Asiakasarvon määrittäminen päivittäistavarakaupassa

Loppuraportti

MS-E2177 Operaatiotutkimuksen projektityöseminaari

Projektiryhmä

Joonas Laihanen (projektipäällikkö)

Aleksi Pasanen

Eero Rantala

Samuli Turunen

Aiheenasettaja: Kesko

---

31. toukokuuta 2017

Sisältö

1 Johdanto

Yritys voi hyödyntää parempaa ymmärrystä asiakkaiden käyttäytymisestä monin tavoin. Päivittäistavarakaupassa tämä voi tarkoittaa esimerkiksi markkinoinnin tehokkaampaa kohdentamista ja yksilöidympää palvelua. Toisaalta on mahdollista ennustaa asiakkaiden käyttäytymistä kerätyn tiedon perusteella toiminnan suunnittelussa. Päivittäistavaralla tarkoitetaan yleisesti elintarvikkeita ja ruokaostosten yhteydessä hankittavia kulutustavaroita.

Aiheen asettajayritys Kesko (“yritys”, "asiakas") omistaa kotimaiset K-kauppaketjut. Tämä projektityö keskittyy K-kaupoissa tehdyistä päivittäistavaraostoksista saatuun dataan. Päivittäistavarakaupassa tavallinen tapa kerätä tietoa asiakkaan käyttäytymisestä on hyödyntää hänen kanta-asiakkuuttaan. Bonuskortin lukeminen ostosten maksun yhteydessä tallentaa hyödyllistä tietoa kauppiaan tietojärjestelmiin. Keskon Plussa-asiakkuus on yrityksen oma bonusjärjestelmä. Plussa-asiakkaat saavat esimerkiksi ajoittain alennusta tietyistä tuotteista, kohdennettuja tarjouksia, sekä ostohyvityksiä niin sanottujen Plussa-pisteiden muodossa. Noin 2,2 miljoonalla taloudella on Plussa-kortti.

Tässä projektissa on keskitytty arvioimaan Keskon asiakkaiden arvoa heille tulevaisuudessa perustuen ostodataan. Yksi mielekäs tapa arvioida asiakkaasta tulevaisuudessa saatavaa tuottoa on CLV (*asiakasarvo*, engl. *Customer Lifetime Value*) [1]. Usein CLV määritellään asiakkaasta saatavana voittona, mutta tässä projektissa CLV:tä käsitellään asiakkaan tekeminä ostoina. Myös CLV:n käsittämä aikahorisontti voi vaihdella.

Yrityksen kanta-asiakkuuksien suuri määrä antaa hyvän pohjan tilastolliselle mallintamiselle. Projektissa kehitetään malli asiakastalouden CLV:n ennustamiselle seuraavan 5-10 vuoden ajalle. Projekti ei sisällä esimerkiksi mallin liiketoimintamahdollisuuksien määrittämistä, asiakastalouksien elämäntilanteen mallintamista, päivittäistavarakaupan tai väestörakenteen kehityksen arviointia tai graafisella käyttöliittymällä varustetun työkalun rakentamista. Mallin on tarkoitus tuottaa odotusarvoon perustuva ennuste, joka saattaa poiketa toteutuneesta ostosten määrästä. Asiakasyritykseltä saatava data koostuu talouskohtaisesta ostosten loppusummista kuukausitasolla; tiedosta asiakastalouden kortinhaltijan iästä, asuinalueesta, sekä yrityksen itse mallintamista ominaisuuksista kuten elämänvaihe ja niin sanottu SoW (*asiakasosuus*, engl. *Share of Wallet*).

Projektin edetessä on päädytty ratkaisuun, että paras tapa ennustaa yksittäistä asiakasta pitkällä aikavälillä, on mallintaa ihmiset sellaisiin luokkakokonaisuuksiin, joiden käytöstä pystytään tilastollisesti ennustamaan. Täten CLV:n ennustamiseen liittyy kaksi osateh-

tävää; talouksien luokittelu tila-avaruuden pienentämiseksi ja tilasiirtymien mallintaminen. Projektissa kehitetään ja vertaillaan useita erilaisia malleja. Luokitteluun sovelletaan esimerkiksi k-prototypes sekä päätöspuualgoritmeja. Tilasiirtymien mallintamiseen puolestaan käytetään Markov-ketjuja. Näin ollen lopulliset tulokset antavat numeerisia ennusteita jokaisen yksittäisen asiakkaan CLV:n arvoksi, mutta myös luokittelun myötä viitekehyksiä, joissa asiakkaat ovat jollain tapaa samanlaisia, mikä auttaa ymmärtämään ostoskäyttäytymistä laajemmin kuin vain pelkkä yksittäisen asiakkaan arvo.

2 Kirjallisuuskatsaus

2.1 Asiakasarvo ja sen määrittämisen ongelma

CLV on numeerinen mitta tietyn asiakassuhteen yritykselle tuottamalle arvolle. Sitä voidaan diskontata, se voi sisältää arvion siitä, kuinka todennäköisesti asiakas lopettaa asioinnin kokonaan, tai se voidaan määrittää katteen huomioon ottaen kuten myös asiakkaaseen liittyvät kulut huomioiden. Kvantitatiivisesti yksi tapa eritellä CLV:hen vaikuttavia tekijöitä on laskea se oheisella tavalla:

$$CLV_i = \sum_{t=1}^{T} \frac{\sum_{j=1}^{J_i}(p_{ijt} - c_{ijt}) - \sum_{k=1}^{K_i} m_{mcikt}}{(1+r)t}, \tag{1}$$

jossa CLV_i on asiakkaan i arvo yritykselle aikajänteellä T, p_{ijt} tuotteen j hinta asiakkaalle i aikana t, c_{ijt} tuotantokustannukset tuotteelle j asiakkaalle i aikana t, m_{mcikt} markkinointikustannukset asiakkaalle j aikana i, J_i tuotteiden määrä, K_i markkinoitavien tuotteiden määrä ja r diskonttauskerroin, joka kuvaa rahan arvon yleistä laskua tulevaisuudessa. [10]

Tässä projektissa CLV on kuitenkin vain yksinkertaisesti kaikkien päivittäistavaroiden ostojen asiakaskohtainen summa ottamatta kantaa kuluihin, joita tuoteiden tarjoamisesta aiheutuu valitulle aikajänteelle 5-10 vuoden ajalle eli matemaattisesti

$$CLV_i = \sum_{t=1}^{T} \sum_{j=1}^{J_i} p_{ijt}, \tag{2}$$

käyttäen edellisen kaavan merkintöjä. Tätä aikaa pidetään kiinnostavan pitkänä aikana ostoskäytöksen muutoksien tapahtumiseen, kuin myös riittävän lyhyenä, jotta ennustamiseen eivät vaikuta sellaiset tulevaisuuden trendit, joita ei voi puhtaasti datan perusteella ennustaa, kuten uusien kauppaketjujen tuleminen markkinoille ja verkkokaupan kehitys.

Myöhemmissä tuloksissa on valittu, että CLV lasketaan kymmenen vuoden aikajänteelle. Aikajänteen muokkaus kuin myös diskonttauksen tai muun monimutkaisemman tavan, joka huomioisi kulut ja näin ehkä paremmin kuvaisi asiakkaan todellista arvoa, huomioiminen laskettaessa CLV:tä, on trivaali toiminpide, joka voidaan tarvittaessa tehdä ilman, että tämän projektin analyysi perustavanlaatuisesti muuttuisi.

Historiadataan pohjautuen on usein mahdollista ennustaa suurempien ryhmien keskimääräistä käytöstä, vaikka yksilötasolla tämä ei onnistuisi. Luvussa 2.2 perehdytään siihen, kuinka yksilöt eli datan havainnot voidaan ryhmitellä potentiaalisesti ennustettaviin kokonaisuuksiin. Tämän jälkeen mietitään, kuinka ryhmien käytöstä voidaan ennustaa. Tämän luokittelun jälkeen on CLV:n laskeminen hyvin suoraviivainen toimenpide. Lisäksi laskennan toteuttaminen jollekin monimutkaisemmalle CLV määrittelytavalle olisi hyvin vaivatonta.

2.2 Luokittelualgoritmit

Asiakkaiden ryhmiin luokittelun on projektin haastavin tehtävä toteuttaa. Luokittelu kannattaa, koska ryhmien käyttäytymistä voidaan ennustaa pitkällä aika välillä paremmin kuin yksittäisen asiakkaan käyttäytymistä. Luokittelussa ovat vaihtoehtoina sellaiset menetelmät, joissa luokittelu on ihmiselle intuitiivinen ja toisaalta sellaiset menetelmät, joissa luokittelun tulosta on vaikea tulkita, mutta ennustevirhe puolestaan saattaa olla pienempi. Yksi kysymys on myös luokkien määrä. Ennustevirhe saadaan erittäin pieneksi kun jokainen havainto on oma luokkansa, mutta tällaisella mallilla ei ole paljoa arvoa yleisen ostoskäyttäytymisen ymmärtämisessä. Menetelmien tarkempi esittely on luvussa 3.

Seuraavaksi esitellään kaksi yleisesti tunnettua algoritmista lähestymistapaa luokitteluun, jotka vaikuttavat lupaavimmilta juuri tähän ongelmaan ja käytössä olevaan dataan. Johtuen datan määrästä ja siitä, että data on monimuotoista eli myös kategorisista muuttujista koostuvaa, esitellään myös perinteisistä algoritmeista muunnellut vähemmän tunnetut mukaelmat.

2.2.1 k-means-pohjaiset menetelmät

Eräs tunnetuimmista ja yksinkertaisimmista algoritmeista on k-means klusterointi, joka jakaa datan yhteensä k:hon eri luokkaan, jossa k on parametrina annettu luokkien mää-

rä. Luokittelun perustana on rajata havainnot luokkiin niin, että luokkien sisäisten vaihteluiden summa on mahdollisimman pieni. Luokkien sisäistä vaihtelua mitataan yleensä laskemalla kaikkien luokkaan kuuluvien havaintojen etäisyyksien neliöiden summa. Etäisyysfuktioksi valitaan jatkuvien muuttujien tapauksessa tyypillisesti euklidinen etäisyys.

Jos taas luokiteltavat havainnot sisältävät kategorisia muuttujia, ei euklidista etäisyyttä voi käyttää havaintojen erilaisuuden mittana. Sekatyyppisistä muuttujista koostuvien havaintojen etäisyyttä voidaan mitata esimerkiksi niin kutsutulla Gower-etäisyydellä. Tämä etäisyys on alunperin esitelty Gowerin artikkelissa [3] ja se voidaan esittää kahdelle havainnolle i ja j painotettuna summana näiden eri muuttujista

$$S_{ij} = \frac{\sum_{k=1}^{N} w_{ijk} S_{ijk}}{\sum_{k=1}^{N} w_{ijk}}, \tag{3}$$

jossa w_{ijk} painotus muuttujalle k, kun havaintoja X_i ja X_j vertaillaan ja S_{ijk} etäisyys muuttujassa k näiden kahden havainnon suhteen. Painotusten määrityksen ongelmaa on käsitelty muun muassa julkaisussa [9]. Helpoin ratkaisu on valita tasapainotus eli että kaikkia muuttujia painotetaan samalla kertoimella. Nyt on luonnollisesti määriteltävä S_{ijk}, niin että havaintojen etäisyyttä yksittäisessä muuttujassa mitataan järkevästi. Määritellään, että

$$S_{ijk} = \frac{|X_{ij} - X_{jk}|}{r_k}, \tag{4}$$

jossa r_k on havaintojen vaihteluväli eli suurimman ja pienimmän arvon erotus. Tämä määrittely toimii kuitenkin vain numeerisiin tapauksiin ja kategoristen muuttujien kanssa määritellään

$$S_{ijk} = \begin{cases} 0, & kun\ X_{ik} = X_{jk} \\ 1, & kun\ X_{ik} \neq X_{jk} \end{cases}. \tag{5}$$

Tämän lisäksi olisi mahdollsita eritellä kategorisista muuttujista ordinaaliset tapaukset eli sellaiset, joilla tapaukset on jonkinlaista järjestystä, mutta ei kuitenkaan sellaista, mikä olisi suoraan numeerisesti vertailukelpoista.

Huomionarvoista on myös, että luokiteltaessa k-means-menetelmällä tulee luokkien määrä k olla ennalta valittu. Huono luokkamäärän valinta voi johtaa heikkoon tulokseen, joten tämän algoritmin yhteydessä lisäongelmaksi saattaa muodostua sopivan luokkamäärän selvitys. Lisäksi algoritmin toiminta ei ole täysin deterministinen siltä osin, että se pitää alustaa k kappaleella itse valittuja tiloja. Muita potentiaalisia ongelmia ovat, että pahimmassa tapauksessa laskennallinen kompleksisuus on superpolynominen tai ekspotentiaalinen, jos iteraatioiden määrä kasvaa havaintojen määrän kasvaessa suureksi. Aineiston muuttujien määrä ja luokkien määrä k vaikuttavat yleisillä toteutuksilla lineaarisesti

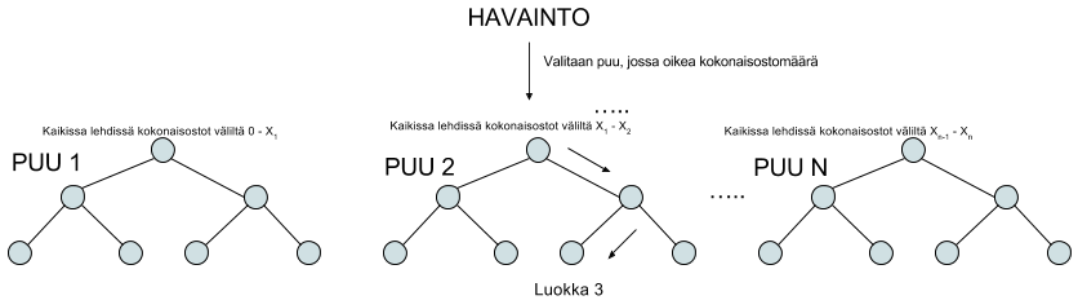
laskenta-aikaan. Myös keskiarvon epärobustisuus saattaa heikentää tulosten laatua.

Yllä olevat ongelmat ovat laajalti tunnistettuja [6], joten niihin löytyy useita valmisratkaisuja; luokkamäärän valintaan X-means klusterointi, k-means++ aloitustilojen ja tyydyttävän nopean suppenemisen takaamiseen ja k-medians tai k-medoids klusterointi keskiarvon aiheuttamaan robustisuusongelmaan. Kategoristisen muutujien aiheuttamaa ongelmaa voi lähestyä myös k-meansin muunnelmalla, joka on nimetty kirjallisuudessa [4] k-prototypes-menetelmäksi.

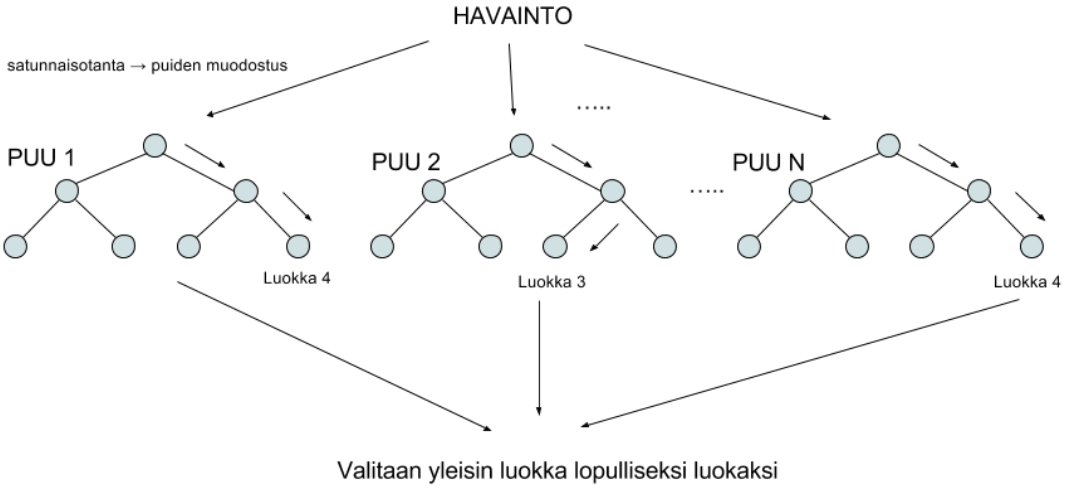
k-prototypes algoritmin periaate on samankaltainen kuin edellä kuvattu idea; määritellään etäisyysfunktio sekatyyppiselle datalle sopivaksi ja sovelletaan perinteistä k-means -algoritmia. Tarkemmin ottaen k-prototypes laskee numeeristen muuttujien etäisyydet ja tähän lisätään kertoimella kerrottu kategoristen muuttujien välinen etäisyys. Tässä numeerinen etäisyys voi olla esimerkiksi euklidinen, kategorinen diskreetin avaruuden yksinkertainen etäisyys ja kerroin valitaan niin, että koetaan näiden kahden muuttujaryhmän painoarvojen olevan sopivat keskenään. Tämän algoritmin on todettu toimivan tutkimuksessa hyvin suurillekin datamäärille ja sille on valmistoteutuksia. [4]

2.2.2 Päätöspuut luokkien ennustamisessa

Päätöspuu koostuu yhteen liitetyistä päätös- ja tulossolmuista, eli lehdistä [2]. Jokainen päätössolmu jakaa havainnot kahtia jonkin muuttujan tila-avaruuden suhteen. Lehdet sisältävät päätöspuun tuottamat tulokset; luokittelupuussa ennustetut luokat ja regressiopuussa selitettävän muuttujan arvot. Päätöspuu voidaan sovittaa ennustamaan selitettävää muuttujaa usealla eri algoritmilla, jotka pyrkivät jokaisen päätössolmun jälkeen arvioimaan, minkä muuttujan suhteen havainnot kannattaisi jakaa puun ennusteen parantamiseksi. Yksi algoritmien parametri onkin yleensä pienin sallittu ennusteen parannus, jonka jälkeen uudesta solmusta tulee lehti päätössolmun sijaan.



(a) Monen puun päätöspuumalli, jossa puiden muodostus ja luokitteluvalinta tehdään t ennustettavan tekijän mukaan.



(b) Random Forrest -menetelmän mukainen puihin jako ja luokittelun valinta.

Kuva 1: Kaksi tapaa tehdä luokittelu päätöspuiden avulla. Jokainen yksittäinen puu on tehty sen perusteella minkälainen jako kussankin vaiheessa on optimaalisin ennusteen parantamiseksi.

Päätöspuut soveltuvat hyvin datalle, jossa on sekaisin jatkuvia ja kategorisia muuttujia. Lisäksi ennustamisen kannalta hyödyllinen ominaisuus on se, että koska tila-avaruus jaetaan aina kahtia, voi sovitetulla puulla ennustaa myös ennennäkemättömiä havaintoja. Päätöspuut ovat kuitenkin herkkiä ylisovitukseen, eli pitkälle "kasvatettu"puu selittää kyllä sovitusdatan vaihtelun erinomaisesti, mutta ei toimi hyvin ennennäkemättömälle datalle. Toinen ongelma on se, että päätöspuu saattaa painottaa sellaisia kategorisia muuttujia, joissa luokkia on runsaammin.

Päätöspuut ovat siis helposti tulkittavia ja laajasti käytettyjä malleja, jotka ovat kyvykkäitä käsittelemään sekä numeerisia että kategorisia muuttujia luontaisesti. Niiden parametrien sovitukseen on useita algoritmeja, jotka antavat mahdollisuuden vaikuttaa esimerkiksi puun syvyyteen. Yhden päätöspuun tarkkuutta voidaan parantaa yhdistämällä useita päätöspuita, esimerkiksi Random Forest -menetelmällä [8]. Menetelmässä muodostetaan useita päätöspuita satunnaisotannalla datasta valituille osajoukoille. Tämän jälkeen

havainto luokitellaan kaikissa puissa ja valitaan lopulliseksi havainnoksi yleisin luokka, johon puut havainnon luokittelevat, kuten kuvassa 1b. Näin vältetään ylisovitusta.

Random Forrest -menetelmän lisäksi useata puuta käyttävää päätöspuuluokittelua voidaan tehdä myös muilla tavoin, esimerkiksi voidaan jakaa aineisto selitettävän muuttujan suhteen eri puihin, niin että samassa puussa selitettävät muuttujat ovat mahdollisimman lähellä tosiaan kuvan 1a mukaisesti. Tätä voidaan pitää hyvin yksinkertaisena tapana toteuttaa niin sanottu regressiivinen ositus, jota myöhemmin tässä työssä hyödynnetään. Luvussa 3 kuvataan tarkemmin, kuinka päätöspuita sovelletaan asiakkaiden luokkitteluun.

2.3 Tilasiirtymien mallinnus Markov-ketjuilla

Markov-ketju on satunnaisprosessi, jossa uusi tila riippuu ainoastaan edellisestä tilasta [**?**]. Tiloista siirrytään toisiin tiloihin tietyillä todennäköisyyksillä joka aika-askeleella tai jatkuvassa tapauksessa satunnaisin aikavälein. Malli on tältä osin varsin yksinkertainen, se on niin sanotusti muistiton - uuteen tilaan vaikuttaa vain ja ainoastaan edellinen tila, ei esimerkiksi tätä edeltävä tila. Luonnolisestikin tämän vuoksi joitain pitkänaikajänteen ilmiöitä saattaa jäädä havaitsematta, mutta etuna on mallin yksinkertaisuus ja näin ollen ymmärrettävyys. Myös tilojen määrä pysyy paremmin halittavissa, koska jo kahden askeleen muistilla varustetussa mallissa olisi periaatteessa neliöllinen määrä tiloja muistittomaan verrattuna.

Aikaisemmin kuvattua klusterointia voidaan käyttää Markov-mallissa ketjun tila-avaruutena. Tällöin ei tarvitse käytää Markovin piilomalleja tai muita monimutkaisempia malleja, jotka voisivat muuten tulla kyseeseen.

