26.5.2017

Aalto University
School of Science

MS-E2177 Operaatiotutkimuksen projektityöseminaari

Väliraportti:

Asiakasarvon määrittäminen päivittäistavarakaupassa

Projektiryhmä Joonas Laihanen (projektipäällikkö)
Aleksi Pasanen
Eero Rantala
Samuli Turunen

Aiheen asettaja Kesko

Valmiit tehtävät

Tilastollinen analyysi ja datan esikäsittely

Yrityksen toimittamaan dataan on tutustuttu sekä yksi- että moniulotteisella tilastollisella analyysilla. Datan muuttujien ominaisuuksista on saatu hyvä käsitys visualisoimalla yksittäisten muuttujien jakaumia mm. histogrammeilla ja muuttujien välisiä riippuvuuksia hajontakuvioilla.. Alkuperäisessä datassa talouksien tiedot ovat kuukausitasolla. Joidenkin muuttujien arvoissa on kuitenkin runsaasti kuukausivaihtelua, koska ne ovat valmiiksi mallinnettuja. Siksi katsottiin parhaaksi muuttaa tiedot vuositasolle ottamalla summa ostoista ja käyntikerroista sekä moodi muista muuttujista kultakin vuodelta.

Projektiryhmän tehtävänä on luoda malli, jossa otetaan talouksien tilojen muuttuminen huomioon usean vuoden aikana. Mallin yleisyyden ja laskennan rajoitteiden vuoksi samankaltaisia talouksia luokitellaan yhteen. Koska käytettävissä oleva data sisältää noin sadantuhannen Plussa-talouden ostot euroissa vuodelta 2015 ja 2016, siitä voidaan laskea todennäköisyyksiä siirtyä jonkin muuttujan vuoden 2015 arvosta johonkin toiseen vuonna 2016. Vaikka siirtymien estimoidut todennäköisyydet todettiin suuruusluokiltaan järkeviksi, niistä myös huomattiin, että datassa on siirtymiä, jotka eivät todellisuudessa olisi mahdollisia. Esimerkiksi jos yritys on luokitellut talouden vuonna 2015 eläkeläistaloudeksi, joka käsittää vain yli 65-vuotiaita pääkorttilaisia, ei ole mahdollista, että siitä tulisi vuonna 2016 aikuistalous, jossa kuuluisi olla vain 35-64 vuotiaita. Tällaisia siirtymiä ei siis oteta huomioon todennäköisyyksien estimoinnissa. Lisäksi joidenkin talouksien kohdalla osa tiedoista, kuten ikä, puuttuvat. Tällaiset taloudet siivotaankin datasta pois ennen mallin laskemista. Datassa on myös mukana harvinaisia arvoja, kuten esimerkiksi määrittelemätön sukupuoli. Nämä ovat yrityksen mukaan riittävän harvinaisia, ettei niitä tarvitse ottaa malliin mukaan. Talouden osoitteen postinumerodata on korvattavissa yrityksen toimittamalla uudella datalla, jossa postinumero on muunnettu todennäköisyydeksi asioida K-kaupassa. Uuden datan uskotaan toimivan paremmin selittävänä tekijänä.

Laskennan implementointi

Laskennan implementointi on toteutettu eli annetun luokittelun perusteella voidaan määrittää Markov-ketju ja laskea CLV (Customer Lifetime Value, suom. asiakasarvo) halutun aikahorisontin yli. Laskenta siis toimii, mutta projektisuunnitelman mukaisesti tehtävälle on allokoitu paljon aikaa. Näin varaudutaan laskennan muokkaamiseen ongelmien ilmaantuessa.

Klusteroinnin tuottamia tai muulla tavalla valittuja kategorioita käytetään Markov-mallissa ketjun tila-avaruutena. Mallissa oletetaan, että todennäköisyys siirtyä tilasta toiseen riippuu ainoastaan nykytilasta. Todennäköisyydet estimoidaan käyttämällä osaa, esimerkiksi 50 %, datasta. Todennäköisyyksien avulla voidaan laskea esiintyvyysmatriisi

