K:n prototyypin ryhmittelymenetelmän ja moni-imputoinnin sovellus työhyvinvointiaineistoon

Suvi Ahtinen

Pro gradu -tutkielma

Tilastotiede

Matematiikan ja tilastotieteen laitos
Jyväskylän yliopisto

18. kesäkuuta 2020

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos
Ahtinen, Suvi: K:n prototyypin ryhmittelymenetelmän ja moni-imputoinnin sovellus työhyvinvointiaineistoon
Tilastotieteen pro gradu -tutkielma (41 sivua) + 5 liitettä, kesäkuu 2020.

Tiivistelmä

Tässä tutkielmassa sovelletetaan k:n prototyypin ryhmittelymenetelmää aineistoon, joka perustuu peruskoulun ja toisen asteen oppilaitosten henkilökunnan mielipidekyselyyn omasta työhyvinvoinnistaan. Menetelmä on valittu, koska sen avulla voidaan ryhmitellä aineistoa yksilöiden välisten vastauksien samankaltaisuuksien perusteella ja huomioida aineiston kategoriset sekä jatkuvat muuttujat. Aineisto sisältää runsaasti puuttuvaa tietoa, joten ryhmittely toteutetaan täydellisesti havaitun aineiston lisäksi moni-imputoituihin aineistoihin.

Moni-imputoinnissa muodostetaan iteratiivisesti viisi eri aineistoa, joihin tehdään ryhmittely ja vertaillaan ryhmille muodostuneita keskustojen keskiarvoja. Imputoitavalle vastemuuttujalle valitaan sopivat selittävät muuttujat, jotka sisältävät vähintään 50 prosenttia havaittuja arvoja ja korreloivat vastemuuttujan kanssa.

Ryhmittelyanalyysiin valitaan 70 mielipidekysymyksestä 22 kysymystä tulosten raportoinnin selkeyttämiseksi. Valinnassa käytettävän algoritmin avulla etsitään muuttujia, joissa voidaan havaita klusteroitumista. Aineistosta on valittu myös kaksi taustamuuttujaa: ikä ja työvuodet. Tällöin voidaan tarkastella myös työhyvinvointiin vaikuttavien taustatekijöiden ryvästymistä.

Ennen ryhmittelymenetelmän suorittamista valitaan, kuinka moneen ryhmään havainnot lajitellaan. Valinta tehdään sisäisten validointikriteerien indeksien avulla. Tässä tutkielmassa esitellään neljä yleisesti käytettyä indeksiä, joista Davies–Bouldin- ja Calinski–Harabasz-indeksien perusteella aineistoon sopii kaksi ryhmää. Lisäksi esitellään ulkoinen Rand-indeksi, jonka avulla voidaan tutkia täydellisesti havaitun ja moni-imputoitujen aineistojen ryhmittelyiden yhtäläisyyksiä.

Kahteen klusteriin ryhmitellystä moni-imputoidusta sekä täydellisesti havaitun aineiston tuloksista voidaan todeta ensimmäisen klusterin sisältävän negatiivisesti työhyvinvoinnistaan ajattelevia, jotka kokevat itsensä myös väsyneeksi ja stressaantuneeksi. Toinen klusteri taas sisältää enemmän positiivisesti työhyvinvoinnistaan ajattelevia, jotka kokevat väsymystä ja stressiä vähemmän, sekä ovat työskennelleet vähemmän aikaa samassa koulussa kuin ensimmäisen klusterin henkilöt.

