi tarbimiskulutuste suhtest kvartiili tarbimiskulutustesse osakaaluna logaritmskaalal, annab ülevaate katkendlik oranž joon joonisel 9 (lk. 63). Segmenti iseloomustavad kvartiili keskmisest enam kui kaks korda kõrgemad kulutused vabale ajale. Samuti on segmendi kulutused mitmesugustele kaupadele ja teenustele; hotellidele, kohvikutele, restoranidele ning haridusele kvartiili keskmisest kõrgemad. Vaid segmendi kulutused transpordile jäävad kvartiili keskmisest madalamateks. Sarnane tulemus saadakse võrreldes segmendi tarbimiskulusid kvartiili keskmisega absoluutväärtustes. Tulenevalt segmendi keskmistest madalamatest kulutustest sundkulutustele, nagu eluase ja toit ning mittealkohoolsed joogid, ja kvartiili keskmisest kõrgematele kulutustest vabale ajale, mitmesugustele kaupadele ja teenustele ning hotellidele, kohvikutele, restoranidele, annab autor segmendile nimeks *„Kulutame meelelahutusele“*.

Täiendavat infot leitud segmentide kohta on võimalik leida uurides tunnuseid, mis on segmendi leibkondade kohta teada, kuid, mida pole kaasatud klastrite moodustamiseks. Neljanda kvartiili segmentide kohta LEU 2012 andmetes olemasolevatest kategooriliste tunnuste esinemissageduste suhtest kvartiili keskmistesse samade tunnuste esinamise sagedusse, annab ülevaate lisa 10. Esinemissageduse alusel kvartiili keskmisest enim erinevad segmentide kategoorilised tunnused on koondatud tabelisse 9.

Tabel 9. Segmente enim iseloomustavad kategoorilised tunnused IV kvartiilis.

Segment	**Keskmisest kõrgema esinemissagedusega tunnused**	**Keskmisest madalama esinemissagedusega tunnused**
Kulutame transpordile	I taseme haridus, üksikvanem, auto.	Üksik, III taseme haridus.
Sööme, joome ja suitsetame	Pensionär, üksik, lasteta paar.	III taseme haridus, auto.
Kulutame transpordile ja kogemustele	Lastega paar, auto, mitteaktiivne, õpilane, Kesk-Eesti.	Üksik.
Kulutame eluasemele	Üüripind, korter, üksiklasteta paar, III taseme haridus, Kirde-Eesti.	Auto.
Kulutame majapidamisele	Mitteaktiivne, Kirde-Eesti.	Õpilane.
Meeldivad riided ja väljas käimine	Mitteaktiivne, üksikvanem, lastega paar, muu kooselu, õpilane.	
Kulutame meelelahutusele		Kirde-Eesti, üüripind, töötu, lastega muu kooselu.

Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Taaskord andis oodatult parima tulemuse kliendisegmentide tuvastamisel tarbimise kogukulu IV kvartiilis K-keskmiste tehnika. Tuvastati seitse küllaltki sarnase suurusega ja üksteisest eristuvat kliendisegmenti. Leitud kliendisegmendid eristuvad üksteisest märgatavalt ka klasteranalüüsi kaasamata jäetud kategooriliste tunnuste alusel.

2.3. Järeldused Eesti leibkondade segmenteerimisest

LEU 2012 andmetel läbiviidud klasteranalüüs kliendisegmentide tuvastamiseks Eesti leibkondade hulgas annab hea võimaluse klasteranalüüsi metoodika praktikas kasutamise uurimiseks ning soovituste andmiseks uurijatele huvipakkuvate kliendisegmentide moodustamiseks. Alljärgnevalt tutvustatakse klasteranalüüsi läbiviimisel tehtud tähelepanekuid erinevatel klasteranalüüsi etappidel ning antakse soovitusi spetsiifilisemate segmenteerimiste läbiviimiseks.

Klasteranalüüs ei sea olulisi eeldusi analüüsi läbiviimiseks kasutatavatele andmetele, kuid eelnev põhjalik töö andmetega aitab kaasa uurimisprobleemile sobilikumate lahendite leidmisele. Uurimisprobleemist lähtuvalt võib osutuda mõistlikuks andmete transformeerimine välistamaks klastrite tekkimist mõne üksiku domineeriva tunnuse alusel, mis välistab andmetest loomulike klastrite tuvastamise uurijat tegelikult huvitavate tunnuste alusel. Käesolevas magistritöös omas kogu leibkonna tarbimiseelarve tugevat korrelatsiooni tarbimiskulu liikidega. Vältimaks klastrite teket vaid leibkondade tarbimise kogukulu alusel ja saamaks tulemuseks leibkondade tarbimiseelistusi väljendavaid klastreid, viidi klasteranalüüs läbi leibkonna tarbimise kogukulu alusel moodustatud neljas kvartiilis ning transformeeriti kulutused kulutuste osakaaluks leibkonna tarbimiseelarvest.

Klasteranalüüsi käigus leitud segmente iseloomustab selge eristumine keskmise tarbimiskulutuste osakaalu alusel. Samas on muutujate väärtuste varieerumine leitud segmentides küllaltki suur ning segmendid seetõttu küllaltki heterogeensed. Ühe tunnuse alusel võib pea igas segmendis leiduda sinna esmapilgul mitte sobivaid leibkondi. Olukord on tingitud analüüsi kaasatud suurest muutujate arvust. Suure muutujate arvuga läbiviidud klasteranalüüsi tulemuste sisulise tõlgendamise ja praktikas kasutamise keerukuse eest, hoiatab klasteranalüüsiks praktilisi kasutusnõuandeid andes

ka Norušis (2011: 378) ning soovitab analüüsi kaasata vähem muutujaid, mis on valitud selle alusel, mille alusel soovitakse, et klastrid omavahel sarnaneksid (*Ibid.*: 376). Käesolevasse analüüsi kaasati kõik tarbimiskulud, leidmaks üldisemat andmetes peituvat struktuuri edasisteks spetsiifilisematest uurimisprobleemidest lähtuvateks segmenteerimisteks. Kliendisuhete juhtimise tarvis Eesti leibkondade tarbimiseelistuste alusel kliente segmenteerides, on organisatsioonidel mõistlik klasteranalüüsi kaasata väiksem arv tarbimiskulutusi kirjeldavaid muutujaid ning valida need konkreetse organisatsiooni vajadustest lähtuvalt.

Leibkondade omavahelise sarnasuse määramiseks kasutati käesolevas magistritöös eukleidilist kaugust ruudus kuna sooviti võrrelda omavahel erinevate klasterdamise tehnikatega saadud tulemusi ning kõik klasteranalüüsi kaasatud tunnused olid ühel pideval skaalal. Kui klasteranalüüsi läbiviija eesmärgiks ei ole tehnikate omavaheline võrdlemine, on soovitatav proovida ka teisi sarnasuse mõõdikuid, mis võivad anda segmenteerimise eesmärgist lähtuvalt sobilikuma klasterdamise lahendi.

Kõikides tarbimiskulutuste alusel moodustatud leibkondade kvartiilides osutuks Calinski-Harabaszi indeks väärtuse alusel parimaks K-keskmiste tehnikaga leitud lahend. See on põhjendatav asjaoluga, et K-keskmiste tehnika koondab klastri keskpunktile sfäärjalt lähimaid vaatlusi ning Calinski-Harabaszi indeks sobib hästi just kompaktsete sfäärjatele klastrite tuvastamiseks. Kompaktsed sfäärjad klastrid kujutavad endast ka parimaid lahendeid kliendisuhete juhtimise seisukohalt, kuna kliendisuhete juhtimise tarvis otsitakse sarnase suurusega ja üksteisele võimalikult sarnase käitumisega klientidest koosnevaid gruppe.

Uurides erinevate tehnikatega leitud klastrite suurusi, leidis kinnitust teoreetiline seisukoht, mille kohaselt üksiku seose tehnikaga leitud lahend annab tulemuseks kõige suurema klastrite suuruste varieeruvusega lahendi ja K-keskmiste tehnika annab tulemuseks kõige sarnasema suurusega klastrid. Tabelisse 10 (lk. 70) on koondatud iga kvartiili parimaks tunnistatud klastrite arvu korral lahendiks saadud klastrite suurused neisse kuuluvate leibkondade arvu järgi kasvavas järjekorras erinevate klasterdamise tehnikate korral. Tabelist näeme, et üksiku seose tehnika leiab tulemuseks väga erineva suurusega klastrid, milledest väiksematesse kuulub ainult üks erandlik leibkond ning suurimatesse klastritesse enamik kvartiili leibkondadest. Tulemus ei muutu oluliselt ka

100 klastrilise lahendi korral. Tegemist on niinimetatud ketistumise efektiga, mille korral tekib üks keskne suur klaster, millele samm-sammult lisatakse väiksemaid erandlikke klastreid ja vaatlusi. K-keskmiste tehnikaga leitud lahendid on seevastu aga kõige väiksema suuruste varieeruvusega, andes tulemuseks küllaltki sarnase suurusega klastrid. Kliendisegmentide tuvastamiseks on ebasovivad väga väiksed ja väga suured klastrid, kuna suurtes klastrites on klastritesisene leibkondade eelistuste varieeruvus väga suur ning väga väikesed klastrid ei paku kliendigrupi kõnetamisel mastaabisäästu. Sobilikuks lahendiks on küllaltki sarnaste suurustega klastrid, mis on klastrisiseselt sarnastest leibkondadest koosnevad ning ülejäänud klastritest selgelt eristuvad. Seetõttu sobib ka klastrite suuruse seisukohalt kliendisegmentide tuvastamiseks kõige paremini K-keskmiste tehnika.

Tabel 10. Leitud klastrite suurused (leibkondade arv klastris) erinevate klasteranalüüsi tehnikate korral.

