Klusterointi hierarkkisilla ja kombinatorisilla menetelmillä - sovelluksena tilastomenetelmien peruskurssiaineisto

Teemu Holopainen

23. maaliskuuta 2012

Tilastotieteen pro gradu -tutkielma

Jyväskylän yliopisto

Matematiikan ja tilastotieteen laitos

Tiivistelmä

Tiedonlouhinta on kehittyvä tieteenala, joka pyrkii helpottamaan tietoa sisältävän aineiston käsittelyä ja antamaan välineitä sen rakenteen ymmärtämiseen. Koska yleisesti saatavilla olevan tiedon määrä on nykyaikana valtava, on tärkeää osata käsitellä suuria tietomääriä tehokkaasti. Yksi tapa helpottaa tiedon käsittelyä on jakaa aineisto osiin eli ryhmiin. Tässä työssä käsitellään klusterointia eli ryhmittelyä, joka on yksi keskeisimmistä tiedonlouhinnan menetelmäkokonaisuuksista.

Tutkielmassa esitellään yleisimpiä klusterointimenetelmiä, joita ovat kombinatoriset menetelmät sekä hierarkkiset menetelmät. Käsiteltäville klusterointimenetelmille on yhteistä se, että ne tarvitsevat hyvin vähän oletuksia ja ennakkotietoa aineistosta. Klusterointia kutsutaankin ohjaamattomaksi oppimismenetelmäksi, jossa aineiston rakennetta tutkitaan ilman ulkoista tietoa.

Kombinatoriset menetelmät liittävät jokaisen aineiston havaintopisteen suoraan yhteen klusteriin ilman aineistoa kuvaavaa todennäköisyysmallia. Klusterointitehtävä voidaan kuvata määriteltävän tavoitefunktion optimointiongelmana, kun havaintojen väliset etäisyydet tunnetaan. Tunnetuin kombinatorinen klusterointimenetelmä on K-means, joka perustuu klusterien keskiarvojen laskentaan. Tässä työssä keskeisempi menetelmä on K-medoids, joka on robusti versio K-means-menetelmästä.

Hierarkkisten menetelmien tapauksessa klusterointi tehdään askel kerrallaan yhdistämällä aina kaikista samankaltaisimmat havaintopisteet toisiinsa, jolloin kokonaisuudeksi saadaan hierarkkinen rakenne. Se voidaan esittää puumaisena kuviona, mikä on selkeä etu kombinatorisiin menetelmiin verrattuna. Hierarkkiset menetelmät jaetaan yhdisteleviin ja jakaviin menetelmiin sen mukaan miten tasolta toiselle edetään.

Itse menetelmien suorittaminen on varsin yksinkertaista, ja se onnistuu tietokoneohjelmilta hetkessä. Tulosten tulkinta ja ennen kaikkea niiden arviointi on paljon haastavampaa. Klusteroinnin validointi tarkoittaa suoritetun klusterointitehtävän onnistumisen arviointia. Tutkielmassa käsitellään validointia useasta näkökulmasta.

Esiteltyjä menetelmiä sovelletaan toteuttamalla klusterointitehtävä tilastomenetelmien kurssiaineistolle. Aineisto on kerätty Jyväskylän yliopistossa opetetulta tilastomenetelmien peruskurssilta kyselylomakkeilla. Lopputuloksena löydetään kolme taustoiltaan ja motivaatiotekijöiltään erilaista oppijaryhmää.

Avainsanat: Hierarkkinen klusterointi, K-medoids, läheisyysmatriisi, ohjaamaton oppiminen, samankaltaisuusindeksi, satunnaisuuden testaus, validointi.

Sisältö

1 Johdanto

Ihmisellä on luontainen taipumus jakaa asioita ryhmiin ja luokkiin. Ajatteluamme ohjaavat erilaiset kokonaisuudet, joita muodostamme tiedostamattamme. Ihminen pystyy käsittelemään suuria tietomääriä paremmin jakamalla ne osiin (Saariluoma 2004). Viime vuosikymmeninä yhteiskunnassamme saatavilla olevan informaation määrä on kasvanut räjähdysmäisesti. Voidaankin ajatella, että nyt haasteena on kehittää tietokoneelle samanlaisia toimintoja tiedon käsittelyyn kuin mitä ihmisaivot suorittavat automaattisesti. Tähän tarpeeseen syntyneitä kasvavia tieteenhaaroja ovat koneoppiminen (*machine learning*), tekoäly (*artificial intelligence*), hahmontunnistus (*pattern recognition*) ja *tiedonlouhinta* (*data mining*). Niissä kaikissa tarvitaan myös tilastollista päättelykykyä (Witten & Frank 2005).

Tiedonlouhinta käsittää aineiston varastointiin, organisointiin ja etsimiseen liittyvät haasteet. Se pyrkii ennen kaikkea tarjoamaan välineitä aineiston rakenteen ja trendien ymmärtämiseen. Hastie, Tibshirani ja Friedman (2009) kutsuvat tätä aineistosta oppimiseksi. He jakavat tilastollisen oppimisen karkeasti ohjattuun (*supervised*) ja *ohjaamattomaan* (*unsupervised*) *oppimiseen.*

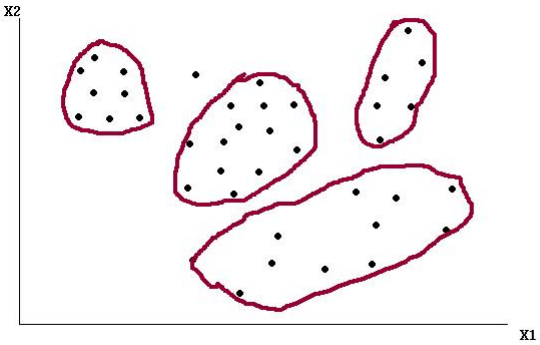
Ohjatussa oppimisessa pyritään ennustamaan vastemuuttujan arvoja muiden selittävien muuttujien avulla. Tilastollinen luokittelu kuuluu ohjattuihin oppimismenetelmiin. Siinä tavoitteena on luoda aineiston perusteella luokitin, jota voidaan käyttää ennustamaan vastaavan uuden aineiston vastemuuttujan arvoja. Ohjaamattomassa oppimisessa aineistossa ei ole havaittua vastemuuttujaa, vaan tavoitteena on kuvailla muuttujien riippuvuuksia sekä aineiston rakennetta. *Klusterointi (data clustering, cluster analysis)* on keskeisin ohjaamattoman oppimisen menetelmäkokonaisuus.

Aineiston klusterointi eli ryhmittely (ryvästely, rypästely) tarkoittaa keskenään mahdollisimman samanlaisien, mutta toisaalta toisistaan mahdollisimman erottuvien, osien etsimistä. Klusteri muodostuu sopivin kriteerein lähellä toisiaan olevista havainnoista. Ongelmana on, että yleistä kriteeriä havaintojenositteluun ei ole olemassa, vaan erilaiset kriteerit tuottavat erilaisia ryhmityksiä. Kriteerin valinta onkin yksi tärkeimmistä klusterianalyysin vaiheista.

Muita vaiheita ovat käytettävien muuttujien, läheisyysmitan sekä algoritmin valinta. Muuttujien valintaan kuuluu aineiston esikäsittely ja klusterointitehtävän kannalta olennaisten muuttujien painottaminen. Läheisyysmitta tarkoittaa sitä, kuinka havaintojen välistä samankaltaisuutta mitataan. Algoritmi puolestaan kertoo, kuinka klusterointi suoritetaan käytännössä.

Klusterointialgoritmeille on tyypillistä, että jonkinlainen rakenne aineistolle löydetään aina riippumatta siitä, onko se todellinen. Olennainen vaihe klusterointitehtävässä onkin algoritmin ehdottaman rakenteen hyvyyden arviointi. Tätä vaihetta kutsutaan klusteroinnin validoinniksi. Se voidaan tehdä vertaamalla eri tavalla tehtyjä klusterointeja toisiinsa tai vertaamalla aineistolle tietyllä algoritmilla saatua rakennetta satunnaiselle datalle saatavaan rakenteeseen. Vertailuihin on käytössä erilaisia tilastollisia tunnuslukuja, joita kutsutaan *samankaltaisuusindekseiksi (similarity indexes).*

Vaikka klusteroinnin käyttömahdollisuudet ja tehokkaampien menetelmien



Kuva 1: Esimerkki kaksiulotteisesta tilanteesta, jossa ryhmittely voi osoittautua vaikeaksi ongelmaksi.

tarve ovat kasvaneet suuriksi vasta viime aikoina, ei aineiston numeerinen osittelu ole uusi asia. Sen juuret ovat taksonomiassa eli tieteellisessä luokittelussa, joka on biologien tapa jakaa eliöt ryhmiin yhtenäisten ominaisuuksien perusteella. Numeeristen taksonomisten algoritmien kehittämisestä ovat vastanneet Sokal & Sneath (1963). Niiden lopputuloksena saadaan *hierarkkinen klusterointi*, joka selittää eliöiden sukulaissuhteita.

Nykyään erilaisia menetelmiä ja algoritmeja mitä erilaisimpiin sovellustarkoituksiin on hyvin paljon. Silti klusteroinnille yleisesti on olemassa hyvin vähän teoreettisia perusteita. Tähän on useita syitä, mutta tärkeimpänä lienee se, että yhtä totuutta klusteroinnista ei ole löydettävissä. Yleisen klusteroinnin teorian kehittäminen onkin osoittautunut erittäin haastavaksi tehtäväksi, eikä ongelmaa käsitellä tässä tutkielmassa. Viime vuosina aihetta ovat käsitelleet ainakin Ben-David ja von Luxburg (2005).

Klusteroinnin teoriaa sivutaan kuitenkin luvussa 2 antamalla klusteroinnille eräs määritelmä, minkä jälkeen käydään läpi klusterointitehtävän kulku ja siihen kuuluvat peruskäsitteet. Lukuisista eri klusterointimenetelmistä tässä tutkielmassa keskitytään eniten käytettyihin kombinatorisiin sekä hierarkkisiin menetelmiin, jotka esitellään luvuissa 3 ja 4. Klusteroinnin validointia käsitellään omassa luvussaan, heti menetelmien esittelyjen perään (luku 5). Sen jälkeen esitellään sovellusaineisto, tilastomenetelmien peruskurssin kurssikyselyaineisto, johon demonstroidaan luvuissa 3, 4 ja 5 esiteltyjä menetelmiä. Oikean aineiston käyttö menetelmien vertailussa on tärkeää, sillä synteettiset datat voivat antaa harhaanjohtavia tuloksia (Ben-David & von Luxburg 2005).

2 Klusteroinnin peruskäsitteet

Klusteroinnilla tarkoitetaan aineiston jakamista ryhmiin siten, että ryhmien sisällä havainnot ovat keskenään mahdollisimman samankaltaisia ja toisaalta ryhmien välillä mahdollisimman erilaisia. Tätä varten täytyy määritellä, mitä samankaltaisuudella ja erilaisuudella tarkoitetaan, ja siksi tarvitaan etäisyysmittoja. Ennen niiden esittelyä annetaan klusteroinnille matemaattinen määritelmä ja luvun lopuksi käydään läpi ryhmittelytehtävän eri vaiheet. Tämä luku pohjautuu pääosin Hastien, Tibshiranin & Friedmanin teokseen (2009).

2.1 Matemaattinen määritelmä

Klusteroinnin määritelmä johtaa samalla suoraan yksittäisen klusterin määritelmään. Useita erilaisia määritelmiä klusteroinnille on esitetty, mutta yhtä universaalisti hyväksyttyä määritelmää ei ole olemassa. Tämä johtuu siitä, että määritelmät sisältävät joko termejä kuten ”samanlainen” ja ”läheinen”, joita ei voi helposti määritellä, tai määritelmät ovat rajoittuneita joidenkin tietynmuotoisten klustereiden etsintään. Nyt esitetään yleinen määritelmä niin sanotulle ”kovalle” klusteroinnille (Theodoridis & Koutroumbas 2006, 488-489).

Kun käytössä on aineisto

$$X = \{x_1, x_2, ..., x_N\},$$

klusteroinnin tavoitteena on löytää X:n ositus $\Re = \{C_1, C_2, ..., C_k\}$ siten, että seuraavat kolme ehtoa toteutuvat:

1. $C_l \neq \emptyset$, $l = 1, ..., k$
2. $\cup_{l=1}^{k} C_l = X$
3. $C_l \cap C_{l'} = \emptyset$, $l \neq l'$, $l, l' = 1, ..., k$.

Nämä ehdot takaavat, että jokaiseen klusteriin kuuluu vähintään yksi piste (1); kaikki klusterit yhdistämällä saadaan koko havaintojoukko (2); ja että jokainen havaintojoukon piste voi kuulua vain yhteen klusteriin (3).

Yllä olevan määritelmän ”kovuus” tarkoittaa sitä, että jokainen piste kuuluu yhteen ja vain yhteen klusteriin. Lievempää versiota, missä havainnon kuuluvuus kuhunkin klusteriin määrätään todennäköisyytenä, kutsutaan *sumeaksi klusteroinniksi (fuzzy clustering)* (Theodoridis & Koutroumbas 2006, 489-490, 600-622). Sumeaa klusterointia ei käsitellä tässä työssä lukuun ottamatta luvussa 3 lyhyesti esiteltävää EM-algoritmin johdannaista.

2.2 Läheisyysmatriisi

Joissain tapauksissa klusteroitava aineisto koostuu pelkistä havaintoparien välisistä suhteista muodostetusta läheisyysmatriisista (*proximity matrix*). Nämä parittaiset läheisyydet voivat olla joko samankaltaisuuksia (*similarities*) tai erilaisuuksia (*dissimilarities*). Samankaltaisuudet mittaavat sitä, kuinka lähellä havaintopisteet ovat toisiaan ja erilaisuudet sitä, kuinka kaukana ne ovat toisistaan.

Useimmat algoritmit olettavat matriisin koostuvan erilaisuuksista, joten mikäli käytössä on samankaltaisuudet, ne muunnetaan erilaisuuksiksi käyttämällä soveltuvaa monotonista vähenevää funktiota. Muita läheisyysmatriisin D ominaisuuksia ovat:

- D on $N \times N$ matriisi, jossa N on havaintojen lukumäärä ja jokainen alkio $d_{ii'}$, i, $i' = 1, ..., N$, kuvaa havaintojen i ja i'välistä etäisyyttä.
- D on symmetrinen ja ei-negatiivinen, $d_{ii'} \geq 0$.
- Diagonaalialkioille pätee $d_{ii} = 0$, $i = 1, 2, ..., N$.

Vaikka läheisyysmatriisin sisältämät erilaisuudet tulkitaan havaintoparien välisiksi etäisyyksiksi, ne eivät kuitenkaan yleensä täytä metriikan ehtoja. Kolmioepäyhtälön $d_{ii'} \leq d_{ik} + d_{i'k}$, $k = 1, ..., N$, ei tarvitse olla voimassa. Yksinkertaisimmillaan läheisyysmatriisi koostuu pelkistä nollista ja ykkösistä. Esimerkkinä tästä on luvussa 2.3 esiteltävä nominaalisten muuttujien läheisyysmatriisi.

2.3 Muuttujien väliset erilaisuudet

Useimmiten käytössä on mittaukset x_{ij} havainnoille $i = 1, ..., N$ ja muuttujille $j = 1, ..., p$. Ensin määritellään yleinen erilaisuusmitta $d_j(x_{ij}, x_{i'j})$ muuttujan j suhteen ja sitten havaintojen i ja i' välinen erilaisuus

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} w_j d_j(x_{ij}, x_{i'j}); \quad \sum_{j=1}^{p} w_j = 1, \tag{1}$$

missä w_j:t ovat painoja. Käytettäväksi erilaisuusmitaksi on useita vaihtoehtoja:

- Kvantitatiivisille muuttujille yleisin valinta erilaisuusmitaksi on neliöity euklidinen etäisyys

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2.$$

- Järjestysasteikollisille eli ordinaalisille muuttujille valitaan usein niin sanottu *Manhattan-etäisyys* (Theodoridis & Koutroumbas 2006, 493)

$$d_j(x_{ij}, x_{i'j}) = |x_{ij} - x_{i'j}|.$$

- Järjestämättömille kategorisille eli nominaalisille muuttujille erilaisuusmitta täytyy määritellä tapauskohtaisesti. Mikäli muuttujan arvot ovat samat, niiden välinen erilaisuus asetetaan nollaksi, ja mikäli ne eroavat, niille annetaan jokin ennalta määrätty positiivinen arvo. Toisin sanoen, jos muuttujalla on M eri arvoa, voidaan muodostaa symmetrinen matriisi, jossa alkiot $L_{rr} = 0$, $L_{rr'} \geq 0$, $r \neq r'$, $r, r' = 1, ..., M$, ja yleensä valitaan $L_{rr'} = 1$.

2.4 Havaintojen väliset erilaisuudet

Seuraavaksi yhdistetään p:n yksittäisen muuttujan väliset erilaisuudet yhdeksi havaintoparia $(x_i, x_{i'})$ kuvaavaksi erilaisuusmitaksi. Kaavassa (1) painoilla w_j säädellään yksittäisen muuttujan vaikutusta määriteltäessä havaintojen erilaisuutta. Painojen valinnan pitäisi perustua aina tapauskohtaiseen harkintaan.

On syytä huomata, että saman painon asettaminen kaikille muuttujille ei tarkoita suoraan muuttujien asettamista samanarvoisiksi. Yksittäisen muuttujan x_j vaikutus havaintojen erilaisuuteen $D(x_i, x_{i'})$ riippuu sen suhteellisesta vaikutuksesta valittuun keskiarvoiseen erilaisuusmittaan yli kaikkien havaintoparien

$$\overline{D} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} D(x_i, x_{i'}) = \sum_{j=1}^{p} w_j \overline{d}_j,$$

$$\overline{d}_j = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_j(x_{ij}, x_{i'j}).$$

Koska j:nnen muuttujan suhteellinen vaikutus erilaisuusmittaan on $w_j \cdot \overline{d}_j$, asettamalla $w_j \sim 1/\overline{d}_j$ saadaan kaikkien muuttujien vaikutukset yhtäsuuriksi.