Tilojen lisäksi siirtymätodennäköisyydet tulee määrittää. Tämän projektin yhteydessä todennäköisyydet voidaan estimoida käyttämällä osaa, esimerkiksi 50 % datasta, ja tutkimalla järjestelmällisesti, kuinka usein mistäkin tilasta siirrytään toiseen. Todennäköisyyksien avulla voidaan laskea esiintyvyysmatriisi

$$M_t = \sum_{s=0}^{t} P^{-s}, \tag{6}$$

missä P on siirtymien todennäköisyysmatriisi ja t vuosien määrä, jonka ajalle CLV halutaan laskea. Esiintyvyysmatriisia käyttämällä saadaan odotettu kertymä ostetulle määrälle eli CLV

$$g_t = M_t c, \tag{7}$$

missä c on pystyvektori, joka sisältää jokaisen eri tilan vuosittaisten ostojen keskiarvon. Markov-ketjuja on aikaisemminkin sovellettu CLV:n ennustamiseen. [11]

3 Menetelmät

Seuraavassa analysoidaan ensin projektissa käytetty ostodata, niin että kuvattavien menetelmien toiminta on ymmärrettävää. Edetään sitten itse CLV:n ennustemallien toteutukseen.

3.1 Datan kuvaus ja tilastollinen analyysi

Käydään tarkemmin läpi datan rakenne ja tilastollinen analyysi, joka tulee suorittaa ennen itse ennustemallien toteutusta. Esitellään lyhyesti myös dataan liittyvät käytännön ratkaisut kuten virheellisten havaintojen käsittely.

3.1.1 Käytetyn K-plussa-asiakkaiden ostodatan kuvaus

Kuten jo johdannon yhteydessä mainittiin, data koostuu talouskohtaisesta ostosten loppusummista kuukausitasolla ja muista muuttujista. Vain ostosten loppusumma vaikuttaa suoraan asiakkaan arvoon, mutta tulevia kuukausiostoja ennustattaessa voivat muutkin tiedot olla hyödyllisiä.

Kuukausiostojen lisäksi data sisältää seuraavat tiedot, joista osa, kuten ikä, perustuu kortin pääomistajaan ja osa kaikkiin kortin käyttäjiin, kuten vierailuiden määrä:

- Ikä vuoden tarkkuudella
- Vierailuiden (ostotapahtumien) määrä kuukaudessa
- Asuinalueen postinumero, mutta postinumeron tulkinnan vaikeuden takia algoritmeissa on päädytty käyttämään alkuperäisen datan ulkopuolista Keskon mallintamaa asiointitodennäköisyyttä, joka kuvaa millä todennäköisyydellä tietyn postinumeron asiakas asioi juuri Keskossa
- Yrityksen mallintama asiakkaan elämänvaihe, josta käytetään lyhennettä *LANSEY*, jossa kukin kirjain kuvaa yhtä ryhmää (L=lapsiperheet, A=aikuistalous, N=nuoret

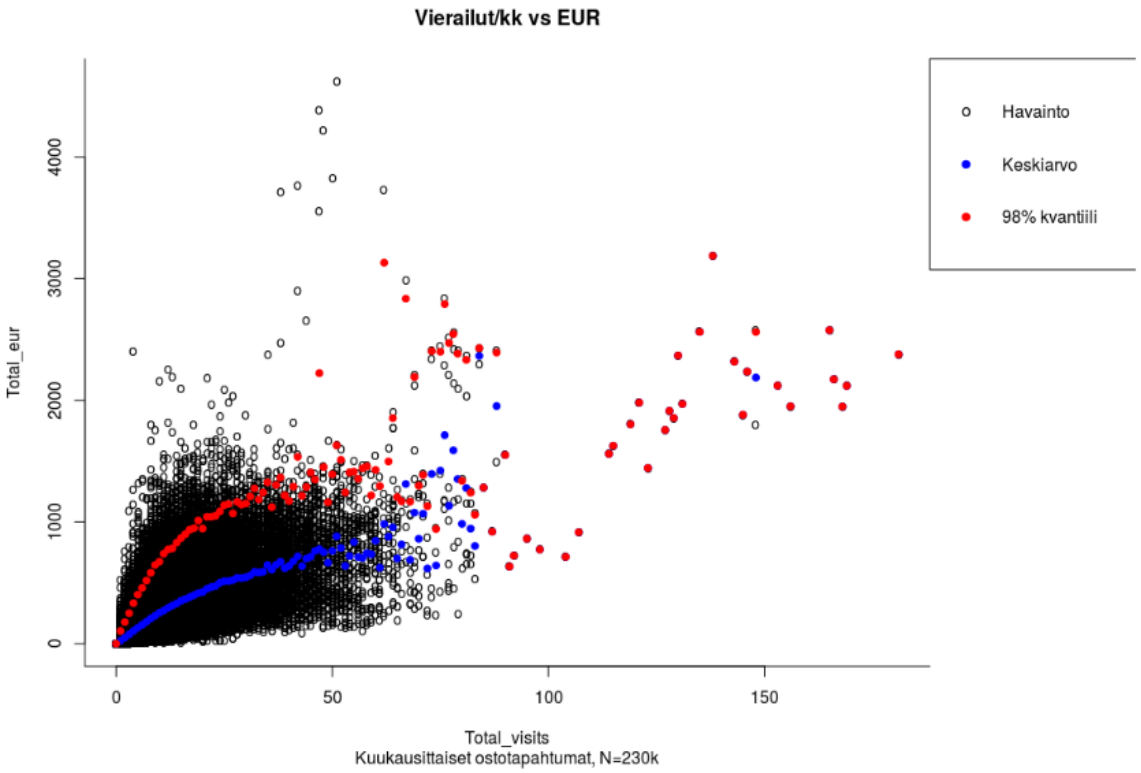
parit, S=sinkut, E=eläkeläiset, Y=yksinelävät)

- SoW (asiakasosuus, engl. *Share of Wallet*), yrityksen mallintama prosenttiluku, joka pyrkii arvioimaan, kuinka suuri osa asiakkaan päivittäistavaraoostoista kohdistuu Keskon kauppaketjuihin

Myös tarkempaa tuotekohtaista ostodataa on olemassa, mutta sitä ei olla koettu tarpeelliseksi käyttää rajallisten aikaresurssien vallitessa.

Projektissa käytetty data perustuu Keskon noin 2,2 miljoonan ihmisen Plussa-kortti -dataan. Laskennallisista syistä kuitenkin vain satunnaisotokselle otettua osaa tästä datasta on käytetty algoritmien työstämisessä, ja lopullisia tuloksiakaan tässä raportissa ei ole tehty kaiken datan pohjalta, sillä tulokset ja erityisesti menetelmät ovat samankaltaisia riippumatta otetaanko koko data mukaan. Tulokset perustuvat noin 100 000 pääkorttilaisen kuukausittaisiin havaintoihin vuosilta 2015 ja 2016.

Toinen syy miksei koko dataa ole käytetty ovat yksittäiset virheelliset havainnot ja ongelmat yrityksen omissa mallintamissa muuttujissa. Virheelliset havainnot, kuten henkilö, jonka ikä on yli 500 vuotta, on yksiselitteisesti poistettu. Näitä on hyvin vähän, mutta alkuperäisessä datassa on useita havaintoja, joissa asiakas ei ole ilmeisesti ilmoittanut yritykselle ikäänsä. Koska näidenkin osuus on marginaalinen, on ne vain poistettu (toinen vaihtoehto olisi ollut mallintaa puuttuva ikä perustuen muihin muuttujiin). Lisäksi erinäisiä marginaalisia äärihavaintoja on karsittu, jotta mallit toimisivat paremmin. Tämä tarkoittaa, että mallit eivät anna näidenkaltaisille havainnoille hyviä ennusteita, mutta näitä karsittuja ääripään havaintoja on vähemmän kuin 2 prosenttia datan kokonaismäärästä.



Kuva 2: Esimerkkihavainnollistus datan muuttujista. Kuvassa kuukausittaiset vierailut ja kokonaisostot. Huomattavaa on, että lähes kaikki havainnot yli 200 tuhannesta tässä esitetystä ovat yhdessä kasassa. Yhtenäisestä joukosta poikkeavat äärihavainnot pystyy melkeimpä laskemaan käsin.

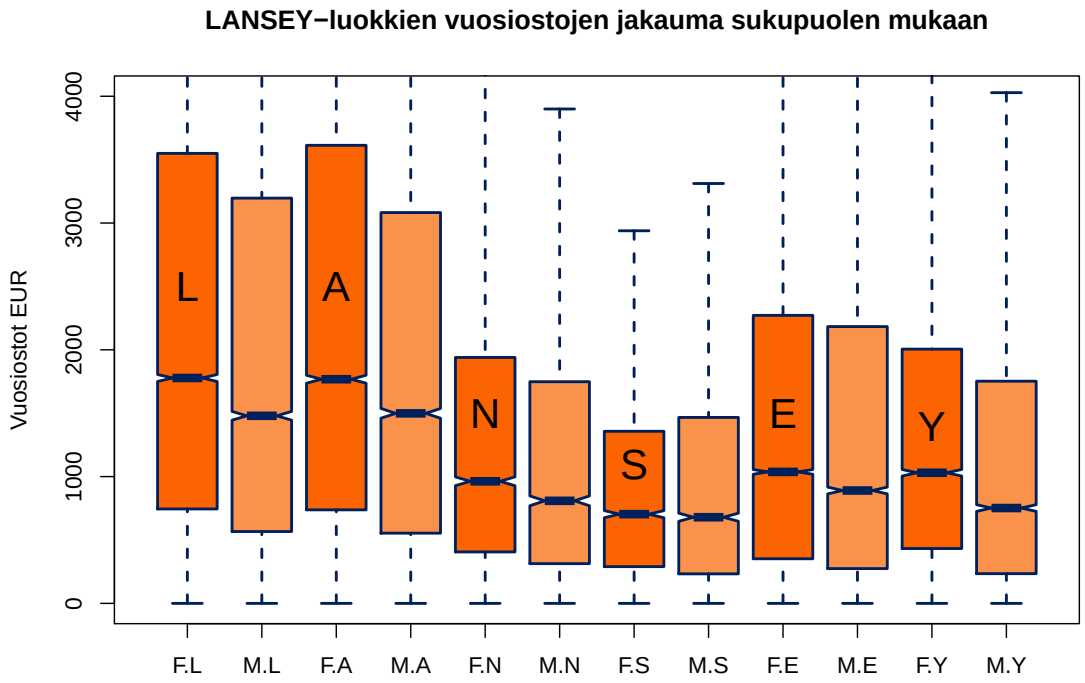
Ongelmat mallinnetuissa muuttujissa liittyvät esimerkiksi siihen, että on rajatapauksia, joista Keskon malli siviilisäädylle ei ole osannut päättää ostoskäytöksen pohjalta, mitä asiakas edustaa. Siviilisääty saattaa muun muassa oskilloida kuukausien välillä, vaikka todellisuudessa asiakkaan siviilisääty ei vaihdu näin nopeasti. Tämä ongelma on ratkaistu siten, että projektissa kuukausikohtainen data on pyöristetty vuosikohtaiseksi dataksi, ottamalla jatkuvissa muuttujissa keskiarvoja tai summia ja kategorisissa moodeja.

3.1.2 Datan tilastollinen perusanalyysi

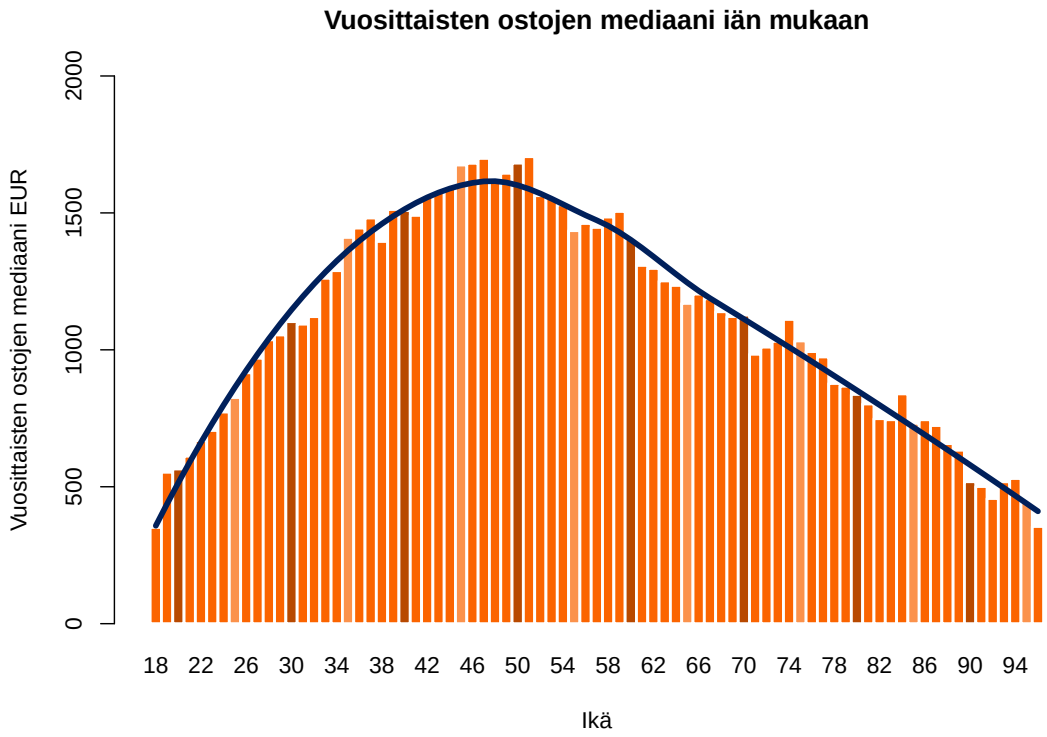
Datan erityispiirteiden ymmärtäminen on olennaista malleja toteutettaessa - parametrien valinta ja koko mallin toimivuus voi perustua esimerkiksi oletukselle muuttujien riippumattomuudesta tai tasajakautuneisuudesta. Tämän vuoksi datalle on tehty laajasti sekä yksiuloitteista kuin myös moniuloitteista tilastollista analyysia. Eri muuttujien tunnusluvut on määritelty, ja erityisesti muuttujien käytöstä toistensa suhteen on havainnoitu kuvilla.

Asiakkaan tarjoama data on varsin hyvin käyttäytyvää. Yksiuloitteinen analyysi osoittaa, että kaikki jatkuvat muuttujat ovat jakautuneet eksponenttijakauman mukaan tai vinoutuneen Gaussin jakauman mukaisesti, kun vain ongelmallisen postinumeronkin tilalle otetaan asiointitodennäköisyys, kuten edellä kerrottiin. Ja LANSEY-luokituksessa, joka esiteltiin edellisessä alaluvussa, ja sukupuolijakaumassakaan, ei ole kertaluokkien suuruisia eroja eri tyyppien määrien välillä.

Moniuloitteinen analyysi paljastaa muutamia kiinnostavia ilmiöitä. Muun muassa sen, että naiset ostavat keskimäärin enemmän kuin miehet kaikissa demografisissa luokissa paitsi sinkuissa miehet aivan hieman enemmän. Tätä on havainnollistettu kuvassa 3, josta näkyy myös, kuinka erilailla eri LANSEY-luokat ostavat. Toinen tässä tarjottava esimerkkikuva on iän vaikutuksesta vuosittaisiin ostoihin kuvassa 4.



Kuva 3: LANSEY-luokkien vuosiostojen jakauma sukupuolen mukaan. Tummemmat palkit edustavat naisia ja vaaleammat miehiä. Vekki palkin keskellä on otoksen mediaani ja palkin alempi puolisko edustaa toista ja ylempi kolmatta neljännestä otoksesta.



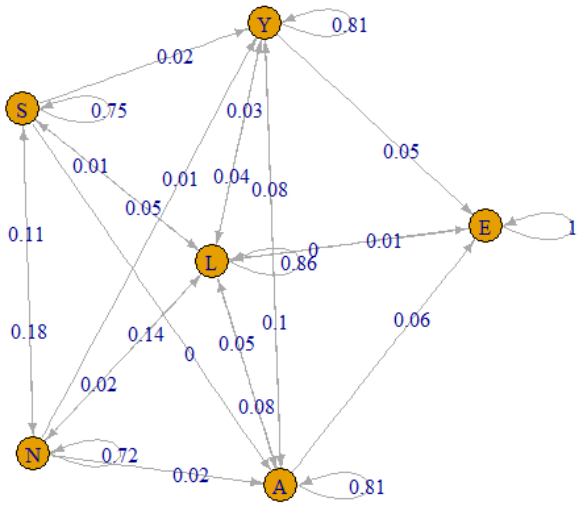
Kuva 4: Vuosittaisten ostojen mediaani iän mukaan. Tasakymmenet tummemmalla ja viitoset vaalealla. Tumman sininen viiva on satunnaisvaihtelu häivyttävä trendisovite, joka perustuu R:n lokaaliin polynomiseen *loess*-sovitefunktioon. Tätä samaa sovitetta on käytetty myös joissain myöhemmissä kuvaajissa.

3.2 Toteutettujen mallien kuvaus ja toiminta

Esitellään tässä kirjallisuuskatsauksen pohjilta ja jatkuvan kehitystyön jäljiltä syntyneet lopulliset mallit. Ensimmäiseksi esitellään lyhyesti triviaaleja baseline-ratkaisuja luokittelulle. Tämän jälkeen edetään k-protypes-algoritmin ja päätöspuumallin toteutukseen käytännön tasolla. Lopuksi esitellään omakehitteinen ratkaisu, joka on nimetty riippumattomien keskiarvojen malliksi, joka poikkeaa merkittävästi muista ratkaisuista. Se antaa CLV-ennusteen perustuen mielivaltaisiin ihmisen ennalta määräämiin luokkiin, eikä algoritminen luokittelu ole tarpeen. Kaikki mallit sisältävät Markov-ketjulla tilojen kehityksen simuloinnin samalla tavalla ja lopuksi suoraviivaisen CLV:n laskemisen.

3.2.1 Yksinkertaisista baseline-malleista

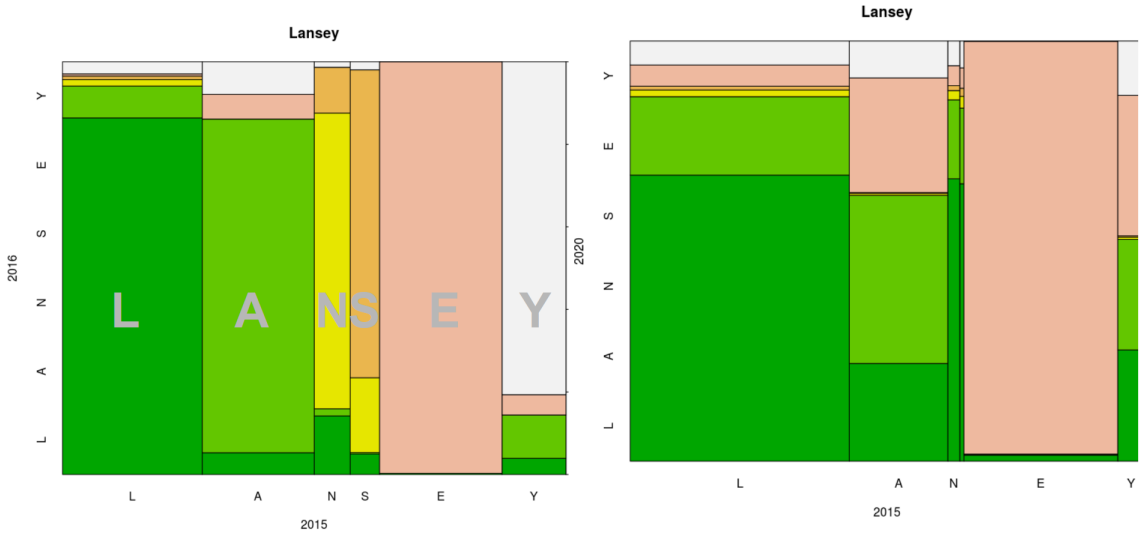
Projektia kehittäessä ensimmäisenä luotiin hyvin yksinkertaisia ja nopeasti toteutettavissa olevia malleja CLV:n ennustamiseen. Näitä ei pidä väheksyä, sillä mallin monimutkaisuus ei takaa hyvää mallia tai hyvää ennustekykyä. Yksinkertaisin baseline-luokittelu, josta on hyvä lähteä laajentamaan luokka-avaruutta, on ottaa LANSEY:n kuusi eri tilaa jokainen omaksi luokakseen. Koko ennalta tehty mallinnus näihin kuuteen luokkaan on osoitus siitä, että aiheenasettaja Keskon mielestä nämä ovat potentiaalisia asiakasryhmiä, joiden CLV-arvot voisivat poiketa toisistaan, mikä pitääkin paikkansa kuten kuvasta 3 näkyy, jos oletetaan, että yhden vuoden ostot korreloivat CLV:n kanssa. Ottamalla tiloiksi LANSEY-luokat ja toteuttamalla kirjallisuuskatsauksessa esitetty todennäköisyyksien määritys tilasiirtymien välille saadaan kuvan 5 mukainen Markov-ketju. Samaa Markov-ketjua on havainnollistettu myös kuvassa 6, jossa lisäksi demonstroidaan, kuinka tilajakauma käyttäytyy, kun otetaan useita askelia. Huomataan muun muassa, kuinka eläkeläisten määrä kasvaa ja yksin elävät pariutuvat. Näiden ilmöiden esiintyminen tukee mallin oikeellisuutta. Itse Markov-ketjun toteutus on hyvin samanlainen kaikissa malleissa ja sen teko käytännössä käydään vielä läpi päätöspuiden toteutuksen yhteydessä.



Kuva 5: Yksinkertainen Markov-ketju graafimuodossa, jossa solmuina LANSEY-luokat.

Muuta huomioitavaa mallista on, ettei siinä varsinaisesti tule uusia asiakkaita tyhjästä tai vastavuoroisesti kuole siinä mielessä, että tiloja katoaisi. Uusien asiakkaiden huomiotta

jättäminen ei ole siinä mielessä ongelma, että lähes kaikki uudet asiakkaat ovat nuoria ihmisiä, eivätkä tilansiirtymät vanhemmasta nuorempaan ole fyysisesti mahdollista. Kuoleminen tai sen suurentunut mahdollisuus ikäihmisillä taas huomioidaan heidän odotusarvoisesti vähenevien ostojen kautta.



(a) Yksinkertainen yhden vuoden Markov-ketju vuodesta 2015 vuoteen 2016

(b) Sama Markov-ketju viidellä askeleella eli viiden vuoden Markov-ketju vuodesta 2015 vuoteen 2020

Kuva 6: Pylväsesitys LANSEY-tilaisesta Markov-ketjusta, jossa pylvään leveys kuvaa ryhmän suhteellista osuutta koko väestöstä ja pystysuuntainen korkeus siirtymätodennäköisyyttä kyseistä väriä vastaavaan tilaan. Jokaista tilaa vastaavan värin voi lukea vasemman kuvan kirjaimien perusteella. Lisäksi oikealle on kuvattu, kuinka Markov-ketju käyttäytyy kun sitä ajetaan viisi vuotta eteenpäin.