Kvartiil	Üksik seos	Keskmine seos	Täielik seos	Ward	K-Keskmine
I	1	5	33	367	411
	1	7	451	447	576
	1	13	590	575	589
	2267	2245	1196	881	694
II	1	3	247	320	598
	2	3	794	580	819
	2267	2264	1229	1370	853
III	1	2	13	211	417
	3	5	59	229	497
	3	11	106	269	499
	2263	2252	2092	1561	857
IV	1	1	18	99	181
	1	2	39	177	197
	1	6	58	183	244
	2	20	149	216	312
	2	20	158	406	329
	2	214	522	479	407
	2261	2007	1326	710	600

Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Leitud eristuvate segmentide arv erinevates leibkondade kvartiilides osutus erinevaks. Läbi erinevate kvartiilide korduvad sarnaste tarbimiseelistustega segmendid. Kokku tuvastas käesoleva magistritöö autor kaheksa üksteisest tarbimiseelistustelt eristuvat

segmenti, millest kolm on esindatud kõikides kvartiilides. Ülevaate segmentide esinemisest erinevates kvartiilides annab tabel 11.

Tabel 11. Segmentide esinemine tarbimiskulutuste kvartiilides

Segment	**I kvartiil**	**II kvartiil**	**III kvartiil**	**IV kvartiil**
Kulutame toidule ja eluasemele	JAH	-	-	-
Kulutame eluasemele	JAH	JAH	JAH	JAH
Sööme, joome ja suitsetame	JAH	JAH	JAH	JAH
Kulutame transpordile ja kogemustele	JAH	JAH	JAH	JAH
Meeldivad riided ja väljas käimine	-	-	JAH	JAH
Kulutame transpordile	-	-	-	JAH
Kulutame meelelahutusele	-	-	-	JAH
Kulutame majapidamisele	-	-	-	JAH

Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Tabelist 11 selgub, et kõigis kvartiilides on esindatud segmendid *„Kulutame eluasemele“*, *„Sööme, joome ja suitsetame“* ning *„Kulutame transpordile ja kogemustele“*. Alates II kvartiilist kasvab segmentide arv igas järgnevas kvartiilis. I Kvartiilis esineb kõigile II kvartiili segmentidele täiendavalt segment *„Kulutame toidule ja eluasemele“*, mis kujutab endast segmenti, mille keskmised kulutused muudele kululiikidele kui toit ja eluase on kvartiili keskmisest oluliselt madalamad. Sisuliselt on tegu segmendiga, kelle sissetulekutest lõviosa kulub sundkulutustele ning kellel pole muule kulutamiseks võimalusi. Segmenti võib pidada segmentide *„Kulutame eluasemele“* ja *„Sööme, joome ja suitsetame“* koosesinemiseks. Ülejäänud kvartiilides, kus tarbimiskulutused leibkonna liikme kohta on kõrgemad, sellist segmenti ei eristu. Alates III kvartiilist lisanduvad segmendid, mida iseloomustab keskmisest märgatavalt kõrgem kulutuste osakaal kulutustele, mis ei ole sundkulutusted. See tähendab, et tarbimiskulutuste kvartiili tõusuga kaasneb leibkonna tarbimisvabaduse suurenemine. Madalamate kvartiilide segmendid esindavad leibkondade jagunemist ellujäämiseks vajalike kulutuste alusel, kõrgema taseme kvartiilides lisanduvad sotsiaalsetest ja vaimsetest vajadustest lähtuvate kulutuste alusel eristuvad segmendid. Jagunemine on seega kooskõlas Maslow inimvajaduste hierarhiaga.

Klasteranalüüsile eelnenud tarbimiskulutuste vaheliste korrelatsioonide uurimisel tekkinud ootus, et klasteranalüüsi tulemusel tekivad segmendid peamiselt sundkulutuste alusel, pidas paika. Kõikides tarbimise kogukulu alusel moodustatud kvartiilides

esinesid peamiselt sundkulutuste alusel eristuvad segmendid ning III ja IV kvartiilis lisandusid segmendid, mida ei saa kirjeldada kui vaid sundkulutuste alusel tekkinuteks. Seetõttu võib öelda, et sundkulutuste alusel läbiviidav segmenteerimine annab laiema katvuse. Muude kululiikide alusel leitavad segmendid kirjeldavad kitsamaid sihtgruppe kõrgemates tarbimise kogukulu kvartiilides. Laiapõhise kliendibaasiga organisatsioonide jaoks võib osutuda seega sobivamaks sundkulutuste alusel segmenteerimine. Kitsa turuniši leidmiseks sobib aga paremini klientide segmenteerimine muude kulutuste alusel.

Klastrite moodustamisse kaasamata jätud tunnuseid uurides selgub, et sama nimega segmendid erinevates kvartiilides eristuvad teistest segmentidest sarnaste tunnuste alusel. Segmendi *„Kulutame eluasemele“* leibkondi iseloomustab kõigis kvartiilides kvartiili keskmisest kõrgem korteris, üüripinnal ja Põhja- või Kirde-Eestis elamise tõenäosus. Kõik tunnused aitavad selgitada kõrgemaid leibkonna kulutusi eluasemele. Kortermajas elamine toob endaga kaasa kõrgemad kulutused, kuna korterelamus elavatel leibkondadel on piiratud võimalused kasutada keskküttest soodsamaid kütteallikaid ning teostada iseseisvalt maja hooldustöid. Põhja- ja Kirde-Eesti on ühed linnastunuimad piirkonnad Eestis ning kinnisvara hinnad linnades on kõrgemad kui linnadest väljaspool. Üüripinnal elamine toob endaga kaasa täiendava kulu üürimakseteks, mida eluaset omavatel leibkondadel ei ole. Võib oletada, et siia segmenti kuuluvad suurema tõenäosusega ka eluasemelaenuga leibkonnad. Eluasemelaenu olemasolu kohta autori käsutuses olevas LEU 2012 andmefailis andmed aga puuduvad. Segment kirjeldab ilmekalt, et eluaset mitte omavad leibkonnad on oma tarbimiskulutuste poolest piiratumad ning seetõttu omavad eristuvat tarbimiskäitumist. Seetõttu väärivad kõrgemate eluasemekulutuste põhjused lähemat uurimist ning võivad anda aluse sisukaks sihtklientide segmenteerimiseks. Kuna segmendi leibkondade tarbimiseelarve on kõrgete eluasemekulude tõttu piiratum kui teistel tuvastatud segmentidel, annab kliendisuhete juhtimisel segmenti rakendada kui sihtklientide rühma, kellele teha madalama hinnatasemega sooduspakkumisi müügimahtude kasvatamiseks hinnatundliku kliendigrupi arvelt.

Segmenti *„Sööme, joome ja suitsetame“* iseloomustab kõikides kvartiilides kõrgem tõenäosus, et uuringu küsimustele on vastanud pensionär, leibkonna näol on tegu üksiku

inimese või lasteta paariga ning haridustase segmendis on keskmisest madalam. Tunnuste esinemist aitab selgitada nendega kaasnev väiksem leibkonna suurus. Väiksem leibkond pakub vähem võimalusi kokku hoida toidu valmistamise mastaabiefektilt, ehk suuremates leibkondades läheb vähem toitu raisku ning suuremas koguses toiduainete ostmisel, pakutakse klientidele mitmeid soodustusi. Samuti on väikestes leibkondades laste osakaal, kes ei tarbi alkoholi ja tubakat, madalam. Täiendavalt on uurimist väärt, kas kõrgem kulutuste osakaal toidule ja mittealkohoolsetele jookidele on segmendis tingitud vaid väiksemast leibkonnast või erinevad selle segmendi tarbimiseelistusted ka indiviidi tasemel. Kliendisuhete juhtimise seisukohalt sobib segment sihtrühmaks, kellele teha huvitavadi toiduainete ja alkohoolsete jookide pakkumisi, kuna segment on valmis kulutama toidule ja alkohoolsetele jookidele teistest tuvastatud segmentidest enam.

Segmenti *„Kulutame transpordile ja kogemustele“* iseloomustab suurem tõenäosus omada leibkonnas lapsi ja autot ning uuringule vastanute hulgas on kvartiili keskmisest enam mitteaktiivseid. Auto omamisega kaasnevad leibkondadele mitmed transpordiga seotud täiendavad sundkulutused (näiteks liikluskindlustus, tehnohooldus, parkimine jne.), mis suurendavad leibkonna transpordikulutusi. Ilmselt suureneb lastega leibkondadel vajadus sõltumatu ja mugava transpordivõimaluse olemasoluks, mis selgitab kõrgemat auto omamise tõenäosust. Lapsed omakorda selgitavad mitteaktiivsete suuremat osakaalu segmendis, kuna mitteaktiivseteks loetakse teiste hulgas lapsega kodus olevad vanemad. Täiendavalt väärivad uurimist auto ning laste olemasolu ning transpordikulutuste seos. Samuti väärib täiendavat uurimist mugavusteenuste ja transpordikulutuste vaheline seos. Kõrge transpordikulude osakaaluga segmenti kirjeldab igas kvartiilis ka kõrgem kulutuste tase hotellidele, kohvikutele, restoranidele ning erinevatele kaupadele ja teenustele. Seetõttu sobib segment sihtkliendiks mugavusteenuseid pakkuvatele organisatsioonidele. Kõrge auto omamise tase segmendis viitab parkimise võimaldamise vajalikkusele mugavusteenuseid pakkuvate organisatsioonide läheduses. Eeltoodule tuginevalt väärib transpordikulutuste osakaal tunnusena kindlat kohta mugavusteenuseid pakkuvate organisatsioonide segmenteerimisülesannetes.

Segmendi *„Meeldivad riided ja väljas käimine“*, mis eristub III ja IV kvartiilis, leibkondi iseloomustab suurem tõenäosus et neis on üksikvanem, õpilane, mitteaktiivne või laps. Kõrgemaid kulutusi riietele ja väljas käimisele võib selgitada sooviga hea partneri leidmise nimel välja paista. Just noortel ja üksikvanematel on vajadus hea partneri leidmise järele üldjuhul kõrgem, kuna vajatakse tuge laste saamiseks ning kasvatamiseks. Muutujatevahelised seosed segmendis väärvivad kindlasti üksikasjalikumat uurimist. Sihtkliendi rühmana sobib segment riidekauplustele ja meelelahutusasutustele, kuna segmendi leibkonnad omavad ülejäänud segmentidest kõrgemat tarbimiseelarvet avaldudes vaid III ja IV kvartiilis ning väärtustavad oma tarbimises kulutusi just ilukaupadele ja toitlustusasutustele.