Vaikka muuttujien arvottaminen yhtä tärkeiksi usein tuntuu järkevältä, se ei klusterointitehtävässä ole aina hyvä valinta. Kun tavoitteena on löytää aineiston luonnolliset rakenteet, joillain muuttujilla voi olla enemmän erottelukykyä kuin toisilla. Tällaisille muuttujille kuuluu antaa suurempi paino havaintojen erilaisuutta määriteltäessä. Yhtäsuurien painojen antamisella kaikille muuttujille on taipumusta tasapainottaa aineistoa niin, että klusterointialgoritmit eivät enää erota oikeita klustereita. Sopivan erilaisuusmitan asettaminen on paljon tärkeämpi tehtävä menestyksellisen klusteroinnin saavuttamiseksi kuin itse käytettävän klusterointimenetelmän valinta.

Puuttuvat havainnot

Reaalimaailman aineistoja käsiteltäessä törmätään usein puuttuviin havaintoihin. Tällöin havaintovektorin arvot kaikkien muuttujien osalta eivät ole tiedossa. Syitä arvojen puuttumiseen voi olla useita, ja suositeltavat toimenpiteet puuttumisen käsittelyyn vaihtelevat tapauskohtaisesti puuttumisen luonteen mukaan.

Jos puuttuvien havaintojen määrä suhteessa koko aineistoon on pieni, voidaan puuttuvia arvoja sisältävät havaintovektorit poistaa aineistosta. Mikäli näin ei ole, voi havaintojen poistaminen muuttaa ongelman luonnetta viemällä

oleellista informaatiota pois. Toinen vaihtoehto puuttuvien havaintojen käsittelyyn on puuttuvien arvojen imputointi. Tällöin puuttuvat arvot korvataan käytettävistä olevista muuttujan arvoista lasketulla tunnusluvulla, yleensä keskiarvolla tai mediaanilla. Muitakin vaihtoehtoja on - moni-imputointi tarjoaa sofistikoidumman vaihtoehdon, jota ei kuitenkaan käsitellä tässä työssä (Allison 2000).

Kun määritellään kahden havaintovektorin välistä erilaisuusmitan (1) arvoa, yleisin tapa toimia puuuttuvien arvojen tapauksessa on laskea erilaisuus yli niiden muuttujien, joiden kohdalla käytettävissä on molemmat arvot x_{ij} ja $x_{i'j}$. Mikäli havaintoparille ei ole yhtään yhteistä havaittua muuttujaa, on valittava jokin edellä kuvatuista vaihtoehdoista.

Kategorisille muuttujille on mahdollista lisätä puuttuville havainnolle oma luokka "puuttuva", jota kohdellaan aivan kuten muita luokkia. Näin voidaan toimia silloin, kun on järkevää pitää kahta havaintoyksikköä samanlaisena, jos molemmista puuttuu arvo saman muuttujan kohdalla.

2.5 Klusterointitehtävän vaiheet

Klusterointi voidaan jakaa osatehtäviin seuraavasti (Theodoridis & Koutroumbas 2006, 484-485):

Muuttujien valinta (*Feature selection*) Muuttujat tulee valita niin, että mahdollisimman paljon informaatiota kiinnostuksen kohteena olevasta asiasta säilytetään. Turha tieto ja toisto pyritään eliminoimaan esikäsittelemällä aineisto. Tähän vaiheeseen kuuluvat poikkeavien havaintojen etsintä ja tunnistaminen sekä puuttuvien havaintojen käsittelystä päättäminen.

Läheisyysmitan valinta (*Proximity measure*) Läheisyysmitta kertoo kohteiden välisistä suhteista samanlaisuuksien tai erilaisuuksien kautta. Yleensä pyritään tasaamaan kaikkien muuttujien vaikutukset läheisyysmittaan niin, ettei yhtä muita dominoivaa muuttujaa synny silloin, kun sille ei ole tulkinnallisia perusteita.

Klusterointikriteerin valinta (*Clustering criterion*) Riippuu siitä, mitä ominaisuuksia klusteroinnissa halutaan painottaa. Yleensä aineistosta etsittävistä klustereista on jonkinlainen aavistus, jonka perusteella voidaan valita sopiva kriteeri. Esimerkiksi hierarkkisessa klusteroinnissa on mahdollista valita kriteeri, jonka mukaan painotetaan klusterien kompaktiutta niitä etsittäessä, tai kriteeri, jonka mukaan painotetaan klusterien yhtenäisyyttä. (Luku 3 & 4)

Algoritmin valinta (*Clustering algorithm*) Algoritmin valinta tarkoittaa klusterointitehtävän toteutusta käytännössä, eli kuinka kahdessa edellisessä kohdassa tehdyt valinnat saadaan suoritettua tehokkaasti käytettävissä olevalle aineistolle. (Luku 3 & 4)

Validointi (*Validation of the results*) Kun tulokset on saatu, niiden hyvyyttä täytyy tutkia ja arvioida. (Luku 5)

Tulkinta (*Interpretation*) Klusteroinnin tulkintavaiheen menestyksellinen suorittaminen vaihtelee sovelluskohtaisesti ja oikeiden johtopäätösten tekemiseksi tarvitaan usein asiantuntijan apua. Tässä työssä tulkintaohjeita ei juurikaan anneta, mutta aihetta on käsitellyt ansiokkaasti Anderberg (1973).

3 Kombinatoriset klusterointimenetelmät

Klusterointimenetelmällä tarkoitetaan tapaa ratkaista kyseessä oleva klusterointiongelma. Nyt käsiteltävät menetelmät ovat kombinatorisia menetelmiä, jotka liittävät jokaisen havainnon suoraan yhteen klusteriin ilman mitään aineistoa kuvaavaa todennäköisyysmallia. Tarvitaan vain jokin tavoitefunktio, jota lähdetään minimoimaan, kun käytössä on havaintojen väliset etäisyydet. Tämän luvun päälähde on Hastie, Tibshirani & Friedman (2009).

3.1 Klusterien hajonnasta

Kombinatoristen menetelmien yhteydessä käytetään koodausmerkintää $C(i) = l$ tarkoittamaan, että havaintopiste $i = 1, ..., N$ liitetään kuulumaan klusteriin $l = 1, ..., k$. Klusterointi voidaan ajatella matemaattisena, niin sanotun *tappiofunktion (loss function)* minimointitehtävänä. Koska tavoitteena on asettaa lähellä toisiaan olevat havaintopisteet samaan klusteriin, luonnollinen minimoitava funktio on

$$W(C) = \frac{1}{2} \sum_{l=1}^{k} \sum_{C(i)=l} \sum_{C(i')=l} d(x_i, x_{i'}). \tag{2}$$

Tämä kriteeri mittaa sitä, kuinka lähellä klusterin pisteet ovat toisiaan, eli se antaa tietoa klusterin kompaktiudesta. Sitä kutsutaan klusterin *sisäiseksi hajonnaksi (within scatter)*, sillä

$$T = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{ii'} = \frac{1}{2} \sum_{l=1}^{k} \sum_{C(i)=l} \left(\sum_{C(i')=l} d_{ii'} + \sum_{C(i')\neq l} d_{ii'} \right),$$

missä $d_{ii'} = d(x_i, x_{i'})$ ja T on klusteroinnista riippumaton havaintoaineiston kokonaishajonta. Yleisemmin määritellään

$$T = W(C) + B(C),$$

missä

$$B(C) = \frac{1}{2} \sum_{l=1}^{k} \sum_{C(i)=l} \sum_{C(i')\neq l} d_{ii'},$$

on klusterien *välinen hajonta (between scatter)*. Kun klusterit ovat hyvin erottuvia, on niiden välinen hajonta suurta ja tavoitefunktion $W(C)$ minimointi on yhtäpitävää $B(C)$:n maksimoinnille, sillä

$$W(C) = T - B(C).$$

Periaatteessa minimointi voidaan tehdä käymällä läpi kaikki mahdolliset tavat jakaa havainnot osajoukkoihin, mutta tällöin on kyseessä niin sanottu

NP-täydellinen *(non-deterministic polynomial-time complete)* optimointiongelma, mikä tarkoittaa laskennallisesti erittäin vaativaa ongelmaa (Garey & Johnson 1979). Käytännössä on tyydyttävä menetelmiin, jotka löytävät lokaalin optimin. Tällainen on seuraavaksi esiteltävä K-means, joka on yksi vanhimmista ja eniten käytetyistä klusterointimenetelmistä (McQueen 1967). Muita kombinatorisia menetelmiä löytyy kirjallisuudesta (Anderberg 1973, Theodoridis & Koutroumbas 2006, Witten & Frank 2005).

3.2 K-means

K-means-menetelmällä on vankka historia ja sillä on sovelluksia useilla aloilla. Signaalinkäsittelyssä ja kuva-analyysissa menetelmää sovelletaan vektorikvantisaatiossa tiedon kompressointiin. Näissä yhteyksissä puhutaan usein yleistetystä Lloydin algoritmista menetelmän idean esittäjän mukaan (Lloyd 1957).

K-means on niin sanottu prototyyppimenetelmä, jossa jokaisella klusterilla on oma edustajansa, perusversiossa keskiarvo. Keskeinen idea prototyyppimenetelmissä on koko ajan parantaa alussa yleensä satunnaisesti muodostettua jaotusta. Tämä tehdään siirtämällä havainnot lähimpään klusteriin ja sen jälkeen päivittämällä prototyyppejä, kunnes on saavutettu lokaali optimi.

K-means-menetelmä käyttää havaintoparin (x_i,$x_{i'}$) välisenä dissimilariteettimittana neliöityä euklidista etäisyyttä

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2.$$

Tällöin minimoitava klusterin sisäinen hajonta voidaan kirjoittaa

$$\begin{aligned} W(C) &= \tfrac{1}{2} \textstyle\sum_{l=1}^{k} \sum_{C(i)=l} \sum_{C(i')=l} ||x_i - x_{i'}||^2 \\ &= \textstyle\sum_{l=1}^{k} N_l \sum_{C(i)=l} ||x_i - \bar{x}_l||^2, \end{aligned}$$

missä $\overline{x}_l = (\overline{x}_{1l}, ..., \overline{x}_{pl})$ on klusteriin l liittyvä keskiarvovektori ja N_l on siihen kuuluvien havaintopisteiden lukumäärä.

Klusterointi C^* saadaan ratkaisemalla optimointiongelma

$$\min_{C, \{m_l\}_1^k} \sum_{l=1}^{k} N_l \sum_{C(i)=l} ||x_i - m_l||^2. \tag{3}$$

Minimointi tehdään seuraavalla algoritmilla:

1. Alusta k klusterikeskiarvoa $m_1, ..., m_k$ satunnaisesti tai mahdollisen ennakkotiedon avulla.

2. Vuorottele seuraavien kahden askeleen välillä.

   - *Asetusaskel:* Liitä jokainen havainto lähimmän keskiarvon klusteriin

   $$C(i) = \underset{1 \le l \le k}{argmin} ||x_i - m_l||^2.$$

- *Päivitysaskel:* Laske uudet klusterikeskipisteet

$$\overline{x}_l = \underset{m}{argmin} \sum_{C(i)=l} ||x_i - m||^2,\ l = 1, ..., k.$$

3. Lopeta, kun klusterit eivät enää muutu.

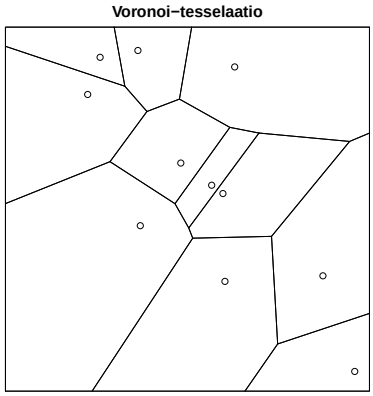
Lopputuloksena saadaan *Voronoi-tesselaatio* eli jaotus, jossa jokainen klusteri edustaa yhtä osaa koko havaintoavaruudesta. Kuvassa 2 on tehty Voronoi-tesselaatio satunnaisille pisteille. Jokainen näkyvä piste rajaa alueen, jonka sisällä kaikki muut pisteet ovat lähempänä kyseistä pistettä, eli alueen edustajaa, kuin mitään muuta pistettä.

Tulkitsemalla asetusaskeleen odotusarvoaskeleena (*expectation*) ja päivitysaskeleen maksimointiaskeleena (*maximization*) saadaan eräs *EM-algoritmin* johdannainen. Edelleen määräämällä klusterikuuluvuudet probabilistisesti deterministisyyden sijaan, sekä käyttämällä klusterikeskuksina monimuuttujaisia normaalijakaumia, eikä pelkkiä keskiarvoja, päästään yleistettyyn EM-algoritmiin (Dempster, Laird & Rubin 1977). Tällöin klusterikuuluvuudet muuttuvat kovista pehmeiksi, eli pisteiden kuuluvuudet kuhunkin klusteriin saadaan todennäköisyyksinä. Tässä tapauksessa kyseessä on sumea klusterointimenetelmä.

Koska K-means-menetelmän idea on suhteellisen yksinkertainen ja sen toteutus helposti tehtävissä, sitä käytetään usein myös muiden monimutkaisempien menetelmien osana ositusvaiheessa. Algoritmi on nopea suorittaa, joten se voidaan ajaa useaan kertaan varsin vaivattomasti. Näin onkin syytä tehdä, sillä vaikka algoritmi löytää lokaalin optimin, niin lopputulos riippuu annetuista alkuarvoista (Peterson, Ghosh & Maitra 2010).

K-means-algoritmi vaatii syötteenään ennakkotietoa klusterien lukumäärästä k. Siten, jos minkäänlaista ennakkotietoa oikeasta määrästä k ei ole, voi olla parempi käyttää jotain muuta menetelmää. Toki keinoja k:n selvittämiseen on olemassa. Yksi mahdollisuus on aloittaa algoritmi tarpeeksi pienellä k ja kasvattaa sitä, eli muodostaa uusi klusteri aina, kun havainnon etäisyys lähimmästä klusterista ylittää jonkun kynnysarvon. Tällöin tarvitaan kuitenkin ennakkotietoa sopivasta kynnysarvosta. Paljon yleisempää onkin ajaa algoritmi useaan kertaan eri arvoilla k ja valita sopivin. Kriteerejä parhaan klusterimäärän valintaan käsitellään luvussa 5.2.3.

K-means-menetelmä on hyvä valinta silloin, kun muuttujat ovat kvantitatiivisia ja aineisto muuten siistiä; eli kun neliöity euklidinen etäisyys on sopiva valinta erilaisuusmitaksi. Usein näin ei kuitenkaan ole. Askel robustimpaan suuntaan on käyttää keskiarvojen sijaan laskentaan mediaaneja, jolloin menetelmänä on K-medians (Bradley, Mangasarian & Street 1997). Seuraavaksi esitellään kuitenkin vielä monikäyttöisempi versio K-medoids.



Kuva 2: Voronoi-jaotus yhdelletoista satunnaisesti tuotetulle pisteelle.

3.3 K-medoids

K-medoids on menetelmä, jossa klusteria edustamaan valitaan oikea havaintopiste. Menetelmän kulku on hyvin samantapainen kuin K-means-menetelmän tapauksessa, mutta klusterien keskipisteitä ei tarvitse laskea, vaan tarvitsee vain ylläpitää klusteria edustavan havainnon indeksiä. Siten menetelmää voidaan käyttää, kun aineistosta on käytössä vain läheisyysmatriisi.

K-medoids-algoritmi

1. Alusta satunnaiset k havaintopistettä klustereiden edustajiksi $m_1, ..., m_k$.
2. Aseta jokainen havainto lähimpään klusteriin
$$C(i) = \underset{1 \leq l \leq k}{argmin}\, D(x_i, m_k).$$
3. Nykyiselle ositukselle etsi havaintopiste i_l^*, jonka etäisyys klusterin muihin pisteisiin on pienin, ja aseta se klusterin uudeksi edustajaksi
$$i_l^* = \underset{\{i:C(i)=l\}}{argmin} \sum_{C(i')=l} D(x_i, x_{i'}),\ m_l = x_{i_l^*},\ l = 1, ..., k.$$
4. Vuorottele askelien 2 ja 3 välillä, kunnes klusterit eivät enää muutu.

4 Hierarkkinen klusterointi

Hierarkkiset klusterointimenetelmät liittyvät kiinteästi numeeriseen taksonomiaan (Sokal & Sneath 1963). Siinä ryhmittely tehdään askel kerrallaan yhdistämällä samankaltaisia pisteitä toisiinsa, jolloin kokonaisuudesta muodostuu hierarkkinen jaotus. Siten tulokseksi ei saada selkeitä, tiettyjä klustereita vaan monitasoinen, sisäkkäinen klusterirakenne: alimmalla tasolla jokainen klusteri koostuu vain yhdestä havainnosta, ylimpänä oleva yksittäinen klusteri sisältää koko datan.

Menetelmien etuihin kuuluu hyvä ymmärrettävyys, sillä klusterointi voidaan kuvata puumaisen rakenteen avulla - syntyvää kuviota kutsutaan dendrogrammiksi. Lisäksi menetelmät tarvitsevat syötteenään vain läheisyysmatriisin, eikä itse havaintoja tarvita. Hierarkkiset menetelmät jakautuvat *yhdisteleviin (merge, agglomerative)* ja *jakaviin (split, divisive)* menetelmiin sen mukaan, miten osittelu tehdään. Tämä osio pohjautuu, niiltä osin kun muuta lähdettä ei mainita, teokseen Hastie, Tibshirani & Friedman (2009).