Yksinkertaista LANSEY-luokista koostuvaa luokittelumalli on helppo laajentaa, esimerkiksi hajottamalla lapsiperheet kolmeen eri ikäluokkaan. Näin saataisiin kahdeksasta luokasta koostuva malli. Myöhemmin esitettävä oma kehitteinen ennustemalli on jossain määrin vielä laajennus tästä.

Eräs yksinkertainen baseline-malli on olettaa ostojen pysyvän vakiona, eli ennustaa seuraavan vuoden ostot edellisvuoden ostoilla. Tässä jokainen havainto on oma luokkansa, eikä luokkasiirtymiä tapahdu. Tämä malli onkin erittäin ennustevoimainen arvioimaan seuraavaa vuotta, ja sen voittaminen yhden vuoden aikajänteen ennustamisessa voi olla haastavaa. Kuitenkin kun ennustetaan pitkän aikavälin kehitystä tämän triviaalimallin ennustevoima katoaa, joten tähän ratkaisuun ei voida tyytyä usean vuoden ostokäytökseen perustuvan CLV:n ennustamiseen. Mutta summittaiseen lyhyen aikavälin ennustettavuuden toimivuuden validointiin menetelmä on varsin käyttökelpoinen työkalu.

3.2.2 k-prototypes toteutus ja lyhesti valmisfunktioiden käytön ongelmista

Heti alkuun mainittakoon, että k-protypes-menetelmää ei ole projektissa toteutettu sellaiseen loppullisen muotoon asti, mikä antaisi mielekkään tarkkoja ennusteita. Näin ollen tämän algoritmin toimintaa ei käsitellä enää tuloksissa, vaan tyydytään tässä kuvaamaan tehty toteutus, ja pohditaan lyhyesti, miksi sen toiminta voisi olla vaillinnaista sekä mitä käytännön ongelmia niin tässä projektissa kuin myös yleisesti niin sanottujen valmisfunktioiden käyttöön liittyy.

k-prototypes klusterointi toteutettiin erinäisten R:n valmisfunktioiden avulla usealla eri tavalla. Käyttäen muun muassa paketin `clustMixType` funktiota `kproto` ja paketin `protoclust` funktiota `protoclust`, jotka siis tekevät k-prototypes luokittelun. Näiden lisäksi kokeiltiin paketin `cluster` funktiota `pam`, joka on robustihko k-means-menetelmän variaatio. Käytettäessä funktioita `pam`, tulee määrittää etäisyysmatriisi havainnoilla. Tämä onnistuu esimerkiksi paketin `cluster` funktiolla `daisy` ja valitsemalla metriikaksi kirjallisuuskatsauksessa mainittu Gower-etäisyys. [3] Kuvataan seuraavaksi lyhyesti toteutettu lähestymistapa algoritmisesti:

Algorithm 1: K-prototypes- ja K-means-algoritmilla parhaan luokittelun etsimisen kuvaus

```
Valitse mahdollisimman suuri N ja jokin joukko erinäisiä painotuksia J
  laskentakapasiteetin puitteissa
for j ∈ J do
    for i ∈ {2, ..., N} do
        Laske etäisyysmatriisi painotuksella j (esim. Gower-etäisyys)
        Luokitus ← Laske k-protypes- tai k-means-luokitus arvolla k = i
        if Virhe(Luokitus) - Virhe(ParasLuokitus) < 0 then
            ParasLuokitus ← Luokitus
        end
    end
end
return ParasLuokitus
```

Valitettavasti kaikki suoraan k-prototypes luokittelun tekevät algoritmit tuottavat ja alustavalle testidatalla kelvottomia tuloksia, kun arvioidaan tarkkuutta perustuen edellisvuoden pohjalta tehtävään ennusteeseen, kuten osiossa 4.2.1 tehdään onnistuneimmille malleille. Sen sijaan k-means Gower-etäisyydellä lähestyy kohti hyviä ennusteita luokkamäärän k kasvaessa, kun hyvänä ennusteena pidetään vertailukohtana olevaa 'edellisvuosi ennus-

teena' -mallia, jonka pohjalta yllä olevassa algoritmissa lasketaan *Virhe*-funktion arvo. Lisäksi gower-etäisyyksiä laskettaessa algoritmi mahdollistaa erilaisten painokertoimien valitsemisen, joita säätämällä ennustetarkkuutta pystyy parantamaan hieman. Kuitenkin myös laskenta-aika kasvaa tahdilla, mikä vaikuttaa eksponentiaaliselta, kun luokkamäärää kasvatetaan kaikilla kokeillulla painokerrointen arvoilla. Esimerkiksi luokkien määrän olessa $k = 50$ on suhteellinen prosentiaalinen virhe noin 69% ja laskenta-aika noin 6 minuuttia käytetyillä laskentaresursseilla. Luokkien määrän olessa $k = 70$ on suhteellinen prosentiaalinen virhe noin 55% ja laskenta-aika noin 10 minuuttia. Mikäli virhe jatkaisi pienenemistään ja käytössä olisivat tehokkaammat laskentaresurssit ja paljon aikaa algoritmin suorituksen odottamiseen, suhteellinen virhe saattaisi tulla yhtä pieneksi kuin muissa enemmän onnistuneissa toteutuksissa.

```
> clusters$clustering
  [1]  1  1  2  2  3  3  4  4  5  5  6  6  7  7  8  8  9 10 11 11 12 12  1  1 13 13 14 15 16
 [30] 16 17 17  2  2 18 19  9 10 20 20  6  6  9  9 15 15 21  8 10 10 22  5 23 23 13 13  5  5
 [59] 22  5 18 18 24 25 26 22 23 23  6  6  5  5 23 18  9  9 14  9  7  7 27  3 26 26 28 15 19
 [88] 19 15  7 14  9 29 29 26 26  9  9 30 30  5  5 22 22 30 22 18 18 12 22 10 10 17 17 19 19
[117] 13 13 23 23 13 13  5 22  5  5 23 31 23 21  9  9 12 12 13 13 10  9 12 12 18 19 13 13  3
[146]  3 10 10 13 13 12 12 30 30  2  2  8 32 33 33 23 23 25 15 13  2 11 11 11 11 19 19 33 33
[175] 33 33  7  7  2  2  2  2 28 28  2  2 30 30 34 35  1 34 21 21 10 10 29 29 35 35 12 12 36
```

Kuva 7: k-prototypes-algoritmin antamia luokittelutuloksia ensimmäisille noin 200 havainnolle, kun luokittelu tehdään $k = 40$ luokkaan.

Heikon ennustekyvykkyyden syy ei ole selvä. Se voi esimerkiksi johtua huonoista lähtöarvoista klusteroinnille tai ketjuuntumisesta, joka voi seurata jollei painoarvoja Gower-etäisyydessä valita oikein. Myöhemmin toteutettavan päätöspuumallinkin yhteydessä malli antoi tietyiltä osin järjettömiä ennusteita, kunnes eräät tämän mallin painokertoimet ymmärrettiin valita sopivasti. Sopivia painokertoimia ja muita parametreja voi yrittää löytää kokeilemalla, tai tehokkaammin jollain heuristisella haulla. Myös muiden kielten kuin vain R toteutuksia voisi kokeilla. Raja toimivan ja toimimattoman algoritmin välillä voi olla yksittäinen parametrisointivirhekin koodissa.

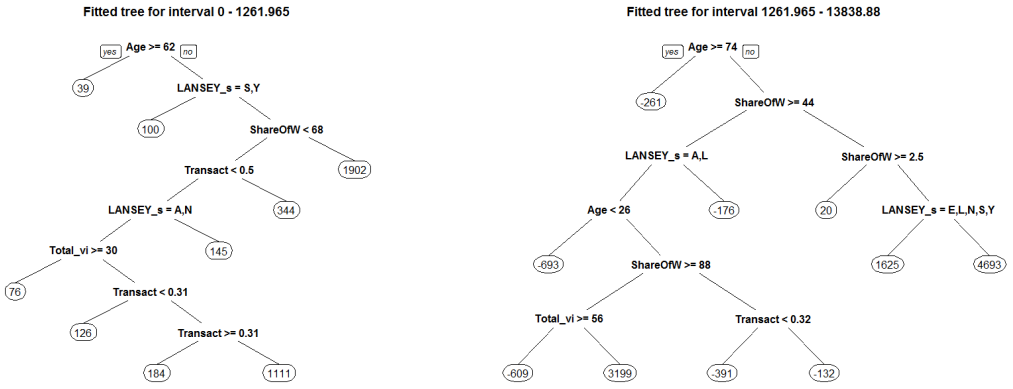
Toisaalta tarkastelemalla luokittelua osajoukolle, kuten kuvassa 7 on esitetty nähdään, että luokittelu toimii ainakin jossain määrin järkevästi, sillä kaksi peräkkäistä havaintoa eli sama henkilö kahtena peräkkäisenä vuotena, pysyy yleensä samassa luokassa, mikä on odotettua. Lisäksi toteutettu Markov-ketju kestää järkevyystarkastelut, kuten että ihmisistä tulee eläkeläisiä pitkällä aikavälillä, elleivät he poistu järjestelmästä ennenaikaisesti.

Valitettavasti tässä projektissa ei olla säädyllisessä ajassa onnistuttu löytämään syytä k-means-pohjaisten menetelmien epätyydyttäville luokitteluille. Lisäksi huomioon ottaen, että k-prototypes-algoritmin tarjoamat luokat eivät ole samalla tavalla intuitiivisia kuin muut tämän projektin algoritmeista ja kolmannen toimivan algoritmin hyödyn jäänessä

rajalliseksi, keskitytään jatkossa muihin ratkaisuihin.

3.2.3 Päätöspuumenetelmän ja Markov-ketjun toteutus

Päätöspuita ja Markov-ketjua hyödyntävässä mallissa lasketaan Markov-mallissa käytettävät tilat rekursiivisella ositusalgoritmilla, jolla muodostetaan regressiopuita. Mallissa sovitetaan käyttäjän valitsema määrä puita siten, että aluksi mallin opettamiseen käytettävän aineiston kotitaloudet jaetaan vuoden kokonaisostojen perusteella väleihin, joissa on yhtä monta kotitaloutta. Jokaiselle välille sovitetaan päätöspuu käyttämällä R:n *rpart*-funktiota ANOVA-ositusmenetelmällä.



(a) Puu, jossa kokonaisostot väliltä 0-1261.965 euroa.

(b) Puu, jossa kokonaisostot väliltä 1261.965-13838.88 euroa.

Kuva 8: Kahden päätöspuun mallin puut. Esitetty kahden eri päätöspuun haarautuminen tapahtuu yhteensä 19 eri luokkaan. Haarautumiskriteerit pystyy lukemaan solmuista, jotka eivät ole lehtiä.

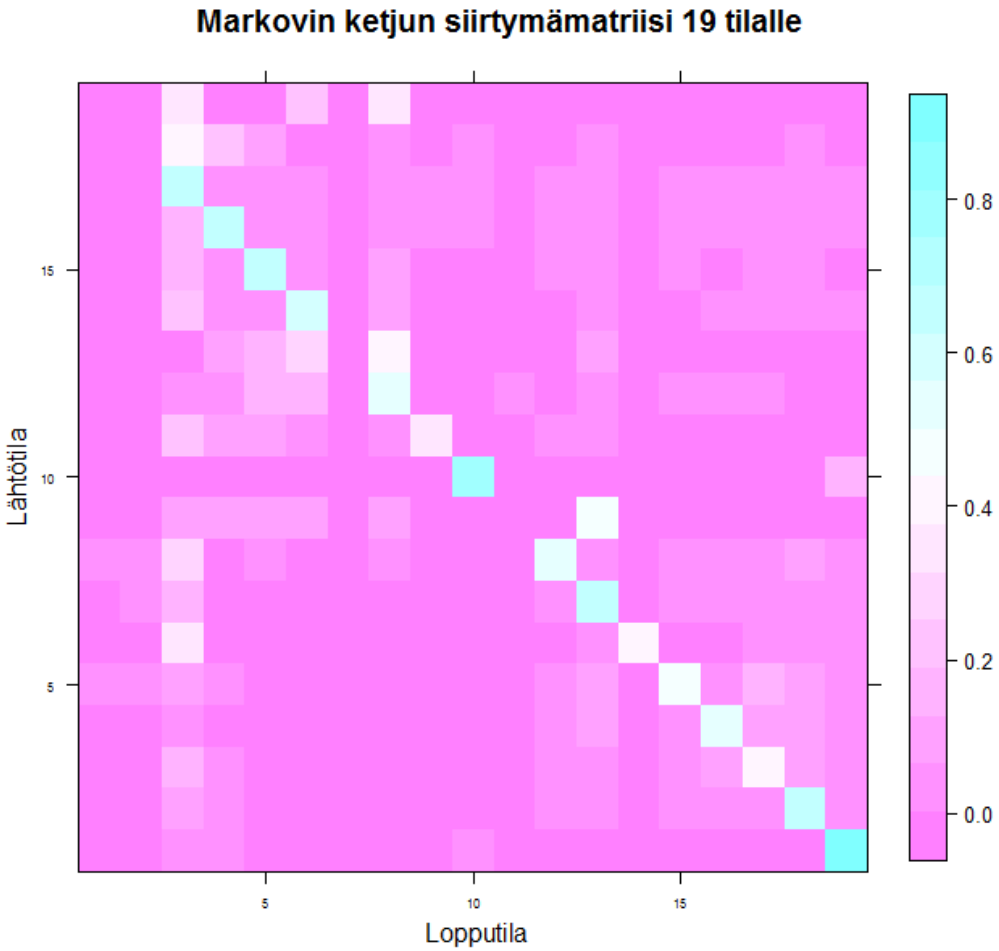
Malli, joka sovitetaan aineistoon on regressiomalli, jossa ennustetaan vuosittaisten ostojen erotusta käyttämällä kaikkia aineiston kotitalouksien muuttujia selittävinä tekijöinä lukuunottamatta postinumeroa, jonka sijaan käytetään asiointitodenäköisyyttä. Päätöspuiden lehtiä käytetään Markov-mallissa tiloina. Lopullinen lehtien määrä riippuu algoritmin suorittamien ositusten määrästä, jota voidaan säätää kompleksisuusparametrilla, joka määrittää, kuinka paljon neliövirheen täytyy parantua, jotta jako kannattaa suorittaa. Osaa selittävistä tekijöistä ei välttämättä käytetä osituksiin yhtä paljon kuin toisia, minkä vuoksi mallissa voidaan asettaa kustannus sille, kun jako tehdään kunkin muuttujan suhteen. Näin voidaan päästä järkevämpään tulokseen, jos muutoin jotakin oleellista selittävää tekijää ei käytettäisi riittävästi. Tämän projektin datan tapauksessa painotetaan erityisesti muuttujia LANSEY ja ikä järkevien puiden saaamiseksi.

Koska päätöspuumallissa ennustetaan vuosittaisten ostojen erotusta, Markov-mallissa lasketaan odotettu kokonaismuutos ja jokaisen talouden CLV saadaan lisäämällä muutokseen näiden alkutilan kokonaisostot. Kuvassa 8 näkyy esimerkki kahdesta puusta, jotka on sovitettu eri ostomääräväleille sellaisella kompleksisuusparametrin arvolla, että lehtiä muodostuu yhteensä 19 kappaletta. Käyttämällä näitä lehtiä Markov-mallin tiloina on laskettu todennäköisyydet tilasiirtymille, jotka näkyvät kuvassa 9. Kuvataan vielä toteutetun päätöspuumenetelmän toiminta kokonaisuudessaan järjestelmällisesti:

Algorithm 2: Usean päätöspuun luokittelu, Markov-ketjun siirtymien määritys ja CLV:n laskeminen

1 Valitse sovitusparametrit: t (aika), n (puiden määrä), c_p (puun haarautumista kontrolloiva kompleksisuusparametri)
2 Jaa havainnot n kappaleeseen järjestyksessä oleviin väleihin I_1, ... , I_n selitettävän kokonaisostot muuttujan mukaan, niin että kullakin välillä suunnilleen yhtä monta havaintoa
3 **for** $I \in \{I_1, ..., I_n\}$ **do**
4 | *dataDiff* $\leftarrow$ *data(2016arvot) - data(2015arvot)*
5 | Sovita puu T_I edellä laskettuun vuosien 2016 ja 2015 väliseen erotukseen *dataDiff*
6 **end**
7 Määritä Markov-ketjun tiloiksi x_1, ... , x_m kaikkien puiden yhteensä m lehteä
8 Laske siirtymämatriisi P eli kuinka usein eri tiloista siirrytään toiseen vuosien 2015 ja 2016 välillä aineistossa $data$ ja skaalaa tilasta siirtymien kokonaismäärällä
9 Lasketaan tilojen keskiostot vektoriin c
10 Luokittele $data$ käyttäen muodostettuja puita T_{I_1} , ... , T_{I_n} ja valitsemalla puu näistä kokonaisostojen mukaan
11 Lasketaan ennustettujen ostojen summa tuleville vuosille kaikille luokille eli CLV $\leftarrow (P^1 + ... + P^t)$ c
12 Määritetään asiakkaan i, josta ollaan kiinnostuneita, arvo CLV_i, tarkastelemalla, mihin luokkaan asiakas kuuluu tai luokittelemalla hänet päätöspuilla johonkin luokkaan, jos kyseessä on uusi havainto
13 **return** CLV_i

Lisäksi, jotta mallia ei validoitasi samalla datalla kuin se on muodostettu, data tulee jakaa satunnaisesti kahteen osaan harjoitteludataksi (80 %) ja testidataksi (20 %). Käytännössä tämä tarkoittaa, että malli muodostetaan harjoitteludatan avulla ja laskettaessa mallin ennusteen virhettä tulee käyttää pelkkää testidataa. Palataan validointiin tarkemmin luvussa 4.2.1 ja muistetaan käyttää tätä validointiperiaatetta myös seuraavaksi esiteltävän menetelmän yhteydessä.



Kuva 9: Markov-siirtymämatriisi 19 tilalla. Värisävy kuvaa siirtymätodennäköisyyttä tilasta toiseen. Selvästi yleisintä on, että pysytään samassa tilassa vuosien välillä. Joillain yksittäisillä luokilla on kuitenkin paljon vaihtuvuutta, esimerkiksi kaksi ensimmäistä tilaa, ja joillakin aivan erityisen vähän, esimerkiksi viimeinen tila, joka koostuu vanhuksista.

3.2.4 Riippumattomien keskiarvojen malli

Riippumattomien keskiarvojen malli tai *jakaumapohjainen ennustaminen* perustuu datan yksiulotteisten jakaumien analysointiin. Tässä esitettävä malli on itse kehitetty, mutta sillä on yhtymäkohtia Bayesilaiseen ennustamiseen [5], jos ajateltaisiin laskettavien suhdelukujen olevan todennäköisyyksinä. Datan tilastollisessa analyysissa huomattiin muun muassa, että naispuoliset kortinhaltijat ostavat yhtä poikkeusta lukuun ottamatta kaikissa LANSEY-luokissa miespuolisia kortinhalitjoita enemmän, kun mittana käytetään vuosittaisten ostojen mediaania. Havainto esitetty kuvassa 3. Myös iän suhteen vuosiostojen mediaani tuotti hyvin säännönmukaisen ja ennustekelpoisen jakauman, joka on esitetty kuvassa 4, esitettävän mallin toimivuus perustuu pitkälti tämänkaltaiseten ilmiöiden

olemassaoloon. Ja mainittakoon, että tässä kuvattava malli on siis rakennettu sen pohjalle, että luokkien ymmärtämiseen käytetään tunnuslukuna mediaania, eikä mallin nimessä esiintyvää keskiarvoa. Malli on oleellisesti samanlainen riippumatta käytetystä tunnusluvusta.

Mallissa oletetaan, että jakaumat ovat toisistaan riittävissä määrin riippumattomat, jolloin jakaumia yhdistämällä saadaan eri jakaumien yhteisvaikutus. Mallissa käytettiin neljää muuttujaa, jotka olivat: LANSEY-luokka, ikä, sukupuoli ja asiointitodennäköisyys. Asiointitodennäköisyys oli väliltä $[0, 1]$ kolmen desimaalin tarkkuudella. Nämä luokiteltiin 0.05 väleihin saadaksemme riittävän suuren otoksen kuhunkin luokkaan. Muut muuttujat pidettiin luokittelun suhteen ennallaan, eli LANSEY-luokkia 6kpl, iät kokonaislukuarvoina $[18, 100]$ ja sukupuoli F=Nainen /M=Mies.

Kunkin muuttujan jokaiselle arvolle laskettiin mediaaniluku vuosiostoille ja tälle mediaanille laskettiin suhdeluku koko datan 'globaalin' mediaanin suhteen, joka on $1\,238.87$ €. Esimerkkinä LANSEY-luokkien kertoimet taulukossa 1.

Taulukko 1: LANSEY-luokkien mediaaniostot ja suhdeluku globaaliin mediaaniin

LANSEY	**Mediaani**	**Suhdeluku**
L	1 663.41	1.343
A	1 667.30	1.346
N	902.57	0.729
S	693.47	0.560
E	988.60	0.798
Y	898.04	0.725

Kertoimien välillä voi olla riippuvuutta, jonka takia laskentaan toteutettiin 'lambda'-kerroin, jolla pystyy voimistamaan tai heikentämään kunkin muuttujan suhdeluku -kerrointa. Sopivimman lambdakertoimen etsimiseen käytettiin menetelmää, jossa verrattiin vuoden 2016 ostojen toteumaa mallin antamaan arvoon ja minimoitiin näiden erotuksen neliösumma lambdakertoimia varioimalla. Variointi toteutettiin automatisoidusti kultaisen leikkauksen optimointialgoritmia hyödyntäen. Tuloksena saatiin lambdakerroin 1.05, joka tarkoitta sitä, että kertoimet ovat parhaimmillaan lähes sellaisenaan, eli tulos ei parane jos suhdelukujen kertoimia voimistetaan tai heikennetään.

Tilasiirtymämatriisit laskettiin jokaiselle LANSEY-tilalle (6x6 matriisi) ja kummallekin sukupuolelle ja kullekin iälle erikseen (0-100), jolloin saatiin 2x100 kokoelma 6x6 matriiseja.