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

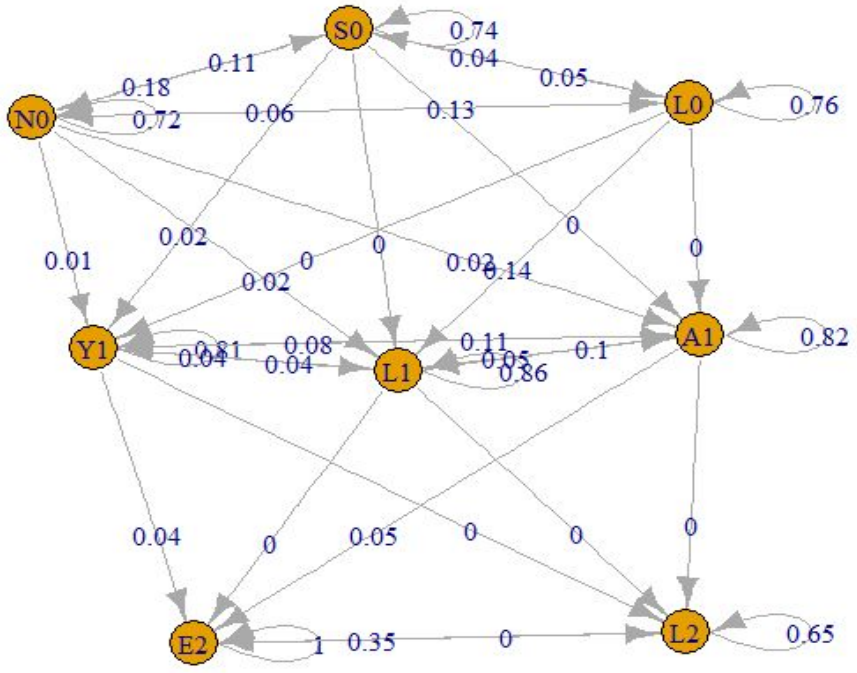
missä P on todennäköisyysmatriisi ja t on vuosien määrä, jonka ajalle CLV halutaan laskea. Esiintyvyysmatriisia käyttämällä saadaan odotettu kertymä ostetulle määrälle eli CLV

$$g_t = M_t c \ ,$$

missä c on pystyvektori, joka sisältää jokaisen eri tilan vuosittaisten ostojen keskiarvon. [1]

Luokittelun implementointi: Baseline-malli

Ensimmäisenä luokittelun ja laskennan yhdistävänä mallina luotiin Markov-malli ilman klusterointialgoritmia. Malliin otettiin mukaan muuttujat LANSEY (eli kuusi eri demograafista ryhmää, esim. L=lapsiperheet) sekä ikä, joka jaettiin kolmeen ryhmään: alle 35-vuotiaat, 35-64-vuotiaat ja yli 65 vuotiaat. Nämä muuttujat muodostivat yhteensä kahdeksan mahdollista tilaa, koska lapsiperheluokkaa lukuunottamatta LANSEY:n muissa luokissa on yksilöiden ikä määritelty johonkin edellä kuvatuista kolmesta ikähaarukasta. Siirtymäkaavio alustavan mallin siirtymistä ja niiden välisistä siirtymätodennäköisyyksistä näkyy kuvassa 1 ja eri tiloille lasketut CLV:t ovat taulukossa 1. Taulukossa ikäluokat 0,1 ja 2 vastaavat alle 35-, 35-64- sekä yli 65-vuotiaita.



Kuva 1: Ensimmäisen mallin siirtymäkaavio.

Taulukko 1: Ensimmäisen mallin eri tilojen CLV:t viiden vuoden ajalta.

Tilan numero	LANSEY	Ikäluokka	CLV (€)
1	A	1	11279.251
2	E	2	8135.232
3	L	0	10025.967
4	L	1	12626.224
5	L	2	9682.633