4.1 Klusterien sisäkkäisyys

Olkoon käytössä aineisto

$$X = \{x_i,\, i = 1, ..., N\},$$

missä x_i:t ovat p-ulotteisia havaintovektoreita, joille on tehty annetun määritelmän mukainen klusterointi

$$\Re = \{C_l,\, l = 1, ..., k\}.$$

Sanotaan että klusterointi $\Re_1$, joka sisältää k_1 klusteria, on sisäkkäinen klusteroinnille $\Re_2$, jossa on $k_2\,(< k_1)$ klusteria, mikäli jokainen $\Re_1$:n klusteri on $\Re_2$:n osajoukko. Tällöin $\Re_1$:stä kutsutaan myös $\Re_2$:n tytärklusteriksi. Esimerkiksi klusterointi $\Re_1 = \{\{x_1, x_3\}, \{x_4\}, \{x_2, x_5\}\}$ on sisäkkäinen klusteroinnille $\Re_2 = \{\{x_1, x_3, x_4\}, \{x_2, x_5\}\}$, mutta $\Re_1$ ei ole sisäkkäinen klusteroinnille $\Re_3 = \{\{x_1, x_2, x_4\}, \{x_3, x_5\}\}$. (Theodoridis & Koutroumbas 2006, 541)

Hierarkkiset klusterointimenetelmät tuottavat sisäkkäisen klusterihierarkian. Tämä tarkoittaa sitä, että jos kaksi havaintovektoria yhdistetään yhdeksi klusteriksi tietyllä tasolla, niin ne pysyvät samassa klusterissa kaikilla seuraavilla tasoilla. Aikaisella tasolla tulleesta huonosta valinnasta ei näin pystytä toipumaan myöhemmissä vaiheissa.

4.2 Yhdistelevät menetelmät

Yhdistelevät menetelmät lähtevät liikkeelle alimmalta tasolta, jossa jokainen havainto on omana klusterinaan. Jokaisella tasolla lähimmät kaksi klusteria yhdistetään uudeksi klusteriksi. Näin seuraavalla tasolla on aina yksi klusteri vähemmän. Yhdistettävällä lähimmällä klusteriparilla tarkoitetaan sitä paria, jonka yhteinen erilaisuus on pienin. Tätä varten täytyy määritellä erilaisuusmitta klustereiden välille.

Erilaiset versiot yhdistelevistä menetelmistä eroavat eniten toisistaan erilaisuuden mittaamisessa ja siinä kuinka erilaisuusmittaa päivitetään aina seuraavalle tasolle siirryttäessä. Useita lähestymistapoja on käytetty tehokkaiden yhdistelevien algoritmien muodostamiseksi (Theodoridis & Koutroumbas 2006). Ne perustuvat joko matriisiteoriaan tai graafiteoriaan. Matriisiteoriaan perustuvat tavat ovat yleisempiä, ja niitä käytetään tässä työssä.

Kahden klusterin tai klusterin ja yksittäisen havaintopisteen välisen etäisyyden mittaamiseen on yleisesti käytössä kolme eri tapaa, joista käytetään nimitystä *linkkifunktiot*.

1. Klusterien lähimpänä toisiaan olevien pisteiden välinen etäisyys *(single linkage)*, eli lähimmän naapurin menetelmä määritellään

$$D_{SL}(C_a, C_b) = min_{x_i \in C_a, x_{i'} \in C_b} d(x_i, x_{i'}).$$

2. Klusterien kaukausimpina toisistaan olevien pisteiden välinen etäisyys *(complete linkage)*, eli kaukaisimman naapurin menetelmä määritellään

$$D_{CL}(C_a, C_b) = max_{x_i \in C_a, x_{i'} \in C_b} d(x_i, x_{i'}).$$

3. Klusterien keskimääräinen etäisyys *(average linkage)* saadaan kaavasta

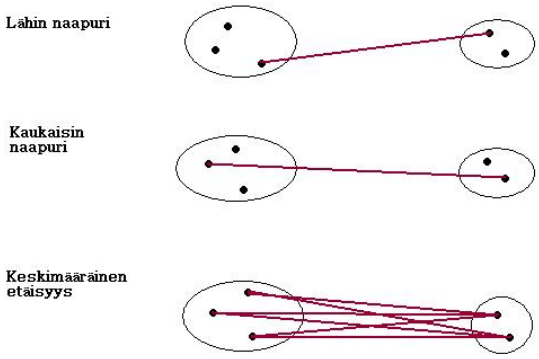
$$D_{AL}(C_a, C_b) = \frac{1}{N_a N_b} \sum_{x_i \in C_a} \sum_{x_{i'} \in C_b} d(x_i, x_{i'}),$$

   missä C_a, C_b ovat klustereita ja N_a, N_b niiden sisältämien alkioiden lukumäärät.

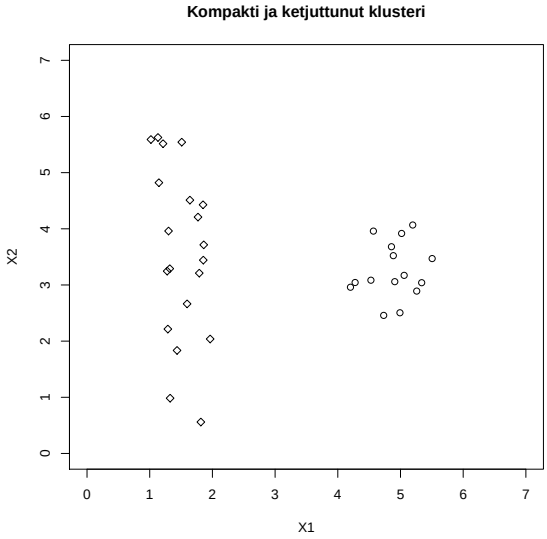
Mikäli käytettävät erilaisuudet kuvaavat aineiston klusterirakennetta hyvin eli tuottavat kompakteja, toisistaan erottuvia klustereja, antavat kaikki kolme linkkifunktiota samat tulokset (kuva 3). Silloin kun näin ei ole, tulokset eroavat.

Lähimmän naapurin menetelmällä on taipumusta venyttää klusterin reunoja, sillä se ei vaadi kuin yhden klusterin havaintopisteistä sijaitsevan lähellä uutta pistettä. Tällöin klusterin *läpimitta* kasvaa ja *kompaktius* kärsii (kuva 4). Klusterin läpimitta *(diameter)* määritellään sen sisältämien havaintopisteiden suurimpana parittaisena erilaisuutena. Läpimitan kasvuun johtavaa ilmiötä, missä uusia havaintopisteitä yhdistetään klusterin reunalle suhteellisen pienellä kynnysarvolla, kutsutaan *ketjuttumiseksi (chaining)*. Yleensä ketjuttumisen ajatellaan olevan menetelmän heikkous, mutta toisaalta se mahdollistaa erimuotoisten *yhtenäisten* rakenteiden löytämisen aineistosta.

Kaukaisimman pisteen linkkifunktio edustaa toista ääripäätä. Se vaatii kaikkien klusterin havaintopisteiden sijaitsevan lähellä toisen klusterin kaikkia pisteitä, jotta yhdistäminen tehdään. Klusterit säilyvät tällöin kompakteina ja läpimitaltaan suhteellisen pieninä. Toisaalta klusterien läheisyysominaisuus kärsii, eli klusteriin yhdistetään havaintopisteitä, jotka voivat olla paljon lähempänä toisen klusterin pisteitä kuin osaa oman klusterinsa pisteistä. Tällöin aineistossa mahdollisesti piileviä yhtenäisiä rakenteita ei löydetä.



Kuva 3: Linkkifunktioiden periaatteet.



Kuva 4: Vasemmalla on esimerkki ketjuttuneesta klusterista, oikealla kompaktista klusterista.

Ryhmäkeskiarvoihin perustuva linkkifunktio on kompromissi näiden kahden ääripään välillä. Se pyrkii tuottamaan suhteellisen kompakteja sekä suhteellisen yhtenäisiä klustereja. Kuitenkin sen antamat tulokset riippuvat havaintojen erilaisuusmitan numeerisesta skaalasta. Minkä tahansa monotonisesti kasvavan muunnoksen soveltaminen keskiarvoisen etäisyyden kaavaan voi muuttaa tuloksia. Lähimmän naapurin menetelmä ja kaukaisimman pisteen linkkifunktio riippuvat vain erilaisuuksien järjestyksestä eli ovat siten invariantteja vastaaville muunnoksille.

Tavallisen keskiarvon version lisäksi kolmannesta linkkifunktiosta on yleisesti käytössä muita versioita, joissa klusterin edustajana yhdistettäessä on jokin muu tunnusluku. Klusterien etäisyyksien määräämiseen on tällöin käytössä erilaisia painotuksia. Tunnetuin näistä eri versioista on Wardin menetelmä tai pienimmän varianssin algoritmi (*The Ward method, minimum variance algorithm*), joka esitellään lyhyesti seuraavassa.

Wardin menetelmä

Wardin menetelmässä klusterien läheisyyttä mitataan klusterin sisäisen hajonnan avulla. Seuraavaksi yhdistettävät kaksi klusteria valitaan niin, että yhdistäminen johtaa pienimpään mahdolliseen lisäykseen klusterien sisäisessä variansissa. Menetelmä on eräänlainen välimuoto keskimääräisen etäisyyden ja kaukaisimman pisteen linkkifunktion välillä, sillä se pyrkii tuottamaan kompakteja klustereja, mutta ottaa huomioon kaikki havaintopisteet.

Wardin menetelmässä edetään seuraavasti: (Anderberg 1973)

1. Alussa jokainen havaintopiste x_i, $i = 1, ..., N$, on omana klusterinaan.

2. Jokaisella askeleella vähennetään klusterien määrää yhdellä niin, että klusterien yhdistäminen lisää tappiofunktion $W(C)$ (kaava 2) arvoa pienimmällä mahdollisella määrällä.

3. Jatketaan yhdistämistä tasolle $N - 1$, jolloin jäljellä on enää yksi klusteri.

4.3 Jakavat menetelmät

Jakavat menetelmät aloittavat ylimmältä tasolta rekursiivisesti jakamalla yhden olemassaolevista klustereista kahdeksi uudeksi klusteriksi. Jaettavaksi valitaan klusteri, jonka tuottamien uusien ryhmien keskinäinen erilaisuus on suurin. Kun aineisto halutaan osittaa suhteelliseen pieneen määrään klustereita, voivat jakavat menetelmät olla yhdisteleviä menetelmiä parempi valinta hierarkkiseen klusterointiin.

Macnauhgton Smith, Williams, Dale & Mockett (1965) ovat esittäneet jakavan perusmenetelmän. Siinä alussa kaikki havainnot ovat samassa klusterissa C_1. Sitten valitaan havaintopiste, jonka keskimääräinen etäisyys kaikkiin muihin havaintopisteisiin on suurin. Tämä havaintopiste on toisen klusterin C_2 ensimmäinen jäsen. Jokaisella tulevalla askeleella se C_1:n havaintopiste, jonka keskimääräinen etäisyys muista C_1:n pisteistä vähennettynä sen keskimääräisellä

etäisyydellä C_2:ssa jo oleviin pisteisiin on suurin, siirretään C_2:seen. Tätä jatketaan, kunnes kyseinen ero muuttuu negatiiviseksi eli C_1:ssä ei ole enää pisteitä, jotka ovat keskimäärin lähempänä klusteria C_2.

Tuloksena alkuperäinen klusteri on jaettu kahteen tytärklusteriin, joista toisen muodostavat klusteriin C_2 siirretyt havainnot ja toisen klusteriin C_1 yhä jääneet havainnot. Näin on saatu hierarkian toinen taso. Jokainen tuleva taso muodostetaan käyttämällä samaa jakomenetelmää yhteen ylemmän tason klusteriin. Aina seuraavana jaettavaksi klusteriksi voidaan valita esimerkiksi suurimman läpimitan klusteri. Rekursiivinen jakaminen jatkuu, kunnes jokainen klusteri sisältää vain yhden havainnon.

4.4 Dendrogrammi

Molemmat yllä kuvatut osittelutavat tuottavat N-1-tasoisen klusterihierarkian. Käyttäjän tulkittavaksi jää, miltä tasolta aineiston "luonnolliset" ryhmät löytyvät eli millä tasolla olevien klusterien sisältämät havaintopisteet ovat keskenään huomattavasti samankaltaisempia kuin mitä klusterien väliset havaintopisteet.

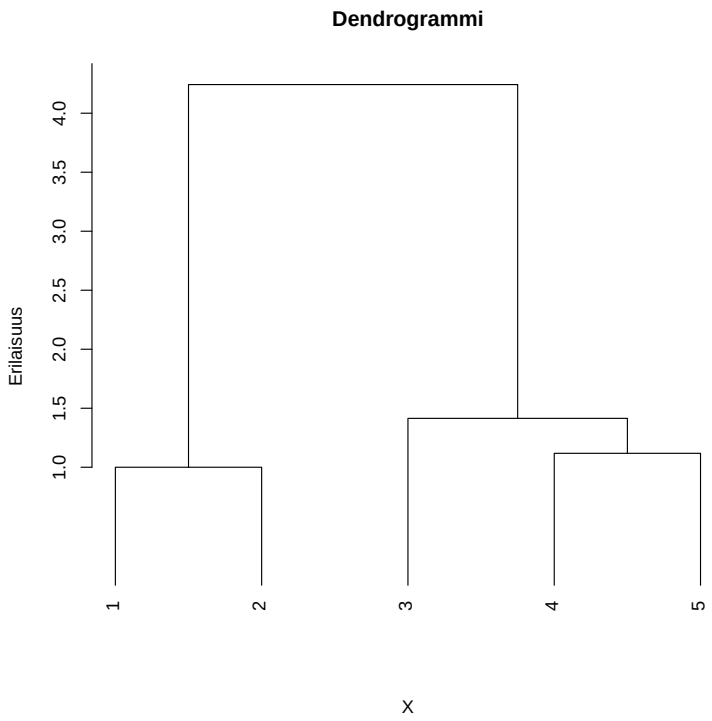
Kaikilla yhdistelevillä ja joillakin jakavilla menetelmillä on monotonisuusominaisuus, eli yhdistettävien klusterien välinen erilaisuus kasvaa monotonisesti jokaisella tasolla. Hierarkia voidaan myös kuvata puuna siten, että jokaista havaintopistettä edustaa solmu ja jokainen yhdistäminen (jakaminen) muodostaa uuden solmun. Solmun korkeus määräytyy suhteessa sitä edeltävän tason solmun erilaisuuden arvoon. Lisäksi yhden havaintopisteen solmuille asetetaan korkeudeksi nolla. Tämäntyyppistä graafista esitystä kutsutaan dendrogrammiksi (kuva 5).

Hierarkkisten menetelmien suosio perustuu suurelta osin dendrogrammien helppoon tulkittavuuteen, mutta asiassa on myös vaaransa. Hierarkkiset menetelmät tuottavat aina hierarkkisen rakenteen, riippumatta siitä, onko aineistossa todellisuudessa ollenkaan kyseistä rakennetta. Eri algoritmit, samoin kuin pienet muutokset aineistossa, voivat tuottaa hyvinkin erilaisia rakenteita. Lisäksi useimmat käytössä olevat algoritmit ovat ahneita, eli ne valitsevat aina jokaisella askeleella parhaalta tuntuvan vaihtoehdon ajattelematta kokonaisuutta. Takaisinpaluuta askeleiden välillä ei yleensä ole.

Dendrogrammia tulee pitää pääasiassa aineiston klusterirakennetta kuvaavana graafisena työkaluna, joka on syntynyt käytettävän algoritmin kautta. Dendogrammi ei ole aineiston graafinen esitys samassa mielessä kuin esimerkiksi hajontakuvio tai histogrammi.

Kofeneettinen matriisi

Dendrogrammin avulla hierarkkisille klusteroinnille voidaan muodostaa niin sanottu kofeneettinen matriisi (*cophenetic matrix*), joka on hyödyllinen työkalu klusteroinnin validointivaiheessa, mitä käsitellään seuraavassa luvussa. Kofeneettinen etäisyys havaintoparille (i, i') tarkoittaa niiden välistä erilaisuuden tasoa, jossa i ja i' liitetään ensimmäistä kertaa samaan klusteriin. (Theodoridis & Koutroumbas 2006, 568-569)



Kuva 5: Dendrogrammi, joka on saatu klusteroimalla esimerkin 4.1 aineisto (X) lähimmin naapurin menetelmällä.

Esimerkki 4.1 Olkoon käytössä aineisto, jossa on viisi havaintoa ja kaksi muuttujaa koottuna seuraavaan matriisiin:

$$X = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 5 & 4 \\ 6 & 5 \\ 6.5 & 6 \end{bmatrix}$$

Sitä vastaava dissimilariteettimatriisi, kun erilaisuuksina käytetään euklidista etäisyyttä, on seuraava:

$$D = \begin{bmatrix} 0 & 1 & 5 & 6.4 & 7.4 \\ 1 & 0 & 4.2 & 5.7 & 6.7 \\ 5 & 4.2 & 0 & 1.4 & 2.5 \\ 6.4 & 5.7 & 1.4 & 0 & 1.1 \\ 7.4 & 6.7 & 2.5 & 1.1 & 0 \end{bmatrix}$$

Klusteroimalla D saadaan dendrogrammi (kuva 5), josta edelleen kofeneettinen matriisi:

$$D_c = \begin{bmatrix} 0 & 1 & 4.2 & 4.2 & 4.2 \\ 1 & 0 & 4.2 & 4.2 & 4.2 \\ 4.2 & 4.2 & 0 & 1.4 & 1.4 \\ 4.2 & 4.2 & 1.4 & 0 & 1.1 \\ 4.2 & 4.2 & 1.4 & 1.1 & 0 \end{bmatrix}$$

Vertaamalla kofeneettista matriisia kuvaan 5 nähdään erilaisuudet huomioivan dendrogrammin ja kofeneettisen matriisin välinen yhteys. Kofeneettiset etäisyydet saadaan suoraan katsomalla dendrogrammista, millä erilaisuuden tasoilla yhdistämiset on tehty. Pitkä etäisyys kolmannen ja viimeisen yhdistämisen välillä kertoo kahdesta klusterista aineistossa.

5 Klusteroinnin validointi

Klusterointimenetelmillä on taipumus löytää klustereita myös, vaikka aineisto olisi oikeasti klusteroitumatonta. Klusteroinnin validointi tarkoittaa suoritetun klusterointitehtävän onnistumisen arviointia. Tavoitteena on selvittää, onko aineiston rakenne todella löydetty klusteroimalla vai onko saatu rakenne sattumasta johtuvaa.

Arvioinnissa käytetään erilaisia kriteerejä ja tunnuslukuja. Niiden avulla voidaan toteuttaa tilastollinen testaaminen, jossa sattuman osuutta klusterointiin selvitetään. Tämän luvun lähdeteoksena on Theodoridis ja Koutroumbas (2006).