Avainsanat: k:n prototyypin ryhmittelymenetelmä, koulun hyvinvointiprofiili, luokitteluvirhe, moni-imputointi, ryhmittelymenetelmien validointikriteerit, työhyvinvointi

Sisältö

1 Johdanto

Ryhmittelymenetelmien käyttäminen aineiston analysoinnissa on perusteltua, jos halutaan pelkistää suuria aineistoja havainnollisempaan muotoon, luoda uusia mahdollisia tutkimushypoteeseja tai testata ja todistaa ennen menetelmän käyttöä päätettyjen tutkimushypoteesien todenperäisyyttä. Menetelmien ideana on jaotella useamman muuttujan perusteella havainnot ryhmiin eli klustereihin: tilastoyksiköiden arvot voidaan jaotella samankaltaisuuden perusteella samaan ryhmään. Ryhmittelymenetelmien avulla voidaan siis havaita aineiston yksilöiden ryvästymistä, mutta menetelmä ei kuitenkaan sovellu muuttujien välisten yhteyksien mallintamiseen. (Theodoridis & Koutroumbas 2008, s. 486-487.)

Tässä tutkielmassa sovelletaan k:n prototyypin ryhmittelymenetelmää (*K-prototypes clustering*) Anne Konun (2016–2019) koulujen työhyvinvointiaineistoon, jossa vastaajat ovat ala- ja yläkoulun sekä toisen asteen oppilaitosten henkilökuntaa. K:n prototyypin ryhmittelymenetelmää käytetään, koska menetelmällä voidaan ryhmitellä jatkuvia ja kategorisia muuttujia tehokkaasti. (Huang et al. 1997.) Ryhmittelymenetelmästä saatujen tulosten avulla voidaan havaita, onko aineistossa selkeitä havainnoista koostuvia toisistaan eroavia klustereita. Tämän lisäksi voidaan tutkia, kuinka suuria nämä ryhmittelymenetelmän avulla saadut klusterit ovat ja miten klustereiden keskeisimmät arvot eroavat toisistaan.

Aikaisemmissa koulujen henkilökuntaa koskevissa tutkimuksissa on havaittu, että 2000-luvulla työhyvinvointi on ollut parempi peruskouluissa kuin toisen asteen oppilaitoksissa. Lisäksi miesten, määräaikaisissa ja osa-aikaisissa työsuhteissa olevien työhyvinvointi on ollut parempi kuin naisten ja vakituisessa työsuhteessa olevilla. (Konu et al. 2010.) Tämän tutkielman päätavoite on tutkia ryhmittelymenetelmän avulla, millaisia ryhmiä aineistosta voidaan muodostaa. Ala- ja yläkoulun sekä toisen asteen oppilaiden vastauksien ryhmittelyyn on aikaisemmin käytetty hierarkkisia ryhmittelymenetelmiä, ja havaittu aineiston klusteroitumista (Kylväjä et al. 2019). Aikaisemmissa tutkimuksissa ryhmittelymenetelmiä ei ole kuitenkaan sovellettu henkilökuntaa koskevaan aineistoon, joten tutkielman tuottamat tulokset ovat uusia.

K:n prototyypin ryhmittelymenetelmän yhtenä ehtona on, että yksikään tilastoyksikkö ei voi sisältää puuttuvaa tietoa. Tutkielmassa käytetty työhyvinvointiaineisto sisältää kuitenkin suhteellisen paljon puuttuvaa tietoa, ja aineistossa on yli 70 kysymystä. Havaintojen poistaminen voi vaikuttaa ryhmittelymenetelmän tuloksiin, koska aineistosta häviää samalla myös havait-

tua tietoa: jos yhteenkin kyselyn kysymykseen on jätetty vastaamatta, koko tilastoyksikkö poistuu aineistosta.

Aikaisemmissa tutkimuksissa imputointia on tehty ryhmittelymenetelmiin soveltuvilla menetelmillä, mutta kyseisissä tutkimuksissa puuttuvuus on usein simuloitua tai määrättyä (Zhang et al. 2006; Somasundaram & Nedunchezhian 2011). Tämä ei vastaa todellista puuttuvan tiedon ongelmaa, jolloin puuttuvien havaintojen oikeita arvoja ei pystytä jäljittämään: tietojen puuttuminen ei yleensä jakaudu aineistoon tiettyjen ehtojen mukaisesti.