6	N	0	7756.154
7	S	0	6510.873
8	Y	1	8532.251

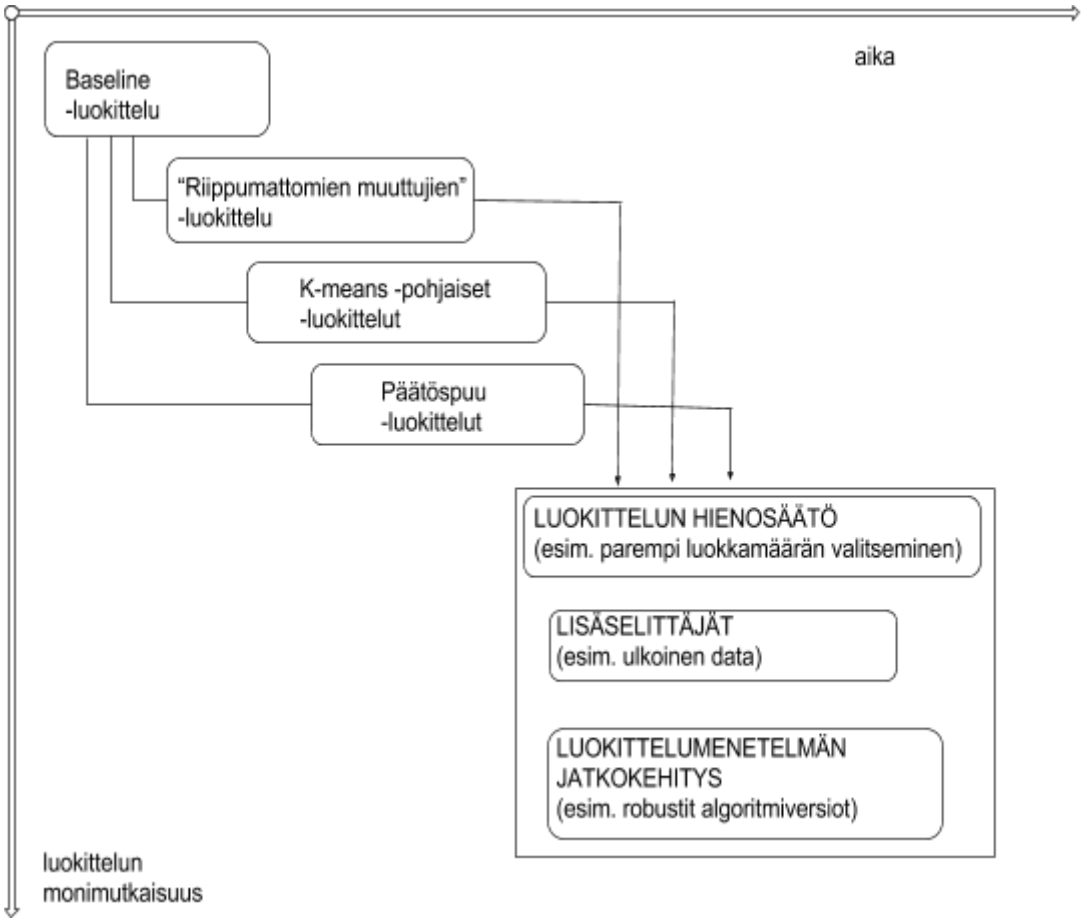
Vastaava malli tehtiin myös käyttämällä em. muuttujien lisäksi sekä sukupuolta että niin kutsuttua Share of Walletia, joka yrityksen mallinnuksen mukaan kertoo, kuinka suuren osan päivittäistavarakulutuksestaan talous käyttää Keskon ketjuissa. Näiden kahden muuttujan lisääminen malliin lisäsi tilojen määrän 1540:aan. Baseline-mallia tuskin käytetään varsinaisten tulosten laskemiseen, mutta se osoittaa Markov-ketjun laskennan toimivuuden ja laskentaresurssien riittävyyden ainakin yksinkertaisilla malleilla.

Mallissa CLV lasketaan kotitalouksille, jotka ovat mukana datassa. Mallin ennusteisiin ei siis synny mukaan uusia kotitalouksia. Kuolemat ja poistuvat asiakkaat oletetaan ilmenevän mallissa klustereiden keskiarvoissa; esimerkiksi vanhojen eläkeläisten keskimääräisiä ostoja alentavana datassa piilevänä tekijänä.

Keskeneräiset tehtävät ja seuraavat askeleet

Edistyneempien luokittelujen kehitys ja kirjallisuuskatsaus

Alustavat kokeilut antavat viitteitä siitä, että muisti- ja laskentakapasiteetti näyttävät vaativan datan esiklusterointia, jotta algoritmeja voidaan käyttää projektin laskentaresursseilla. Laskentaresurssien rajoitteita voidaan kiertää ottamalla suuresta datamäärästä pienempi satunnaisotos. Yksi lähestymistapa on olettaa, että datassa on riippumattomuuksia tai riippuvuuksia. Esimerkiksi eri muuttujien riippumattomuusoletus mahdollistaa luokittelun erikseen kunkin tai vain osan muuttujien suhteen. Muuttujien riippuvuuksia on tutkittu tilastollisessa analyysissä. Kehitettävistä luokitteluista K-means- ja päätöspuupohjaiset menetelmät eivät oleta muuttujien riippumattomuutta, joten on mahdollista, että niiden käyttäminen koko dataa käyttäen vaatii edellä mainittua esiklusterointia tai että koko dataa ei käytetä . Kuvassa 2 kuvataan edellä esitelty toteutettu baseline-malli ja suuntaviivat luokittelun kehitykselle. Eritellään kuvan luokittelumallit tarkemmin seuraavaksi.