5.1 Lähtökohta

Klusterointi on ohjaamattoman oppimisen menetelmä, jossa ennakkotietoa aineiston rakenteesta ei tarvita. Silti minkä tahansa klusterointimenetelmän käyttö edellyttää jonkinlaisia tulokseen vaikuttavia valintoja. Näitä ovat erilaisten parametrien valinta tai jonkinlaisten klusterien muotoa rajaavien rajoitteiden määrääminen.

Esimerkiksi hierarkkisten menetelmien tapauksessa on määriteltävä etäisyysmitta ja käytettävä linkkifunktio. Muiden kuin hierarkkisten menetelmien tapauksessa on yleensä määritettävä ainakin etsittävien klusterien lukumäärä. Parametrien ja rajoitteiden huono valinta voi johtaa vääriin johtopäätöksiin aineiston rakenteesta, joten klusterointialgoritmien tulosten huolellinen arviointi on tarpeen.

Koska klusterointialgoritmit tuottavat ennalta tuntemattomia klustereita, algoritmin suorittamisen jälkeen esille nousee seuraavia kysymyksiä:

1. Ovatko löydetyt klusterit todellisia vai sattuman aiheuttamia?
2. Kuinka monta klusteria aineistossa on?
3. Onko olemassa parempaa ryhmittelyä aineistolle kuin löydetty klusterointi?

Näihin kysymyksiin vastaamiseksi tarvitaan klusterointitulosten validointityökaluja.

Suurin osa validointityökaluista käsittelee yksittäisiä klusterointeja eli arvioitavana on jonkin tietyn algoritmin tuottama yksikäsitteinen ositus koko aineistolle. Toisaalta on tärkeää huomata, että myös hierarkkinen klusterointi voidaan esittää yksittäisenä klusterointina katkaisemalla se jollakin tietyllä korkeudella. Tällöin kuitenkin menetetään hierarkkisuuden antama lisäinformaatio, joten joitakin validointimenetelmiä pelkästään hierarkkisille klusteroinneille on kehitetty.

Lisäksi on mahdollista lähestyä klusteroinnin validointia suoraan yksittäisiä klustereita tarkastelemalla. Tällöin tarkoituksena on tutkia yksittäisen klusterin hyvyyttä. Kyseinen lähestymistapa voi olla käyttökelpoinen esimerkiksi suuria aineistoja käsiteltäessä, jolloin yhdenkin erottuvan klusterin löytäminen

voi osoittautua arvokkaaksi informaatioksi. Tässä työssä ei käsitellä yksittäisen klusterin validointia. Bailey & Dubes (1982) esittävät graafiteoreettisen lähestymistavan aiheeseen.

5.2 Arviointikriteerit

Klusteroinnin kelvollisuutta voidaan tarkastella arvioimalla klusterointituloksia jonkin kriteerin suhteen. Erilaisia kriteerejä on kolmenlaisia: ulkoinen, sisäinen ja suhteellinen. Jokaiseen kriteeriin liittyy joukko tilastollisia tunnuslukuja, joita käyttäen ryhmittelyn numeerinen arviointi toteutetaan.

Kirjallisuudessa näitä tunnuslukuja kutsutaan usein myös *samankaltaisuusindekseiksi (similarity indexes)*. Niitä on kolmea tyyppiä: havaintoparien laskemiseen perustuvat mitat, joukkojen vastaavuuteen perustuvat mitat (Meilã 2005) sekä informaatioteoriaan perustuvat mitat (Vinh, Epps & Bailey 2009). Tässä työssä keskitytään näistä vanhimpaan ja sitä kautta tunnetuimpaan tyyppiin, havaintoparien laskemiseen perustuviin indekseihin, joista löytyy hyvä kokoelma kirjallisuudesta (Albatineh, Niewiadomska-Bugaj & Mihalko 2006).

5.2.1 Ulkoinen kriteeri

Tässä tapauksessa aineiston rakenteesta on käytössä klusterointimenetelmästä riippumatonta ulkoista tietoa. Käytännössä ulkoinen tieto esitetään ennaltamäärättynä aineiston osituksena, johon klusterointimenetelmän tuottamaa ositusta voidaan verrata.

Jos $\wp_1$ ja $\wp_2$ ovat ositukset (ryhmien lukumäärä voi vaihdella), tunnuslukujen määrittelyä varten lasketaan havaintojen klusterikuuluvuudet seuraavasti:

SS - havaintoparien lukumäärä, jotka kuuluvat samaan klusteriin molemmissa osituksissa,

SD - havaintoparien lukumäärä, jotka kuuluvat samaan klusteriin osituksessa $\wp_1$, mutta eri klusteriin osituksessa $\wp_2$,

DS - havaintoparien lukumäärä, jotka kuuluvat samaan klusteriin osituksessa $\wp_2$, mutta eri klusteriin osituksessa $\wp_1$,

DD - havaintoparien lukumäärä, jotka molemmat kuuluvat eri klusteriin molemmissa osituksissa.

Kun lisäksi $M = SS + SD + DS + DD$ eli havaintoparien kokonaismäärä $M = N(N-1)/2$, missä N on havaintopisteiden määrä aineistossa, päästään yleisimpien ulkoisten tunnuslukujen määritelmiin.

Ehkä tunnetuin samankaltaisuusindeksi on *Randin tunnusluku* (*Rand statistic*, Rand 1971)

$$R = \frac{SS + DD}{M}.$$

Helposti nähdään, että $R \in [0, 1]$ ja arvo 1 saadaan, kun ositukset ovat identtiset. Arvo 0 saavutetaan, kun yhtään havaintoparia ei ole samassa klusterissa eikä

eri klusterissa kummassakaan osituksessa. Tämä tapahtuu vain siinä harvinaisessa ja epäkäytännöllisessä tapauksessa, kun toinen ositus sisältää vain yhden klusterin ja toisen osituksen kaikki havaintopisteet ovat omina klustereinaan.

Koska ulkoisen kriteerin käyttöön liittyy läheisesti satunnaisuuden testaus, mistä enemmän luvussa 5.3, samankaltaisuusindeksin toivotaan tuottavan kahdelle satunnaiselle ositukselle arvon läheltä nollaa. Randin tunnusluvulle näin ei ole, eikä kyseinen arvo ole vakio, mikä on ongelmallista. Hubert & Arabie (1985) esittivät satunnaiskorjatun version Randin tunnusluvulle, joka saadaan kaavalla

$$q' = \frac{q - E(q)}{max(q) - E(q)}, \tag{4}$$

missä $q = R$ ja $E(q)$ on tunnusluvun odotusarvo sekä $max(q) = 1$.

Kaksi muuta samantapaista tunnuslukua ovat *Jaccardin kerroin* (*Jaccard coefficent*, Jaccard 1912)

$$J = \frac{SS}{SS + SD + DS}$$

ja *FM-indeksi (Fowlkes and Mallows index*, Fowlkes & Mallows 1983)

$$FM = \sqrt{\frac{SS}{SS + SD}\frac{SS}{SS + DS}}.$$

Nämä tunnusluvut saavat arvoja väliltä $[0, 1]$, ja mitä suurempi on tunnusluvun arvo, sitä lähempänä ositukset ovat toisiaan. Samanlaisia ongelmia kuin Randin tunnuslukuun liittyy myös Jaccardin kertoimeen ja FM-indeksiin. Satunnaiskorjauksen kaavaa (4) voidaan soveltaa niistä molempiin. Ongelmalliseksi korjattujen arvojen laskemisessa voi muodostua odotusarvojen $E(q)$ laskenta. Tätä aihetta käsittelevät Albatineh, Niewiadomska-Bugaj & Mihalko (2006).

Lisäksi kahden riippumattoman läheisyysmatriisin välisen assosiaation mittaamiseen on käytössä ulkoisen kriteerin tunnusluku *Hubertin* Γ (*Hubert's* Γ, Hubert & Arabie 1985):

$$\Gamma = (1/M) \sum_{i=1}^{N-1} \sum_{i'=i+1}^{N} Z(i, i')Y(i, i'), \tag{5}$$

missä $Z(i, i')$ ja $Y(i, i')$ ovat matriisien Z ja Y alkiot (i, i'). Suuret arvot tarkoittavat vahvaa assosiaatiota.

Normalisoimalla Hubertin Γ päästään muotoon, joka vastaa läheisyysmatriisien välistä otoskorrelaatiokerrointa:

$$\hat{\Gamma} = \frac{(1/M) \sum_{i=1}^{N-1} \sum_{i'=i+1}^{N} (Z(i, i') - \mu_Z)(Y(i, i') - \mu_Y)}{\sigma_Z \sigma_Y}, \tag{6}$$

missä

$$\mu_Z = (1/M) \sum_{i=1}^{N-1} \sum_{i'=i+1}^{N} Z(i, i'),$$

$$\sigma_Z = \sqrt{(1/M)\sum_{i=1}^{N-1}\sum_{i'=i+1}^{N}(Z(i,i')^2 - \mu_Z^2)},$$

sekä μ_Y ja σ_Y määritellään vastaavasti. Normalisoidun Hubertin $\hat{\Gamma}$ arvot ovat välillä $[-1, 1]$. Itseisarvoltaan suuret arvot viittaavat vahvaan riippuvuuteen.

5.2.2 Sisäinen kriteeri

Tässä tapauksessa tehtyä klusterointia arvioidaan vertaamalla sitä johonkin aineistosta laskettuun, sen rakennetta kuvaavaan ominaisuuteen. Tavoitteena on selvittää klusteroinnin hyvyyttä käyttämällä ainoastaan datasta johdettua sisäistä informaatiota. Aineisto oletetaan esitetyksi läheisyysmatriisin avulla. Sisäinen kriteeri soveltuu käytettäväksi niin yksittäisen klusteroinnin kuin klusterihierarkian tapauksessa.

Klusterihierarkian validointi

Aiemmin määriteltyä hierarkkisen klusteroinnin kofeneettista matriisia D_c voidaan käyttää validoinnissa. Nyt määritellään tilastollinen tunnusluku, joka mittaa D_c:n ja läheisyysmatriisin D välistä riippuvuutta. Koska molemmat matriisit ovat symmetrisiä ja pelkkiä nollia diagonaalilla sisältäviä, voidaan vertailuihin käyttää vain matriisien yläkolmiossa olevia alkioita, joita on $M = N(N-1)/2$ kappaletta.

Olkoon $d_{ii'}$ ja $c_{ii'}$ D:n ja D_c:n alkiot (i, i'). Tällöin voidaan laskea matriisien välinen kofeneettinen korrelaatiokerroin (*cophenetic correlation coefficient*, CPCC), jota käytetään, kun matriisit ovat välimatka- tai suhdeasteikollisia (Rolph 1970). Se määritellään kuten Pearsonin tulomomenttikorrelaatiokerroin, siis

$$CPCC = \frac{(1/M)\sum_{i=1}^{N-1}\sum_{i'=i+1}^{N} d_{ii'}c_{ii'} - \mu_D\mu_c}{\sqrt{((1/M)\sum_{i=1}^{N-1}\sum_{i'=i+1}^{N} d_{ii'}^2 - \mu_D^2)((1/M)\sum_{i=1}^{N-1}\sum_{i'=i+1}^{N} c_{ii'}^2 - \mu_c^2)}},$$

missä

$$\mu_D = (1/M)\sum_{i=1}^{N-1}\sum_{i'=i+1}^{N} d_{ii'}$$

ja μ_c on vastaava keskiarvotermi D_c:lle. Korrelaatiokertoimena CPCC saa arvoja väliltä $[-1, 1]$, ja mitä lähempänä ykköstä se on, sitä paremmin kofeneettinen matriisi ja läheisyysmatriisi vastaavat toisiaan.

Hierarkkisten menetelmien taipumus tuottaa hierarkkinen rakenne aineistolle myös silloin, kun sitä ei siinä ole, näkyy kofeneettisen korrelaatiokertoimen arvoissa. Täysin satunnaiselle aineistolle voidaan saada varsin korkeitakin arvoja, valitusta linkkifunktiosta riippuen jopa luokkaa 0.9 (Rolph 1970). Siten satunnaisuuden testauksen rooli CPCC:n yhteydessä on korostetun tärkeä.

Kofeneettiseen korrelaatiokertoimeen vaikuttaa moni ongelmaan liittyvä parametri, kuten aineiston koko, käytetty klusterointialgoritmi ja käytetty läheisyysmitta. Tämän takia CPCC:n tarkan tiheysfunktion laskenta nollahypoteesin

vallitessa on haastavaa. Siksi kofeneettisen korrelaatiokertoimen merkitsevyystestauksessa käytetään simulointitekniikoita.

Yksittäisen klusteroinnin validointi

Hubertin assosiaatiomittaa Γ (5) tai sen normalisoitua versiota $\hat{\Gamma}$ (6) voidaan käyttää myös sisäisen kriteerin tapauksessa. Matriisi Y määritellään

$$Y(i,i') = \begin{cases} 1, & \text{kun havaintopisteet } i \text{ ja } i' \text{ kuuluvat samaan klusteriin,} \\ 0, & \text{muulloin,} \end{cases}$$

kaikille i, $i' = 1, ..., N$. Nyt Γ:a sovelletaan mittaamaan Y:n ja läheisyysmatriisin $D(= Z)$ välistä vastaavuutta. Jälleen (itseisarvoltaan) suuret arvot kertovat vahvasta assosiaatiosta.

5.2.3 Suhteellinen kriteeri

Ulkoisen ja sisäisen kriteerin käyttö klusteroinnin validoinnissa perustuu tilastollisiin testeihin. Seuraavassa luvussa esitellään, kuinka tunnuslukujen arvojen satunnaisuutta voidaan testata simuloimalla. Sen sijaan tämän luvun suhteelliseen kriteeriin ei vastaavia testejä liity, vaan ideana on verrata eri tavalla tehtyjä klusterointeja toisiinsa.

Lähtokohtana on joukko ryhmittelyjä, joista pyritään valitsemaan sopivin. Mikäli A on tiettyyn algoritmiin liittyvä parametrien joukko, voidaan ongelma muotoilla seuraavasti:

"Tietyn algoritmin tuottamien klusterointien joukosta, eri A:n parametrien arvoilla, valitse se, joka parhaiten sopii aineistoon X." (Theodoridis & Koutroumbas 2006, 747)

Tarkastellaan erikseen kahta tapausta parametrien valitsemiseksi:

1. Parametrijoukkoon A ei sisälly klusterien lukumäärää k. Menettely perustuu oletukseen, että mikäli aineisto sisältää klusterirakenteen, kyseinen rakenne löydetään useammilla eri A:n parametrien arvoilla. Nyt algoritmi ajetaan useaan kertaan eri parametrien arvoilla ja valitaan niistä suurin väli, jossa k pysyy vakiona. Etsityt parametrien arvot valitaan löydetyn välin keskikohtaa vastaaviksi. Samalla on löydetty klusterien lukumäärä aineistossa.

2. Klusterien lukumäärä k sisältyy parametrijoukkoon A, kuten on kaikkien tässä työssä esitettyjen klusterointimenetelmien kohdalla. Ensin valitaan suoritusindeksiksi sopiva tilastollinen tunnusluku q. Sopivin klusterointi etsitään q:n avulla menetellen seuraavasti:

    - Klusterointialgoritmi ajetaan kaikilla $k \in [k_{min}, k_{max}]$, kun pienin k_{min} ja suurin k_{max} klusterien lukumäärä on valittu etukäteen.
    - Kaikille k algoritmi suoritetaan r kertaa, vaihdellen muiden A:n parametrien arvoja.

- Parhaat suoritusindeksin q arvot, jokaiselle k, piirretään k:n funktiona kuvaajaksi, josta paras klusterointi tunnistetaan.

Päättely riippuu suoritusindeksin ominaisuuksista. Mikäli q ei sisällä kasvavaa tai vähenevää trendiä, kun k kasvaa, etsitään q:n arvojen kuvaajan maksimia (minimiä). Jos q:n arvot kasvavat (vähenevät) k:n mukana, etsitään kuvaajasta suurinta lokaalia muutosta. Kyseinen muutos näkyy polvekkeena kuvaajassa ja kertoo aineistosta löytyvien klusterien lukumäärän. Toisaalta mikäli selvää polveketta ei kuvaajasta löydy, on syytä epäillä, että aineisto ei sisällä klusterirakennetta. Suoritusindeksiksi soveltuu esimerkiksi seuraavaksi esiteltävä tunnusluku.

Muokattu Hubertin Γ (*The modified Hubert* Γ)

Klusteroinnin validoinnissa tavallista Hubertin assosiaatiomittaa Γ enemmän käytetty muokattu versio perustuu geometriseen ajatteluun (Bezdek & Pal 1998). Muokattu Hubertin Γ lasketaan aivan kuten (5) tai (6), mutta matriiseina ovat aineistoon liittyvä läheisyysmatriisi D ja samaa etäisyysmittaa käyttäen saatu matriisi Q.

Laskentaa varten muistetaan koodausmerkintä $C(i) = l$, kun havaintopiste i kuuluu klusteriin l. Olkoon m_l klusterin l edustaja, se voi olla keskimmäinen havainto tai jotenkin muuten muodostettu. Nyt Q määritellään $N \times N$ matriisina, jonka elementti $Q(i, i') = d(m_l, m_{l'})$, kun $C(i) = l$ ja $C(i') = l'$. Toisin sanoen Q sisältää parittaisten havaintojen klusterien edustajien väliset etäisyydet. Siten suurten klusterien tapauksessa matriisissa on paljon samoja arvoja.

Muokattu Hubertin Γ etsii erityisesti kompakteja klustereja, josta suuret arvot ovat indikaatioita. Klusterien lukumäärille $k = 1$ ja $k = N$ indeksiä ei ole määritelty. Satunnaiselle datalle se omaa nousevan trendin k:n kasvaessa, joten kun haetaan sopivinta klusterien lukumäärää, etsitään suurinta paikallista muutosta. Mikäli verrataan eri algoritmeilla saatuja klusterointeja toisiinsa kiinnitetyllä ryhmien lukumäärällä, valitaan suurimman indeksin arvon tuottava klusterointi.