Moni-imputointia ja ryhmittelymenetelmää on kuitenkin sovellettu epidemiologisessa tutkimuksessa, joka vastaa tämän tutkielman tutkimusongelmaa: puuttuvia havaintoja ei ole simuloitu vaan puuttuvat havainnot ovat todellisia, ja aineistoa täydennetään moni-imputoinnin avulla (Basagaña et al. 2013). Tässä tutkielmassa vertaillaan täydellisesti havaitun ja moni-imputoitujen aineistojen ryhmittelyssä muodostuneita klustereiden keskustoja toisiinsa. Lisäksi täydellisesti havaittua aineistoa ja moni-imputoitujen aineistojen klusterirakenteiden samanlaisuutta voidaan vertailla ulkoisen validointikriteerin Rand-indeksin avulla.

Moni-imputoinnissa aineistoja muodostetaan viisi kappaletta, joihin k:n prototyypin ryhmittely voidaan suorittaa ja vertailla ryhmiä klustereiden keskustojen keskiarvojen avulla. Jokainen puuttuvia tietoja sisältävä muuttuja tarvitsee imputointimallin. Imputointimallissa on vastemuuttuja, jota täydennetään, sekä selittävät muuttujat, joissa esiintyy mahdollisimman vähän puuttuvuutta, ja jotka korreloivat imputoitavien muuttujien kanssa. Lisäksi aineistosta voidaan poistaa muuttujia, jotka eivät ole hyödyllisiä imputoinnissa tai oleellisia aineiston ryhmittelyn kannalta. (van Buuren 2018, luku 9; Pfaffel 2019.)

Tässä tutkielmassa tehdään myös sopivien muuttujien valintaa, koska kaikki mielipidekyselyn kysymykset eivät vaikuta merkittävästi ryhmittelyyn. Muuttujat valitaan k:n prototyypin ryhmittelymenetelmään soveltuvan algoritmin avulla, missä muuttujien arvoja sekoitetaan viisi kertaa ja lasketaan luokitteluvirheen arvoja: Mitä suuremman luokitteluvirheen arvon muuttuja saa, sitä enemmän muuttujan sisältämät havainnot poikkeavat toisistaan. Jos havainnot poikkeavat toisistaan, niin muuttuja soveltuu ryhmittelymenetelmään käytettäväksi. (Pfaffel 2019.)

K:n prototyypin ryhmittelymenetelmän luotettavuuden lisäämiseksi klusterien lukumäärä valitaan sisäisten validointikriteerien indeksien perusteella, jotka lasketaan eri klusterien lukumäärillä ryhmitellyille aineistoille. Tässä tutkielmassa esitellään neljä yleisesti käytettyä sisäistä validointikriteerien

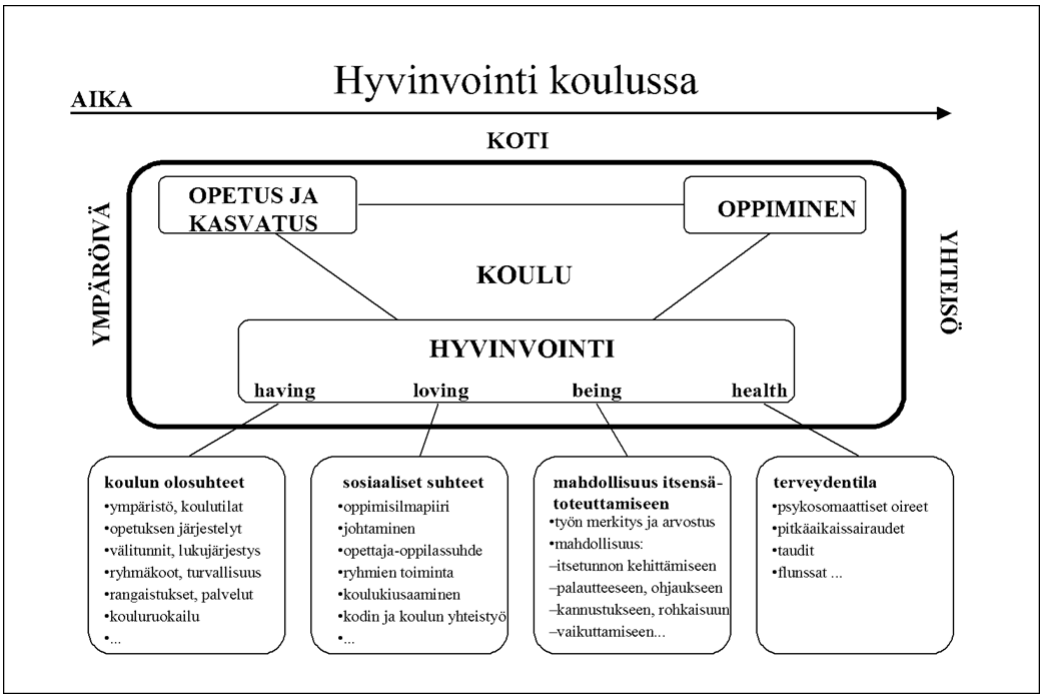
indeksiä, jotka soveltuvat k:n prototyypin ryhmittelymenetelmälle: Dunn-indeksi, Calinski–Harabasz-indeksi, Davies–Bouldin-indeksi ja C-indeksi.