Kokkuvõtvana võib öelda, et kasumit taotleva organisatsiooni tarvis LEU andmete struktuuri uurides, on mõistlik uurimisprobleem täpsustada ning kaasata uuringusse vaid uurimisprobleemist tulenevalt otseselt huvipakkuvad tarbimiskulu liigid ning neid uurimise eesmärgist lähtuvalt transformeerida. Täiendavalt on soovitav klasteranalüüsi kaasata lisaks tarbimiskulutustele täiendavaid leibkondi kirjeldavaid tunnuseid. Käesolevas uurimuses osutusid tekkinud segmente hästi kirjeldavateks läbi kõikide kvartiilide auto olemasolu leibkonnas, perekonnaseis ja elamispinna kasutamise alus. Leibkonna geograafilise asukoha alusel erinesid segmendid oluliselt oma tarbimiseelistustes I kvartiilis. Teisisõnu selgitavad geograafilised tunnused leibkondade tarbimiseelistusi olulisel määral vaid madalama sissetulekuga leibkondade hulgas. Kui uurimise eesmärk ei ole tehnikate omavaheline võrdlemine, on mõistlik proovida klasteranalüüsi läbiviimiseks erinevaid sarnasusmõõdikuid. Klasteranalüüsi peatamise reeglitest sobib loomulike kliendisegmentide arvu tuvastamiseks andmetes Calinski-Harabaszi indeks. Sobilikuimaks klasteranalüüsi tehnikaks kliendisegmentide leidmisel osutus käesolevas magistritöös K-keskmiste tehnika.

KOKKUVÕTE

Aina tiheneva konkurentsi ja parema info ning teadmiste leviku tingimustes, ei piisa organisatsioonidele edukaks olemiseks enam vaid efektiivsest tootmise korraldamisest. Organisatsioonid peavad suutma täita klientide üha personaalsemaid soove ning olema võimelised reageerima klientide eelistuste muutumisele. Selleks peavad organisatsioonid oma klientidega teadlikult suhtlema. Lahenduse pakub kliendisuhete juhtimine – juhtimisstrateegia, mis kasvatab organisatsiooni väärtust läbi kliendisuhtlusest kogutud informatsioonile reageerimise.

Kliendisuhete juhtimine koosneb neljast üksteisega tihedalt seotud etapist: klientide identifitseerimisest, klientide meelitamisest, klientide hoidmisest ja klientide arendamisest. Kliendisuhete juhtimise edukaks juurutamiseks organisatsioonis ja seeläbi maksimaalse täiendava väärtuse loomiseks, peab kliendisuhete juhtimist vaatama kui tervikstrateegiat. Paraku rakendavad organisatsioonid praktikas kliendisuhete juhtimise erinevaid etappe valikuliselt ja jätavad pahatihti strateegiale, kliendi identifitseerimata jätmise näol, aluse loomata. Nii eksisteerib paradoksaalne olukord, kus organisatsioonid üritavad säilitada kliente keda nad ei tunne. Kliendisuhete juhtimise osad, mis on organisatsioonide poolt alarakendatud, on valdkonnad, kus on võimalik luua enim täiendavat väärtust. Seetõttu peitub kliendi identifiseerimises läbi turu segmenteerimise suur potentsiaal väärtusloomeks.

Organisatsioonid koguvad kliendisuhtlusest aina kasvaval hulgal informatsiooni, mistõttu muutub üha keerulisemaks olulise eristamine mitteolulisest. Relevantse info tuvastamisel on organisatsioonidele abiks andmekaeve – analüüsimeetod, mis võimaldab olulise info eraldamist andmetest ilma eelneva konkreetse probleemispüstituseta. Erinevad andmekaeve tehnikad võimaldavad otsustajatel hoomata suuremaid andmehulkasid ja fokuseerida oma tähelepanu olulisele, võimaldades saada vastuseid koguni küsimustele, mida küsidagi ei osata.

Kliendisegmentide tuvastamiseks sobivaim andmekaeve meetod on klasteranalüüs. Kasteranalüüs on meetod heterogeense andmekogu jagamiseks väiksematesse homogeensematesse gruppidesse ehk klastritesse. Inimkäitumise osaks on grupeerumine enda sarnastega ning eristumine endast erinevatest. Klasteranalüüsi käigus grupeeritaksegi objektid klastritesse omavahelise sarnasuse alusel. Sarnasus on aga subjektiivne mõiste ja võimaldab uurijatel klasteranalüüsi sooritada lugematul hulgal erinevatel viisidel ning saada sellest tulenevalt erinevaid tulemusi. Erinevatel viisidel läbiviidud klasterdamiste erinevad tulemused pole aga probleemiks, kuna klasteranalüüsi eesmärk on suuresti hüpoteeside genereerimine, mitte nende paikapidavuse testimine. Klasteranalüüsi kasutataksegi harva eraldiseisvana, kuna klastrite leidmine pole tihtipeale uurimise lõppeesmärk. Kui klastrid on välja selgitatud, kasutatakse teisi meetodeid mõistmaks, mida saadud tulemused endast kujutavad ning millised on klastrite tekkepõhjused. Seetõttu on kliendisegmentide moodustamisel mõistlik klasteranalüüsi abil uurida, millised loomulikud sarnase käitumisega klientide grupid eksisteerivad ning seejärel süveneda nende põhjalikumasse uurimisesse.

Klasteranalüüs ei sea olulisi eeldusi analüüsi läbiviimiseks kasutatavatele andmetele, kuid eelnev põhjalik töö andmetega aitab kaasa uurimisprobleemile sobilike lahendite leidmisele. Uurimisprobleemist lähtuvalt võib osutuda mõistlikuks andmete transformeerimine välistamaks klastrite tekkimist mõne üksiku domineeriva tunnuse alusel, mis välistab andmetest loomulike klastrite tuvastamise uurijat tegelikult huvitavate tunnuste alusel. Käesoleva magistritöö läbiviimisel kasutatud andmetes on kogu leibkonna tarbimiseelarve tugevat korreleerunud erinevate tarbimiskulu liikidega. Vältimaks klastrite teket vaid leibkondade tarbimise kogukulu alusel ja saamaks tulemuseks leibkondade tarbimiseelistusi väljendavaid klastreid, viidi klasteranalüüs läbi leibkonna tarbimise kogukulu alusel moodustatud neljas kvartiilis ning transformeeriti kulutused kulutuste osakaaluks leibkonna tarbimiseelarvest.

Magistritöö käigus leitud segmente iseloomustab selge eristumine üksteisest erinevate tarbimiskulu liikide keskmise väärtuse alusel. Samas on muutujate väärtuste varieerumine segmentides küllaltki suur ning segmendid seetõttu küllaltki heterogeensed. Ühe tunnuse alusel leidub pea igas segmendis sinna esmapilgul mitte sobivaid leibkondi. Olukorra tingis analüüsi kaasatud suur muutujate arv. Magistritöös

kaasati analüüsi kõik teada olevad tarbimiskulu liigid, leidmaks ülist andmetes peituvat struktuuri. Konkreetse organisatsiooni huvidest lähtuva segmenteerimise läbiviimisel, on homogeensemate segmentide leidmise eesmärgil, mõistlik klasteranalüüsi kaasata väiksem arv tarbimiskulutusi kirjeldavaid muutujaid ning valida muutujad konkreetsest uurimisprobleemist lähtuvalt.

Erinevate tehnikaga leitud klasterdamise tulemuste omavahelise võrreldavuse eesmärgil kasutati magistritöö empiirilises osas sarnasuse mõõdikuna eukleidilist kaugust ruudus. Kui klasteranalüüsi läbiviimise eesmärgiks on konkreetsete kliendisegmentide tuvastamine, on mõistlik proovida ka alternatiivseid sarnasuse mõõdikuid.

Klasteranalüüsi tehnikad jagunevad hierarhilisteks ning mittehierarhilisteks. Käesolevas magistritöös võrreldi kliendisegmentide tuvastamisel hierarhilistest klasterdamise tehnikatest üksiku seose, keskmise seose, täieliku seose ja Wardi tehnikaid ning ainsa mittehierarhilise tehnikana K-keskmiste tehnikat. Kõikidel klasterdamise tehnikatel on nii omad tugevused kui ka puudused. Seetõttu sobivad erinevad klasterdamise tehnikad erinevate probleemide lahendamiseks. Kliendisegmentidena kasutamiseks on sobilikud klastrid, mis on piisavalt suured, et pakkuda mastaabisäästu, kuid samas piisavalt väikesed, et pakkuda sinna kuuluvatele klientidele võimalikult nende personaalsetele eelistustele sobivaid lahendusi. Seega otsitakse kliendisegmentidena klastreid, mille suurused on võimalikult võrdsed ning milledesse kuuluvad võimalikult sarnaste eelistustega kliendid. Võrreldud tehnikatest annab tulemuseks väiksema klastrite suuruste ja kujude varieeruvusega klastrid K-keskmiste tehnika.

Kliendisegmentide moodustamisel otsitakse loomulikult eksisteerivaid kliendigruppe, ilma nende kohta eeldusi tegemata. Seetõttu ei määranud magistritöö autor eelnevalt otsitavate kliendisegmentide arvu vaid kasutas kliendisegmentide arvu määramiseks matemaatilisi teste. Kõikide magistritöös kasutatud klasterdamise tehnikate korral on andmetes eksisteerivate klastrite hulga määramiseks võimalik kasutada Calinski-Harabaszi indeksit. Calinski-Harabaszi indeks sobib hästi loomulike kliendisegmentide tuvastamiseks, kuna eelistab lahendeid, millel on suur klastrite vaheline varieeruvus ja väike klastrite sisene varieeruvus**,** mis iseloomustab kompaktseid sfäärjaid klastreid**.** Hierarhiliste tehnikate korral kasutati täiendavalt Duda-Hart indeksit. Kuna otsiti

sfäärijaid klastreid, piisab segmentide kirjeldamiseks neid iseloomustavate muutujate keskmiste väärtuse võrdlemisest andmekogumi keskmisega.