Lopulliset asiakasarvot laskettiin alkutilan perusteella kertomalla kunkin tilamuuttujan suhdelukukerroin yhteen ja kertomalla tämä tulo globaalilla mediaanilla. Ennustessa käytettiin tilasiirtymämatriiseita, jolloin jokaiselle LANSEY-tilalle laskettiin mallin mukainen asiakasarvo, joka kerrottiin tilasiirtymän todennäköisyydellä. Kuten aikaisemmin esitellään taas menetelmän toiminta tiivistetysti vaiheittain:

Algorithm 3: Riippumattomien keskiarvojen mallin toteutuksen kuvaus

1 Suhdelukujen laskeminen jokaiselle muuttujalle, joita käytetään mallinnuksessa (eli LANSEY, ikä, sukupuoli ja asiointitodennäköisyys)
2 Lasketaan lambda-kerroin $\lambda \leftarrow \arg\min_\lambda f(\lambda, data)$, jossa f *ennusteen* ja ostojen toteutuman erotuksen neliö (minimointi voidaan tehdä esimerkiksi kultaisen leikkauksen optimoinnilla)
3 *Ennuste* on asiakkaan niin sanottu *luokka-arvo*, joka lasketaan mediaanisuhdelukujen pohjalta kerrottuna lambdalla, ja jos $\lambda < 1$, niin tällöin suoraan kertomisen sijaan niille muuttujille, joiden suhdeluku pienempi kuin 1, kerrotaan näiden käänteislukua jollain mielekkäällä kertoimella joka on > 1. Tarkka toteutus on tässä suhteellisen mielivaltainen, mutta oleellista on, että niiden muuttujien, joilla on keskenään päällekkäisyyttä keskenään selittävinä tekijöinä, vaikutusta pienennetään jossain määrin jollain tapaa
4 Lasketaan tilasiirtymämatriisit perustuen siirtymiin eri LANSEY-luokkien välillä kaikissa ikä- ja sukupuoliluokissa
5 Asiakkaan CLV on kaikkien vuosien tilajakaumien (iän ja sukupuolen mukaan lasketun) summan ja viimeisen vuoden *luokka-arvon* laskettuna kaikille LANSEY-luokille tulo. Eli asiakkaan CLV:n lasku tapahtuu suoraan sellaiselle asiakkaalle, jolla on samat ominaisuudet (LANSEY, ikä, sukupuoli ja asiointitodennäköisyys) kuin datassa jo löytyvällä havainnolla, mutta jos uudella havainnolla on jokin muuttujan arvo, mitä ei löydy alkuperäisessä datassa (kuten ei-kokonaisluku ikä), tulee havainnon muuttujia interpoloida lähimpää muuttujaa vastaavaksi

4 Tulokset

Käydään läpi menetelmien tarjoamat tulokset ja pyritään myös validoimaan niiden oikeellisuus.

4.1 Mallien antamat tulokset: luokittelusta CLV:hen

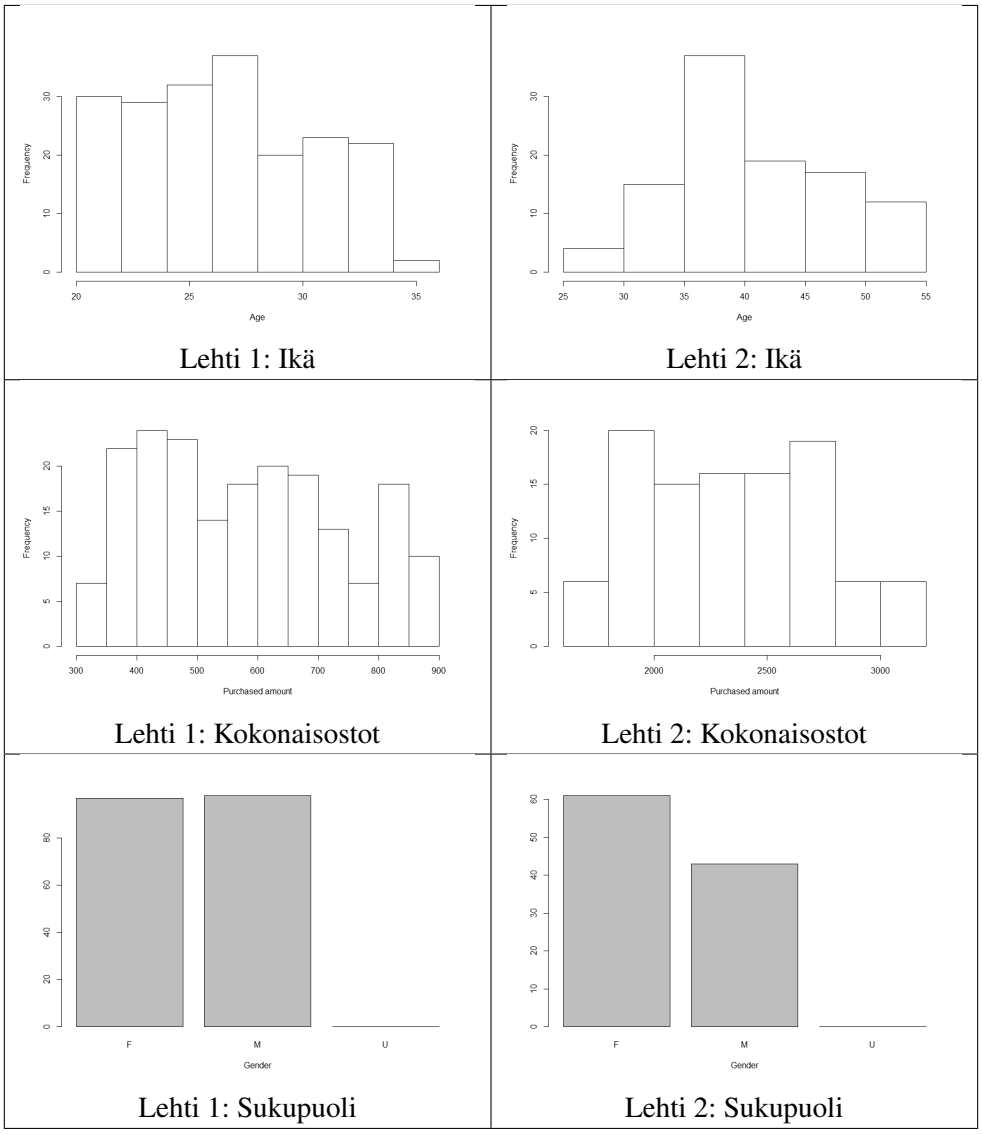
Esitellään tässä minkälaisia luokkia mallit muodostavat, kun luokkien määrä valitaan käytännön tarkoituksiin sopivan kokoisiksi eli että ennustetarkkuus on hyvä, mutta toisaalta

että laskentateho riittää ja että luokissa on useita havaintoja, jotta vältyttäsiin ylisovitukselta ja muilta ongelmilta, joita liittyy ennusteen uskottavaksi saamiseksi pitkälle aikavälille. Tarkastellaan myös lyhyesti muodostuvaa Markov-ketjua ja lopuksi lasketaan esimerkinomaisesti CLV:itä eri yksilöille ja täten myös ryhmille.

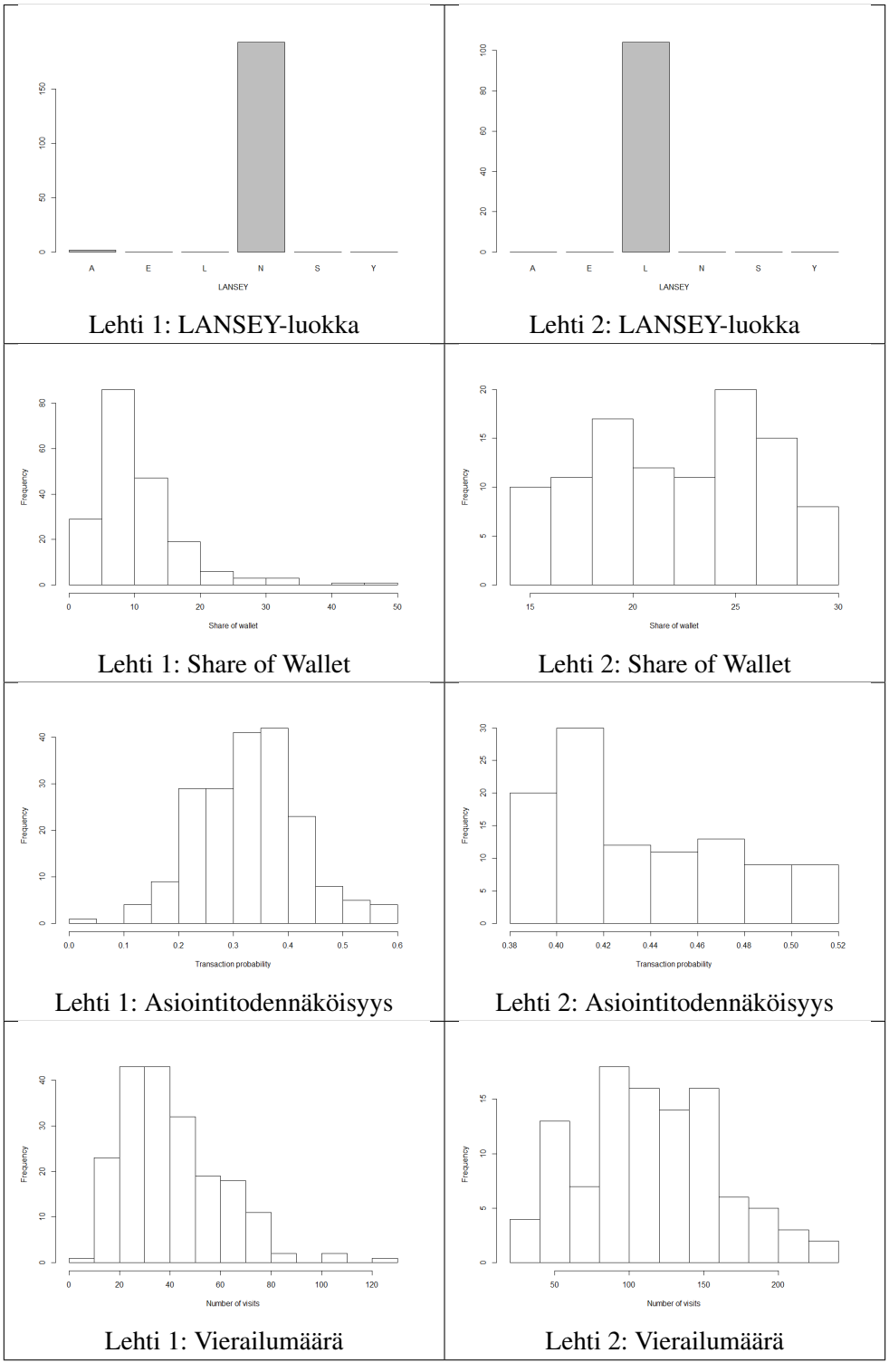
4.1.1 Luokittelualgoritmien antamat luokittelut

Tarkastellaan nyt päätöspuumallin antamia luokitteluja. Toisin kuin aikaisemmin läpi käyty ja hylätty k-prototypes, saatavat luokat ovat jokseenkin ymmärrettäviä kokonaisuuksia, ja erityisesti päätöspuusta voidaan lukea halutessa haara haaralta, kuinka lopulliseen lehteen eli luokkaan päädytään. Jokaisessa luokassa on useita erilaisia havaintoja, joilla on ainakin joitain yhteisiä ominaisuuksia. Päätöspuumallin tulokset laskettiin käyttämällä viittä puuta, kompleksisuusparametrin arvolla 0,0005. Kustannusvektori asetettiin siten, että LANSEYn ja iän suhteen osituskustannukset olivat kumpikin 1 ja loput 10. Neutraaleilla jakokustannuksilla ennustettu ostetun määrän vuosittainen kehitys iän ja LANSEY-luokan suhteen olivat epäintuitiivisia, sillä esimerkiksi eläkeläisillä vuosittaiset ostot kasvoivat, mikä on odottamatonta, sillä eläkeläisten säästöt ja menot yleensä pienenevät iän myötä kuin myös todennäköisyys kuolla kasvaa pienentäen odotusarvoisia ostoja. Järjettömät tulokset johtuivat siitä, ettei puussa ollut tehty osituksia riittävästi iän eikä LANSEY-luokan suhteen. Asettamalla iän ja LANSEYn perusteella tehtävien osituksien kustannuksia pienemmiksi suhteessa muihin muuttujiin, ne saadaan otettua mallissa paremmin huomioon.

Tarkastellaan seuraavaksi kahta satunnaisotoksella valittua luokkaa, kun luokittelu tehdään oikealle ostodatalle. Kumpaankin luokkaan liittyviä jakaumia on vierekkäin käyty läpi kuvassa kuvissa 10 ja 11. Kuvassa 10 näkyvä kokonaisostojen jakauma edustaa seuraavan vuoden ennustettua kokonaisostoa. Kuvista nähdään, että lehdessä 1 on selvästi nuorempaa väkeä, jotka asioivat huomattavasti harvemmin kuin lehden 2 ihmiset. LANSEY-luokatkin ovat eksklusiivisia keskenään näissä lehdissä. Lehteä 1 voisi karakterisoida esimerkiksi termillä säästeliäät nuoret parit, jotka tekevät vain pakollisia ostoja ja lehteä 2 termillä keski-ikäisistä vanhemmista koostuvat lapsiperheet, jotka tekevät pakollisten ostojen lisäksi myös vaihtelevasti satunnaisostoja.



Kuva 10: Verrataan kahta satunnaistalehteä eli päätöspuumallin antamaa luokkaa keskenään tutkimalla eri muuttujien jakaumia. Ensimmäisen lehden jakaumat vasemmalla ja toisen oikealla. Käydään tässä läpi muuttujat ikä, kokonaisostot, sukupuoli ja loput muuttujista seuraavassa kuvassa.



Lehti 1: LANSEY-luokka | Lehti 2: LANSEY-luokka

Lehti 1: Share of Wallet | Lehti 2: Share of Wallet

Lehti 1: Asiointitodennäköisyys | Lehti 2: Asiointitodennäköisyys

Lehti 1: Vierailumäärä | Lehti 2: Vierailumäärä

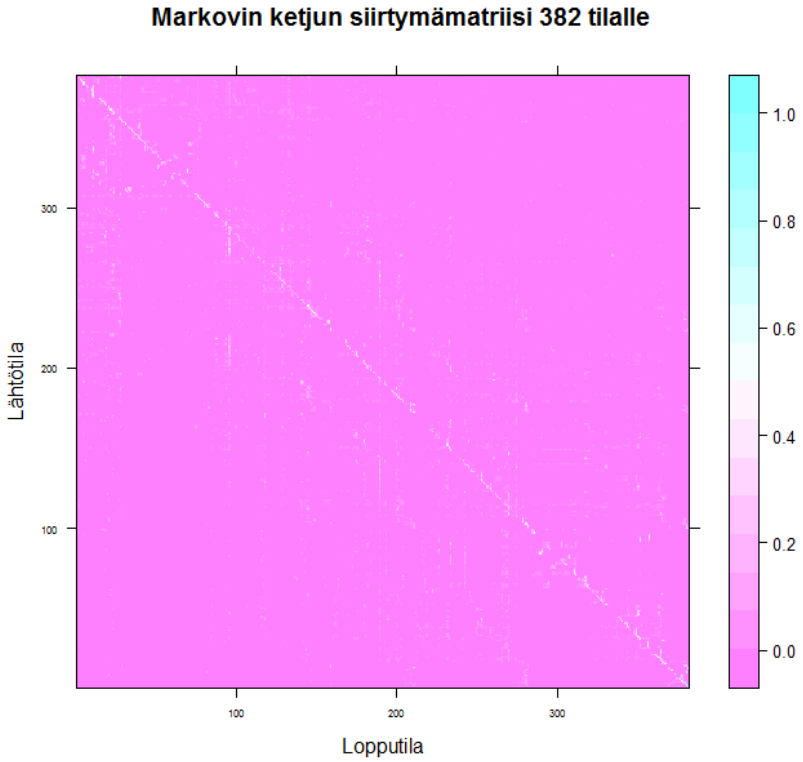
Kuva 11: Jatkoa edelliselle kuvalle. Verrataan tässä puolestaan muuttujista LANSEY-luokan, Share of Wallet:in, asiointitodennäköisyyden ja vierailumäärän jakaumia.

Edellä kuvattu päätöspuu luokittelee ongelmitta myös ennennäkemättömät havainnot. Valitaan vain haara haaralta uuden havainnon muuttujien arvojen avulla. Riippumattomien keskiarvojen mallissa luokat puolestaan valitaan itse, joten mitään luokittelun ongelmaa

ei samalla tavalla ole, mutta malli ei osaa luonnostaan luokitella täysin uusi havaintoja. Tähän ongelmaan on monta ratkaisua, kuten etsiä vanhojen havaintojen joukosta mahdollisimman paljon uutta havaintoa muistuttava havainto, esimerkiksi pienin mahdollinen Gower-etäisyys tai muuta vastaavaa, ja luokitella ne sitten samalla tavalla.

4.1.2 Markov-ketjun toiminta

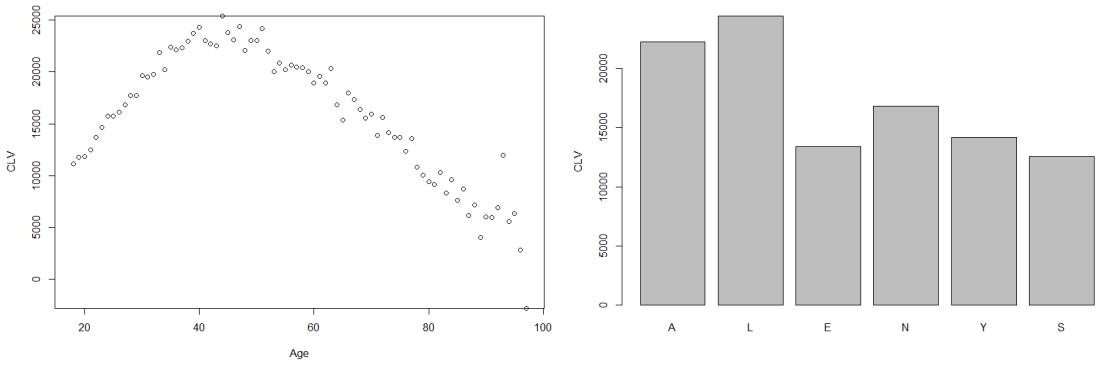
Markov-ketjut toimivat tässä kuten menetelmien kuvan 9 alustavassa esimerkissäkin. Nyt koska päätöspuumallissa on yhteensä 382 lehteä, tilojakin on Markov-ketjussa yhtä monta. Jokainen lehti eli tila sisältää samankaltaisia yksilöitä, mutta kuten kuvassa 10 nähtiin, on luokkienkin sisälläkin jonkin verran hajontaa. Tämän kokoista Markov-ketjua vastaavan tilamatriisin pyörittely ei vielä ole laskennallisesti työlästä. Tilasiirtymämatriisi näkyy kuvassa 12. Kuvasta nähdään vaaleampi poikkiviiva eli että tilasta päädytään usein samaan tilaan takaisin. Yksittäisiä voimakkaita siirtymiä tilasta toiseen myös hahmottuu satunnaisina kirkkaina pisteinä poikkiviivan ulkopuolella, mutta pääosin satunnaisesta tilasta toiseen siirrytään erittäin pienellä tai olemattomalla todennäköisyydellä.



Kuva 12: Aikaisemmin esitetyn kaltainen siirtymämatriisi, jossa tiloja yhteensä 382.

4.1.3 Ennusteet luokkien CLV-arvoiksi

Simuloimalla edellä esitettyä Markov-ketjua eteenpäin ja summaten vuosittaiset odotusarvoiset ostot saadaan CLV siis määriteltyä päätöspuumallien yhteydessä. Vastaavasti riippumattomien keskiarvojen malli laskee valituille luokille CLV-ennusteet, jotka ovat arvoltaan melko yhteneviä. Esitetään päätöspuumallin laskemia CLV-arvoja ryhmäkohtaisesti. Kuvassa 13 on eritelty keskimääräinen CLV-arvo jokaiselle ikäryhmälle sekä jokaiselle LANSEY-ryhmälle, kun CLV lasketaan 10 vuoden aikajänteelle.



(a) Ennustetut CLV-arvot eri ikäryhmille. (b) Ennustetut CLV-arvot eri LANSEY-luokille.

Kuva 13: Päätöspuumallin antamat ennusteet 10 vuoden CLV:ksi erinäisille tarkasteltaville ryhmille.

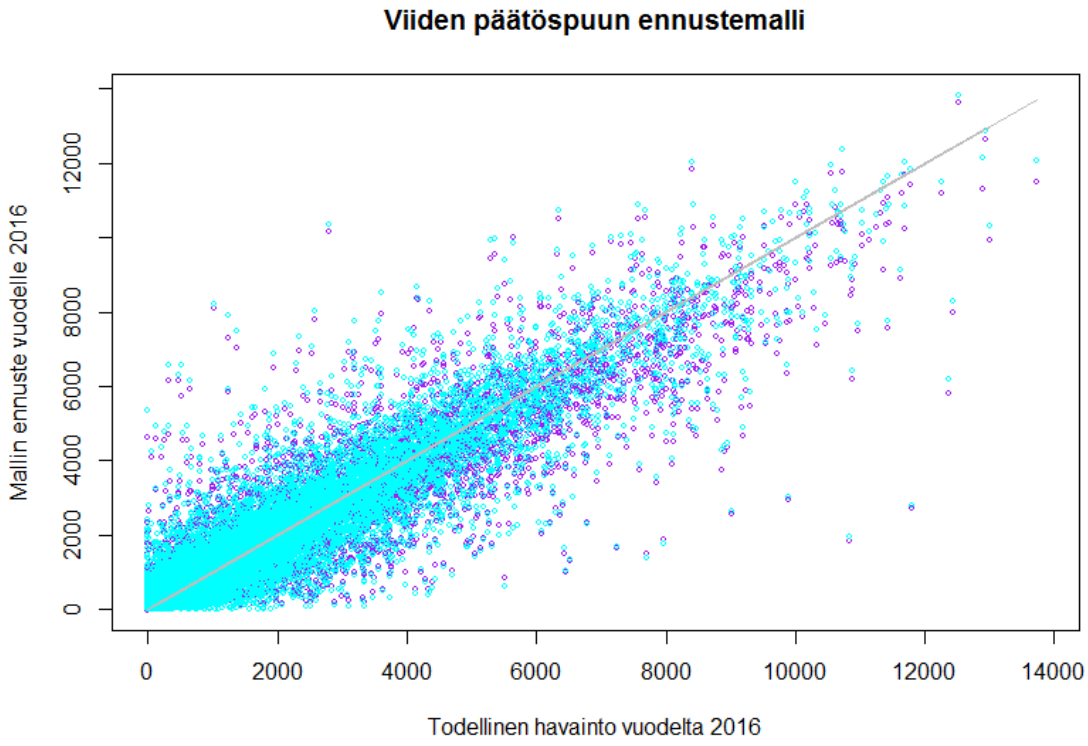
CLV on korkeimmillaan noin 45-vuotiailla ollessaan jopa $25\,000$ euroa ja mitä nuoremmaksi taikka vanhemmaksi tästä mennään sitä pienemmäksi CLV tulee. Huomattavaa on, että hyvin vanhoilla CLV on enää luokkaa $5\,000$ euroa sen vuoksi, että nämä ihmiset ovat lähellä elinikänsä loppua. Nuorilla taas jatkuva kasvu CLV:n arvossa selittyy niin kasvavilla tuloilla kuin myös lasten hankinnan takia. LANSEY-luokissakin näkyy runsasta vaihtelua. Suurimman CLV-luokan talous on lähes kaksinkertainen alhaisimpaan CLV-luokaan verrattuna.