Tutkielman rakenne on seuraava: Moni-imputointimallin selittävien muuttujien valintaa perustellaan luvussa 3.1 sekä imputointialgoritmin taustalla olevaa laskennallista teoriaa luvussa 3.2. K:n prototyypin menetelmää esitellään luvussa 4.2, ja menetelmään soveltuvaa muuttujien valintaa luvussa 4.3. Ryhmien lukumäärän valintaan käytettyjä sisäisten validointikriteerien indeksien laskentaa esitellään luvussa 5. Täydellisesti havaitun ja moni-imputoitujen aineistojen klusterirakenteiden eroja vertaillaan luvussa 5.1 esiteltävän Rand -indeksin avulla. Luvussa 6 esitellään tuloksia ja luvussa 7 pohditaan menetelmien toteutusta sekä jatkotutkimusten mahdollisuuksia.

2 Aineiston esittely

Aineistona käytetään Anne Konun (2016–2019) Koulun hyvinvointiprofiilin aineistoa, joka on tarkoitettu käytettäväksi laajempiin tutkimuksiin havaitsemaan opiskelu- ja työhyvinvoinnissa tapahtuvia muutoksia. Aineisto ei kuitenkaan ole pitkittäistutkimus, vaan vastaajat oletetaan olevan riippumattomia toisistaan. Tällöin ei tutkita tiettyjen henkilöiden tai koulujen ajassa tapahtuvia muutoksia vaan selvitetään yleistä kokonaiskuvaa koulujen henkilökunnan kokemasta työhyvinvoinnista vuosien 2016–2019 aikana. (Konu 2016–2019.)

Koulun hyvinvointiprofiilin aineiston taustalla on teoreettinen kouluyhteisölähtöinen hyvinvointimalli, jonka avulla on tehty henkilökunnalle tarkoitettuja mielipidekysymyksiä ja terveydentilaa arvioivia kysymyksiä. Malli koostuu neljästä osiosta, jotka yhdistävät hyvinvointia, opetusta, kasvatusta sekä oppimista (kuva 1). Kolme osiota liittyy kouluympäristöön: olosuhteet, sosiaaliset suhteet ja itsensätoteuttamisen mahdollisuudet. Lisäksi on kysytty terveydentilaa, kuten erilaisten oireiden esiintymistä. (Konu 2010.)



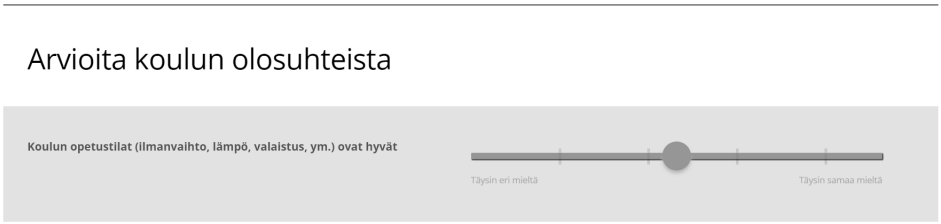
Kuva 1: Konun (2002) esittämä hyvinvointimalli koulussa.