Magistritöös läbiviidud klasteranalüüsis osutus kõikides tarbimise kogukulu alusel moodustatud leibkondade kvartiilides Calinski-Harabaszi indeks väärtuse alusel parimaks K-keskmiste tehnikaga leitud lahend. K-keskmiste tehnika parimat sobivust kliendisegmentide tuvastamiseks kinnitas ka erinevate tehnikatega leitud klastrite suuruste uurimine, mille käigus tuvastati, et K-keskmiste tehnika andis tulemuseks kõige sarnasema suurusega klastrid. Seega leidis kinnitust teoreetilises osas tekkinud ootus, et parimaks tehnikaks kliendisegmentide tuvastamisel magistritöös võrreldud tehnikatest on K-keskmiste tehnika.

Tuvastatud kliendisegmentide arv erinevates kvartiilides kasvab koos leibkondade tarbimise kogukulu kasvuga. Madalamate kvartiilide segmendid esindavad leibkondade jagunemist ellu jäämiseks vajalike kulutuste alusel, kõrgema taseme kvartiilides lisanduvad sotsiaalsetest ja vaimsetest vajadustest lähtuvate kulutuste alusel eristuvad segmendid. See tähendab, et tarbimiskulutuste kvartiili tõusuga kaasneb leibkonna tarbimisvabaduse suurenemine. Jagunemine on kooskõlas Maslow inimvajaduste hierarhiaga. Sellest järeldub, et sundkulutuste alusel läbiviidav segmenteerimine annab laiema kliendibaasi katvuse. Muude kulutuste alusel leitavad segmendid kirjeldavad kitsamaid sihtgruppe kõrgemates tarbimise kogukulu kvartiilides. Laiapõhise kliendibaasiga organisatsioonide jaoks võib osutuda seega sobivamaks sundkulutuste alusel segmenteerimine. Kitsa turuniši leidmiseks sobib aga paremini klientide segmenteerimine muude kulutuste alusel.

Läbi erinevate kvartiilide korduvad sarnaste tarbimiseelistustega segmendid, millele autor andis lähtuvalt kulutuste struktuurist neid iseloomustava nime. Kokku tuvastas käesoleva magistritöö autor kaheksa üksteisest tarbimiseelistustelt eristuvat segmenti, millest kolm: *„Kulutame eluasemele“*, *„Sööme, joome ja suitsetame“* ning *„Kulutame transpordile ja kogemustele,* on esindatud kõikides kvartiilides. III ja IV kvartiilis on kordub segment *„Meeldivad riided ja väljas käimine“.*

Tuvastatud kliendisegmente iseloomustati täiendavalt läbi tunnuste, mida klasterdamiseks ei kasutatud, kuid leiduvad LEU 2012 andmetest. Selgus, et sama

nimega segmendid eristuvad teistest segmentidest samade tunnuste alusel. Segmendi *„Kulutame eluasemele“* leibkondi iseloomustab kõigis kvartiilides kvartiili keskmisest kõrgem korteris, üüripinnal ja Põhja- või Kirde-Eestis elamise tõenäosus. Segmenti *„Sööme, joome ja suitsetame“* iseloomustab kõikides kvartiilides kõrgem tõenäosus, et uuringu küsimustele on vastanud pensionär, leibkonna näol on tegu üksiku inimese või lasteta paariga ning haridustase segmendis on keskmisest madalam. Segmenti *„Kulutame transpordile ja kogemustele“* iseloomustab suurem tõenäosus omada leibkonnas autot ja lapsi ning uuringule vastanute hulgas on kvartiili keskmisest enam mitteaktiivseid. Segmendi *„Meeldivad riided ja väljas käimine“* leibkondi iseloomustab suurem tõenäosus olla lastega leibkond. Samuti on segmendi leibkondades keskmisest enam üksikvanemaid, õpilasi ja mitteaktiivseid. Edasist uurimist vajavad segmente kirjeldavate tunnuste esinemise põhjused. Nende põhjuste tundmine võimaldab leitud kliendisegmente paremini mõista ning sellest lähtuvalt töötada välja organisatsiooni strateegia nende kliendisegmentidega suhtlemisel.

Klasteranalüüs sobib ideaalselt kasumit taotlevale organisatsioonile loomulikult eksisteerivate kliendisegmentide tuvastamiseks ja aitab seeläbi korraldada tulemuslikumalt kliendisuhete juhtimist ja viib seega täiendava väärtusloomeni. LEU andmete alusel kliendisegmentide loomisel on mõistlik uurimisprobleem täpsustada ning kaasata klasterdamisse vaid uurimisprobleemist tulenevalt uurijale otseselt huvipakkuvad tarbimiskulu liigid ning neid uurimise eesmärgist lähtuvalt transformeerida. Täiendavalt on soovitav klasteranalüüsi kaasata lisaks tarbimiskulutustele ka teisi leibkondi kirjeldavaid tunnuseid. Käesolevas magistritöös osutusid tekkinud segmente hästi kirjeldavateks läbi kõikide kvartiilide auto olemasolu leibkonnas, perekonnaseis ja elamispinna kasutamise alus. Kliendisegmentide tuvastamiseks parimad tulemused saavutatakse K-keskmiste tehnikat kasutades ja klastrite arvu määramisel matemaatiliste testide abil.

VIIDATUD ALLIKAD

1. **Abbott, A.** The Casual Devolution. – Cluster Analysis, Vol.1, Logic and Classics, SAGE Benchmarks in Social Research Methods, Edited by Byrne, D. and Uprichard, E., Los Angeles, London, New Delhi, Singapore and Washington DC, 2012, pp. 15–42.

2. **Anderson, J. L., Jolly, L. D. and Fairhurst, A. E.** Customer relationship management in retailing: A content analysis of retail trade journals. – Journal of Retailing and Consumer Services, 2007, Vol 14, pp. 394–399.

3. **Ashdown, S. P.** An Investigation of the Structure of Sizing Systems: A Comparison of Three Multidimensional Optimized Sizing Systems Generated from Anthropometric Data. – International Journal of Clothing and Technology, 1998, Vol. 10, 5, pp. 324–341.

4. **Bae, S. M., Park, S. C. and Ha, S. H.** Fuzzy web ad selector based on web usage mining. – IEEE Intelligent Systems, 2003, pp. 62–69.

5. **Bahrami, M, Ghorbani, M. and Arabzad, M.** Information Technology (IT) as An Improvement Tool For Customer Relationship Management (CRM). – Procedia – Social and Behavioral Sciences, 2012, Vol. 41, pp. 59–64.

6. **Bailey, K. D.** Sociological Classification and Cluster Analysis. – Cluster Analysis, Vol. 2, (Useful) Key Texts, SAGE Benchmarks in Social Research Methods, Edited by Byrne, D. and Uprichard, E., Los Angeles, London, New Delhi, Singapore and Washington DC, 2012, pp. 83–100.

7. **Berry, M. J.A. and Linoff, G. S.** Data Mining Techniques For Marketing, Sales, and Customer Relationship Management. Second Edition. Indianapolis, Indiana: Wiley Publishing, Inc., 2004, 643p.

8. **Berson, A., Smith, S. and Thearling, K.** Building data mining applications for CRM. – McGraw-Hill, 2000.

9. **Boulding, W., Staelin, R., Ehret, M. and Johnston, W. J.** A Customer Relationship Management Roadmap: What Is Known, Potential Pitfalls. And Where to Go. – Journal of Marketing, 2005, Vol. 69, No. 4, pp. 155–166.

10. Business banking: The face of the future – Will SMEs point the way back to profitability? EFMA, Paris, March 2012, 55p.

11. **Calinski, T. and J. Harabasz** A dendrite method for cluster analysis. – Communications in Statistics, 1974 Vol. 3, No. 1, pp. 1–27.

12. **Chalmeta, R.** Methodology for customer relationship management. – The Journal of Systems and Sorfware, 2006, Vol. 79, pp. 1015–1024.

13. **Dennis, C., Marsland, D. and Cockett, T.** Data mining for shopping centers-customer knowledge management framework. – Journal of Knowledge Management, 2001, Vol 5. pp. 368-374.

14. **Hormazi, A. M and Giles, S.** Data Mining: A Competitive Weapon for Banking and Retail Industries. – Information Systems Management, 2004, vol.21, Issue 2, pp. 62–71.

15. **Jain, A. K., Murty, M. N. and Flynn, P.J.** Data Clustering: A Review. – Cluster Analysis, Vol.1, Logic and Classics, SAGE Benchmarks in Social Research Methods, Edited by Byrne, D. and Uprichard, E., Los Angeles, London, New Delhi, Singapore and Washington DC, 2012, pp. 305–381.

16. **King, S. F. and Burgess, T. F.** Understanding success and failure in customer relationship management. – Industrial Marketing Management, 2008, Vol. 37, pp. 421–431.

17. **Kracklauer, A. H., Mills, D. Q. and Seifert, D.** Customer management as the origin of collaborative customer relationship management. – Collaborative Customer relationship Management – taking CRM to the next level, Edited by

Kracklauer, A. H., Mills, D. Q. and Seifert, D., Heidelberg: Springer, 2004, pp. 3–6

18. **Kuo, R. J., Ho, L. M. and Hu, C. M.** Integration of self-organizing feature map and K-means algorithm for marketing segmentation. – Computers and Operations Research, 2002, 29(11), pp. 1475–1493.

19. **Lee, J. H. and Park, S. C.** Intelligent profitable customers segmentation system based on business intelligence tools. – Expert Systems with Applications, 2005, Vol. 29, pp. 145–152.

20. **Lee, S. C., Suh, Y. H., Kim, J. K. and Lee, K. J.** A cross-national market segmentation of online game industry using SOM. – Expert Systems with Applications, 2004, Vol. 27, pp. 559–570.