4.2 Tulosten arviointi

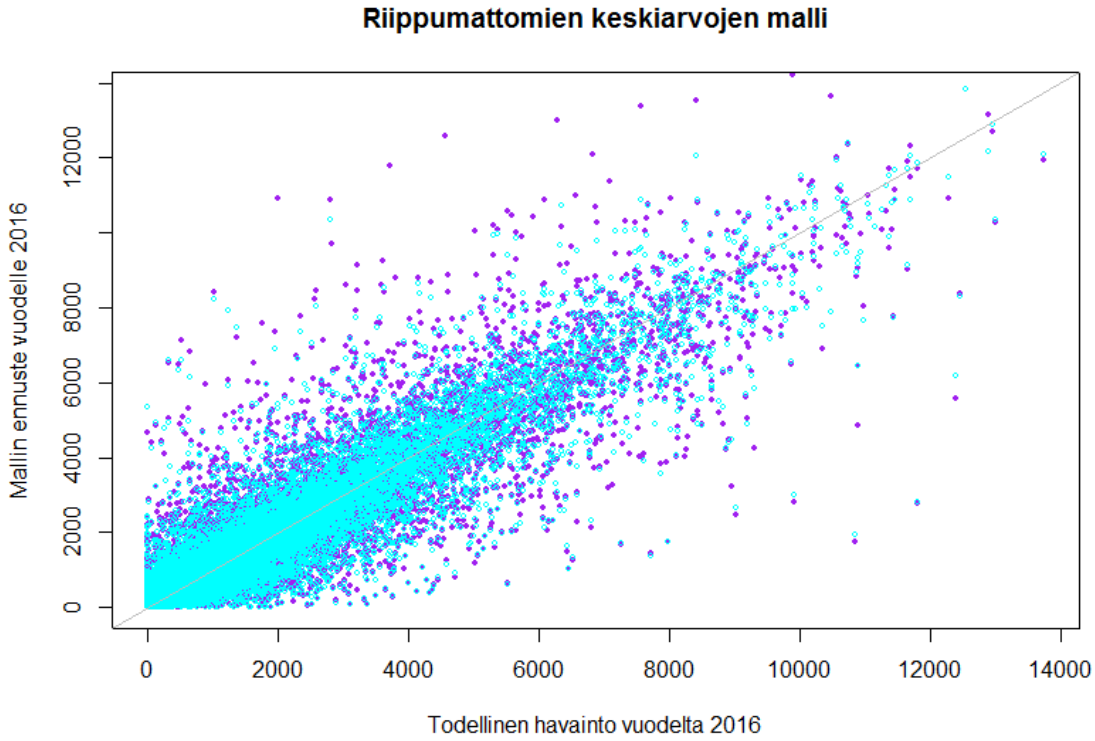
Arvioidaan, ovatko toteutetut mallit toimivia perustuen lähinnä tulosten tarkkuuteen ja yleiseen järkevyystarkasteluun.

4.2.1 Tulosten ja mallien validointi

Mallien ennustetarkkuutta hyvin lyhyellä aikavälillä voidaan tarkastella tutkimalla, mitä malli ennustaa seuraavan vuoden olevan perustuen edellisvuoden dataan. Tässä tunnetaan vuosien 2015 ja 2016 kokonaisostot, joten tutkitaan kuinka paljon vuoden 2015 pohjalta tehty ennuste poikkeaa todellisista 2016 vuoden arvoista ja lasketaan erotusten neliösummat. Otetaan mukaan tarkasteluun myös baseline-malleissa esitetty periaate, jossa käytetään edellisvuotta seuraavan vuoden ennusteena.



Kuva 14: Päätöspuumallin ennusteet, kun ennustetaan vaaka-akselia eli vuotta 2016 vuoden 2015 dataan pohjautuen purppuralla värillä olevissa pisteissä päätöspuumallilla ja syaanin värisissä pisteissä edellisvuoden ostojen avulla pystyakselilla. Viistolla suoralla ennusteet vastaavat todellisuutta täysin.



Kuva 15: Vastaava kuin kuva 14, mutta purppurien pisteiden ennustemallina nyt riippumattomien keskiarvojen malli.

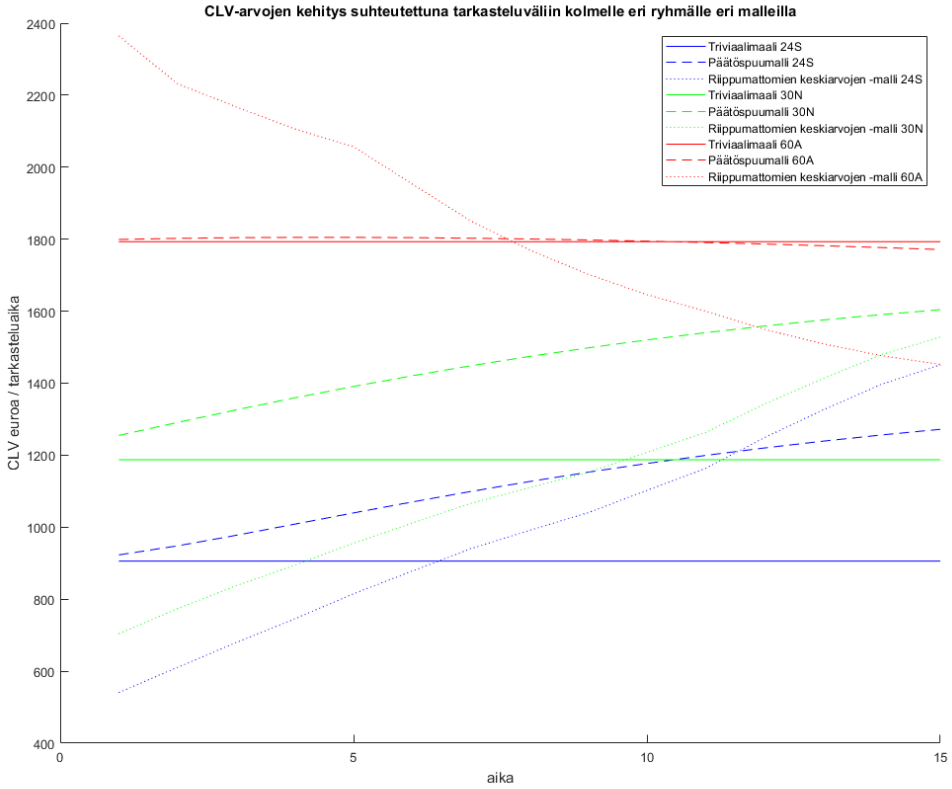
Kuvissa 14 ja 15 näkyy, kuinka mallit ennustavat molemmat päällisin puolin melko samanlaisesti seuraavan vuoden ostoja. Virheet todellisesta arvosta ovat symmetrisesti osa liian suuria ja osa liian pieniä. Lisäksi nähdään, että yleensä kun edellisvuoden baseline-ennuste on huono, ei myöskään mallin ennuste ole kovin hyvä. Laskemalla virheet saadaan tulos, että päätöspuumallin virheiden keskihajonta 3–4 % pienempi kuin baseline-mallin ja riippumattomien keskiarvojen mallin tapauksessa noin 16 % suurempi. On tärkeää, ettei mallin ennuste ole merkittävästi baseline-mallia huonompi yksittäisen askeleen ennustamisessa, mutta mallin hyvyys mitataan tässä projektissa pidemmän aikavälin ennustamisessa. Vaikka tämä projekti on toteutettu kahden vuoden ostodataan perustuen, on silti olemassa keinoja tutkia, onko pidemmän aikavälin ennuste hyvä.

Mallien antamien tulosten realistisuutta voi tarkastella tutkimalla sellaisten ryhmien CLV:n arvon kehitystä, joille kehityksen kulku on intuitiivisesti selvää. Esimerkiksi lapsettomissa vähän alle kolmekymmentä vuotiailla on kohtalaisen suuri todennäköisyys jälkikasvun hankintaan. Tämän vuoksi heidän tekemiensä ostojen arvon tulisi odotusarvoisesti nousta lähivuosien aikana. Vastaavasti hyvin vanhat eläkeläiset siirtyvät käyttämään esimerkiksi ateriapalveluita ja tulot yleensä muutenkin vähenevät, joka johtanee pieneneviin ostoihin.

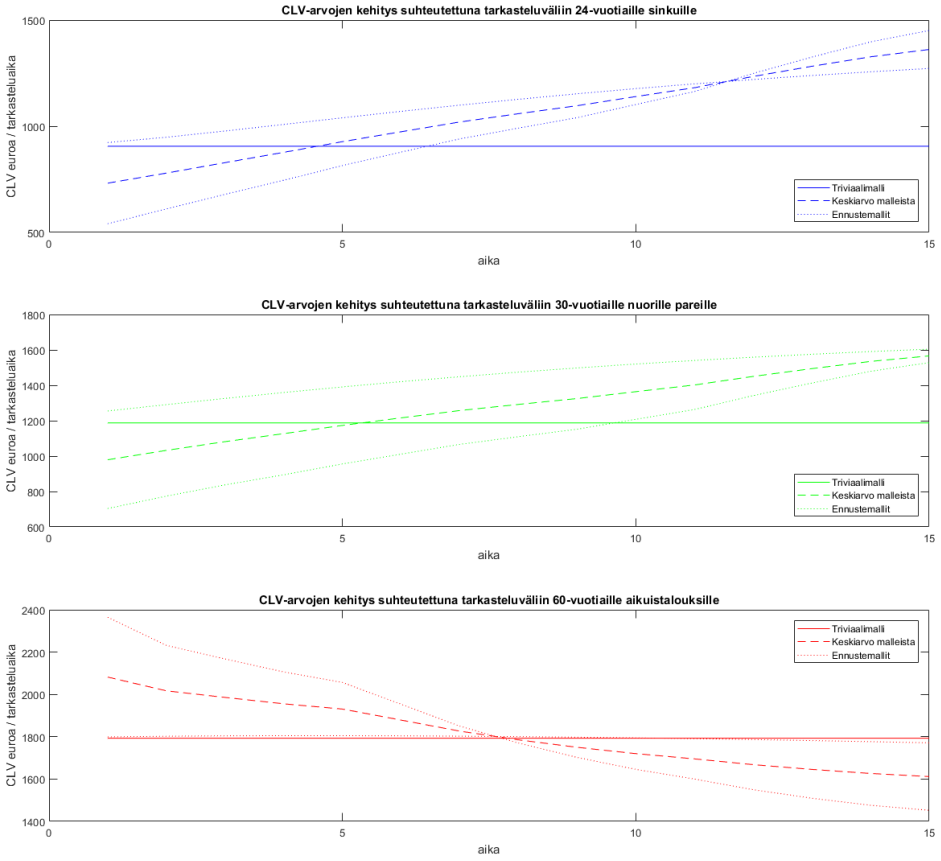
Tarkastellaan liitteen B kuvisssa, kuinka erilaiset ryhmät tuovat yritykselle arvoa eri vuosina. Tarkastelu lähtee liikkeelle valitun ryhmän iästä ja sitten tutkitaan, kuinka tämän ryhmän arvo kehittyy lähivuosina. Saadut kuvaajat yhtyvät hyvin intuitiivisiin käsityksiin siitä, kuinka tilanteen tulisikin kehittyä, ja koska tätä ihmisen kehittämää intuitiota ei varsinaisesti mitenkään huomioitu algoritmien kehittämisvaiheessa, voidaan tällaista validointia pitää pätevänä, ja voidaan todeta mallien toimivan uskottavasti yhtä vuotta pidemmälläkin aikavälillä. Vastaavasti esimerkiksi pelkkä edellisvuosi ennusteena väittää asiakkaan ostojen pysyvän vakiona, mikä on selvästi huono oletus monissa tapauksissa.

4.2.2 Mallien loppullinen vertailu ja arviointi

Verrataan tässä tarkemmin, kuinka eri mallit ennustavat CLV-arvoja samoille valituille ihmisryhmille. Valitaan kolme ryhmää vertailtavaksi: 24-vuotiaat sinkut, 30-vuotiaat nuoret parit ja 60-vuotiaat aikuistaloudet. Lisäksi on valittu, että asiointitodennäköisyys on väliltä 0.25–0.35 ja henkilöt ovat naisia, koska riippumattomien keskiarvojen mallissa nämä tulee täsmentää. Kuvassa 16 nähdään mallien ennusteet kaikkille ryhmille. Jotta mallien toimintakyky tulisi ilmi eri aikajänteille, kuvassa on laskettu CLV jaettuna tarkasteluajalla, joka vaihtelee yhdestä vuodesta aina viitentoista vuoteen. Yleisesti ottaen huomataan, että luonnollisesti edellisvuosi ennusteena antaa vakioennusteen kaikille ryhmille ja edistyneemmät mallit ennustavat kahdelle ensimmäiselle tarkasteluryhmälle CLV-arvon kasvua ja viimeiselle ryhmälle laskua.



Kuva 16: Kolmen eri ryhmän vertailu triviaalilla edellisvuosi ennusteena mallilla, päätöspuumallilla ja riippumattomien keskiarvojen mallilla viidentoista vuoden ajalta. CLV on aina laskettu tarkasteltavan aikajänteen mukaan ja jaettu aikajänteen pituudella kuvaajien luettavuuden vuoksi.



Kuva 17: Tarkemmat kuvat CLV:n kehityksesta ryhmittäin. Mallien keskiarvoa voidaan pitää yksittäisen mallin ennustetta luotettavampana ja tässä tapauksessa yksittäisten mallien ennusteita voi lisäksi pitää tämän ennusteen väljinä luottamusväleinä.

Malleissa on kuitenkin merkittävää epävarmuutta. Ja muitakin yksittäisiä ongelmia esiintyy, kuten että tässä päätöspuumalli ennustaa vanhahkon aikuistalouden arvon pysyvän melko vakiona, mikä ei ole intuitiivista. Toisaalta jos lopullisena ennusteena käytetään kahden hyvän mallin keskiarvoa, kuten kuvassa 17, saadaan lähes poikkeuksetta hyviä tuloksia, kun ennustetaan riittävän pitkälle aikavälille. Jos oletetaan, että mallit on kehitetty niin sanotusti toisistaan riippumattomasti, voidaan ajatella, että mallit ennustavat kaksi toisistaan riippumatonta havaintoa tarkasteluaikajänteelle. Tällöin perustuen luottamusvälien laskentaan hyvin pienen havaintokoon vallitessa, tulee mallien keskiarvolle 50 % luottamusväliksi mallien ennusteiden ja keskiarvon erotus eli ylärajaksi se malli, joka ennustaa suurempaa arvoja ja vastaavasti pienempää arvoa ennustava alarajaksi, kuten kuvassa 17 on esitetty. Muita tapoja arvioida mallin epätarkkuutta olisi laskea ennusteita useasti vaihtelemalla sitä osajoukkoa koko datasta, jonka pohjalta CLV on laskettu. Tässä kuitenkin oletettaisiin, että mallit ennustavat täydellisesti, kunhan data on riittävän

kattava, mikä ei kovin ole perusteltua. Mallit vaikuttavat olevan kykeneviä ennustamaan, missä ryhmissä tulee tapahtumaan arvon laskua tai nousua, ja eri ihmisryhmien käytöksen yleiseen vertailuun.

Siispä kehitetyt mallit molemmat läpäisevät erinäisiä lyhyen aikavälin kuin myös pitkän aikavälin validointiin liittyviä testejä, ja lisäksi ne toimivat riittävän nopeasti käytännön tarpeisiin ja auttavat hahmottamaan, mikä tekee asiakkaista erilaisia ja erityisesti miksi jotkin asiakkaat ovat vähemmän arvokkaita kuin toiset eli ymmärtämään kuinka CLV pinnan alla määräytyy sen lisäksi että saadaan konkreettisia lukuarvoja CLV:iksi. Malleja on myös helppo laajentaa tarvittaessa uuden datan tarpeisiin. Tätä väitettä tukee se, että malleja kehittäessä suoritettiin jatkuvaa herkkyysanalyysia siinä mielessä, että mallien perustuomivuudesta varmistuttiin kun käytettiin vain osaa datan muuttujista tai havainnoista ennustamiseen. Yksittäinen huono puoli on se, että malli ei osaa ennustaa äärihavaintojen käyttäytymistä, mikä ei ole kuitenkaan ongelma siinä mielessä, ettei näiden yksilöiden käytöstä edes pystyne ennustamaan millään prediktiivisella analytiikalla tarkasti. Jotta voitaisiin oikeasti varmistua menetelmien toimivuudesta pitkän aikavälin ennustamisessa, tarvittaisiin validointidataa enemmän kuin kahden vuoden ajalta.

Verrataan tehtyjä malleja vielä lopuksi taulukossa 2, jossa on listattu mallien ominaisuuksia. Mallit näyttävät yleisilmeeltään varsin suotuisilta Keskon liiketoiminnan kehitykseen CLV:n ymmärtämisen kautta.

Taulukko 2: Kahden eri mallin vahvuuksien ja heikkouksien vertailua.

Päätöspuu pohjainen malli	**Riippumattomien keskiarvojen malli**
Mallin monimutkaisuus käyttäjälle:	
Mallin käyttö vaatii mallin parametrien ymmärtämistä sekä näille sopivien arvojen löytämistä. Parametrien löytämiseen on valittava sopivan kokoinen kokoelma havaintoja. Puiden sovituksen jälkeen riittää syöttää data.	Yksinkertainen. Funktio, johon syötetään asiakasluokan parametrit - asiointitodennäköisyys, LANSEY, ikä, sukupuoli
Ennustetarkkuus vuoden päähän:	
Hiukan parempi kuin edellisen vuoden ostot selittäjänä. Virhetermien keskihajonta 3–4 % pienempää kuin edellisen vuoden selittäjällä.	Huonompi kuin edellisen vuoden ostot selittäjänä. Virhetermien keskihajonta 16 % suurempaa kuin edellisen vuoden selittäjällä.
Laskenta-aika, laskennallinen kompleksisuus :	
3,5 s 20 vuoden ennusteelle, kun puita on 5 ja tiloja 382. Rajoittava osuus on oikean alkutilan löytäminen koko datalle sovitetun mallin avulla. Tämän aikakompleksisuus on pahimmassa tapauksessa $\mathcal{O}(mn)$, m = suurimman puun solmujen määrä, n = kotitalouksien määrä.	0,4 s 20 vuoden ennusteelle, kun tiloja 1200 Laskenta-aika lineaarinen ennusteen pituuden kuten myös havaintojen määrän suhteen. Eli kun ennusteen pituus on vakio 5 tai 10 vuotta aikakompleksuus on $\mathcal{O}(n)$, n = kotitalouksien määrä.
Algoritmin monimutkaisuus:	
Kompleksisin osuus on regressiopuun muodostus, joka sekin hoituu yhdellä R-komennolla. Muuten yksinkertainen.	Lambda-kertoimen määrittämiseen käytetty kultaisen leikkauksen optimointialgoritmi on kompleksisin. Muuten yksinkertainen.
Muita etuja:	
Regressiopuumallin parametrit mahdollistavat säädettävyyden esimerkiksi sopivan Markov-mallin tila-avaruuden saavuttamiseksi. Lisäksi uusia selittäviä tekijöitä on helppo lisätä malliin.	Toteutuksen ja ylläpidon kompleksisuus on matalaa tasoa. Läpinäkyvä malli, jonka toiminta on helppo ymmärtää ja siten mallin puutteet pystyy huomioimaan kokonaisvaltaisessa tarkastelussa.
Muita haittoja:	
On mahdollista, että kaikki Markov-malliin päätöspuumallista tuodut tilat eivät sisällä informaatiota joistakin oleellisista selittävistä tekijöistä riippuen siitä, minkä selittävien tekijöiden perusteella sovitusalgoritmi on tehnyt osituksia puuhun.	Epätasainen käytös siirtymissä - esimerkiksi talous, joka on siirtymässä esimerkiksi nuoresta parista lapsiperheeksi käyttäytyy siten, että juuri ennen siirtymää nuoren parin ostot ovat suuremmat kuin nuorilla pareilla keskimäärin, mutta välittömästi siirtymän jälkeen heidän ostot ovat pienemmät kuin lapsiperheillä keskimäärin.

5 Pohdinta

Projektissa määriteltiin melko vahvasti tehtävänanto heti alkuun ja aiheenasettaja suositti melko voimakkaasti Markov-ketjujen käyttöä malllintamiseen, joten projektia lähestyttiin ehkäpä turhankin hanakasti oletuksella, että ensin havainnot luokitellaan konkreettisiin luokkiin, joita simuloidaan Markov-mallilla eteenpäin. Yhtenä ratkaisuna olisi ollut

käyttää jotain ratkaisua, jossa luokkia ei muodosteta tai ainakaan niitä ei niin sanotusti selvitettäisi, vaan ne voisivat toimia jotenkin taustalla. Myöskin Markov-ketjut olisi voinut korvata joillain dynaamisella polkuriippuvaisella menetelmällä, joka saattaisi huomioida pitkien aikojen ilmiöitä asiakkaissa paremmin kuin muistiton Markov-ketju.