Kuva 2: Luokittelualgoritmivaihtoehdot ja projektin eteneminen. Kuvasta näkyy, kuinka ajan kuluessa vaakasuunnassa ja luokittelun monimutkaistuessa pystysuunnassa siirrytään vaiheesta toiseen. Luokittelun monimutkaisuus ei ole välttämättä sama asia kuin ennusteen tarkkuus. Lähtökohtana on jo toteutettu baseline-malli, josta siirrytään osittain rinnakkain toteuttamaan kolmea eri luokitteluvaihtoehtoa. Tämän jälkeen keskitytään parhaaksi havaitun menetelmän parantamiseen muun muassa itse algoritmia kehittämällä tai laajemman (ulkoisen) datan käyttöönotolla ajan salliessa.

K-means -pohjaiset menetelmät

Edellä kuvattu yksinkertainen baseline-luokittelu on hyvä lähtökohta luokittelulle, mutta parempien tuloksien saamiseksi kehitetään algoritminen luokittelutapa. Eräs tunnetuimmista ja yksinkertaisimmista algoritmeista on k-means klusterointi, joka jakaa datan yhteensä k:hon eri luokkaan, jossa k on parametrina annettu luokkien määrä. Yleisesti luokitteluvirhe pienenee luokkien määrän kasvaessa. CLV:n ennustamisessa pienempi määrä luokkia voi kuitenkin olla tehokkaampi, jos se aiheuttaa vähemmän virheellisiä siirtymiä Markov-ketjussa ja on laskettavissa nopeammin. Lisäksi laskenta ja tulosten tulkitseminen on voi olla ongelmallista liian monella luokalla. K-means -pohjaisten algoritmien luokkien määrä valitaan validoimalla eri k:n arvot.

Perinteistä k-means -algoritmia ei voi soveltaa suoraan projektin dataan, sillä se on sekoitus jatkuvia ja kategorisia muuttujia. Tähän ongelmaan on löydetty kaksi lähestymistapaa.

Ensimmäinen on muokata k-means -algoritmin etäisyysfunktiota niin, että se antaa järkevän arvon myös kategorisille muuttujille. Muun muassa Gower-etäisyys [4] soveltuu tähän, kun huomioidaan myös, että R:ssä olevat valmisfunktiot mahdollistavat muuttujien halutun painotuksen. Jos valmistoteutus on puutteellinen, voidaan parempi etäisyysfunktio toteuttaa itse. Näiden lisäksi voidaan tutkia, parantaako robustien k-means-variaatioiden käyttö tuloksia.

Toinen tapa on käyttää ennestään kehitettyä k-prototypes -menetelmää [2]. Tämän algoritmin idea on samankaltainen kuin edellä kuvattu eli etäisyysfunktio määritellään sekatyyppiselle datalle sopivaksi ja sovelletaan perinteistä k-means -algoritmia. Sen on todettu toimivan hyvin suurillekin datamäärille ja sille on valmistoteutuksia.

Päätöspuut

K-means -pohjaisten mallien lisäksi kehitetään päätöspuihin perustuva malli, jossa jälkimmäisen vuoden ostoja selitetään ensimmäisen vuoden perusteella. Sovitetusta päätöspuusta voidaan lukea luokittelu, joka ottaa huomioon tilojen vaihtelun vuosien välillä, eli sen odotetaan toimivan tilojen dynamiikan huomioivana luokittimena.

Päätöspuut ovat helposti tulkittavia ja laajasti käytettyjä malleja. Niiden parametrien sovitukseen on useita algoritmeja, jotka antavat mahdollisuuden vaikuttaa esimerkiksi puun pituuteen. Yhden päätöspuun tarkkuutta voidaan parantaa yhdistämällä useita päätöspuita, esimerkiksi Random Forest -menetelmällä [3].