Tutkielmassa käytettävässä aineistossa vastanneita on yhteensä 1643 henkilöä 71 eri koulusta ja muuttujia on yhteensä 76, joista 6 kappaletta ovat taustamuuttujia ja loput 70 työhyvinvointiin liittyviä kysymyksiä. Joidenkin osioiden kysymyksiin on saatettu jättää kokonaan vastaamatta, joten aineistossa esiintyy runsaasti puuttuvaa tietoa: jos aineistosta poistetaan tilastoyksiköt, jotka sisältävät puuttuvaa tietoa, jäljelle jää 927 havaintoa.

Kyselyn ensimmäinen osio eli olosuhteet koulussa sisältää 18 mielipidekysymystä koulujen fyysisistä tiloista ja niiden ominaisuuksista. Toinen osio on sosiaaliset suhteet, joka on tärkeässä osassa tutkittaessa työyhteisöjen hyvinvointia. Osiossa kysytään 21 eri kysymystä työpaikkakiusaamisesta sekä suhtautumisesta esimieheen, työtovereihin ja koulun oppilaisiin. Kolmannessa osiossa on 18 kysymystä työntekijöiden itsensätoteuttamisen mahdollisuuksista, esimerkiksi mielipidekysymyksiä omista tuntemuksistaan työtään kohtaan. Terveydentila on viimeisin eli neljäs osio, joka sisältää yhteensä 13 kysymystä. Osiossa kysytään erilaisten oireiden esiintymistä, kuten niska- ja hartiakipuja, vatsakipuja, päänsärkyä, väsymystä sekä jännittyneisyyttä ja ärtyneisyyttä. (Konu 2010.)

Helsingin opetusviraston aloitteesta ja rahoituksen myötä Likert -asteikollinen kyselylomake on muutettu suhdeasteikolliseksi verkkolomakkeeksi vuodesta 2016 lähtien. Tällöin kyselyyn on voinut vastata Koulun hyvinvointiprofiilin sivuilla olevaan lomakkeeseen, jossa vastaukset saavat arvoja asteikolla 0–100, jolloin ääripäissä olevat arvot eli 0 tarkoittaa ”Täysin eri mieltä” ja 100 tarkoittaa ”Täysin samaa mieltä” sekä oireiden kuvaamiseen tarkoitetuissa kysymyksissä minimiarvo 0 tarkoittaa oireiden esiintymistä harvoin ja maksimiarvo 100 tarkoittaa oireiden esiintymistä usein.

Kuvassa 2 on havainnollistava esimerkki kyselylomakkeen ensimmäisestä mielipidekysymyksestä. Vastaus voidaan asettaa interaktiivisesti eli lukujen 0–100 välille vetämällä tietokoneen hiiren avulla, tai kosketusnäytön tapauksessa sormen avulla, viivan päällä olevaa ympyrää. (Konu 2016–2019.)



Kuva 2: Mielipidekyselyn ensimmäinen kysymys esimerkkinä kyselylomakkeeseen vastaamisesta (Konu 2016–2019).