Yksi ongelma oli, että pitkän aikavälin validoinnille ei ollut kvantitatiivista menetelmää. Esimerkiksi ostodata useammalta vuodelta olisi jossain määrin ratkaissut tämän ongelman. Oltaisiin voitu havaita, etteivät luodut mallit olekaan yhtä hyviä kuin nyt luulemme, mutta toisaalta tällaisten mallien kehittämisen työmäärä olisi ollut suurempi ja ehkä liiankin suuri tämän projektin puitteisiin. Myös itse mallin kehitys olisi voinut olla systemaattisempaa.

Auki jää myös mallien pätevyysalue. Kehitetyt mallit ovat selvästi triviaalimallia parempia keskipitkällä aikavälillä (5-10 vuotta) ja lyhyellä aikavälillä melko yhdenveroisia. Toisaalta kun mennään kauemmas tulevaisuuteen, nousee varmaan esiin yleisiä suuntauksia, jotka vaikuttavat päivittäistavarakauppaan ja joita mallit eivät osaa ennustaa.

Toimivat mallit tarjoavat ymmärrystä, kuinka ihmiset jakautuvat eri arvoisiin ryhmiin asiakkaalle. Uskomme, että mallejamme pystytään hyödyntymään esimerkiksi markkinoinnin kohdentamisessa ja siinä, että aiheenasettaja pystyy panostamaan itselleen pitkällä aikavälillä arvoikkaimpiin asiakkaisiinsa taikka niihin asiakkaihin, joilla on suurin kasvupotentiaali ja toisaalta panostamaan vähemmän asiakkaihinsa, joiden arvo on laskussa. Malleja tulee kuitenkin laajentaa nykyisestään, esimerkiksi markkinoinnin näkökannalta [1], jotta ne olisivat muutakin kuin vain vahvistamassa jo olemassa olevaa intuitiota. Pitää miettiä, mitä kuluja ostojen aikaansaamiseen liittyy, miten asiakkaat vuorovaikuttavat keskenään ja keskittyä ehkä oleellisempaan kysymykseen, eli siihen miten asiakkaan arvoa lisätään, eikä siihen, mikä se on nyt tai tulevaisuudessa.

A Itsearviointi

Ryhmän sisäisestä toiminnasta: Heti projektin alussa viikoittaiset tapaamiset oli hyvä ja toimiva idea. Ryhmä perusti myös viipymättä joukkoviestiryhmän, jonka avulla tehtävienjakoa oli helppo koordinoida. Viikoittaisista tapaamisista olisi saanut enemmänkin tehoja irti, jos tapaamisissa olisi pidetty tiukemmin ennalta sovitusta sapluunasta. Projektin työstämistapa oli lainattu ketterässä ohjelmistokehityksessä käytetystä tapaamisperinteestä: näissä tapaamisissa jokainen aluksi kertoo lyhyesti, mitä on tehnyt edellisen tapaamisen jälkeen ja sovitaan mitä jokaisen tulee saada tehtyä seuraavaan tapaamiseen mennessä. Tapaamiset ovat myös tiukasti aikarajattuja. Yleinen sääntö kaikille palavereille ja kokouksille on, että jokaisen osallistujan tulee tietää tarkalleen miksi kokous pidetään, mitä asioita siellä käsitellään ja mitkä tavoitteet tulee täyttää, jotta kokous voidaan päättää.

Todellisuudessa lyhyen aikavälin tavoitteet määriteltiin liian löyhästi ja raporttien määräajat ja asiakkaan kanssa sovitut tapaamiset saivat projektin etenemään parhaiten. Lisäksi alkuun ja vauhtiin pääsemisessä oli lievää kankeutta, mutta tällaisen alkukankeuden suhteellista osuutta on vaikea pienentää. Projektin kokonaiskestoa pystyisi toki lähes mielivaltaisesti säätämään allokoimalla projektille kalenterista enemmän aikaa. Projektin kesto oli kurssin puitteissa ennalta määrätty, joten projektinhallinnassa ei ollut tarvetta puristaa projekti läpi minimiajassa. Koska projektin määräaikaa ei pystynyt muuttamaan, niin jousto tapahtuu silloin joko projektin hetkellisestä työmäärästä tai työn laadusta. Ryhmän jäsenet olivat työmäärän suhteen kuitenkin riittävän joustavia, jotta työn laadusta ei tarvinnut projektin missään vaiheessa räikeästi tinkiä. Toki työn laadun ja käytetyn ajan suhteen vallitsee useissa projekteissa, kuten tässäkin, jokin piste, jonka jälkeen käytetyn ajan rajahyöty painuu pieneksi.

Hyvänä puolena tässä kaikessa oli se, että tietynlainen luovuus ja mahdollisuus muutoksiin säilyi läpi projektin. Tästä esimerkkinä oli se, että projektissa päästiin toteutuksen tasolla oikeastaan neljään eri malliin, joista yksi (baseline-malli) muodostui pohjaksi jatkokehitykselle ja toinen jäi kesken ajan puitteissa (k-prototypes). Nämä opettivat, mikä on luokittelussa ja muussa haastavinta, ja jopa auttoivat ymmärtämään, miksi onnistuneet mallit eivät toimineet kovin hyvin vielä ennen kuin ne lopulta saatiin toimimaan ongelmitta.

Pienessä neljän hengen ryhmässä kommunikointi ei noussut ongelmaksi. Kaikki ryhmän jäsenet olivat samalla osaamisen ja tietämyksen tasolla - kukaan ryhmän jäsenistä ei ollut merkittävästi toisia kokeneempi. Vapaasti organisoituva ryhmä olisi todennäköisesti ollut

tehokkaampi kuin perinteisen muodollisesti johdettu ryhmä, joskin vastuualueita jaettiinkin melko dynaamisesti. Projektijohtaja olisi välttämätön tilanteissa, joissa yksi ihminen omaa huomattavaa kokemusta tai jos projektin luonteesta johtuen työmäärä luonnostaan kasautuisi yhdelle henkilölle. Pieni ryhmä samantaustaisia henkilöitä sopii hyvin suhteellisen kapean alan teknisen asiantuntijaprojektin tekemiseen.

Vuorovaikutus asiakkaan kanssa: Asiakasorganisaation kanssa pidettiin tiiviisti yhteyttä. Projektiryhmä ja Keskon edustajat tapasivat projektin alussa säännöllisesti. Keskon edustajien panos projektityön alussa oli merkittävä, kun projektiryhmälle ei ollut vielä muodostunut selkeää visiota projektin toteuttamisesta tai tavoitteista. Projektin edetessä välituloksia ja suunnitelmia esiteltiin Keskon edustajille yhteisissä tapaamisissa ja heillekin tarkentui projektin edetessä minkälainen lopputuote tuottaisi heille eniten arvoa. Keskustelu käytiin kasvotusten ja kumpikin osapuoli koki keskustelun hengen olevan hyvin rakentava. Sähköpostia käytettiin vain käytännön järjestelyiden hoitamiseen.

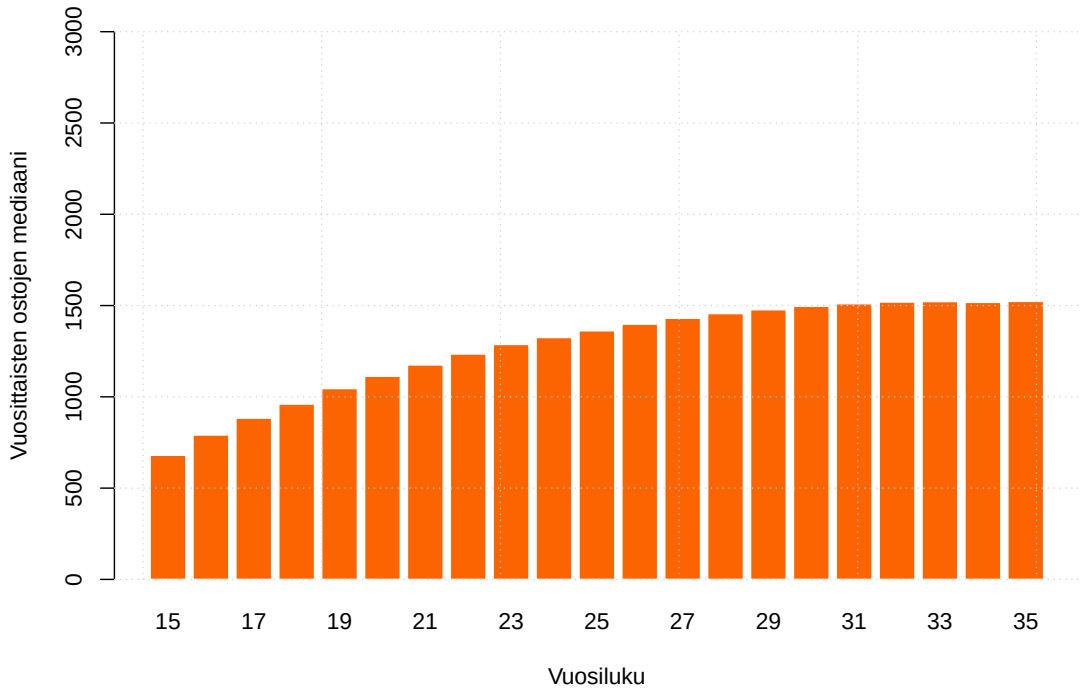
Yksi kiinnostava asiaseikka, jonka käsittely on jäänyt tässä raportissa vähemmälle, mutta käytännössä ryhmää askarrutti, on Keskon itsensä mallintavat muuttujien, kuten LANSEYN, tarkkuus. Nämä muuttujat perustuvat Keskon tekemään aikaisempaan mallinnukseen. Ilmeisesti perustuen pitkälti tuotekohtaisiin ostoihin iän ja muun ilmeisen ohella. Olisimme voineet hyötyä omassa projektissamme heidän aikaisemmin käyttämistä menetelmistään, mutta näistä ei puhuttu niin paljoa. Lähinnä heidän mallinnuksensa tekemät järjettömuudet kuten nopea siviilisäädyn vaihtelu tuottivat pään vaivaa. Projektissa olisi voinut olla mahdollisuus oppia heidän virheistää suoraan ja kehittää tuotoksia vielä nopeammalla tahdilla, näin tuottaen suurimman mahdollisen arvon asiakkaalle projektista. Toisaalta asioiden oppiminen kantapään kautta opetti varmaan hyödyllisempiä projektin hallintaan, tiedonhaukuun, todellisen maailman ei-oppikirjan mukaisen datan ja ilmiöiden käsittelyyn liittyviä taitoja juuri meille opiskelijoille.

Loppusanat: Projektin onnistumisen tärkein mittari on asiakkaan, eli Keskon asiakastyytyväisyys. Keskon edustajat olivat hyvin tyytyväisiä projektin tuotoksiin ja erityisesti siihen, että saivat kaksi mallia kun alunperin odottivat vain yhtä. Toivottavasti Kesko hyötyy mallien tarjoamasta ymmärryksestä yhtä paljon kuin me opimme projektin parissa.

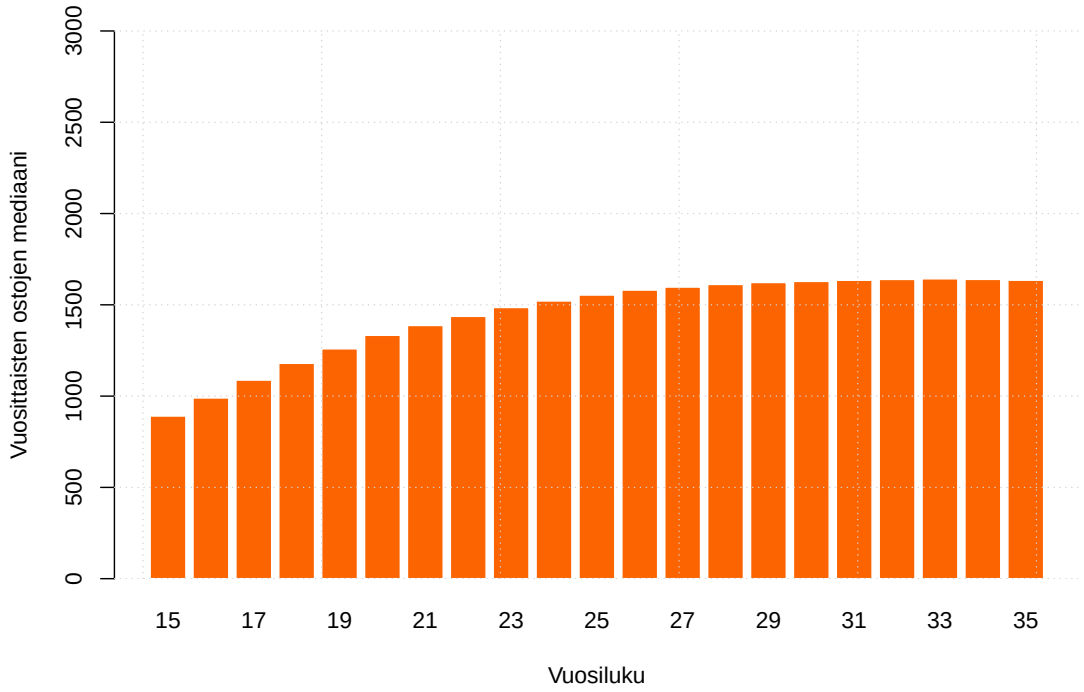
B Mallien ennustekuvaajia kokonaisostojen kehityksestä

Esitetään tässä liitteessä kuvaajia, joista käy ilmi ostojen ennustettu kehitys eri ihmisryhmille keskimäärin kahdenkymmenen vuoden päähän. CLV olisi summa ennustusvälin aikaisista ostoista. Päätöspuumalliin liittyvissä kuvaajissa vaaka-akselilla on kalenterivuosi ja riippumattomien keskiarvojen mallissa ikä. Esitetään ensin päätöspuumallin ennusteet mediaaneina tarkasteltavasta ryhmästä.

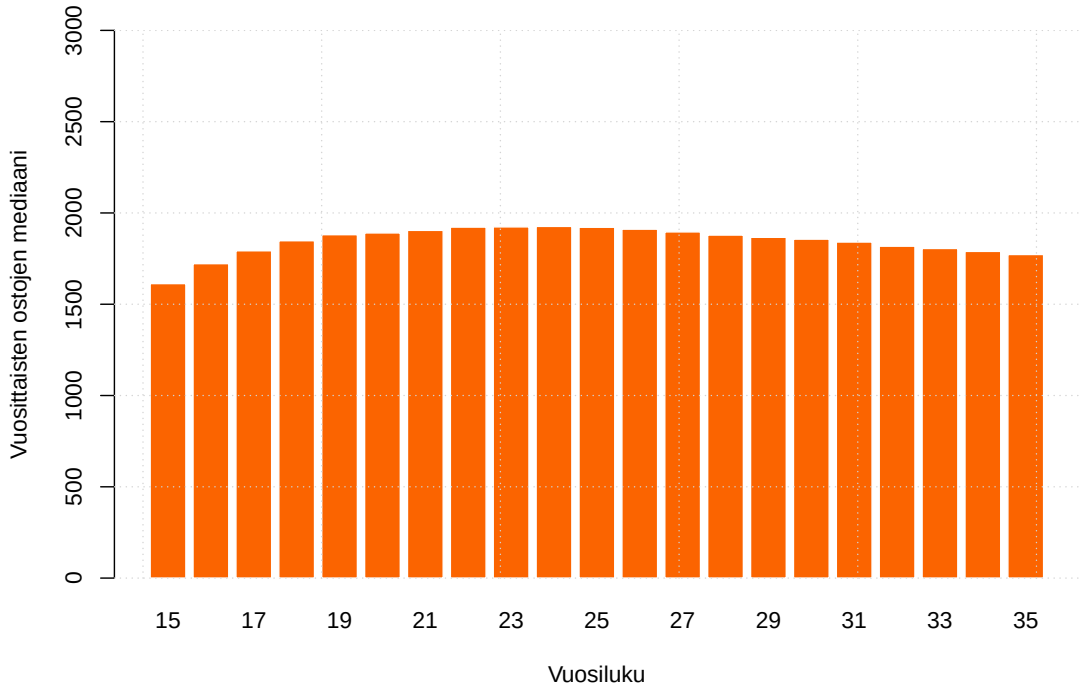
B.1 Päätöspuumallin ennusteet



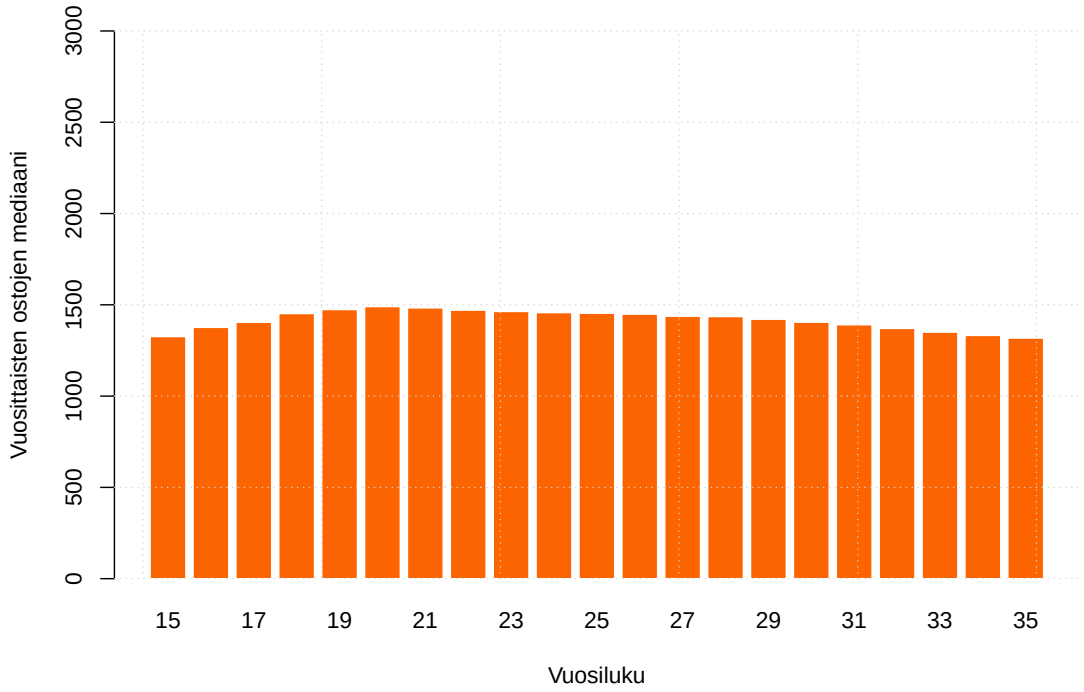
Kuva 18: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle kotitalouksille, joiden pääkorttilainen oli vuonna 2015 0-24 vuotias.



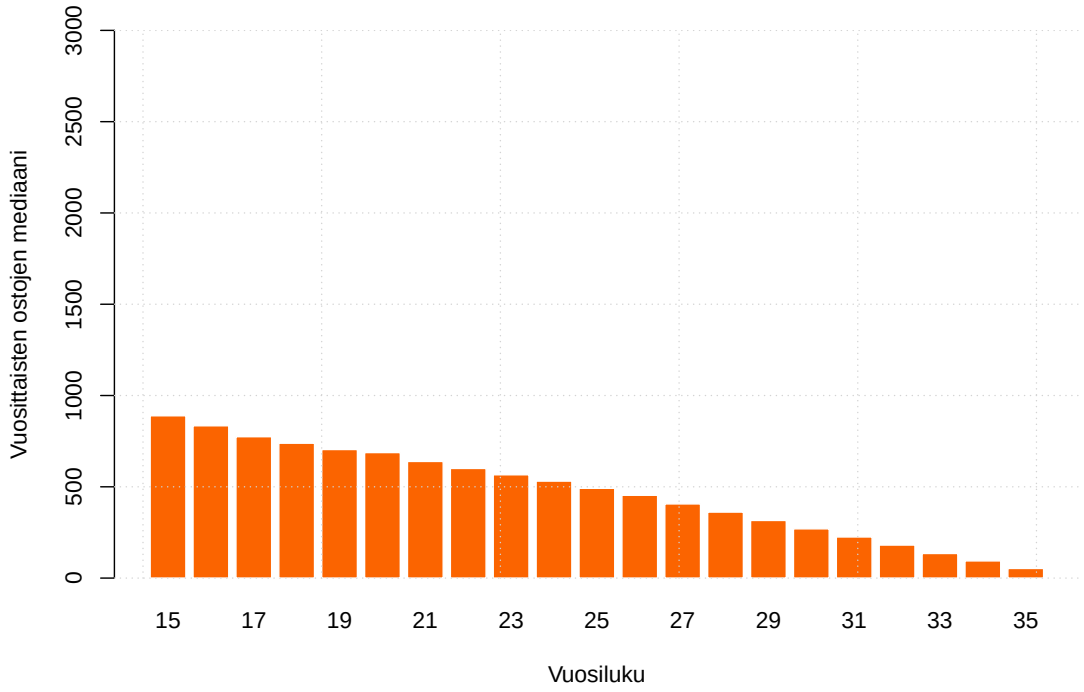
Kuva 19: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle kotitalouksille, joiden pääkorttilainen oli vuonna 2015 25-29 vuotias.