Riippumattomien muuttujien malli

Tilastollisessa perusanalyysissä havaittiin, että eri muuttujien suhteen ostojen keskiarvoilla on merkittäviä eroja. Esimerkiksi lapsiperheet ostavat enemmän kuin sinkut, 40-vuotiaat ostavat enemmän kuin 20- tai 70-vuotiaat, naiset ostavat keskimäärin hieman enemmän kuin miehet ja eri postinumeroiden välillä on myös merkittävää vaihtelua mediaaniostojen suhteen.

Kehitetään malli, joka perustuu yksinkertaiseen oletukseen, että eri kategorian muuttujien välillä ei ole merkittävää riippuvuutta, eli toisin sanoen oletetaan, että naiset ostavat kautta linjan aina hieman enemmän kuin miehet, 40-vuotiaat ostavat aina enemmän kuin 20-vuotiaat riippumatta asuinalueesta tai sukupuolesta. Jos nämä olettamukset pitävät paikkansa, voidaan jokaisen kategorian muuttujalle laskea keskiarvojen suhdeluku. Esimerkiksi osajoukon kaikkien naisten ostojen keskiarvo on 174 ja koko osajoukon ostojen keskiarvo 167, jolloin suhdeluku 174/167 on noin 1,04. Näin saadaan jokaisen kategorian jokaiselle muuttujalle suhdeluku, joka kertoo missä suhteessa kyseinen muuttuja kasvattaa tai pienentää keskiostoa. Kun kaikki suhdeluvut kerrotaan yhteen niin saadaan arvio asiakasryhmän keskiostosta.

Mallin vahvuuksia ovat sen yksinkertaisuus ja se, ettei esimerkiksi postinumeroita tarvitse mitenkään ryhmitellä. Mallin toimivuus saadaan selville kokeilemalla ja tutkimalla mallin tuottamia residuaaleja.