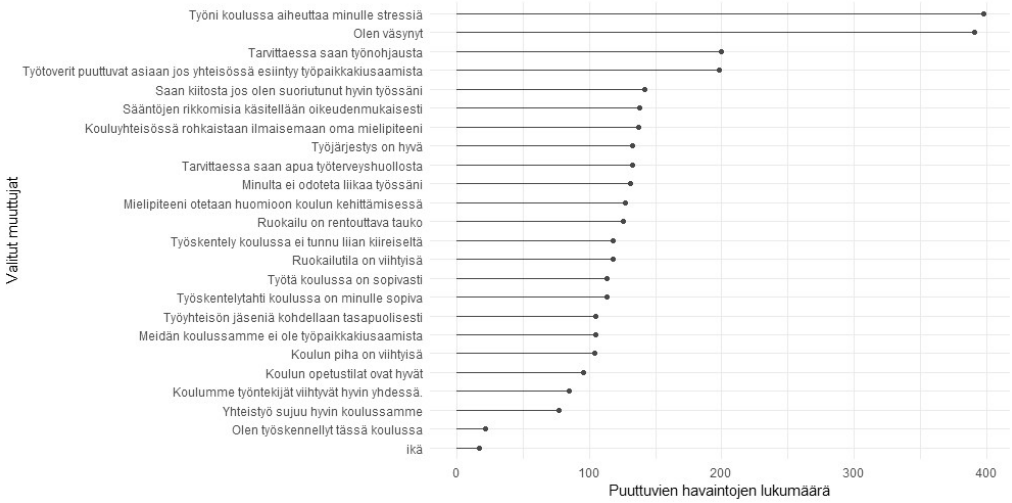
Osallistuminen kyselytutkimukseen on ollut täysin vapaaehtoista, joten kaikki osallistuvan koulun henkilökunnan jäsenet eivät ole välttämättä vastanneet mielipidekyselyyn. Helsingin alueen kouluja on osallistunut kyselyyn enemmän kuin muun maan, koska alueen kouluille tutkimukseen vastaaminen on ollut ilmaista. Tällöin kyselytutkimuksen otanta ei täysin vastaa koko Suomen kouluhenkilökunnan vastauksia. (Konu 2016–2019.)

Kaikkia 76 muuttujaa ei ole käytetty tämän tutkielman lopullisissa ryhmittelyissä, vaan aineistosta on valittu yhteensä 22 mielipidekysymysten muuttujaa, joiden valintaa perustellaan muuttujien valinnan algoritmin avulla. Muuttujien klusteroituneisuus voidaan selvittää aineiston pohjalta luvussa 4.3 esitellyllä tavalla: Jos muuttujan sisällä havaitaan vastauksissa selkeitä eroavaisuuksia, se sopii hyvin ryhmittelymenetelmiin käytettäväksi. Lopullisiin ryhmittelyihin valitut jatkuvat mielipidekysymykset tunnuslukuineen esitellään liitteessä A olevassa taulukossa A1. Kysymyksiin on saatu vastauksia asteikon molemmista ääripäistä.

Mielipidekysymysten lisäksi on valittu kaksi järjestysasteikolliseksi luokiteltua taustamuuttujaa: Ikä ja samassa koulussa kertyneet työvuodet. Ikämuuttuja on jaoteltu jo kyselylomakkeessa 10 luokkaan neljän vuoden välein: 25-vuotiaat tai alle, 26–30-vuotiaat, 31–35-vuotiaat, 36–40-vuotiaat, 41–45-vuotiaat, 46–50-vuotiaat, 51–55-vuotiaat, 56–60-vuotiaat, 61–65-vuotiaat ja 66-vuotiaat tai yli. Sen sijaan työvuodet samassa koulussa ovat jaoteltu kyselylomakkeessa neljään kategoriaan: työskennellyt alle vuoden, työskennellyt 1–5 vuotta, työskennellyt 6–10 vuotta, työskennellyt enemmän kuin 10 vuotta.

Luvussa 4.3 esiteltävän muuttujien valinnan algoritmin perusteella taustamuuttujia ei pitäisi sisällyttää ryhmittelyyn. Ikä- ja työvuosi-muuttujat ovat kuitenkin valittu jälkianalyysiin, jotta voidaan tutkia, millaisia henkilöitä klustereissa olevat vastaajat ovat taustaltaan, vaikka aineistosta ei pyritä estimoimaan muuttujien välisiä yhteyksiä.