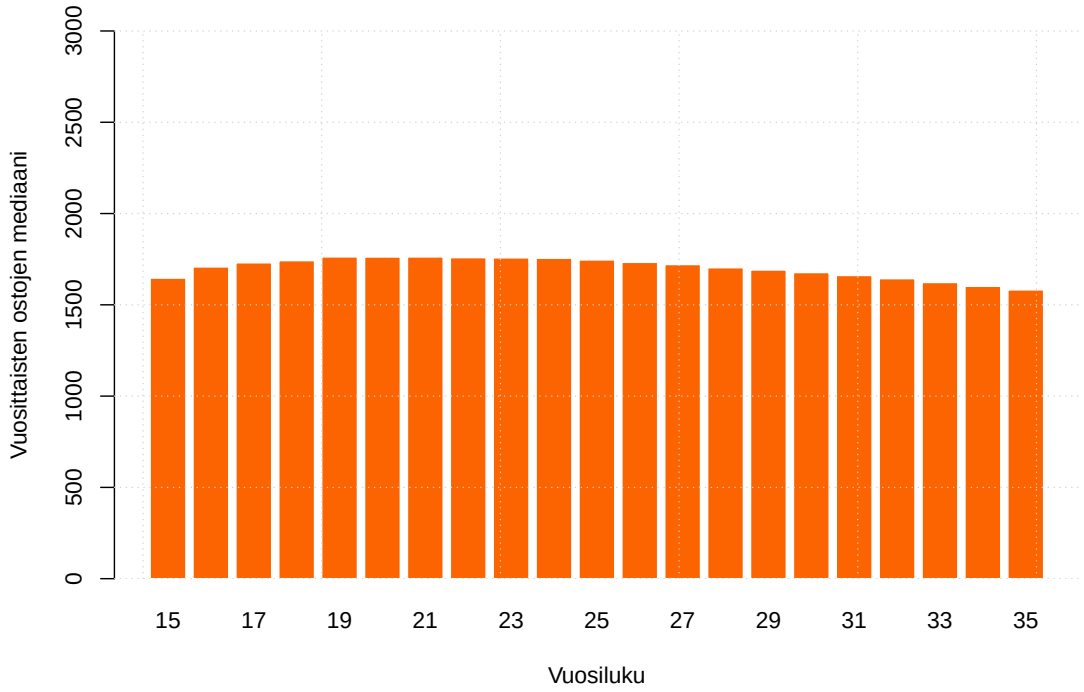
Kuva 20: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle kotitalouksille, joiden pääkorttilainen oli vuonna 2015 40-44 vuotias.



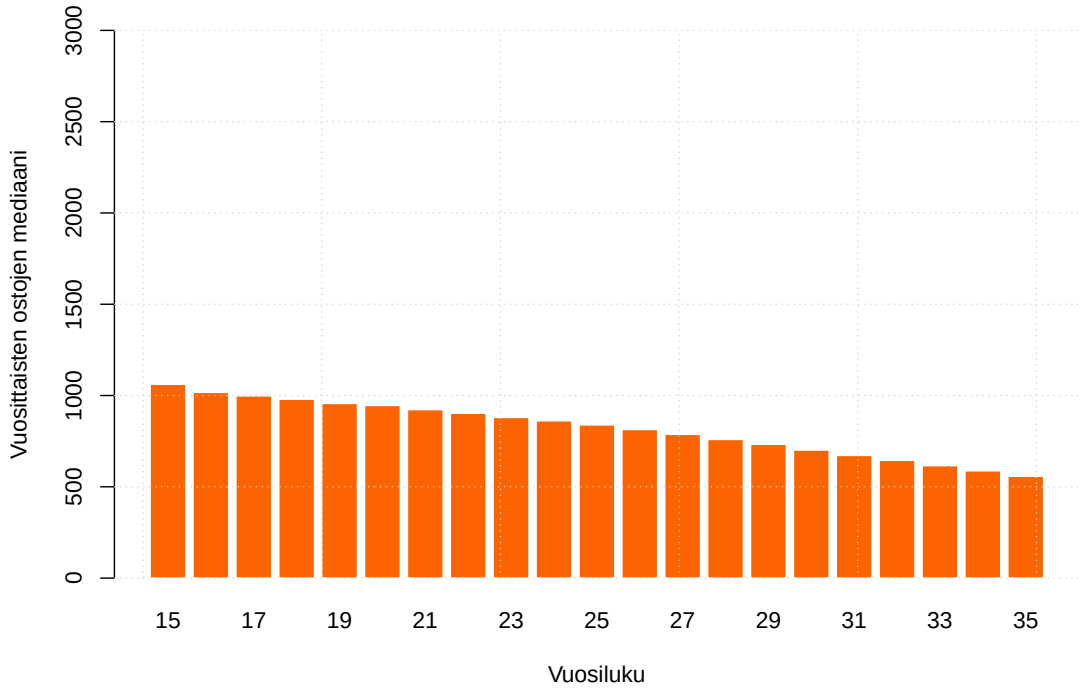
Kuva 21: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle kotitalouksille, joiden pääkorttilainen oli vuonna 2015 60-64 vuotias.



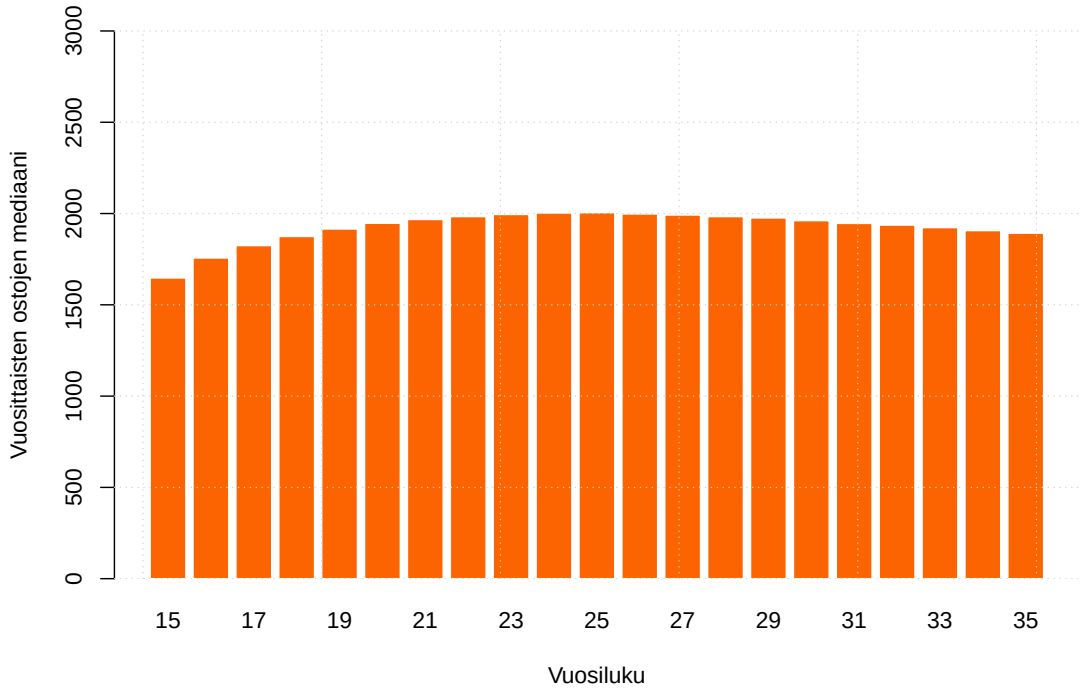
Kuva 22: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle kotitalouksille, joiden pääkorttilainen oli vuonna 2015 ainakin 75 vuotias.



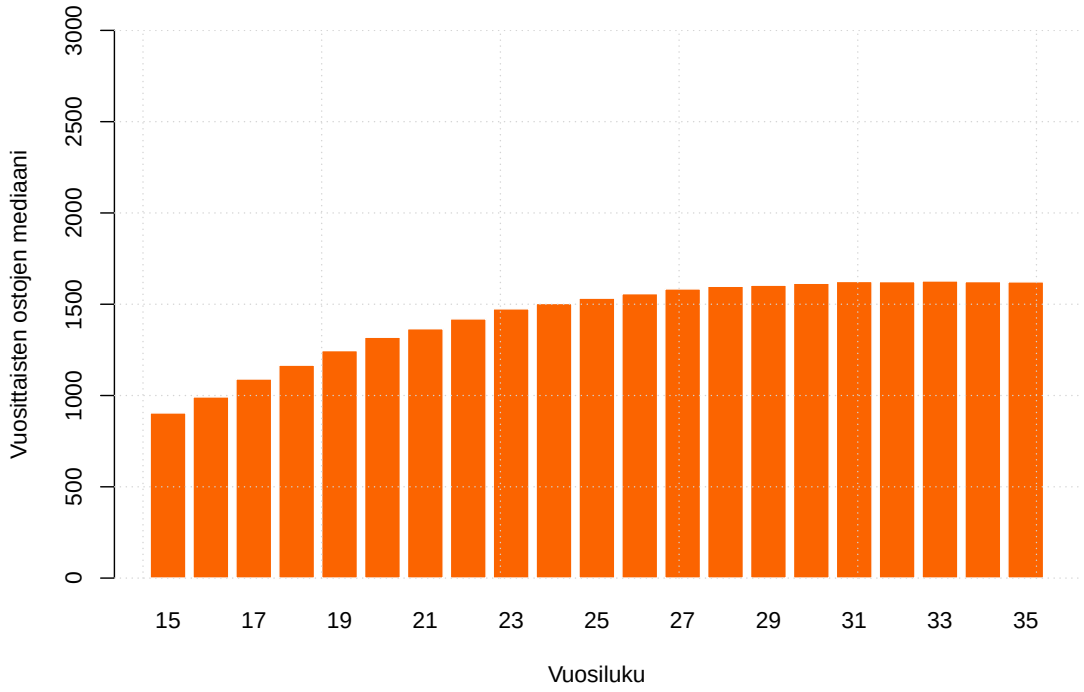
Kuva 23: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle vuoden 2015 aikuistalouksille.



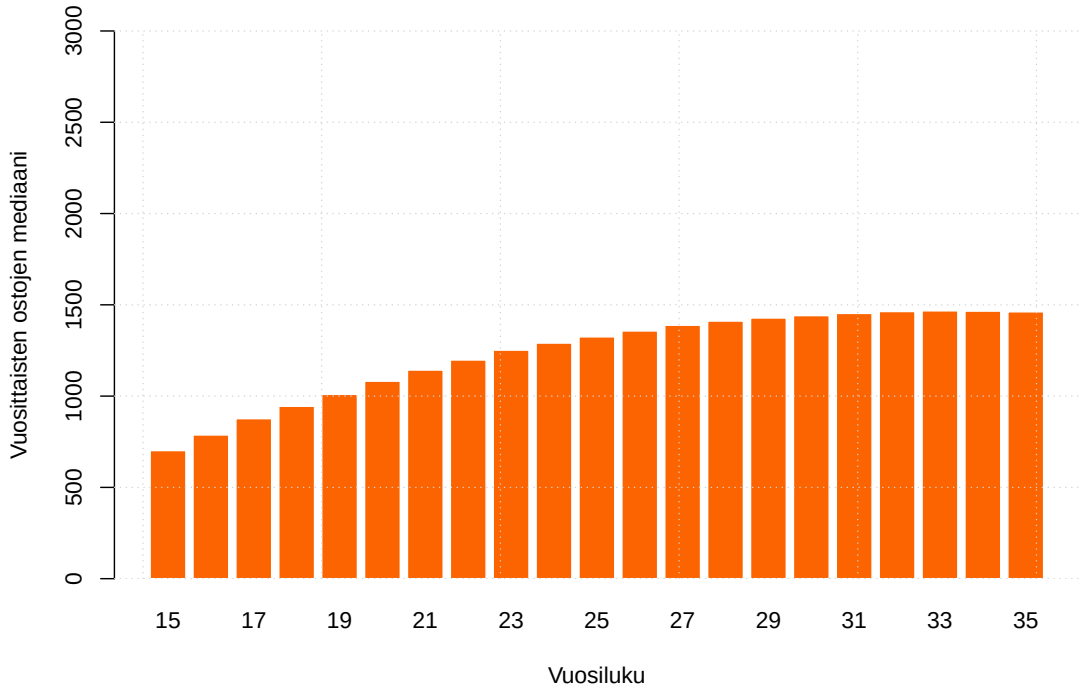
Kuva 24: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle vuoden 2015 eläkeläisille.



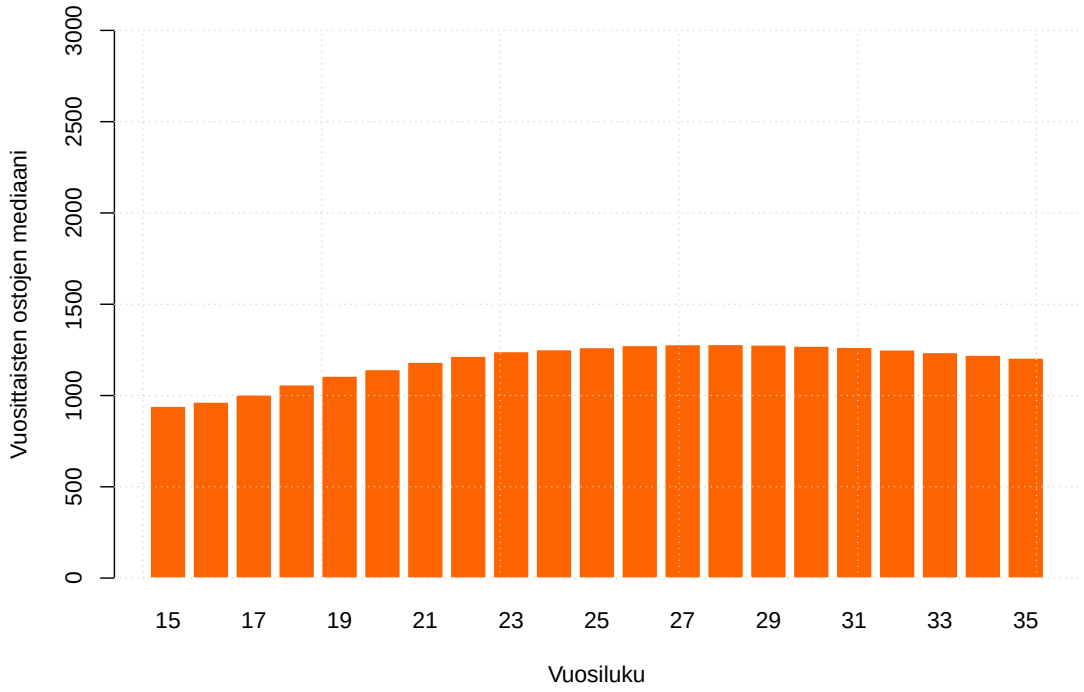
Kuva 25: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle vuoden 2015 lapsiperheille.



Kuva 26: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle vuoden 2015 nuorille pareille.



Kuva 27: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle vuoden 2015 sinkuille.

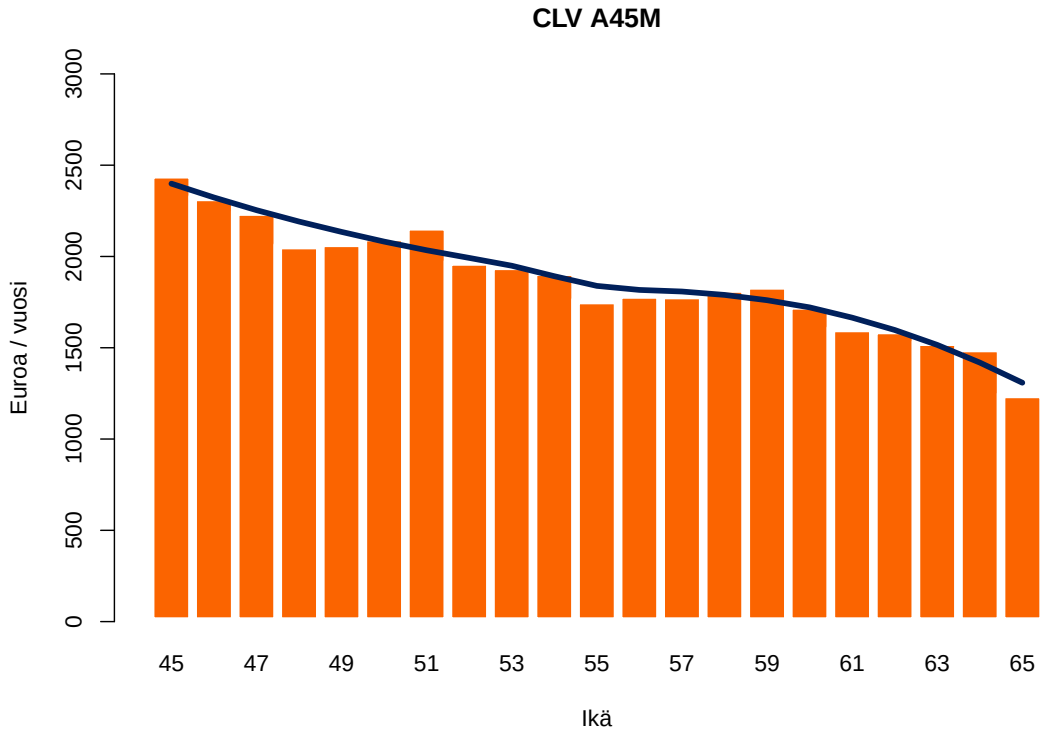


Kuva 28: Vuoden 2015 ostetun määrän mediaani sekä ennustetun määrän mediaani 10-vuoden ajalle vuoden 2015 yksinasuville.

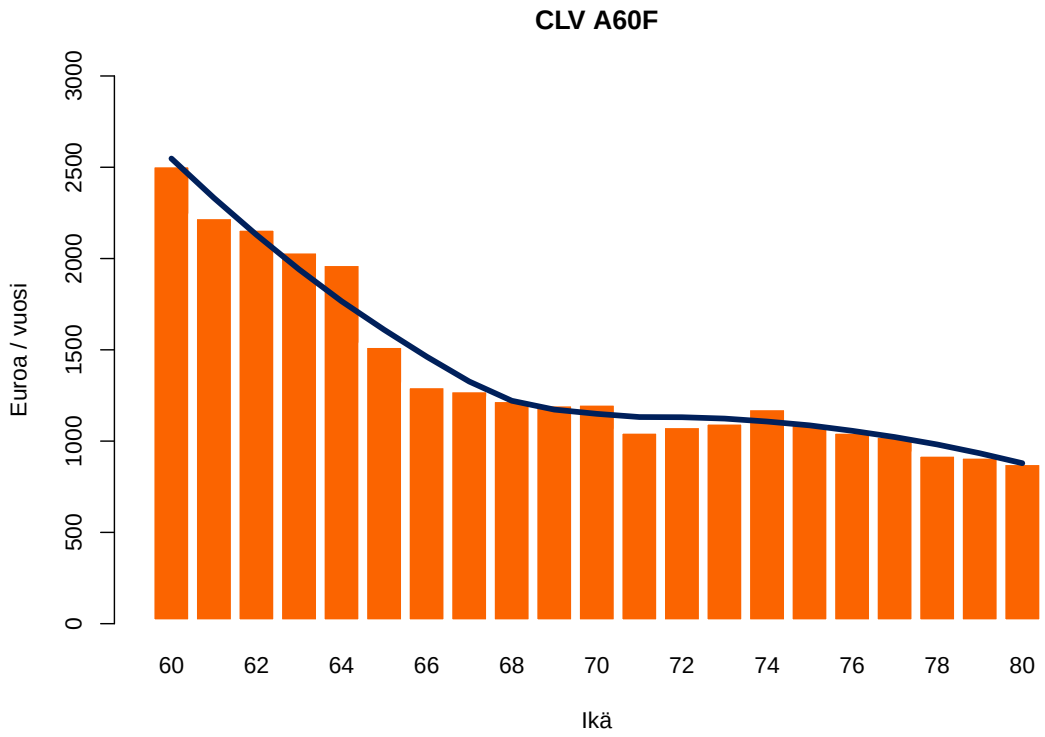
B.2 Riipumattomien keskiarvojen mallin ennusteet

Tässä on esimerkkikuvia siitä, kuinka riippumattomien keskiarvojen malli ennustaa ostojen käyttäytyvän ajan kuluessa. Vuosittaiset ostot ovat tässä yksikertaisesti luokan yhteinen ennuste. Jotta yleinen kehitys tulisi esiin mahdollisimman hyvin ja koska kehitys ei ole aivan yhtä tasaista kuin päätöspuumallissa, tässä on piirretty erillinen trendiviiva kuvaajiin satunnaisvaihtelua häivyttämään.

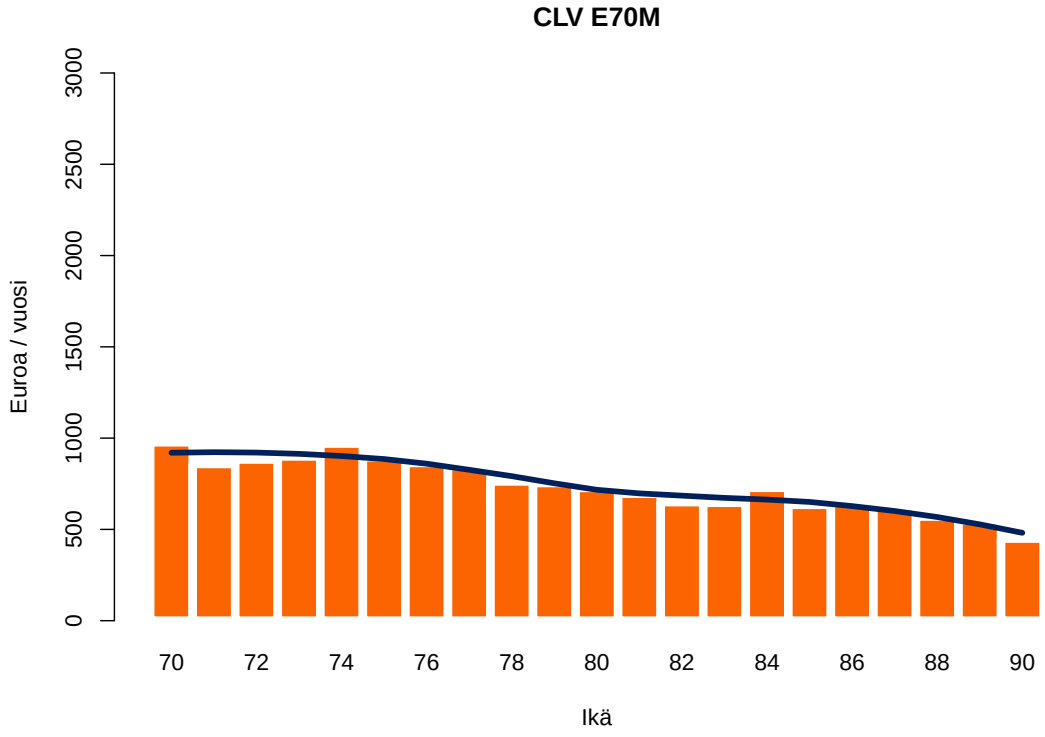
Merkintojen selitys: kuvan ylälaidassa tai kuvauksessa on kirjainlyhenteitä, jotka kuvaavat tarkasteltavan luokan. Kirjaimet L, A, N, S, E ja Y viittaavat luonnollisesti LANSEY-luokkiin, M miehiin, F naisiin ja numerot ikään.