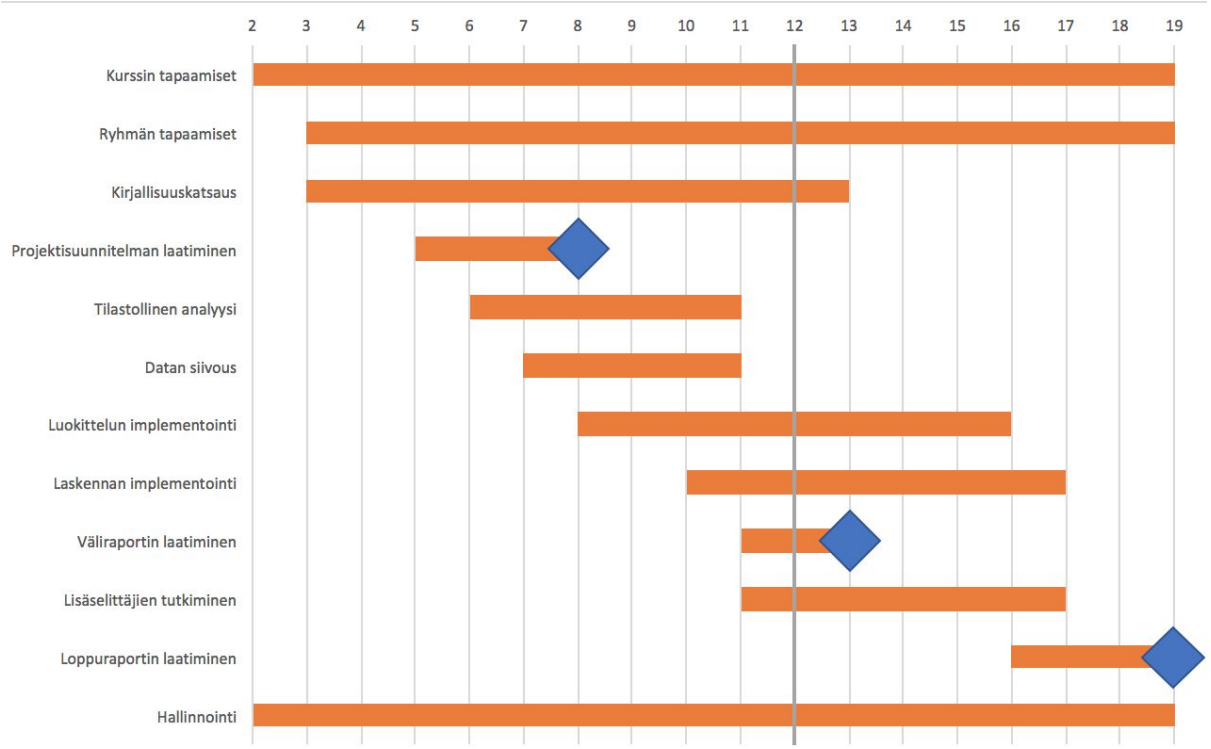
Mallien vertailu

Erityyppisiä malleja vertaillaan validoimalla. Validointi tehdään soveltamalla malleja vuoden ensimmäisen vuoden validointidatalle, jota ei käytetä mallin rakentamiseen, ja ennustamalla seuraavan vuoden ostoja. Talouksille ennustetuista ostoista voidaan laskea virhe esimerkiksi keskimääräisenä virheen neliönä. Validointi voidaan tehdä vain yhdelle aika-askeleelle, koska dataa on vain kahdelta vuodelta. Tilojen kehittymisen mallintaminen Markov-ketjuilla sisältää oletuksen siitä, että tilasiirtymät ovat muistittomia, eli validointia yhdelle aika-askeleelle voidaan pitää sopivana tapana validoida myös useampi aika-askel. Kehitettävät mallit koostuvat kahdesta osasta: luokittelu ja CLV:n ennustaminen. Luokittelu on mallin ainoa “opittava” eli optimoitava osa. Huomattavaa on, että validoimalla luokittelu erillisenä osana, tuloksena saatu luokitin ei välttämättä ole paras valinta koko mallille, sillä se saattaa vääristää Markov-ketjun siirtymiä jotain toista luokitinta enemmän. Näin ollen luokittelun valinnassa on käytettävä koko mallin tuloksena saatavaa CLV:tä ja siitä laskettua virhettä.

Muutokset projektisuunnitelmaan

Aikataulu

Projektin aikataulua on muokattu hieman. Tilastolliselle analyysille on annettiin 2 viikkoa lisäaikaa ja datan siivoukselle viikko, sillä tehtävät vaativat datan määrästä ja kompleksisuudesta johtuen odotettua enemmän keskustelua asiakasyrityksen kanssa. Toisaalta laskennan implementointi on nyt toimiva, eli tehtävä on periaatteessa valmis. Sille on kuitenkin allokoitu aikaa projektin myöhempiinkin vaiheisiin, jos esimerkiksi uudet luokittelut tuottavat odottamattomia ongelmia. Kirjallisuuskatsaukseen voidaan joutua palaamaan vielä viikon 13 jälkeen, mutta siihen perustuva suunnitelma seuraavien mallien kehityksestä vaikuttaa riittävän kattavalta ja perustellulta. Kuvassa 3 on esitetty vastaavilla muutoksilla päivitetty Gantt-kuvaaja viikon 12 lopun tilanteesta.



Kuva 3:. Päivitetty projektin aikataulu.

Riskit

Projektisuunnitelmassa esitetty riskien erittely arvioidaan edelleen asianmukaiseksi, eikä merkittäviä päivityksiä nähdä tarpeellisiksi. Matalaksi arvioitu riski laskennan implementointiin liittyen näyttää tällä hetkellä erittäin epätodennäköiseltä, sillä Markov-ketjun ja CLV:n laskenta toimii yksinkertaisella mallilla suurellakin tilojen määrällä. Tapaamiset asiakasyrityksen edustajien kanssa ovat vahvistaneet käsitystä siitä, että heillä on hyvä käsitys projektin tavoitteista, laajuudesta ja teknisestä toteutuksesta. Asiakkaan kanssa kommunikointiin liittyvä riski on kuitenkin edelleen olemassa, sillä projektin edetessä käsiteltävät asiat ovat yhä uudempia myös asiakkaalle.

Lähteet

[1] Leskelä, L. Stokastiset prosessit, 2015, s.32-34. Verkkodokumentti. Saatavissa: http://math.aalto.fi/%7Elleskela/papers/Leskela_2015-10-14_Stokastiset_prosessit.pdf.

[2] Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, 1998 Verkkodokumentti. Saatavissa: http://arbor.ee.ntu.edu.tw/~chyun/dmpaper/huanet98.pdf

[3] Liaw, A. ja Wiener, M. Classification and Regression by randomForest, 2002. Verkkodokumentti. Saatavissa: http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf

[4] R Documentation, Dissimilarity Matrix Calculation. Internet-sivusto. Saatavissa: https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/daisy.html