21. Leibkonna eelarve uuring 2012. Statistikaamet [http://www.stat.ee/leu-2012-andmefail]. 03.12.2013

22. **Lia, S., Chen, Y. and Deng, M.** Mining customer knowledge for tourism new product development and customer relationship management. – Expert Systems with Applications, 2010, Vol. 37, pp. 4212–4223.

23. **Ling, R. and Yen, D. C.** Customer relationship management: An analysis framework and implementation strategies. – Journal of Computer Information Systems, 2001, Vol. 41, pp. 82–97.

24. **Mendoza, L. E., Marius, A., Pérez, M. and Grimán, A. C.** Critical success factors for a customer relationship management strategy. – Information and Software Technology, 2007, Vol. 49, pp. 913–945.

25. **Milligan, G. W. and Cooper, M. C.** An Eximination of Procedures for Determing the Number of Clusters in a Data Set. – Cluster Analysis, Vol. 2, (Useful) Key Texts, SAGE Benchmarks in Social Research Methods, Edited by Byrne, D. and Uprichard, E., Los Angeles, London, New Delhi, Singapore and Washington DC, 2012, pp. 235–259.

26. Mining Business Databases, Communications of the ACM, November 1996, Vol. 39, No.11, pp. 42–48.

27. **Moin, K. I. and Ahmed, Q. B.** Use of Data Mining in Banking. – International Journal of Engineering Research and Applications, 2012, Vol.2, Issue 2, pp. 738–742.

28. **Ngai, E.W.T., Xiu, L. and Chau, D.C.K.** Application of data mining techniques in customer relationship management: A literature review and classification. – Expert Systems with Applications, 2009, Vol. 36, pp. 2592–2602.

29. **Norušis, M. J.** IBM SPSS Statistics 19 Statistical Procedures Companion. New Jersey: Prentice Hall, 2011, 672 p.

30. **Phan, D. D. and Vogel, D. R. A** model of customer relationship management and business intelligence systems for catalogue and online retailers. – Information and Management, 2010, Vol. 47. pp. 69–77.

31. **Ramassastry, A.** Web sites change prices based on customers' habits, Law Center, [http://edition.cnn.com/2005/LAW/06/24/ramasastry.website.prices/], June 24, 2005 (accessed 29.12.2013)

32. **Rigby, D. K., Ledingham, D.** CRM done right. – Harvard Business Review, 2004, 82 (11), pp. 118–129.

33. **Rust, R. T., Lemon, K. N. and Zeithaml, V. A.** Return on Marketing: Using Customer Equity to Focus Marketing Strategy. – Journal of Marketing, 2005, Vol. 69, pp. 252–261.

34. **Rygielski, C., Wang, J. and Yen, D. C.** Data mining techniques for customer relationship management. – Technology in Society, 2002, Vol. 24 pp. 483–502.

35. SAS Institute Ice. SAS/STAT 13.1 User's GuideCary, NC: SAS Institute Inc., 2013, 9480 p.

36. StataCorp. Stata multivariate Statistics Reference Manual, Release 11, A Stata Press Publication, College Station, TX: StataCorp LP., 2009, 694 p.

37. **Tan, P. N., Steinbach, M. And Kumar, V.** Introduction to Data Mining, 1st Edition, Addison-Wesley, 2006, 769 p.

38. **Tikva, P. ja Arnik, K.** Leibkonna eelarve uuring 2010. Metoodika ülevaade, Tallinn: Statistikaamet, 2012, 211 lk.

39. Turu segmenteerimine. Ettevõtluse Arendamise Sihtasutus [http://www.eas.ee/-et/alustavale-ettevotjale/aeriarendus-ja-uute-klientide-voitmine/turundus-ja-mueuek/turu-segmenteerimine]. 12.03.2014

40. **Wang, S.** Cluster Analysis using a validated self-organizing method: Case of problem identification. – Intelligent Systems in Accounting, Finance and Management, 2001, 10(2), pp. 127–138

41. **Wang, Y.** A clustering method mased on fuzzy equivalence relation for customer relationship management. – Expert Systems with Applications, 2010, Vol 37, pp. 6421–6428.

42. **Wei, J., Lee, M., Chen, H., Wu, H.** Customer relationship management in the hairdressing industry: An application of data minging techniques. – Expert Systems with Applications, 2013, Vol. 40, pp. 7513–7518.

43. **Verdú, S. V., Gracía, M. O., Senabre, C., Marín, A. G. and Franco, F. J. G.** Classification, filtering and identification of electrical customer load patterns through the use of self-organizing maps. – IEEE Transactions on Power Systems, Vol. 21, pp. 1672–1682.

44. **Vesanto, J. and Alhoniemi, E.** Clustering of the self-organization map. – IEEE Transactions on Neural Networks, 2000, 11(3), pp. 568–600.

45. **Yang, Y. and Padmanabhan, B.** A hierarchical pattern-based clustering algorithm for grouping web transactions. – IEEE Transaction on Knowledge and Data Engineering, Vol. 17, pp. 1300–1304.

LISAD

Lisa 1. Tarbimiskulu liikide paariskorrelatsioonid

Muutuja		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	Tarbimiskulutused kokku	1												
2	Toit ja mittealkohoolsed joogid	0,6093*	1											
3	Alkohoolsed joogid ja tubakas	0,3447*	0,2791*	1										
4	Riided ja jalanõud	0,4207*	0,1835*	0,0914*	1									
5	Eluase	0,4094*	0,2158*	0,0496*	0,0819*	1								
6	Majapidamiskulud	0,4973*	0,2377*	0,1510*	0,1557*	0,1217*	1							
7	Tervishoid	0,4079*	0,2241*	0,0401*	0,0903*	0,1459*	0,1221*	1						
8	Transport	0,7111*	0,2335*	0,1605*	0,1505*	0,0847*	0,2337*	0,2000*	1					
9	Side	0,4198*	0,2244*	0,1506*	0,1502*	0,2101*	0,1686*	0,0810*	0,2233*	1				
10	Vaba aeg	0,6619*	0,3244*	0,1822*	0,2503*	0,1805*	0,3237*	0,1946*	0,3337*	0,2803*	1			
11	Haridus	0,2164*	0,0261*	0,0188	0,0934*	0,0638*	0,0747*	0,0742*	0,1113*	0,1155*	0,1332*	1		
12	Hotellid, kohvikud, restoranid	0,5692*	0,1953*	0,1627*	0,2534*	0,1437*	0,1927*	0,1794*	0,4059*	0,2179*	0,3814*	0,1547*	1	
13	Mitmesugused kaubad ja teenused	0,5633*	0,2448*	0,1166*	0,2091*	0,1202*	0,2084*	0,2981*	0,3502*	0,2175*	0,3400*	0,1530*	0,3568*	1

Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Märkused: * $p < 0,05$

Lisa 2. Tarbimiskulu liikide osakaalude paariskorrelatsioonid.

Muutuja		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	Tarbimiskulutused kokku	1												
2	Toit ja mittealkohoolsed joogid	-0,3707*	1											
3	Alkohoolsed joogid ja tubakas	-0,0275*	-0,0188	1										
4	Riided ja jalanõud	0,1620*	-0,2273*	-0,0760*	1									
5	Eluase	-0,3374*	-0,0811*	-0,1567*	-0,2114*	1								
6	Majapidamiskulud	0,1747*	-0,1964*	-0,0707*	-0,0123	-0,1956*	1							
7	Tervishoid	0,0902*	-0,0969*	-0,1148*	-0,0335*	-0,0963*	-0,0246*	1						
8	Transport	0,3989*	-0,4153*	-0,0584*	-0,0336*	-0,3826*	-0,0154	-0,0633*	1					
9	Side	-0,3222*	-0,0294*	-0,0435*	-0,1245*	0,1757*	-0,1368*	-0,1414*	-0,2037*	1				
10	Vaba aeg	0,2172*	-0,2449*	-0,0709*	0,0244*	-0,2597*	0,0437*	-0,0598*	-0,0258*	-0,0966*	1			
11	Haridus	0,0983*	-0,1367*	0,0188	0,0201	-0,0663*	0,0138	-0,0082	0,0176	-0,0359*	0,0356*	1		
12	Hotellid, kohvikud, restoranid	0,3209*	-0,2593*	-0,0354*	0,0819*	-0,2002*	-0,0064	-0,0378*	0,1431*	-0,1020*	0,1048*	0,0599*	1	
13	Mitmesugused kaubad ja teenused	0,1230*	-0,1890*	-0,0710*	0,0628*	-0,2128*	0,0084	-0,0301*	-0,0008	-0,0569*	0,0379*	0,0240*	0,0831*	1

Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Märkused: * $p < 0,05$

Lisa 3. Tarbimiskulu liikide osakaalude paariskorrelatsioonid I kvartiilis.

Muutuja		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	Toit ja mittealkohoolsed joogid	1											
2	Alkohoolsed joogid ja tubakas	-0,0919*	1										
3	Riided ja jalanõud	-0,1630*	-0,0247	1									
4	Eluase	-0,4614*	-0,1976*	-0,1499*	1								
5	Majapidamiskulud	-0,1640*	-0,0607*	0,0312	-0,1775*	1							
6	Tervishoid	-0,1317*	-0,0952*	-0,0220	-0,1227*	-0,0084	1						
7	Transport	-0,2445*	-0,0213	0,0001	-0,2132*	0,0234	0,0034	1					
8	Side	-0,2682*	-0,0867*	-0,0406	0,0107	-0,0735*	-0,0863*	-0,0555*	1				
9	Vaba aeg	-0,1876*	-0,0811*	0,0436*	-0,2063*	0,0446*	-0,0870*	0,0057	-0,0395	1			
10	Haridus	-0,1147*	-0,0301	0,0084	-0,0019	0,0172	0,0439*	0,0183	0,0174	0,0106	1		
11	Hotellid, kohvikud, restoranid	-0,1117*	-0,0157	0,0825*	-0,0316	-0,0313	-0,0165	0,0480*	-0,0073	-0,0164	-0,0174	1	
12	Mitmesugused kaubad ja teenused	-0,1319*	-0,0284	-0,0147	-0,1940*	0,0435*	-0,0506*	-0,0032	0,0272	0,008	0,0188	0,0656*	1

Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Märkused: * $p < 0,05$

Lisa 4. Tarbimiskulu liikide osakaalude paariskorrelatsioonid II kvartiilis.