5.3 Satunnaisuuden testaaminen

Klusteroinnin validoinnin yhteydessä ei pyritä normaalin tilastollisen testaamisen tapaan testaamaan jonkin parametrin merkitsevyyttä, vaan nollahypoteesi muotoillaan hieman eri tavalla: halutaan testata sitä, onko aineiston rakenne satunnainen vai ei.

Käytännössä testejä ei ole mahdollista määrittää analyyttisesti, vaan testisuureiden jakaumat on laskettava empiirisesti käyttämällä simulointi- tai satunnaistoistomenetelmiä eli *Monte Carlo* - tai *Bootstrap*-tekniikoita. Tässä tutkielmassa käytetään Monte Carlo -menetelmiä. Bootstrap-tekniikoiden avulla voidaan esimerkiksi tutkia yksittäisen klusterin uudelleen muodostumista (Shimodaira 2002).

Nyt nollahypoteesi H_0 muotoillaan tarkoittamaan, että aineiston rakenne on satunnainen. Tavoitteen voi saavuttaa vertaamalla klusteroinnin tulosta satunnaisesti tuotettuun dataan tai satunnaisesti tuotettuun klusterointiin. Tätä varten täytyy määritellä sopiva tilastollinen tunnusluku q, joka kuvaa aineiston rakennetta ja jonka avulla vertailu voidaan tehdä. Kaikki edellä arviointikriteerien yhteydessä esitellyt tunnusluvut ovat tähän tarkoitukseen sopivia testisuureita.

Varsinaisessa testaamisessa edetään seuraavasti:

1. Klusteroi aineisto X ja laske tunnusluku q^* käyttäen läheisyysmatriisia D ja saatua jaotusta C.

2. Toista iteraatioille $t = 1, ..., T_{max}$:

   Generoi satunnainen data X_t, muodosta sille läheisyysmatriisi D_t, klusterointi C_t ja laske tunnusluku q_t.

3. Vertaa tunnusluvun q^* arvoa saatuun q_t:n empiirisen otantajakaumaan. Mikäli q^* on itseisarvoltaan epätavallisen suuri eli sijaitsee generoitujen q_t arvojen histogrammin kriittisellä alueella, ennalta valitulla merkitsevyystasolla α, satunnaisuuden hypoteesi hylätään.

5.3.1 Satunnaisen datan generointi

Satunnaisen sijainnin hypoteesi. (*Random position hypothesis*)

Vaatimuksena on, että kaikki datan sijainnit, tietyllä alueella p-ulotteisessa avaruudessa, ovat yhtä todennäköisiä. Yksi tapa tälläisen tilanteen luomiseen on asettaa jokainen piste satunnaisesti tälle alueelle tasajakauman mukaan. Satunnaisen sijainnin hypoteesia voidaan käyttää sisäisen tai ulkoisen kriteerin kanssa.

Sisäisen kriteerin tapauksessa valitaan tunnusluku q, esimerkiksi kofeneettinen korrelaatiokerroin, mittaamaan missä määrin klusteroinnin tuottama rakenne vastaa aineistolle laskettua läheisyysmatriisia. Tätä testataan suhteessa uudelle datalle muodostettuihin läheisyysmatriiseihin, ja niitä vastaaviin klusterointeihin kuten yllä.

Ulkoisen kriteerin tapauksessa määritellään tilastollinen tunnusluku q, esimerkiksi Randin tunnusluku, mittaamaan yhtenevyyden tasoa aineistolle X ennalta määrätyn rakenteen ja tietyn klusterointialgoritmin X:lle tuottaman osituksen välillä. Yllä olevassa algoritmissa läheisyysmatriisin tilalla on tunnetuksi oletettu jaotus $\wp$ ja generoiduilla datoilla jaotukset $\wp_t$. Muuten algoritmi etenee yhtenevästi, satunnaisuuden hypoteesi hylätään, mikäli q^* on epätavallisen suuri.

Satunnaisen graafin hypoteesi. (*Random graph hypothesis*)

Tätä käytetään yleensä, kun aineistosta on vain sisäistä informaatiota ja kun havaintovektorien väliset läheisyydet on määritelty järjestysasteikollisina. Tässä tapauksessa vertailuaineistoksi generoidaan järjestysasteikollisia symmetrisiä $N \times N$ matriiseja A_t. Generointi tehdään asettamalla kokonaisluvut väliltä

$[1, N(N-1)/2]$ satunnaiseen järjestykseen matriisin yläkolmioon, takaisin palauttamatta.

Olkoon D aineistoon X liittyvä järjestysasteikollinen läheisyysmatriisi ja C saatu klusterirakenne. Kun C_t:t ovat saman klusterointialgoritmin A_t:lle tuottamia klusterointeja, voidaan toimia kuten yllä. Määritellään tunnusluku q, esimerkiksi Hubertin Γ, joka mittaa läheisyysmatriisin ja klusteroinnin välistä assosiaatiota. Aineistoon liittyvän q^* arvoa testataan generoitujen vertailuaineistojen avulla, ja epätavallisen suuret arvot aiheuttavat satunnaisuushypoteesin hylkäämisen.

Satunnaisen osituksen hypoteesi. (*Random label hypothesis*)

Tässä oletetaan, että kaikki mahdolliset aineiston X ositukset k:hon ryhmään ovat yhtä todennäköisiä. Satunnaisen osituksen hypoteesi vaatii ennalta määrätyn ulkoisen osituksen, eli sitä käytetään vain ulkoisen kriteerin tapauksessa. Testattava tunnusluku q voidaan määritellä mittaamaan sitä, kuinka hyvin aineistosta laskettu läheisyysmatriisi vastaa tiettyä ulkoisesti määrättyä ositusta. Tähän sopii esimerkiksi Hubertin Γ. Merkitsevyystestauksessa q^*:n arvoa verrataan satunnaisen osituksen mukaan generoidun aineiston vastaaviin q_t:n arvoihin. Jälleen H_0 hylätään, mikäli q^* on epätavallisen suuri.

6 Menetelmien soveltaminen aineistoon

6.1 Aineiston ja tutkimusongelman esittely

6.1.1 Taustaa

Yliopistojen tilastotieteen opetusta ja oppimista sivuaineena pidetään haasteellisena, etenkin yhteiskunta- ja kasvatustieteissä. Varsinkin monet opettajaksi opiskelevat kokevat tilastollisten menetelmien opiskelun vain maisterintutkinnon saamista vaikeuttavana esteenä (Väisänen & Ylönen 2004). Väisänen ja Ylönen ovat tutkineet opiskelijoiden matemaattisten taitojen ja minäkäsityksen merkitystä tilastotieteen peruskurssin tenttimenestykseen. Heidän aineistonsa on kerätty Savonlinnan opettajankoulutuslaitoksen tilastotieteen peruskurssin opiskelijoista (120 opiskelijaa, joista 83.3% naisia) syksyllä 2002.

Analyysimenetelminä Väisänen ja Ylönen ovat käyttäneet regressio- ja ryhmittelyanalyysia. Klusteroinnissa he ovat soveltaneet K-means-menetelmää ja löytäneet neljä erilaista oppijaryhmää. Hyvin menestyvien ja heikommin menestyvien ryhmiä on molempia kaksi.

Hyvin menestyviä ovat opiskelijat, joilla on hyvä matemaattinen tausta, myönteinen matemaattinen minäkäsitys ja myönteinen minäkuva itsestään tilastotieteen oppijana. Toisena hyvin menestyvien ryhmänä on tunnistettu opiskelijat, joilla hyvästä matemaattisesta taustastaan huolimatta on edelleen kielteisempi käsitys itsestään matematiikan ja tilastotieteen oppijana sekä heikko motivaatio.

Heikosti menestyvien ensimmäisen ryhmän oppilaat ovat heikoimpia kaikilla tutkimuksen mittareilla, Väisänen ja Ylönen kuvaavat heitä ”epätoivoisiksi ja äärimmäisen heikoiksi oppijoiksi”. Viimeistä ryhmää luonnehditaan melko heikot taidot omaaviksi, mutta asennetekijöiltään keskitasoisiksi oppijoiksi.

Sukupuolella on havaittu olevan tilastollisesti merkitsevä vaikutus matemaattiseen minäkäsitykseen ja luottamukseen omiin kykyihin. Niin Väisäsen ja Ylösen kuin aikaisemmissakin tutkimuksissa on huomattu, että miehet luottavat enemmän itseensä tilastotieteen ja matematiikan oppijoina. Lisäksi Väisäsen ja Ylösen aineistossa kaikki miesopiskelijat (13) sijoittuvat joko ensimmäiseen tai viimeiseen ryhmään, eli heidän minäkuvansa ja asennetekijänsä ovat hyvää tasoa tai keskitasoa. (Väisänen & Ylönen, 2004)

6.1.2 Aineisto ja muuttujien valinta

Käytettävissä oleva tutkimusaineisto on peräisin 2000-luvulla Jyväskylän yliopistossa opetetulta tilastomenetelmien peruskurssilta. Kurssilta on haluttu selvittää, millaisia oppijaryhmiä Jyväskylän yliopistossa on. Tätä varten on toteutettu kurssikysely, joka pohjautuu osittain Väisäsen ja Ylösen (2004) kyselyn osioihin.

Suoritettu kysely jakautuu ajankohdan mukaan kahteen erilliseen kyselyyn, joissa yhteensä on ollut 68 kysymystä. Alkukyselyssä on selvitetty opiskelijoiden taustaa sekä käsitystä ja uskomuksia itsestään tilastotieteen oppijoina. Loppu-

kyselyssä on keskitytty kurssin arviointiin sekä mitattu opintomenestystä tenttiarvosanan avulla.

Alkukyselyyn vastanneita on ollut huomattavasti loppukyselyyn vastanneita enemmän, toisaalta loppukyselyyn vastanneiden joukossa on ollut myös alkukyselyyn vastaamattomia opiskelijoita. Kaikkiaan kyselyyn vastanneita opiskelijoita kurssilla on ollut yhteensä 216. Alkukyselyn kysymyksiin on vastannut yhteensä 188 opiskelijaa, joista naisten osuus on 58.5%. Loppukyselyyn vastanneita on 110, mutta heistäkään osa ei ole vastannut kaikkiin osioihin.

Aineistosta halutaan selvittää ensisijaisesti, millaisiin ryhmiin opiskelijat jakautuvat matemaattisten taitojen, uskomusten ja minäkäsityksen perusteella. Toisaalta kiinnostavaa olisi selvittää, miksi loppukyselyyn vastanneiden lukumäärä on selvästi pienempi kuin alkukyselyyn vastanneiden; jättävätkö heikot taidot tai huonon motivaation omaavat oppilaat vastaamatta loppukyselyyn muita enemmän?

Loppukyselyn reilusti pienemmän vastausten lukumäärän vuoksi valitaan klusteroitavaksi pelkkä alkukysely. Se koostuu kuudesta eri asioita mittaavasta osiosta. Osioiden yhdistettävissä olevista muuttujista muodostetaan uusia summamuuttujia. Osioiden kysymysten lukumäärä ja vastausten asteikko vaihtelevat sangen paljon, joten summamuuttujat ovat keskenään heterogeenisiä. Kaikki niistä eivät välttämättä kuvaa tutkittavaksi tarkoitettua ominaisuutta kovin hyvin.

Summamuuttujia muodostettaessa huomataan lisäksi, että aineistossa on muutamia vastaajia, jotka ovat jättäneet vastaamatta johonkin summattavista kysymyksistä. Näissä tapauksissa käytetään imputointia - puuttuva vastaus korvataan kysymysmuuttujan mediaanilla.

Muuten puuttuvien vastausten käsittely havaintoparin erilaisuusmittaa laskettaessa hoidetaan luvussa 2.4 kuvatulla tavalla: toimitaan siten, että mikäli toisen opiskelijan vastaus kysymykseen puuttuu aineistosta, kyseistä kysymystä ei huomioida opiskelijaparin erilaisuuden laskussa. Kun loppukysely poistetaan aineistosta, jäljelle ei jää yhtään havaintoparia, missä yhtään yhteistä vastattua kysymystä ei olisi saatavissa.

Seuraavaksi jatkuvalla asteikolla mitatut muuttujat luokitellaan. Siten analyyseihin on käytettävissä täysin diskreetti aineisto, jossa muuttujat ovat joko ordinaalisia tai binäärisiä. Luonnollinen valinta erilaisuusmitaksi tämäntyyppiselle aineistolle on Manhattan-etäisyys (luku 2.3). Aineiston yhteensä 15 klusteroinnissa käytettävää muuttujaa ja viisi taustamuuttujaa on esitelty liitteessä A.

Lähdetään tarkastelemaan esille nousseita kysymyksiä soveltamalla klusterointimenetelmiä alkukyselyyn. Tämän jälkeen tutkitaan loppukyselyyn vastaamista mahdollisesti löydetyissä ryhmissä. Tätä tarkoitusta varten lisätään alkukyselyyn muuttuja, joka kertoo onko opiskelija vastannut loppukyselyyn. Aineiston analysointi tehdään R-ohjelmistolla (2010).

6.2 Aineiston klusterointi

6.2.1 Klusterointimenetelmien valinta ja toteutus

Klusteroidaan aineisto soveltamalla kaikkia luvun 4.2 yhdisteleviä hierarkkisia menetelmiä sekä luvun 4.3 jakavaa hierarkkista menetelmää, kun erilaisuusmittana on Manhattan-etäisyys. Tämän jälkeen suoritetaan tulosten alustava tarkastelu käyttämällä ainoastaan saatuja dendrogrammeja. Niistä voidaan päätellä mahdollisesti sopiva klustereiden lukumäärä K-medoids-algoritmin parametriksi k (luku 3.3).

Klusterointien tuottamat dendrogrammit on esitetty kuvissa 6, 7 ja 8. Kuvassa 6 alhaalla olevassa lähimmän naapurin menetelmän tuottamassa dendrogrammissa nähdään voimakasta ketjuttumista, eli siinä yksittäisiä havaintopisteitä liittyy edeltävään klusteriin lähes jatkuvasti. Näin selkeitä klustereita ei kuvan perusteella erotu, vaan kaikki havainnot ovat suhteellisen lähellä toisiaan. Siten klustereita olisi vain yksi.

Kuvan 6 ylhäällä on keskiarvomenetelmän tuottama dendrogrammi. Sen alaosista voi ehkä löytää jonkinlaista toisistaan erottuvaa rakennetta, mutta ylemmillä tasoilla erot ovat pieniä ja ketjuttumista näkyy. Useamman kuin yhden klusterin olemassaolosta ei näytä olevan näyttöä.

Kuvassa 7 kaukaisimman naapurin menetelmän dendrogrammi näyttäisi jo sisältävän selkeää klusterirakennetta. Kuvasta erottuu ylhäällä kaksi haaraa, joten sen perusteella aineistosta löytyisi ainakin kaksi klusteria.

Kuvan 7 alhaalla Wardin menetelmän tuottama dendrogammi näyttää tähänastisista rakenteista tasapainoisimmalta. Se jakautuu useaan otteeseen melko erottuvasti, joten myös jaot useampaan kuin kolmeen klusteriin näyttävät mahdollisilta.

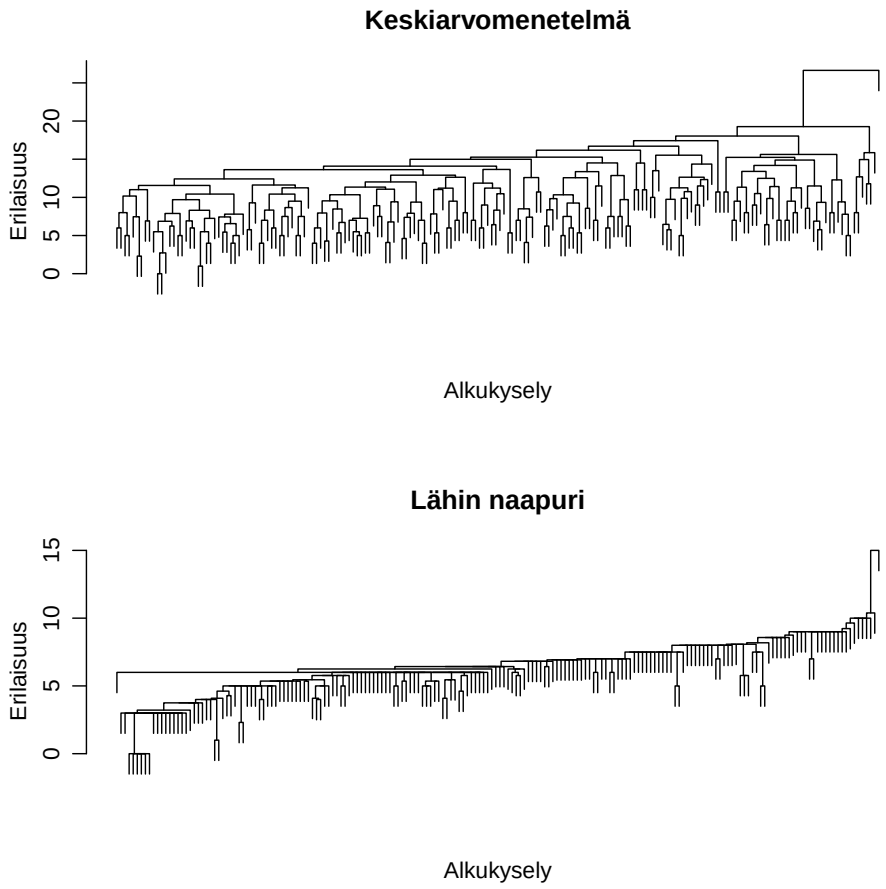
Kuvassa 8 on esitetty jakavan menetelmän tuottama dendrogrammi. Siitä nähdään, että ensimmäiset kaksi jakoa tehdään samalla tasolla, jolloin klustereita muodostuu heti kolme. Mielenkiintoisesti jako kolmeen klusteriin on varsin selkeä, sillä alemmilla tasoilla jakautumiset eivät ole selvästi erottuvia.