Tutkielmassa käytettävä aineisto sisältää puuttuvaa tietoa: liitteessä E on kuvaajia jokaisen aineiston muuttujan puuttuvien havaintojen lukumääristä, jotka ovat jaettu kuvassa 1 esitettyihin neljään osioon. Sen sijaan kuvassa 3 esitellään puuttuvan tiedon lukumääriä Koulun hyvinvointiprofiilin aineiston muuttujissa, jotka ovat valittu tutkielman ryhmittelyyn. Kuvasta havaitaan, että neljä eniten puuttuvia tietoja sisältävät kysymykset ovat ”Olen väsynyt”, ”Työni koulussa aiheuttaa minulle stressiä”, ”Työtoverit puuttuvat asiaan, jos yhteisössä esiintyy työpaikkakiusaamista” ja ”Tarvittaessa saan työnohjausta”. Lisäksi kysymyksissä ”Työtoverit puuttuvat asiaan, jos yhteisössä esiintyy työpaikkakiusaamista” ja ”Tarvittaessa saan työnohjausta” puuttuvia havaintoja on enemmän verrattaessa muihin saman osion kysymyksiin, mutta terveydentilaa kuvaavissa kysymyksissä on jokaisessa muuttujassa lähes yhtä paljon puuttuvaa (liite E).



Kuva 3: Puuttuvien havaintojen lukumäärät muuttujittain graafisesti kuvattuna siten, että y-akselilla on ryhmittelyissä käytettävä muuttuja ja x-akselilla puuttuvien havaintojen lukumäärä.

3 Puuttuvan tiedon käsittely

Puuttuvan tiedon rakenne voidaan jakaa kolmeen: MCAR, MAR ja MNAR. MCAR (*missing completely at random*) tarkoittaa täysin satunnaista puuttumista, jolloin todennäköisyys sille, että havainto puuttuu on samanlainen jokaiselle havainnolle. Tällöin ehdollinen odotusarvo puuttuvuusmatriisille M on $P(M|\phi)$, jossa matriisin M arvo on $m_{ij} = 0$, kun havainto on havaittu ja $m_{ij} = 1$, kun havainto puuttuu. Lisäksi kaavassa oleva ϕ ilmaisee puuttuvuuden todennäköisyyden tuntematonta vakioparametria, jolloin $P(m_{ij} = 1|\phi) = \phi$. MAR (*missing at random*) taas tarkoittaa rakennetta, jossa puuttumisen todennäköisyys riippuu havaituista muuttujista vakioparametrin ϕ lisäksi. Tällöin ehdollinen todennäköisyys kuvataan kaavalla $P(M|Y_{obs}, \phi)$, jossa Y_{obs} tarkoittaa havaittuja muuttujia aineistossa. MNAR (*missing not at random*) tai NMAR (*not missing at random*) tarkoittaa, että puuttumisen todennäköisyyttä ei voida päätellä aineiston puuttuvien havaintojen jakautuneisuuden perusteella. Tällöin puuttuvuuden ehdollista todennäköisyyttä kuvataan kaavalla $P(M|Y_{miss})$, jossa Y_{miss} tarkoittaa muuttujien puuttuvia havaintoja. (Little & Rubin 2002, s. 11–12.) Esimerkiksi tässä tutkielmassa käytettävässä aineistossa kaikki ne koulun työntekijät ovat saat-

taneet jättäneet vastaamatta, jotka kokevat terveydentilansa huonoksi tai hyväksi.

3.1 Imputoinnissa käytettävien muuttujien valinta

Imputointimalliin voidaan valita sopivia selittäviä muuttujia, koska tällöin imputointi toimii nopeammin. Valinta voidaan suorittaa tiettyjen ehtojen avulla: imputointimallin selittäjiksi voidaan valita niitä muuttujia, joissa on paljon käytössä olevia tilastoyksiköitä ja imputoitavan vastemuuttujan sekä selittävien muuttujien välillä on korrelaatiota. Korrelointi imputoitavan ja selittävän muuttujan välillä vähentää MNAR-puuttumisrakennetta, johon moni-imputointi ei sovellu, koska imputointi saattaa aiheuttaa tulosten yli- tai aliestimoitumista. (van Buuren 2018, luku 6.2.) Lisäksi imputointi perustuu enemmän todellisuudessa havaittuihin arvoihin, jos muuttujassa on paljon havaittuja arvoja eli käytettävissä olevia tilastoyksiköitä. Imputoinnin selittävien muuttujien valintaan on käytetty R -ohjelmiston version 3.4.4. `quickpred` -funktiota, joka sijaitsee `mice` -kirjastossa (van Buuren & Groothuis-Oudshoorn 2011).