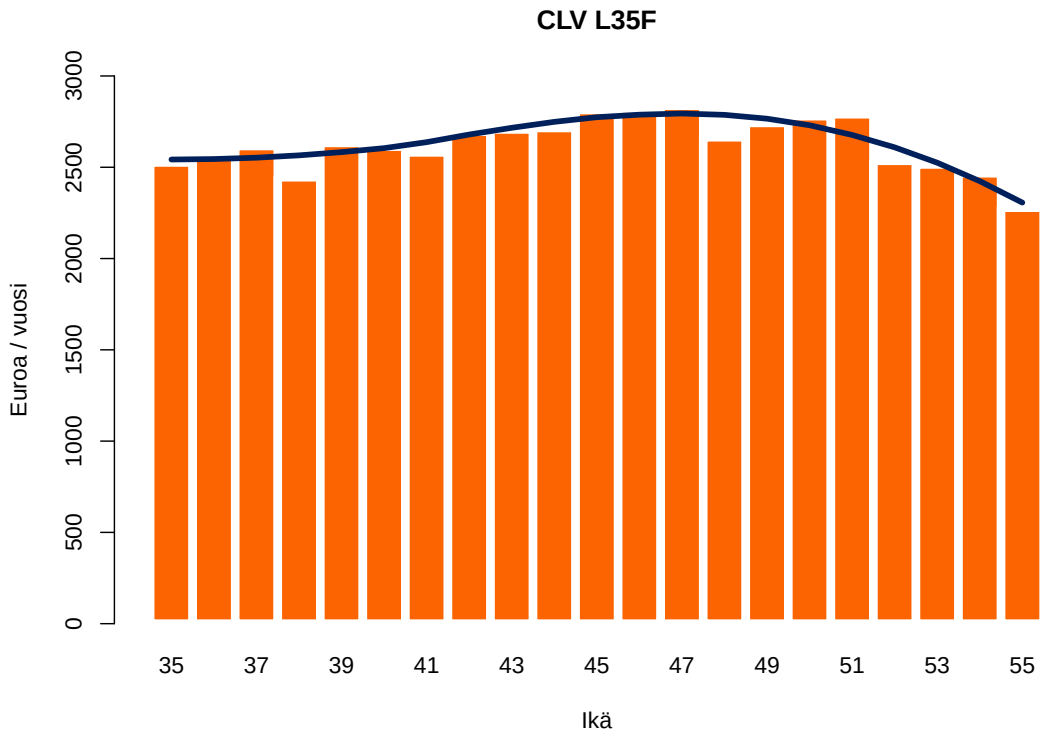
Kuva 29



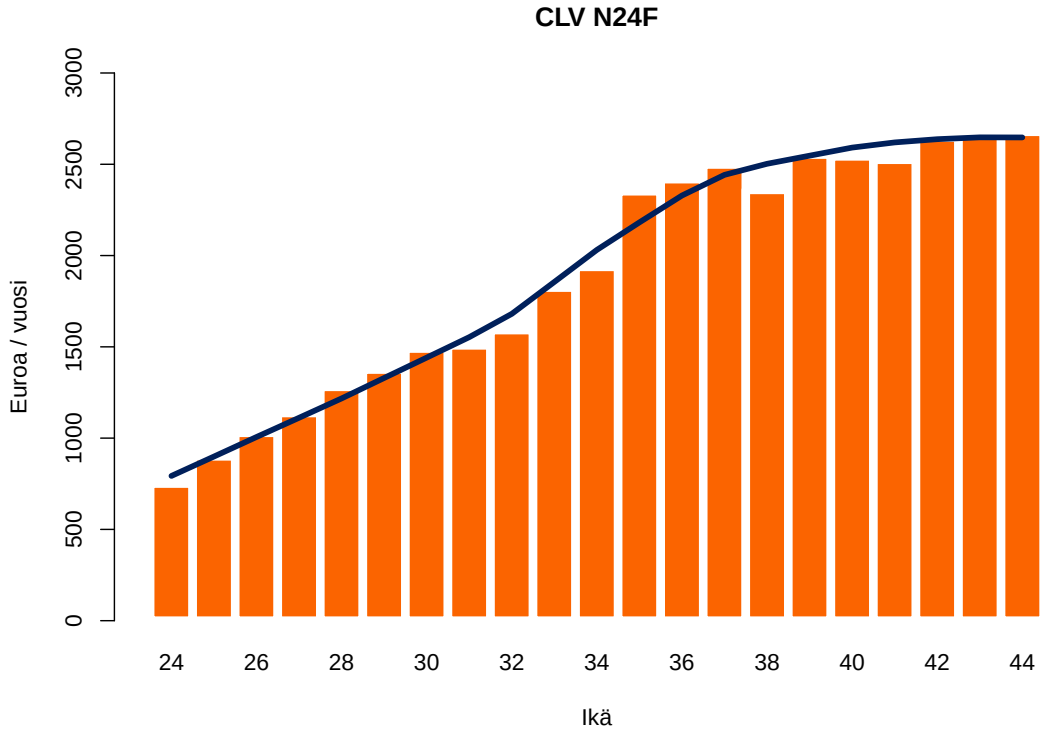
Kuva 30



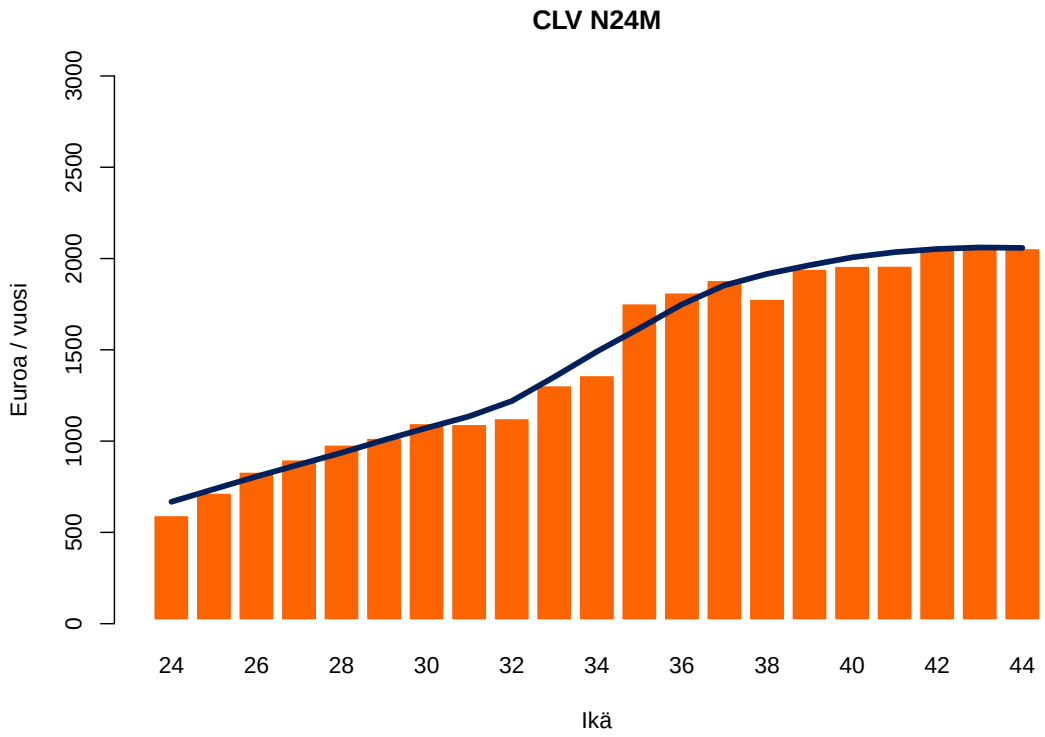
Kuva 31



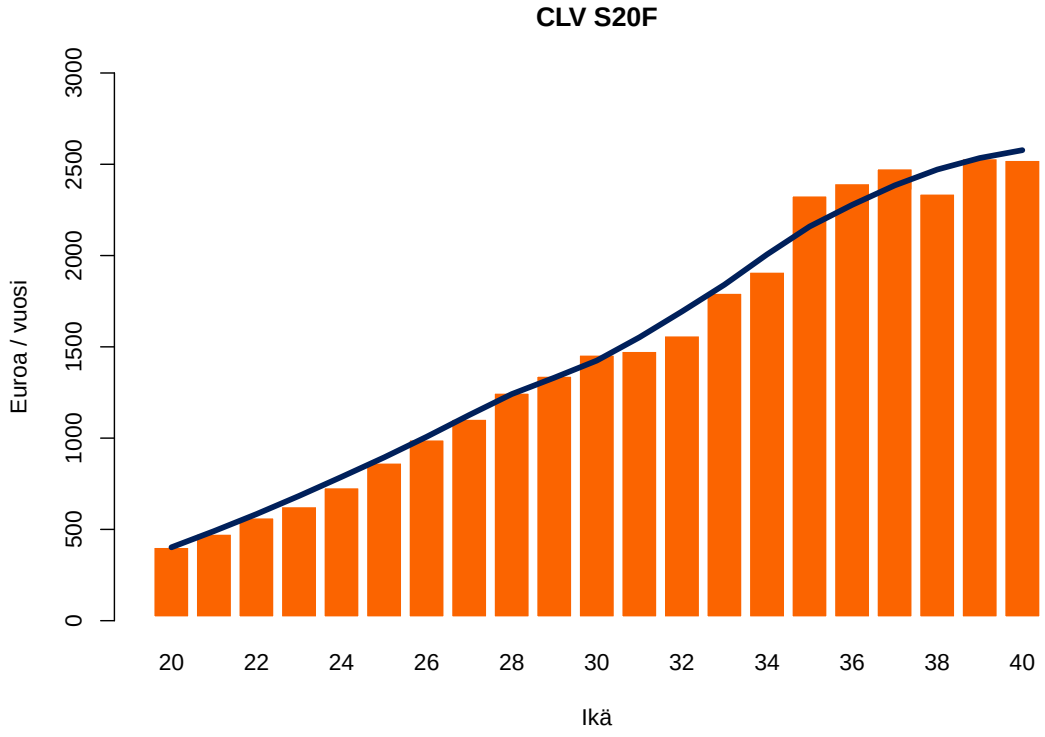
Kuva 32



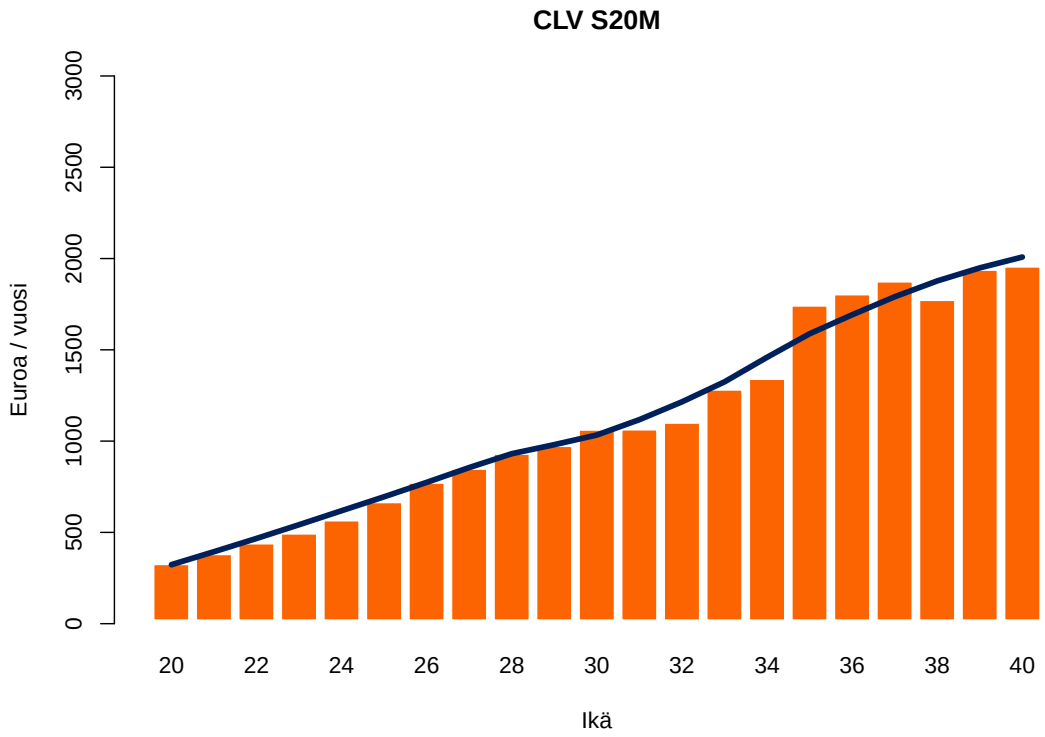
Kuva 33



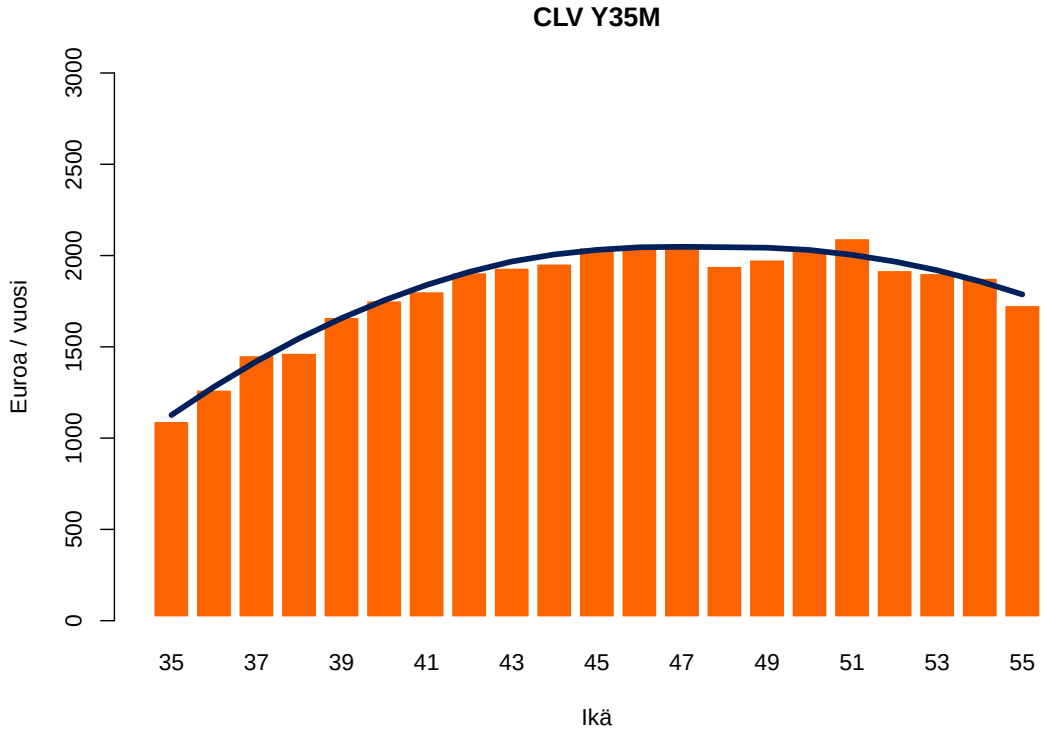
Kuva 34



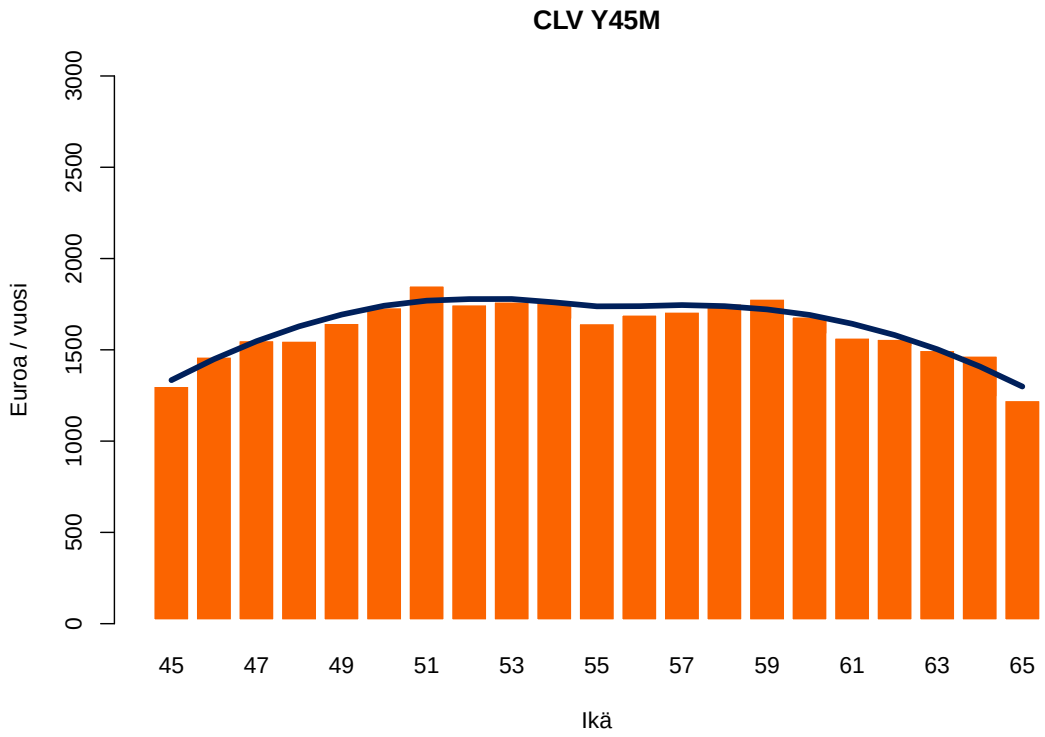
Kuva 35



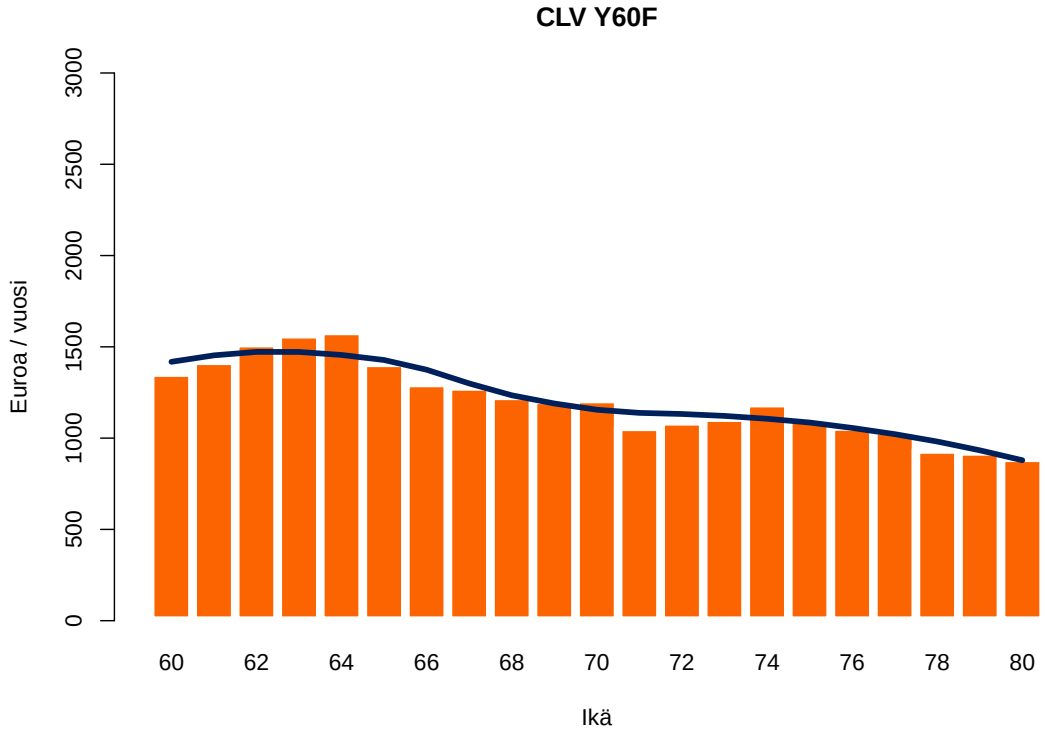
Kuva 36



Kuva 37



Kuva 38



Kuva 39

Viitteet

[1] P. Berger ja N. Nasr. *Customer Lifetime Value: Marketing Models and Applications*, Journal of Interactive Marketing 12.1 (1998): 17-30.

[2] C. Kaewchinporn, N. Vongsuchoto ja A. Srisawat. *A Combination of Decision Tree Learning and Clustering for Data Classification*, Computer Science and Software Engineering (JCSSE), 2011 Eighth International Joint Conference on. IEEE, 2011.

[3] J. C. Gower. *A General Coefficient of Similarity and Some of Its Properties*, Biometrics (1971): 857-871.

[4] Z. Huang. *Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values*, Data Mining and Knowledge Discovery 2.3 (1998): 283-304.

[5] J. Geweke ja C. Whiteman. *Bayesian Forecasting*, Department of Economics, University of Iowa, vuosi 2006. Saatavilla: http://www.ruxizhang.com/uploads/4/4/0/2/44023465/bayesian_forecasting.pdf

[6] K. A. A. Nazeer ja M. P. Sebastian. *Improving the Accuracy and Efficiency of the k-means Clustering Algorithm*, Proceedings of the World Congress on Engineering. Vol. 1. 2009.

[7] L. Leskelä. *Stokastiset prosessit*, luentomoniste, Aalto-yliopisto, vuosi 2015. Saatavilla: https://math.aalto.fi/~lleskela/papers/Leskela_2015-10-14_Stokastiset_prosessit.pdf

[8] A. Liawa ja M. Wiener. *Classification and Regression by randomForest*, R news 2.3 (2002): 18-22.

[9] J. van den Hoven. *Clustering with Optimised Weights for Gower's Metric*, tutkielma, University Amsterdam, vuosi 2016. Saatavilla: http://www.math.vu.nl/~sbhulai/papers/thesis-vandenhoven.pdf

[10] W. Shang, C. Chang ja Q. Li. *Customer Lifetime Value: A Review*, Social Behavior and Personality: An International Journal 40.7 (2012): 1057-1064.

[11] Y. Ekinci, F. Ülengin, N. Uray ja B. Ülengin. *Analysis of Customer Lifetime Value*

and Marketing Expenditure Decisions Through a Markovian-based Model, European Journal of Operational Research 237.1 (2014): 278-288.