Sitten toteutetaan vielä K-medoids klusterointi antaen ryhmien lukumäärän vaihdella välillä $k = 2, ..., 8$. Tulosten visualisointiin moniulotteiselle aineistolle ei löydy hierarkkisten menetelmien dendrogrammeja vastaavaa havainnollista tapaa, joten tässä vaiheessa K-medoids-menetelmän tuottamat ositukset ainoastaan talletetaan validointia varten, jota edetään suorittamaan seuraavaksi.

6.2.2 Validointi

Nyt käytössä ei ole ulkoisesti saatua ositusta, eikä täten luvun 5.2.1 ulkoisen kriteerin menetelmiä voida käyttää. Sisäisen kriteerin tarkastelut klusterihierarkioille voidaan tehdä laskemalla kofeneettiset korrelaatiot. Ne on koottuna taulukkoon 1.

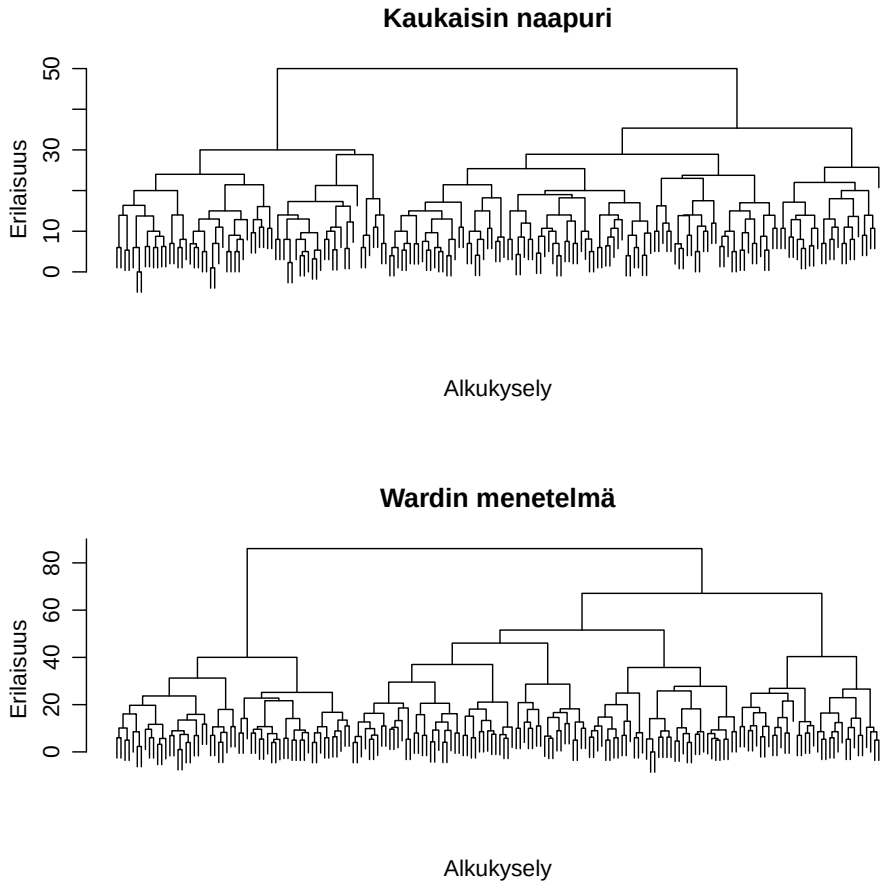
Kaikki saadut kofeneettisen korrelaatiokertoimen arvot näyttävät suurehkoilta korrelaatioilta, mutta kuten luvussa 5.2.2 on todettu, suuret $CPCC$:n arvot ovat varsin yleisiä. Satunnaisuuden testauksen rooli on niiden yhteydessä tärkeää.



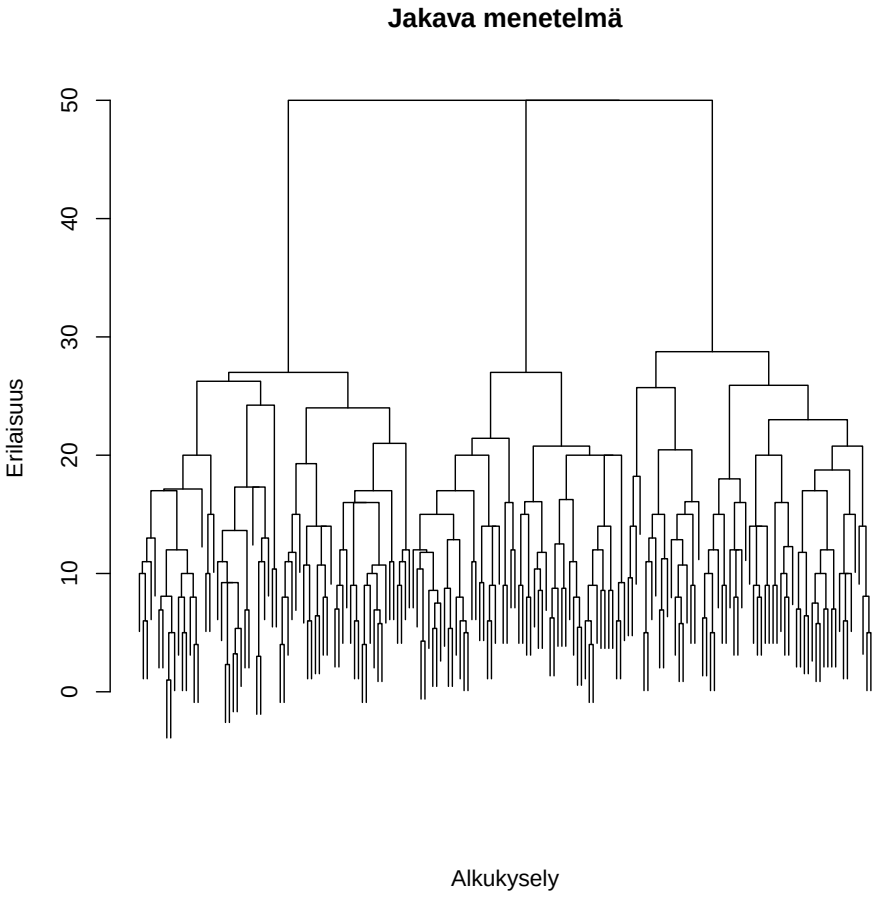
Kuva 6: Keskiarvomenetelmän ja lähimmän naapurin menetelmän tuottamat dendrogrammit, erilaisuusmittana Manhattan-etäisyys.

Taulukko 1: Kofeneettiset korrelaatiokertoimet.

Menetelmä	CPCC
Keskiarvo	.741
Lähin naapuri	.667
Kaukaisin naapuri	.941
Ward	.892
Jakava	.860



Kuva 7: Kaukaisimman naapurin ja Wardin menetelmän tuottamat dendrogrammit, erilaisuusmittana Manhattan-etäisyys.



Kuva 8: Jakavan menetelmän tuottama puurakenne, erilaisuusmittana Manhattan-etäisyys.

Koska dendrogrammien perusteella ei ole nähtävissä kunnollista rakennetta keskiarvomenetelmälle eikä lähimmän naapurin menetelmälle, ne pudotetaan tässä vaiheessa pois jatkoanalyyseista. Siten niiden tuottamien kofeneettisten korrelaatiokertoimien satunnaisuutta ei nyt testata. Mikäli jatkossa osoittautuisi, että aineistossa on yksi tai ei yhtään klusteria, voitaisiin mahdollisesti ottaa keskiarvomenetelmän ja lähimmän naapurin menetelmän tuottamat hierarkiat uudelleen tarkasteltavaksi.

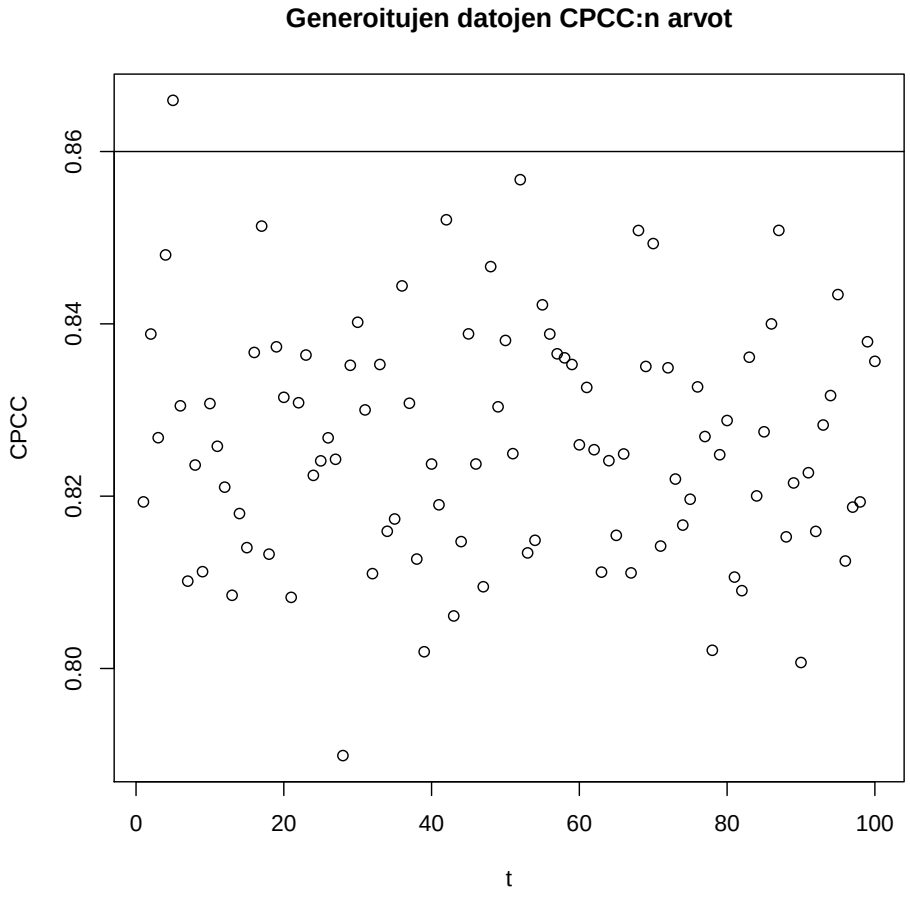
Kaukaisimman naapurin kofeneettisen korrelaatiokertoimen arvo 0.941 kertoo vahvasta riippuvuudesta siihen liittyvien läheisyysmatriisin ja kofeneettisen matriisin välillä. Wardin menetelmän tuottaman klusteroinnin $CPCC = 0.892$ on niinikään varsin suuri, mutta tulkinnallisesti selkeimmän rakenteen tuottanut jakava menetelmä saa $CPCC$:n arvon 0.860, jonka satunnaisuutta testataan ensimmäiseksi.

Satunnaisuuden testaus tehdään luvun 5.3 mukaan, kun käytössä on sisäisen kriteerin tunnusluku $q = CPCC$. Satunnainen aineisto generoidaan käyttäen satunnaisen sijainnin hypoteesia niin, että luodaan T_{max}-kertaa 188 viisitoistaulotteista vektoria peruskurssiaineistoa vastaaviksi. Toisin sanoen generoidaan satunnaisia vastauksia alkukyselyyn: jokaiseen 188 vastausta ja toistetaan T_{max}-kertaa. Jokaiselle kyselylle tehdään klusterointi jakavalla menetelmällä ja lasketaan $CPCC$. Testauksessa tarvittaviksi parametreiksi valitaan yleisesti käytetyt toistojen määrä $T_{max} = 100$ ja merkitsevyystaso $\alpha = 0.05$ (Theodoridis & Koutroumbas 2006) .

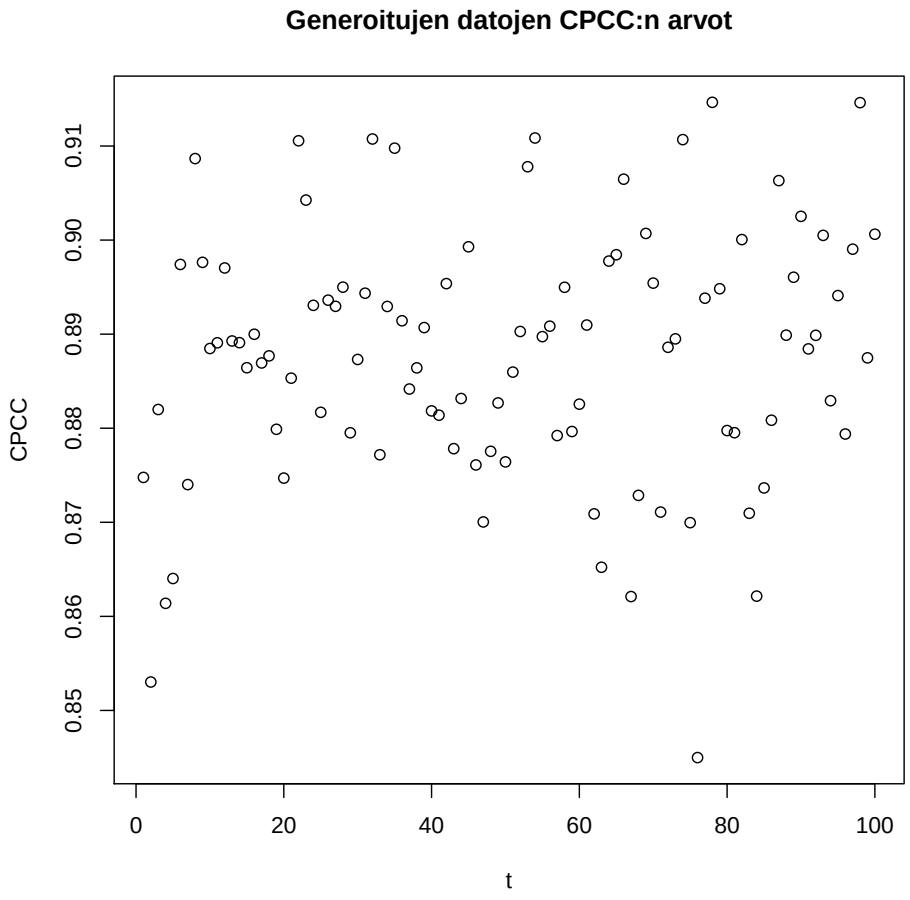
Tulokset on esitetty kuvassa 9. Nähdään, että tämäntyyppisille aineistoille jakava menetelmä näyttää tuottavan varsin kapealle välille asettuvia kofeneettisen korrelaatiokertoimen arvoja. Yksi arvo on suurempi kuin oikealla aineistolla, mutta muut 99 ovat pienempiä. Näin ollen satunnaisuushypoteesi hylätään valitulla merkitsevyystasolla $\alpha = 0.05$. Jakavan menetelmän tuottama rakenne ei ole satunnainen.

Seuraavaksi suoritetaan vastaavasti satunnaisuuden testaus kaukaisimman naapurin sekä Wardin menetelmän tuottamille hierarkioille kofeneettista korrelaatiokerrointa käyttäen. Kuvasta 10 löytyvät kaukaisimman naapurin menetelmällä generoidulle aineistolle saadut $CPCC$:n arvot. Ne asettuvat välille $[0.840, 0.920]$ eli ovat kaikki pienempiä kuin sovellusaineiston tuottama 0.941. Siten kaukaisimman naapurin menetelmää käyttäen saatu klusterirakenne ei ole satunnainen.

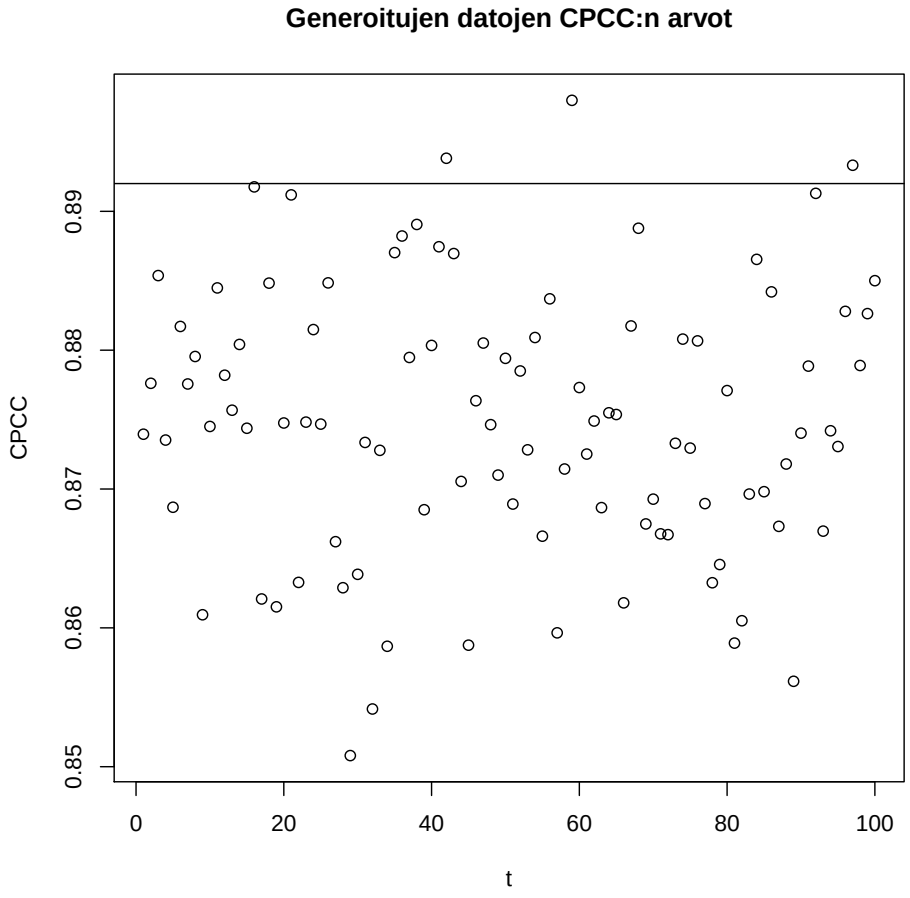
Kuvassa 11 ovat kofeneettiset korrelaatiokertoimet satunnaiselle aineistolle, kun klusterointiin on käytetty Wardin menetelmää. Vertailuarvoa 0.892 suurempia arvoja löytyy kolme kappaletta, joten satunnaisuushypoteesi hylätään myös tässä tapauksessa. Kuitenkin kuvan 97:stä vertailuarvoa pienemmästä $CPCC$:n arvosta kolme on hyvin lähellä sitä. Johtopäätös satunnaisuuden hylkäämisestä ei siten ole niin selvä. Otetaan tämä huomioon jatkossa niin, että mikäli Wardin menetelmän hierarkia osoittautuu suhteellisen kriteerin tarkasteluissa parhaaksi, tulkinnat tehdään varoen.