Muutuja		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	Toit ja mittealkohoolsed joogid	1											
2	Alkohoolsed joogid ja tubakas	-0,0601*	1										
3	Riided ja jalanõud	-0,2055*	-0,0966*	1									
4	Eluase	-0,2140*	-0,1979*	-0,1327*	1								
5	Majapidamiskulud	-0,1297*	-0,0825*	-0,0384	-0,1330*	1							
6	Tervishoid	-0,0595*	-0,1296*	-0,0983*	-0,0808*	-0,0046	1						
7	Transport	-0,3525*	-0,0479*	-0,0495*	-0,3183*	-0,1265*	-0,0915*	1					
8	Side	-0,2219*	-0,0339	-0,0506*	0,0310	-0,0446*	-0,1020*	-0,0398	1				
9	Vaba aeg	-0,2445*	-0,0363	-0,0197	-0,2166*	0,0077	-0,0574*	-0,0307	0,0041	1			
10	Haridus	-0,0921*	-0,0307	-0,0122	-0,0675*	0,0188	0,0055	-0,0162	-0,0098	0,0109	1		
11	Hotellid, kohvikud, restoranid	-0,1542*	-0,0242	0,002	-0,1096*	-0,0493*	-0,0694*	0,0641*	0,0402	0,0441*	0,0108	1	
12	Mitmesugused kaubad ja teenused	-0,1967*	-0,0798*	0,0808*	-0,1770*	-0,0258	0,0051	-0,0258	-0,0531*	0,0255	0,039	0,0668*	1

Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Märkused: * $p < 0,05$

Lisa 5. Tarbimiskulu liikide osakaalude paariskorrelatsioonid III kvartiilis.

Muutuja		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	Toit ja mittealkohoolsed joogid	1											
2	Alkohoolsed joogid ja tubakas	0,0065	1										
3	Riided ja jalanõud	-0,2319*	-0,0907*	1									
4	Eluase	-0,1045*	-0,1563*	-0,1965*	1								
5	Majapidamiskulud	-0,1757*	-0,0905*	-0,0591*	-0,1155*	1							
6	Tervishoid	-0,0176	-0,1231*	-0,0433*	-0,0326	-0,0762*	1						
7	Transport	-0,3727*	-0,0742*	-0,1167*	-0,3090*	-0,0313	-0,1426*	1					
8	Side	-0,0736*	-0,0602*	-0,0905*	0,0817*	-0,1159*	-0,1567*	-0,1174*	1				
9	Vaba aeg	-0,1838*	-0,0796*	-0,0007	-0,2401*	-0,0367	-0,1192*	-0,1191*	0,0237	1			
10	Haridus	-0,1527*	-0,0259	0,0099	-0,0333	0,0012	-0,0708*	0,0052	0,0166	0,0614*	1		
11	Hotellid, kohvikud, restoranid	-0,1923*	-0,0463*	0,0390	-0,1588*	-0,0770*	-0,0994*	0,0587*	0,0117	0,1113*	0,0123	1	
12	Mitmesugused kaubad ja teenused	-0,1461*	-0,0732*	0,0456*	-0,2048*	-0,0117	-0,0730*	-0,0850*	-0,0012	0,0305	-0,024	0,0207	1

Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Märkused: * $p < 0,05$

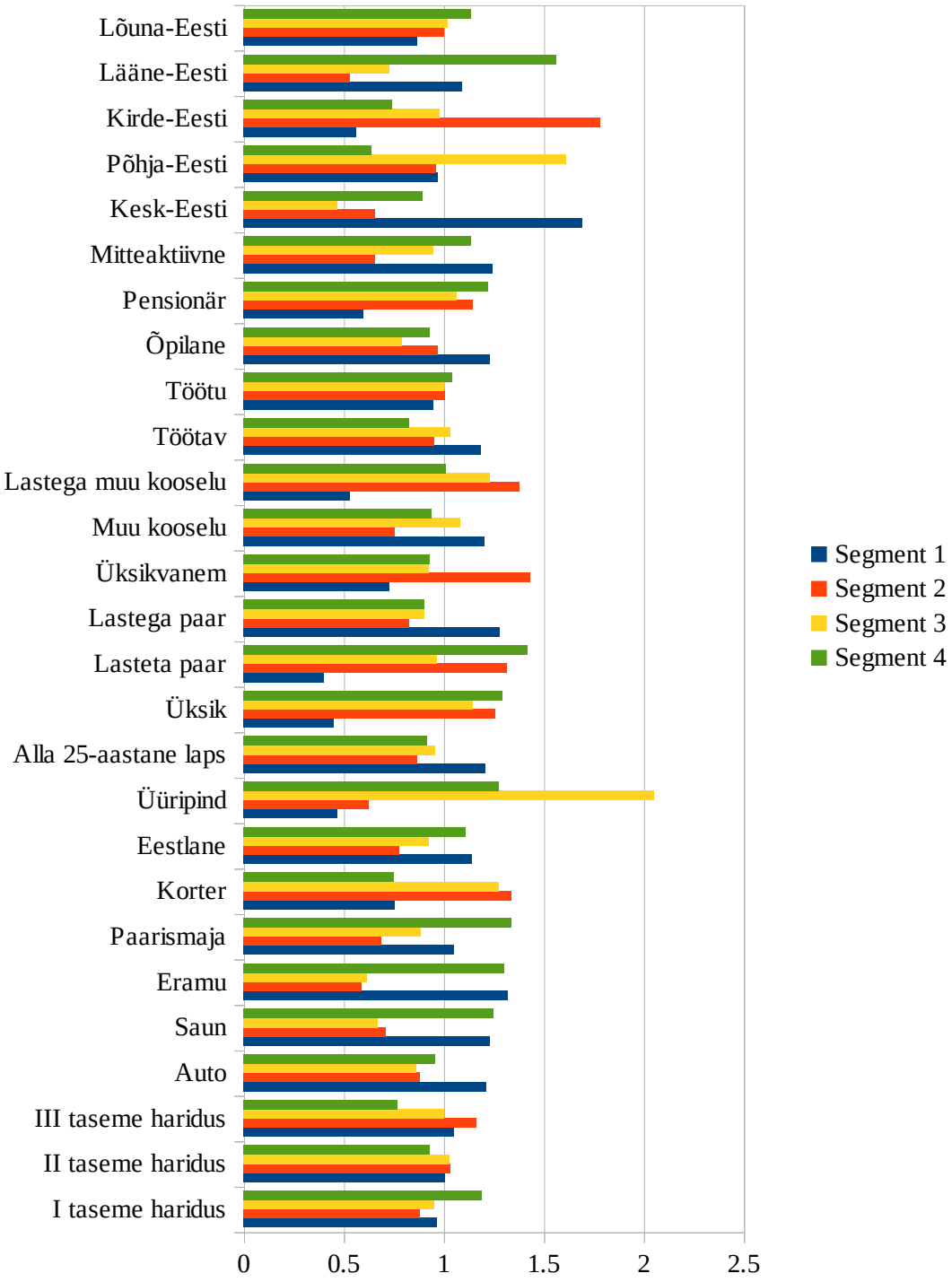
Lisa 6. Tarbimiskulu liikide osakaalude paariskorrelatsioonid IV kvartiilis.

Muutuja		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	Toit ja mittealkohoolsed joogid	1											
2	Alkohoolsed joogid ja tubakas	0,1283*	1										
3	Riided ja jalanõud	-0,1285*	-0,1047*	1									
4	Eluase	0,0190	-0,0861*	-0,1439*	1								
5	Majapidamiskulud	-0,1155*	-0,0545*	-0,0934*	-0,1326*	1							
6	Tervishoid	-0,0087	-0,1311*	-0,0760*	0,0488*	-0,0872*	1						
7	Transport	-0,3999*	-0,0962*	-0,2012*	-0,3561*	-0,1702*	-0,1834*	1					
8	Side	0,0065	0,0697*	-0,0508*	0,0709*	-0,0797*	-0,0927*	-0,1293*	1				
9	Vaba aeg	-0,1292*	-0,0901*	-0,0690*	-0,1368*	-0,0028	-0,0847*	-0,2392*	-0,0583*	1			
10	Haridus	-0,1130*	-0,0711*	-0,0020	-0,0245	-0,0327	-0,0265	-0,0532*	-0,0168	-0,0163	1		
11	Hotellid, kohvikud, restoranid	-0,2702*	-0,0448*	0,0189	-0,1361*	-0,0826*	-0,0930*	0,0047	-0,0232	0,0166	0,0609*	1	
12	Mitmesugused kaubad ja teenused	-0,1507*	-0,1226*	0,0190	-0,1158*	-0,0634*	-0,0686*	-0,0866*	0,0018	-0,0234	0,0164	0,0564*	1

Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

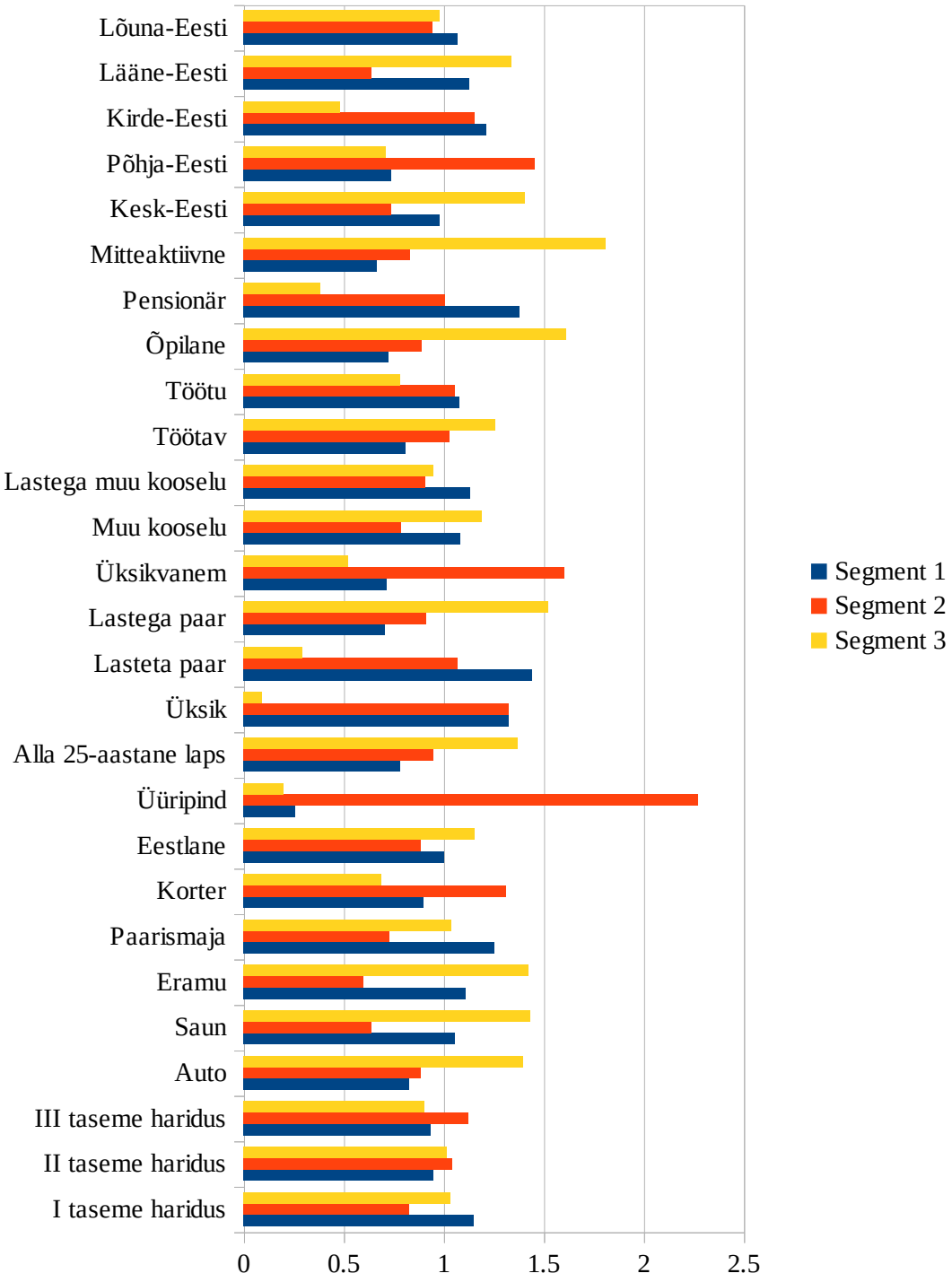
Märkused: * $p < 0,05$

Lisa 7. Kategooriliste tunnuste esinemine I kvartiili segmentides



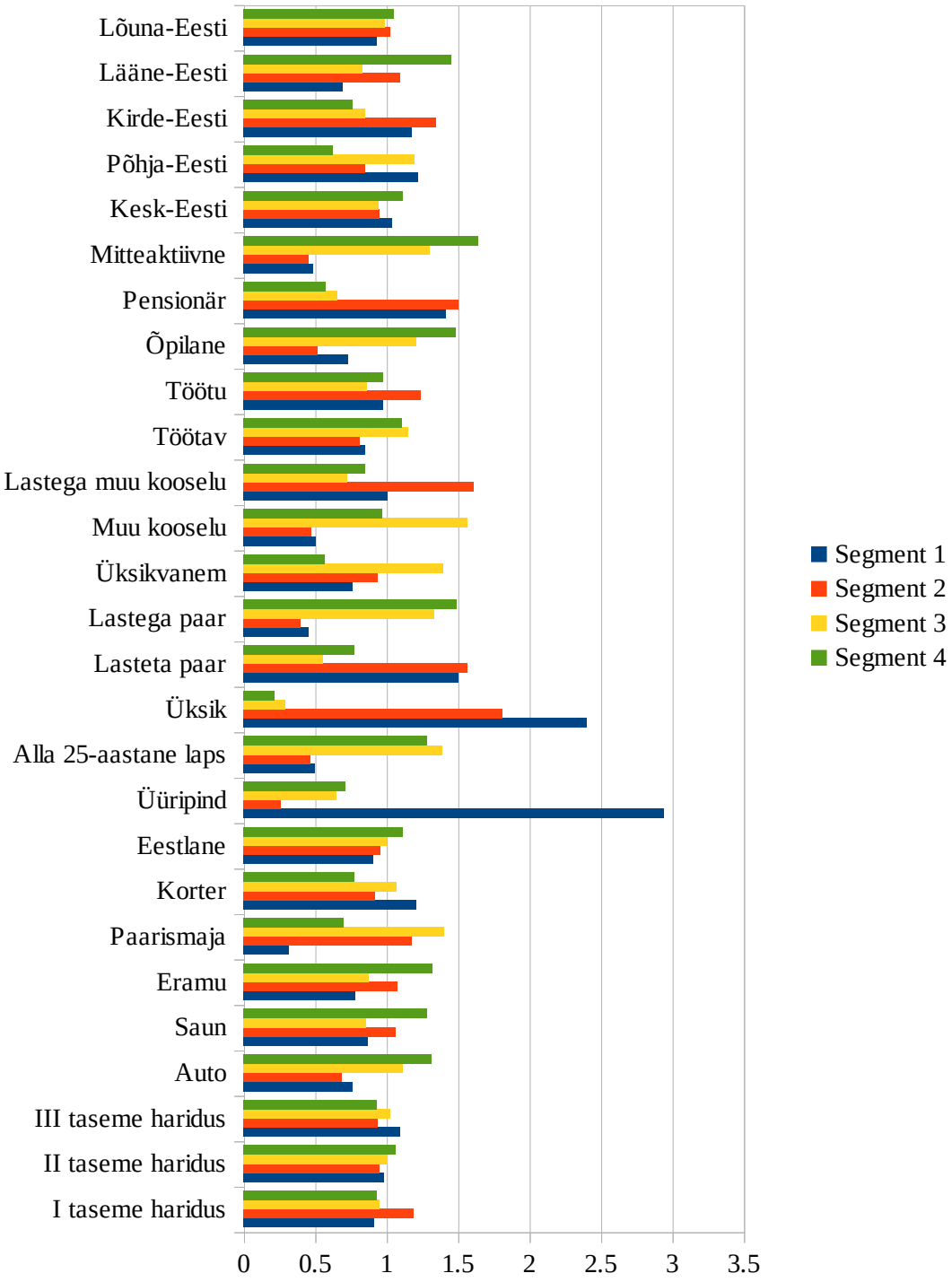
Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Lisa 8. Kategooriliste tunnuste esinemine II kvartiili segmentides



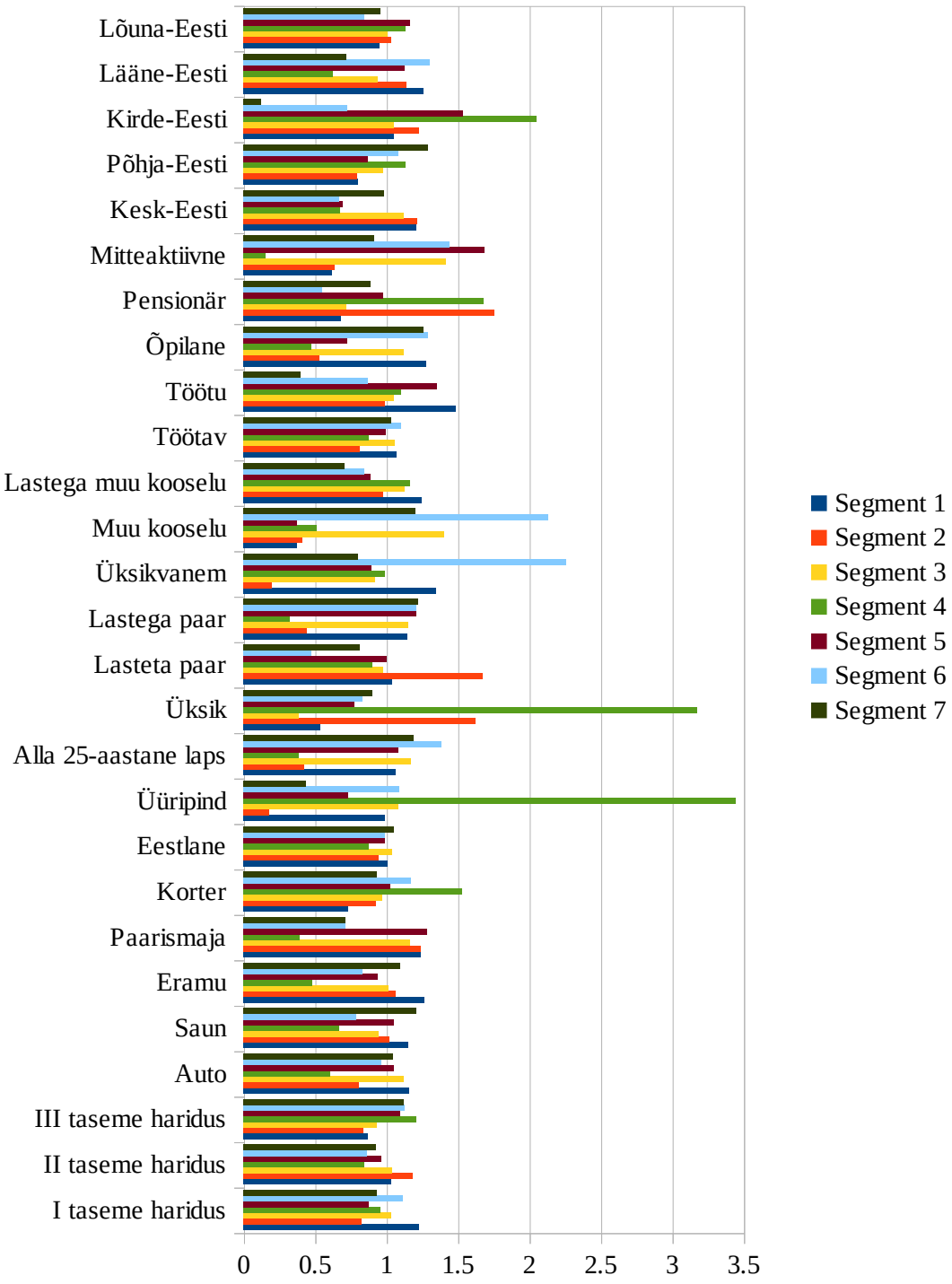
Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Lisa 9. Kategooriliste tunnuste esinemine III kvartiili segmentides.



Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

Lisa 10. Kategooriliste tunnuste esinemine IV kvartiili segmentides.



Allikas: (Leibkonna eelarve ... 2013); autori koostatud.

SUMMARY

USING CLUSTER ANALYSIS FOR IDENTIFYING NATURAL CLIENT SEGMENTS AMONG ESTONIAN HOUSEHOLDS

Kaarel Aus

In a growingly competitive environment, organizations need to do more than just efficiently manage their production in order to be successful. Organizations have to be able to meet their clients' more personal needs and react promptly to changes in clients' preferences. To do that, organizations need to know who their clients are and communicate systematically with them to understand their needs.

Expectations clients have towards organisations depend on the products and services the organization offers and on the personal preferences the clients have. To meet clients' expectations organizations can offer solutions that are exactly the same for all customers or special solutions for each client, as two extremes. Offering special solutions for each client is the most expensive solution and therefore usually economically unreasonable. One universal solution for all clients may not be the most preferred solution by the clients and organizations can lose their clients to competitors who are able to offer solutions that meet client expectations more. Therefore, it is reasonable to compromise by offering solutions developed specifically for client segments that consist of clients who have similar preferences. Primary mission for organizations that want to be successful is to identify such client segments and adjust customer services according to the needs of those client segments. That way value is created simultaneously for the organization as well as its clients.

Clients can be divided into client segments with similar behaviour from the organization's perspective simply by classifying them on the basis of subjectively chosen variables. Alternatively organizations can use data mining to explore the

structure in data and identify natural client segments. Natural structure of data can reveal even such segments that the organizations have not been aware of before. Those client segments represent clients who have so far been ignored and are a perfect target for organizations interested in conquering a market share. On the other hand, awareness of such client segments helps organizations to defend their market position and grow profitability.

Identifying natural client segments creates a strong basis for building up client relationship management (CRM) system. The objective of the current master thesis is to introduce the possibilities cluster analysis offers for identifying natural client segments on the basis of households' consumption preferences. The master thesis gives recommendations for using cluster analysis for client segmentation and guides researchers new to cluster analysis through the cluster analysis process.

For reaching the objective five research tasks have been set for the master thesis:

1) to give an overview of CRM and data mining,
2) to introduce the methodology of cluster analysis,
3) to determine the most suitable cluster analysis technique for client segmentation,
4) to describe the natural client segments found among Estonian households,
5) to give recommendations for using cluster analysis for client segmentation.

To introduce the relevance of identifying natural client segments, overview of CRM and data mining is given. CRM is defined as a management strategy that increases the value of organization through reacting to the information collected from clients. CRM enables organizations to regulate their communication with their clients and helps organizations to react and fulfil the changing and ever growing wishes of their clients.

CRM consists of four closely bounded phases: client identification, client attraction, client retention and client development. For successful implementation, CRM has to be seen as a comprehensive strategy. Unfortunately, in practice, the different phases of CRM are implemented selectively and too often the strategy lacks a foundation as client identification phase has been omitted. A paradoxical situation exists - organizations try to maintain clients whom they do not even know. The inadequately exploited areas of

CRM are the fields that have the most potential for value creation. Therefore, client identification deserves more attention.

From interaction with their clients, organizations collect huge amounts of data. This data is hard to analyse using traditional methods and special methods need to be used. One option is to use data mining. In the current master thesis data mining is defined as an analysis method that deals with extracting relevant information from data without limiting to concrete investigation tasks. Different data mining techniques enable decision makers to comprehend bigger data sets and focus their attention on the important characteristics, enabling to come up with answers to questions that decision makers have not even realized to ask.

The most suitable data mining method for identifying natural client segments is cluster analysis. Cluster analysis is a method for dividing a heterogenic data set into smaller more homogeneous groups called clusters. Part of human nature is to group with similar and to contrast from different. Cluster analysis forms clusters based on similarity and acts therefore quite similarly to human nature. As similarity is a subjective experience, then there are countless different ways to conduct cluster analysis. Different methods can result in different solutions. This, however, is not a problem as the aim of cluster analysis is to generate new hypotheses, not to test them. Therefore, cluster analysis is rarely used in isolation as cluster identification is not usually the end goal for researchers. When clusters are identified, other methods are used to understand what those clusters represent, how they can be used and what are the reasons for their existence. From numerous different cluster analysis techniques single linkage, average linkage, complete linkage, Ward's and K-means techniques, as the most popular, are compared in this master thesis.

For identifying natural client segments in this master thesis households are used instead of individuals as households are found to describe client preferences better. On household level consumption decisions are made also for individuals who cannot make these decisions by themselves (e.g. children, the elderly, etc.). Also the most important consumption decisions are made on the household level – e. g. decisions about where to live and which durable goods to buy. Data of the Estonian Household Budget Survey 2012 (LEU 2012) collected by Statistics Estonia is used to carry out the analysis. LEU

2012 is used as the latest data set available describing household consumption preferences throughout Estonia.

Cluster analysis was performed using squared Euclidean distance as similarity metrics, to enable the comparison of different clustering techniques. When performing cluster analysis for specific segmentation needs, it is recommended to test also alternative similarity measures.

The client segments identified in the master thesis are well distinguishable from each other by the average consumption expenditure pattern. However, in each segment the variability of variables is wide and therefore the segments heterogenic. By examining only one selected variable at a time, there are households in each segment, which seem not to fit there. This situation is caused by the high number of variables that are included to cluster the data. All 12 variables describing different components of consumption expenditure available in LEU2012 were included into forming clusters to find the general structure in data. When carrying out segmentation for specific needs of an organization, it is advisable to include only the consumption expenditures of interest to the organization.

All clustering techniques have their strengths and weaknesses. Different techniques are suitable for solving different segmentation tasks. For identifying client segments, the most suitable clusters are big enough to offer economies of scale, but at the same time, small enough to offer clients solutions best fitting to their needs. Most suitable clusters for this need are equal in size and consist of clients with similar preferences. According to theory, from the techniques compared in current master thesis, the K-means technique gives as a result segments that vary the least in size and shape.

As the focus of the master thesis is to identify natural client segments, the number of client segments to be found was not set prior to the cluster analysis. Instead mathematical tests were used to find out the best fitting number of client segments in the data. Calinski–Harabasz index can be used to assess the best fitting number of clusters in data for all the clustering techniques compared in the master thesis. Calinski–Harabasz index suits well for client segmentation as it favours solutions which have wide between-cluster variability and narrow within-cluster variability – characteristic of

compact spherical clusters. For hierarchical clustering techniques, also Duda-Hart index is used. As the segments formed are spherical, it is sufficient to describe resulting clusters by comparing the average value of variables in a segment to the average values of variables in the data.

Cluster analysis is performed in four quartiles formed by household total consumption expenditure. In all quartiles the best clustering solution according to Calinski–Harabasz index is found by using K-means technique. The best fit of K-means technique is confirmed also by comparing the sizes of clusters formed by different clustering techniques. The theoretical expectation of K-means performing best in forming client segments is confirmed.

The number of client segments identified in different quartiles increases together with total consumption expenditure. The segments in lower quartiles are formed mainly by consumption necessary for survival, in higher quartiles segments that are formed by social and mental needs are adding. Thus, together with the growth of total consumption expenditure also the freedom of consumption grows. This is in accordance with Maslow's hierarchy of human needs. Conclusion for client segmentation from the relationship discovered is that using consumption expenditure necessary for survival gives wide coverage of clients from poor to rich. Using variables describing other consumption expenditures gives narrower client segments in higher consumption expenditure quartiles.

Throughout all quartiles segments with similar consumption preferences repeat themselves and are therefore named similarly by the author. Eight different client segments are identified, out of which three are repeated in all quartiles. Those are named as *„Spend on housing"*, *„Eat, drink and smoke*" and *„Spend on transport and experiences"*. In the III and IV quartile the segment *„Likes clothes and going out"* is detected. When the identified client segments are described using categorical variables that are included in LEU 2012 data set, but not used for conducting clustering, the segments named similarly differ from other segments by the same variables. The relationships between the variables that characterize segments need to be studied further to understand the segments better and develop communication strategies for organizations to relate with those segments.

Cluster analysis suits well for profit seeking organizations for identifying naturally existing client segments contributing thereby into more efficient customer relationship management. Through CRM extra value is created both for the organization and its clients. When using LEU 2012 data set for client base segmentation through clustering, it is reasonable to include only those consumption expenditures to the clustering that are of interest to the researcher. Additionally it is recommended to include other variables that describe household consumption preferences. In current master thesis it was found that car ownership, marital status and the basis of housing possession describe household expenditures well through all quartiles. Transforming the variables prior to clustering could help to get results better fitting to the clustering objective. The best fitting results for client segmentation will be achieved using K-means clustering technique and finding out the number of different client segments through mathematical tests.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kaarel Aus

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Klasteranalüüsi kasutamine loomulike kliendisegmentide tuvastamiseks Eesti leibkondade hulgas“ mille juhendaja on Kaia Philips,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **20.05.2014**