Kuva 9: CPCC:n arvot laskettuna simuloiduille satunnaisille aineistoille jakavalla menetelmällä, poikkiviiva merkkinä vertailuarvosta 0.860.



Kuva 10: Kofeneettisen korrelaatiokertoimen arvot kaukaisimman naapurin menetelmää käyttäen simuloidulle aineistolle. Vertailuarvo 0.941 ei mahdu kuvaan.



Kuva 11: Wardin menetelmää käyttäen saadut CPCC:n arvot generoidulle aineistolle. Sovellusaineistosta saatu vertailuarvo 0.892 näkyy poikkiviivana.

Suhteellisen kriteerin tarkastelut

Edellä tutkittiin kokonaisten klusterihierarkioiden hyvyyttä sisäistä kriteeriä käyttäen. Seuraavaksi siirrytään tarkastelemaan katkaistuja hierarkioita. Kuvien 6, 7 ja 8 dendrogrammien perusteella parhaiten useaan ryhmään jaettavissa on Wardin menetelmän tuottama puurakenne.

Tarkastellaan klusterien lukumäärän valintaa Wardin menetelmän hierarkialle luvun 5.2.3 keinoin. Tehdään katkaisut tasoilla $k = 2, ...8$; suoritusindeksinä käytetään normalisoitua muokattua Hubertin gammaa. Muita vaihdeltavia parametreja klusterien lukumäärän lisäksi ei tässä tapauksessa ole.

Tulokset löytyvät kuvasta 12. Koska Hubertin gammalla on taipumusta kasvaa klusterien lukumäärän mukana, etsitään suurinta paikallista muutosta. Sellainen löytyy kolmen klusterin kohdalta. Ennen kuin vertaillaan eri menetelmien tuottamia kolmen klusterin ratkaisuja, tehdään vastaava tarkastelu vielä kombinatorista K-medoids-klusterointimenetelmää käyttäen.

Kuvasta 13 nähdään, että myös keskimmäisiä havaintoja käyttävän ositusmenetelmän soveltaminen johtaa samaan johtopäätökseen ryhmien lukumäärästä tälle aineistolle. Nyt hyppäys indeksin arvossa siirryttäessä kahdesta kolmeen klusteriin on vielä suurempi kuin Wardin menetelmällä, jonka jälkeen arvot tasaantuvat. Kolme klusteria vaikuttaa selvästi parhaalta ratkaisulta, joten verrataan käytettyjä menetelmiä käyttämällä suoritusindeksinä normalisoitua muokattua Hubertin gammaa.

Suoritusindeksin arvot ovat taulukossa 2. Jakavan menetelmän kolmen ryhmän tuottama arvo $\hat{\Gamma} = 0.337$ on selvästi suurin, joten kuvassa 8 hyvin näkyvä rakenne vaikuttaa parhaalta. Kaukaisimmalle naapurille katkaisu kolmeen klusteriin näyttää sopivan huonommin.

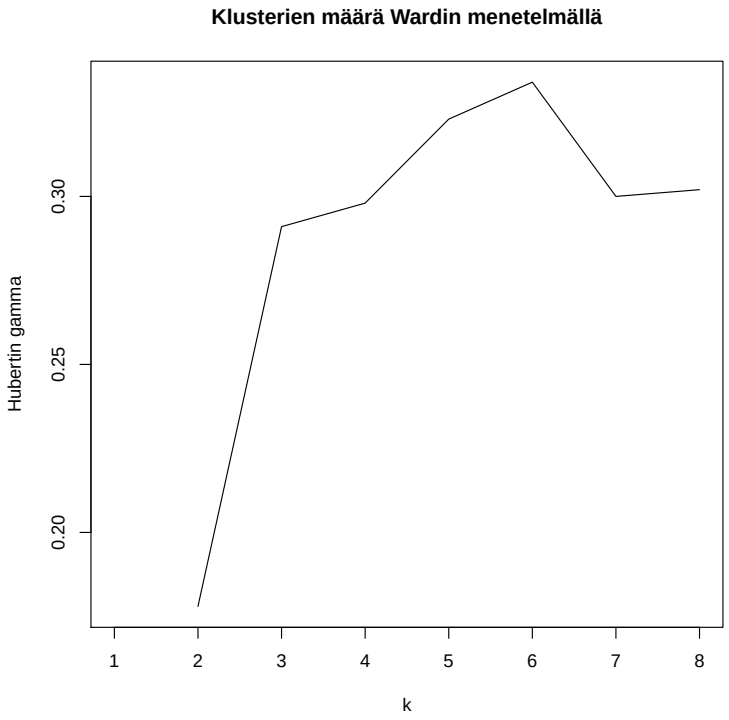
6.2.3 Klusteroinnin tulkinta

Tarkastellaan jakavan menetelmän tuottamaa jakoa kolmeen klusteriin. Ryhmien koot ovat 70 (37.2 %), 55 (29.3 %) ja 63 (33.5 %). Liitteestä B.1 löytyvät klustereiden jakaumat jokaisen muuttujan suhteen. Nyt tehtävät tulkinnat ovat kuvailevia, varsinaista tilastollista testausta ei tässä yhteydessä tehdä.

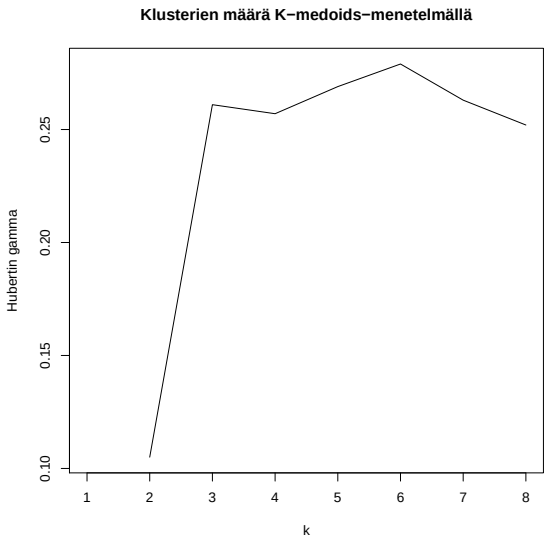
Liitettä B.1 tarkastelemalla nähdään, että kaksi ensimmäistä klusteria sisältävät matemaattisesti lahjakkaita opiskelijoita ja kolmas matemaattisilta taidoiltaan heikompia opiskelijoita. Ensimmäisen klusterin oppilaiden matemaattiset taidot sekä uskomukset omista kyvyistään ovat parhaita. He eivät välttämättä tarvitse tilastotiedettä opinnoissaan, mutta suhtautuvat siihen positiivisesti. Heidän matemaattiset taustansa ovat vahvat ja ajankäyttö opiskelun suhteen keskitasoa.

Toisen klusterin oppilaat vaikuttavat melko hyvät matemaattiset taidot omaavilta, tilastotiedettä eniten tarvitsevilta sekä omiin kykyihinsä uskovilta. Asennetekijöiltään ja suhtautumisessaan tilastotieteeseen he ovat keskitasoisia, mutta käyttävät eniten aikaa opiskeluunsa.

Kolmanteen klusteriin kuuluu eniten heikot taidot omaavia sekä heikoimmin itseensä uskovia oppilaita. He suhtautuvat tilastotieteeseen negatiivisesti ja



Kuva 12: Normalisoidun muokatun Hubertin gamman arvot eri klusterien lukumäärillä k.



Kuva 13: Klusterien lukumäärän valinta K-medoids-menetelmälle normalisoidun Hubertin gamman avulla.

Taulukko 2: Normalisoidun muokatun Hubertin gamman arvot.

Menetelmä, $k = 3$	$\hat{\Gamma}$
Kaukaisin naapuri	.262
Ward	.291
Jakava	.337
K-medoids	.261

yli puolet heistä pelkää, ettei läpäise kurssia. Kolmannen klusterin opiskelijoiden taustat ovat lisäksi heikoimmat ja enemmistö lyhyen matematiikan lukiossa suorittaneista kuuluu tähän ryhmään. Heidän joukossaan on eniten töitä opiskelujen ohella tekeviä.

Liitettä B.1 tarkastelemalla nähdään lisäksi se, että joidenkin muuttujien kohdalla erot klusterien välillä ovat pieniä, mikä voi selittää keskiarvomenetelmän ja lähimmän naapurin menetelmien dendrogrammeissa esiintyvää voimakasta ketjuttumista. Erityisesti sellaisten muuttujien kohdalla, joiden arvoasteikko on leveä, vastaukset asettuvat melko keskelle, huonosti erottuen. Näin ollen ”turhien” luokkien yhdistämiseen olisi voinut olla aihetta.

Tutkitaan seuraavaksi klustereihin jakautumista taustamuuttujien suhteen. Taulukoista 3, 4 ja 5 löytyvät taustamuuttujien; sukupuoli, tiedekunta ja loppukyselyyn vastaaminen, luokkien prosenttiosuudet klustereittain. Liitteessä B.2 on esitetty jokaisen viiden taustamuuttujan esiintyvyydet frekvenssitaulukkona.

Taulukosta 3 nähdään, että naisten osuus kasvaa, kun siirrytään ensimmäisestä ryhmästä kolmanteen ryhmään. Ensimmäisessä klusterissa naisia on 48.5 %, toisessa 60.0 % ja kolmannessa 68.3 %, kun koko aineistossa naisten osuus on 58.5 %. Siten sukupuolella näyttäisi olevan merkitystä opiskelijoiden matemaattisiin taitoihin ja luottamukseen omiin kykyihinsä. Naisten osuus heikoimmin menestyvien sekä vähiten itseensä uskovien ryhmässä on suurempaa kuin muissa ryhmissä, mikä on yhdenmukaista aiempien tutkimusten kanssa (Väisänen & Ylönen 2004).

Taulukosta 4 löytyy opiskelijoiden tiedekunnat klustereittain. Kasvatustieteiden tiedekunnasta opiskelijoita on vain 2.7 %, joten aineisto eroaa vahvasti Väisäsen ja Ylösen aineistosta. Toinen pienen edustuksen osajoukko on Humanistinen tiedekunta. Keskisuuria tiedekuntia kyselyssä ovat Liikunta- ja terveystieteiden sekä Matemaattis-luonnontieteellinen tiedekunta. Liikunta- ja terveystieteiden tiedekunnan opiskelijat jakautuvat tasaisesti kaikkiin klustereihin, mutta Matemaattis-luonnontieteellisen tiedekunnan opiskelijoista valtaosa kuuluu ensimmäiseen klusteriin, kuten oli odotettavissa.

Loput kolme tiedekuntaa ovat vahvasti edustettuina aineistossa yli 20 prosentin osuuksilla. Näistä Taloustieteiden tiedekunnan opiskelijat jakautuvat jokseenkin tasaisesti kaikkiin ryhmiin niin, että toinen klusteri on yleisin. Informaatioteknologian tiedekunnasta valtaosa kuuluu ensimmäiseen, parhaiten menestyvien, myönteisen matemaattisen minäkuvan sekä positiivisesti tilastotieteeseen suhtautuvien ryhmään. Tämäkään ei yllätä, vaan lienee yhteydessä sukupuolen vaikutukseen, sillä informaatioteknologia on tunnetusti varsin miesvaltainen ala. Viimeisimpänä Yhteiskuntatieteellisen tiedekunnan opiskelijoista suurin osa kuuluu kolmanteen heikoimmat taustat ja negatiivisen asenteen omaavien ryhmään.

Yhteenvetona opiskelijoiden tiedekunnista klustereittain todetaan vielä tyyppiarvot, jotka tulevat suurimman edustuksen tiedekunnista. Ensimmäisen klusterin moodi on Informaatioteknologian tiedekunnan opiskelija, joita on 40 % klusterin opiskelijoista. Toisessa klusterissa liki 30 % kuuluu Taloustieteiden tiedekuntaan. Kolmannen klusterin opiskelijoista lähes 50 % tulee Yhteiskuntatieteellisestä tiedekunnasta.

Taulukossa 5 on esitetty loppukyselyyn vastaamisen jakauma klustereittain. Hieman yllättäen kolmas, eli heikoiten menestyvien ryhmä on ainoa, jossa loppukyselyyn vastanneita on hieman enemmän kuin vastaamattomia. Ensimmäisessä ryhmässä vastaamattomia on eniten. Voidaan todeta, että opiskelijoiden heikot matemaattiset taidot sekä heikko motivaatio eivät näyttäisi ainakaan heikentävän halukkuutta vastata loppukyselyyn.

Taulukko 3: Sukupuolen {muuttuja SP} jakautuminen klustereissa prosentteina: $1 = nainen$, $2 = mies$.

C_l	1	2	Σ
C_1	48.5	51.5	100
C_2	60.0	40.0	100
C_3	68.3	31.7	100
X	58.5	41.5	100

Taulukko 4: Opiskelijan tiedekunnan {TDK} jakautuminen klustereissa prosentteina: *1=Humanistinen, 2=Informaatioteknologian, 3=Kasvatustieteiden, 4=Liikunta-* ja *terveystieteiden, 5=Matemaattis-luonnontieteellinen, 6=Taloustieteiden, 7=Yhteiskuntatieteellinen.*

C_l	1	2	3	4	5	6	7	Σ
C_1	2.9	40.0	4.3	10.0	20.0	14.3	8.6	100
C_2	0	21.8	1.8	10.9	9.1	29.1	27.3	100
C_3	4.8	14.3	1.6	6.3	3.2	22.2	47.6	100
X	2.7	26.0	2.7	9.0	11.1	21.3	27.1	100

Mikäli vielä verrataan löydettyjä opiskelijaryhmiä Väisäsen ja Ylösen opettajankoulutuslaitokselle tekemään klusterointiin, niin nähdään, että sekä taidoiltaan että motivaatiotekijöiltään hyvien ja heikkojen opiskelijoiden ryhmät löytyivät molemmista. Joskin luonnehdinta ”epätoivoiset ja äärimmäisen heikot oppijat” olisi selvästi liioiteltu kuvaamaan nyt löydettyä heikoimpien opiskelijoiden ryhmää.

Sen sijaan Väisäsen ja Ylösen löytämiä kahta muuta ryhmää vastaavia klustereja ei löydetty. Kahden ääripään väliin jäävä klusteri on nyt varsin hyvät taidot omaava, mutta myös taitoihinsa uskova eikä päinvastoin, kuten Väisäsen ja Ylösen tutkimuksen toisen menestyvän ryhmän kohdalla. Erot löydettyyn, taidoiltaan vahvimpien opiskelijoiden klusteriin tulevat nyt selvimmin esiin asennetekijöissä sekä suhtautumisessa tilastotieteeseen.

Taulukko 5: Loppukyselyyn vastaamista kuvaavan muuttujan {LK} jakautuminen klustereissa prosentteina: $0 = ei$, $1 = kyllä$.

C_l	0	1	Σ
C_1	60.0	40.0	100
C_2	56.4	43.6	100
C_3	46.0	54.0	100
X	54.3	45.7	100

7 Yhteenveto

Tässä tutkielmassa on käsitelty ryhmittelyä eli klusterointia. Saatavilla olevista lukuisista menetelmistä on esitelty vanhimpia, mutta edelleen paljon käytettyjä, hierarkkisia menetelmiä. Lisäksi on tarkasteltu yleisellä tasolla suurinta eri menetelmien kokonaisuutta, kombinatorisia menetelmiä. Kombinatorisista menetelmistä on esitelty tarkemmin ehkä tunnetuimpana klusterointimenetelmänä pidetty K-means-keskiarvomenetelmä sekä sen robusti versio K-medoids.

Kyseiset menetelmät on valittu tähän tutkielmaan ennen kaikkea niiden laajojen sovellusmahdollisuuksien takia. Lisäksi niiden idea ja toiminta on helppo ymmärtää. Niiden laaja käyttö yleisesti tutkimuksissa osoittaa, että menetelmät toimivat käytännössä, ja niitä voi käyttää ratkaisemaan monenlaisia ryhmittelyongelmia.

Tilastotieteilijän näkökulmasta mahdollisesti mielenkiintoisempia tai mielekkäämpiä vaihtoehtoja olisivat olleet probabilistiset menetelmät eli menetelmät, joihin liittyy todennäköisyysmalli. Menetelmiin, joita nyt käsiteltiin liittyy paljon epävarmuutta, jota esimerkiksi monet kyseiset menetelmät sisältävät tilasto-ohjelmat eivät suoraan käsittele mitenkään. Tämän vuoksi tässä työssä on panostettu vahvasti klusterointimenetelmien esittelyn lisäksi klusteroinnin validointiin.

Klusteroinnin validointi tarkoittaa suoritetun klusterointitehtävän onnistumisen arviointia. Sen toteuttamiseen on olemassa monia tapoja. Tässä työssä on tarkasteltu arviointia kolmen erilaisen kriteerin suhteen. Niihin liittyen on esitelty joitakin sopivia tilastollisia tunnuslukuja eli samankaltaisuusindeksejä sekä käsitelty satunnaisuuden testausta, jonka avulla sattuman osuutta klusterointiin pyritään selvittämään.

Monia käyttökelpoisia samankaltaisuusindeksejä on jäänyt validointiluvun ulkopuolelle, esimerkiksi informaatioteoriaan perustuvat mitat. Lisäksi satunnaisuuden testauksen yhteydessä on keskitytty Monte Carlo - eli simulointitekniikoihin ja jätetty Bootstrap- eli satunnaistoistomenetelmät vähälle. Bootstrap-tekniikoiden avulla voitaisiin tarkastella yksittäisten klustereiden uudelleenmuodostumiskykyä. Yksittäisen klusterin validointiongelmaa ei ole muutenkaan tutkielmassa käsitelty, vaan on keskitytty yhden parhaan klusterikokonaisuuden etsintään.

Menetelmien soveltamiseen käytössä on ollut tilastomenetelmien peruskurssiaineisto. Aineiston laatu ei ole ollut paras mahdollinen. Kysely sisältää puuttuvia havaintoja, ja joidenkin osioiden kysymyksiin on jäänyt parantamisen varaa. Lisäksi aineiston esikäsittelyvaiheessa kun heterogeenisiä summamuuttujia on muodostettu, luokkien lukumäärien valintaan olisi ollut syytä käyttää enemmän harkintaa. Näin kahden toteutetun klusterointimenetelmän kohdalla näkynyt voimakas ketjuttuminen olisi saattanut vähentyä. Toisaalta kyseisen aineiston käsittely on tarjonnut erilaisia reaalimaailman aineistojen kanssa vastaan tulevia haasteita ja oppia niiden ratkaisemiseksi.

Aineiston jakautuminen kahteen eri suureen osaan, on johtanut loppukyselyn analyyseistä pudottamiseen. Tällöin on menetetty informaatio tenttimenestyksestä. Opiskelijan matemaattisesta osaamisen mittaaminen on nyt tehty ainoas-

taan lukion arvosanan perusteella, joka on kuitenkin osoittautunut aiemmissa tutkimuksissa hyväksi osaamisen mittariksi (Väisänen & Ylönen 2004).

Väisäsen ja Ylösen tekemä työ on antanut sovelluksen klusterointitehtävän tulkinnalle hyvää vertailupohjaa. Nyt löydetyistä kolmesta ryhmästä kahdelle löytyy vastaavuudet aiemmasta tutkimuksesta. Yhden ryhmän kohdalla sen sijaan on löydetty päinvastainen tulos. Sukupuolten väliset erot vaikuttavat samansuuntaisilta molemmissa aineistossa siitä huolimatta, että naisten lukumäärässä on suuri ero. Tosin sukupuolen vaikutuksen tarkemmalle tutkimiselle voisi olla aihetta.

Nyt löydetyt kolme klusteria ovat kaikki sen verran suuria, että niitä olisi mahdollista vielä jakaa osiin ja katsoa laajemminkin, millaisia sisäkkäisiä klustereita aineistosta mahdollisesti löytyy. Mutta kuten jo edellä on todettu, yksittäisen klusterin validointi ja tulkinnallisen sisäkkäisen hierarkian etsintä ulkoisen tiedon avulla on jätetty tutkielman ulkopuolelle.

Klusterianalyysi ei koskaan tarjoa valmiita vastauksia ongelmiin, mutta asiantuntijan käytössä se on mainio työkalu uuden tiedon etsintään ja uusien hypoteesien muodostamiseen. Usein klusterointia käytetäänkin eksploratiivisessa data-analyysissa. Kuitenkin edelleen pätee se, mitä Anderberg on vuonna 1973 teoksessaan todennut ”mikään ei korvaa tutkijan aineiston syvällistä tuntemusta ja ymmärrystä”.

Kiitokset

Kiitos tutkielman ohjaaja FT Salme Kärkkäiselle kaikesta avusta tämän prosessin aikana. Hänen positiivisuutensa ja innokkuutensa tarttuivat ajoittain myös tutkielman kirjoittajaan.

Kiitos FT Jouni Kuhalle (London School of Economics) ehdotuksista, millaisia valintoja kannattaa tehdä klusteroitaessa aineistoa, jossa on puuttuvaa dataa.

Lisäksi kiitos professori Jukka Nyblomille, professori Antti Penttiselle sekä TkT Pasi Koikkalaiselle. He kaikki ovat joko suoraan tai epäsuorasti auttaneet tutkielman kirjoituksessa.

Viitteet

[1] Albatineh, A. N., Niewiadomska-Bugaj, M. & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification 23*, 301-313.

[2] Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research 28*, 301-309.

[3] Anderberg, M. R. (1973). *Cluster Analysis for Applications.* Academic Press, New York.

[4] Bailey, T. A. & Dubes, R. C. (1982). Cluster validity profiles. *Pattern Recognition 15*, 61-83.

[5] Ben-David, S. & von Luxburg, U. (2005). Towards a statistical theory of clustering. *PASCAL Workshop on Statistics and Optimization of Clustering*, London. URL: http://www.cs.uwaterloo.ca/~shai/LuxburgBendavid05.pdf (viitattu 20.1.2012).

[6] Bezdek, J. C. & Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics - Part B Cybernetics*, 301-315.

[7] Bradley, B. S., Mangasarian, O. L. & Street, W. N. (1997). Clustering via concave minimization. *Advances in Neural Information Processing Systems 9, 368-374.*

[8] Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*, 1–38.

[9] Fowlkes, E. B. & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association 78*, 553-569.

[10] Garey, M. R. & Johnson, D. M. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman, San Francisco.

[11] Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning.* 2nd ed. Springer, New York.

[12] Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 193-218.

[13] Jaccard, P. (1912). The distribution of flora in the alpine zone. *New Phytologist 11*, 37-50.

[14] Lloyd, S. P. (1957). Least square quantization in PCM. *Bell Telephone Laboratories Paper.*

[15] McQueen, J. P. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

[16] Mcnaughton Smith, P., Williams, W., Dale, M. & Mockett, L. (1965). Dissimilarity analysis: a new technique of hierarchical subdivision. *Nature 202*, 1034-1035.

[17] Meilã, M. (2005). Comparing clusterings: an axiomatic view. *Proceedings of the 22th International Conference on Machine Learning*, 577-584, New York.

[18] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

[19] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*, 846-850.

[20] Rolph, F. J. (1970). Adaptive hierarchical clustering schemes. *Systematic Zoology 19*, 58-82.

[21] Peterson, A. D., Ghosh, A. P. & Maitra, R. (2010). A systematic evaluation of different methods for initializing the K-means clustering algorithm. *IEEE Transactions on Knowledge and Data Engineering.*

[22] Saariluoma, P. (2004). *Käyttäjäpsykologia: ihmisen ja koneen vuorovaikutuksen uusi ajattelutapa.* WSOY, Helsinki.

[23] Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology 51*, 492-508.

[24] Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy.* W.H. Freeman, San Francisco.

[25] Theodoridis, S. & Koutroumbas, K. (2006). *Pattern Recognition.* 3rd ed. Academic Press, Orlando.

[26] Väisänen, P. & Ylönen, S. (2004). Matemaattiset taidot ja matemaattinen minäkäsitys tilastollisten menetelmien oppimisessa. *Kasvatus 4/2004*, 365-378.

[27] Vinh, N. X., Epps, J. & Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? *Proceedings of the 26th International Conference on Machine Learning*, 1073-1080, Montreal.

[28] Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* 2nd ed. Morgan Kaufmann, San Francisco.

A Aineiston osiot ja muuttujat

A.1 Alkukysely

Osio 1 a Yleiset taustatiedot

- Sukupuoli {SP}, ei käytetä klusteroinnissa.
- Syntymävuosi {SV}, ei käytetä klusteroinnissa.
- Lukiomenestys, kolme muuttujaa, yksi binäärinen {MATPIT} ja kaksi järjestysasteikollista {MATARV, AIKARV}.
- Tutkinnon aloitusvuosi {AV}, ei käytetä.
- Tiedekunta {TDK}, kategorinen [1,7], ei käytetä.
- Työssäkäynti opiskelujen ohessa {TYOD}, kategorisoitiin järjestysasteikolliseksi välille [1,4].

Osio 1 b Taustat koskien tilastotieteen opintoja

- Kuusi binääristä muuttujia, jotka yhdistettiin yhdeksi summamuuttujaksi {STAUSTAT}.

Osio 2 Opiskelutavat yleisesti

- Kaksi neliasteikollista muuttujaa yhdeksi summamuuttujaksi {STAPA}.
- Kaksi ajankäyttöä koskevaa kokonaislukumuuttujaa {OHOPD, OMOPD}, kategorisoitiin järjestysasteikolliseksi välille [1,4].

Osio 3 Tilastotieteen kokeminen

- Kolme neliasteikollista muuttujaa summamuuttujaksi {SKOKEM}.

Osio 4 Matemaattiset taidot

- Kolme neliasteikollista matematiikan helppoutta kuvaavaa muuttujaa summamuuttujaksi {STAIDOT}.
- Kaksi laskusääntöjen hallintaa koskevaa muuttujaa summamuuttujaksi {SHALL}.

Osio 5 Matemaattisten taitojen riittävyys

- Neliasteikollinen järjestysasteikollinen taitojen riittävyyttä koskeva muuttuja {OMAKYKY}.
- Binäärinen läpipääsyä koskeva muuttuja {PELKO}.

- Viisi ennakkokäsityksiä koskevaa kolmiasteikollista muuttujaa summamuuttujaksi {SENN}, arvoalue [-5,5].

Osio 6 Tilastotieteen tarpeellisuus

- Kolme menetelmien tarpeellisuutta koskevaa muuttuja summamuuttujaksi {SJATKO}.

Loppukysely

- Binäärinen loppukyselyyn vastaamisesta kertova muuttuja {LK}.

B Muuttujien jakautuminen klustereissa

B.1 Klusteroinnissa käytetyt muuttujat

Muuttuja 1 {MATPIT}

C_l	1	2
C_1	18	51
C_2	12	38
C_3	41	18
$\sum$	71	107

taulukossa riveillä C_l klusterit $l = 1, 2, 3$ ja sarakkeilla vastausten frekvenssit, kun 1 tarkoittaa, että vastaaja on suorittanut lukiossa lyhyen matematiikan ja 2 että vastaaja on suorittanut pitkän matematiikan.

Muuttuja 2 {MATARV} eli matematiikan arvosana lukiossa.

C_l	5	6	7	8	9	10
C_1	0	1	10	22	27	8
C_2	0	7	13	14	13	3
C_3	2	13	22	16	4	1
$\sum$	2	21	45	52	44	12

Muuttuja 3 {AIKARV} eli äidinkielen arvosana lukiossa.

C_l	5	6	7	8	9	10
C_1	0	1	14	27	22	4
C_2	1	1	3	13	24	8
C_3	0	0	4	21	31	2
$\sum$	1	2	21	61	77	14

Muuttuja 4 {OMAKYKY}, joka kuvaa vastaajan uskomusta omien matemaattisten taitojensa riittävyyteen kurssilla.

C_l	1	2	3	4
C_1	44	20	2	1
C_2	34	21	0	0
C_3	5	32	21	5
$\sum$	83	73	23	6

Taulukossa 1 tarkoittaa vahvaa uskoa taitoihin, 2 heikkoa uskoa, 3 heikkoa epäuskoa ja 4 vahvaa epäuskoa.

Muuttuja 5 {PELKO}, joka kertoo, pelkääkö vastaaja, ettei läpäise kurssia.

C_l	1	2
C_1	11	57
C_2	1	54
C_3	33	30
$\sum$	45	141

Taulukossa 1 tarkoittaa vastaajan pelkäävän, ettei läpäise kurssia ja 2 että vastaaja ei pelkoa tunne.

Muuttuja 6 {STAUSTAT} summamuuttuja, joka kuvaa vastaajan edeltäviä ja tulevia tilastotieteen opintoja sekä tilastotieteen pakollisuutta.

C_l	6	7	8	9	10	11	12
C_1	4	31	27	3	2	0	0
C_2	0	15	12	17	8	2	1
C_3	4	15	20	10	8	4	1
$\sum$	8	61	59	30	18	6	2

Suuret arvot tarkoittavat tilastotieteen tarpeellisuutta ja pakollisuutta vastaajalle, pienet päinvastaista.

Muuttuja 7 {TYOD} työhön käytetty aika opiskelujen ohessa.

C_l	1	2	3	4
C_1	39	14	6	2
C_2	33	13	2	1
C_3	24	14	10	7
$\sum$	96	41	18	10

Työhön käytetyt tunnit h on jaettu seuraavasti neljään luokkaan:

1, kun $h = 0$
2, kun $0 < h < 11$
3, kun $11 < h \leq 20$
4, kun $h > 20$.

Muuttuja 8 {STAPA} summamuuttuja kahdesta opiskelutapaa kuvaavasta muuttujasta.

C_l	2	3	4	5	6	7	8
C_1	10	25	21	11	0	1	0
C_2	1	8	24	21	1	0	0
C_3	1	22	19	19	1	0	0
$\sum$	12	55	64	51	2	1	0

Pienet arvot tarkoittavat vastaajan pyrkivän opiskelemaan hyvin ja tehokkaasti.

Muuttuja 9 {SKOKEM} summamuuttuja kolmesta tilastotieteen kokemista koskevasta kysymyksestä.

C_l	3	4	5	6	7	8	9
C_1	1	3	15	38	9	3	0
C_2	0	0	2	22	26	5	0
C_3	0	0	6	26	27	3	1
$\sum$	1	3	23	86	62	11	1

Pienet arvot tarkoittavat vastaajan kokevan tilastotieteen mielenkiintoiseksi ja käytännölliseksi, kun taas suuret arvot tarkoittavat päinvastaista.

Muuttuja 10 {STAIDOT} summamuuttuja, joka kuvaa vastaajien matemaattista lahjakkuutta sekä asennetta matematiikkaan.

C_l	3	4	5	6	7	8	9	10	11	12
C_1	1	1	3	23	29	9	1	1	0	0
C_2	0	0	0	6	15	25	8	1	0	0
C_3	0	0	0	4	16	16	18	8	0	1
$\sum$	1	1	3	33	60	50	27	10	0	1

Pienet arvot kertovat matemaattisesta lahjakkuudesta ja positiivisesta asenteesta.

Muuttuja 11 {SHALL} summamuuttuja kahdesta numeroiden ja matemaattisten laskusääntöjen hallintaa koskevasta kysymyksestä.

C_l	2	3	4	5	6	7	8
C_1	26	9	22	4	6	1	0
C_2	12	10	24	5	4	0	0
C_3	0	4	11	20	22	2	4
$\sum$	38	23	57	29	32	1	4

Pienet arvot kertovat laskusääntöjen hyvästä hallinnasta.

Muuttuja 12 {SENN}, summamuuttuja opiskelijan ennakkokäsityksiin positiivisesti tai negatiivisesti vaikuttaneista asioista.

C_l	-5	-4	-3	-2	-1	0	1	2	3	4	5
C_1	0	0	0	0	1	3	10	19	21	9	3
C_2	0	0	1	0	2	7	13	16	9	5	1
C_3	1	1	3	13	15	15	9	2	2	0	0
$\sum$	1	1	4	33	18	25	32	37	32	1	4

Negatiiviset arvot tarkoittavat vastaajan tilastotieteestä saaman ennakkokäsityksen olevan kielteinen, positiiviset vastaavasti myönteinen.

Muuttuja 13 {SJATKO} summamuuttuja, joka kuvaa tilastotieteen tarpeellisuutta jatkossa.

C_l	2	3	4	5	6
C_1	3	30	27	6	1
C_2	0	6	23	25	1
C_3	1	16	32	13	1
$\sum$	4	52	82	44	3

Suuret arvot tarkoittavat opiskelijan tarvitsevan tilastotiedettä jatkossa, pienet päinvastaista.

Muuttuja 14 {OHOPD} luentoihin ja harjoituksiin osallistumiseen käytetty aika viikossa tunteina.

C_l	1	2	3	4
C_1	4	42	18	3
C_2	1	17	25	9
C_3	5	38	13	3
$\sum$	10	97	56	15

Tunnit h on jaoteltu seuraavasti:

1, kun $h < 10$
2, kun $10 < h < 20$
3, kun $20 < h \leq 30$
4, kun $h > 30$.

Muuttuja 15 {OMOPD} omaehtoiseen opiskeluun käytetty aika viikossa tunteina.

C_l	1	2	3	4
C_1	20	36	10	1
C_2	16	27	6	3
C_3	13	26	16	3
$\sum$	49	89	32	7

Tuntien jako on tehty kuten edellisen muuttujan {OHOPD} tapauksessa.

B.2 Taustamuuttujat

Taustamuuttuja 1 eli sukupuoli {SP}.

C_l	1	2
C_1	34	36
C_2	33	22
C_3	43	20
$\sum$	110	78

Taulukossa 1 kertoo naisten ja 2 miesten frekvenssit.

Taustamuuttuja 2 {SV} eli opiskelijan syntymävuosi.

C_l	1963	1980	1982	1984	1986
C_1	6	6	21	28	9
C_2	8	7	9	19	12
C_3	17	10	9	17	10
$\sum$	31	23	39	64	31

Vuodet on jaettu niin, että ensimmäiseen luokkaan kuuluvat vuodet $1963-1979$, toiseen 1980 ja 1981, kolmanteen 1982 ja 1983, neljänteen 1984 ja 1985 sekä viimeiseen luokkaan $1986, 1987$ ja 1988.

Taustamuuttuja 3 {AV}, joka tarkoittaa yliopistossa aloitusvuotta.

C_l	2001	2002	2003	2004	2005	2006
C_1	3	4	6	10	25	21
C_2	3	2	0	3	12	35
C_3	6	3	4	6	9	34
$\sum$	12	9	10	19	46	90

Ensimmäiseen luokkaan sisältyvät vuodet $1995-2001$, muuten luokat sisältävät vain kyseisen vuoden.

Taustamuuttuja 4 {TDK}, joka kertoo opiskelijan tiedekunnan.

C_l	1	2	3	4	5	6	7
C_1	2	28	3	7	14	10	6
C_2	0	12	1	6	5	16	15
C_3	3	9	1	4	2	14	30
$\sum$	5	49	5	17	21	40	51

Tiedekunnat on koodattu seuraavasti:

$$
\begin{aligned}
1 &= Humanistinen \\
2 &= Informaatioteknologian \\
3 &= Kasvatustieteiden \\
4 &= Liikunta - ja terveystieteiden \\
5 &= Matemaattis - luonnontieteellinen \\
6 &= Taloustieteiden \\
7 &= Yhteiskuntatieteellinen.
\end{aligned}
$$

Taustamuuttuja 5 {LK} eli loppukyselyyn vastaaminen.

C_l	0	1
C_1	42	28
C_2	31	24
C_3	29	34
$\sum$	102	86

Taulukossa arvo 1 kertoo loppukyselyyn vastanneiden ja 0 vastaamattomien frekvenssit.