Aineiston muuttujille voidaan laskea myös puuttumisen todennäköisyyden suhdelukuja, joiden avulla voidaan päätellä muuttujan hyödyllisyys imputoitavan muuttujan selittäjänä. Tässä tutkielmassa on käytetty suhdelukuna `outflux` -kerrointa. Kertoimen avulla voidaan päätellä, kuinka hyödyllinen valittu muuttuja Y_j on muita muuttujia imputoitaessa: Jos `outflux`-kertoimen arvo on lähellä nollaa, niin muuttujassa on paljon puuttuvaa, mutta kertoimen ollessa 1 muuttujassa ei ole puuttuvia havaintoja. Kertoimen laskennassa muodostetaan muuttujista havaintopareja, joista lasketaan yhteen ne parit, kun toinen pareista on puuttuva ja toinen on havaittu. Yhteenlasketut parit jaetaan summalla, jossa puuttuvan ja ei-puuttuvan parien lukumäärään lisätään puuttuvien havaintoparien yhteenlaskettu määrä. (van Buuren 2018, luku 4.1.3.) Suhdeluvun laskentaan on käytetty R -ohjelmiston version 3.4.4. `flux` -funktiota, joka sijaitsee `mice` -kirjastossa (van Buuren & Groothuis-Oudshoorn 2011).

3.2 Moni-imputointi MICE-algoritmin avulla

Moni-imputoinnin keskeisenä ideana on täydentää puuttuvia havaintoja siten, että imputoituja aineistoja muodostetaan useampi määrä. Tällöin havaitut arvot pysyvät samoina, mutta imputoinnissa saadut arvot muuttuvat eri aineistoissa. Moni-imputointiin on käytetty R -ohjelmiston version 3.4.4. `mice` -funktiota, joka sijaitsee `mice` -kirjastossa. (van Buuren & Groothuis-Oudshoorn 2011.)

Yleistettynä moni-imputoinnin MICE-algoritmin vaiheet muodostuvat seuraavasti (van Buuren 2018, luku 4.5.2):

1. Ensin lasketaan imputointimalli yhdelle muuttujalle Y_j ja valitaan alkuimputoinnit satunnaisesti havaittujen Y_j^{obs} joukosta. Laskentaa suoritetaan $t = 1, ..., M$ iteroinnin verran ja toistetaan kaikille muuttujille $j = 1, .., p$. Tällöin saadaan imputointimallin aineisto $\dot{Y}_{-j}^t = (\dot{Y}_1^t, ..., \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, ..., \dot{Y}_p^{t-1})$, joka voidaan olettaa nykyiseksi aineistoksi.

2. Aineiston imputointiparametri $\dot{\phi}_j^t$ voidaan generoida havaittujen arvojen todennäköisyyksien ja imputointimallin avulla eli kaavalla $P(\phi_j^t | Y_j^{obs}, \dot{Y}_{-j}^t, R)$, jossa muuttuja R kuvaa puuttumista: $R = 0$ tarkoittaa havainnon puuttumista ja $R = 1$ havaittua aineiston arvoa.

3. Tämän jälkeen voidaan laskea uusi imputoitujen muuttujien aineisto $\dot{Y}_j^t$ kaavalla $P(Y_j^{miss} | Y_j^{obs}, \dot{Y}_{-j}^t, R, \dot{\phi}_j^t)$.

Algoritmissa esiintyvä ϕ_j^t tarkoittaa imputointiparametria, joka voidaan generoida posteriorijakaumasta $P(\phi_j^t | Y_j^{obs}, \dot{Y}_{-j}^t, R)$ tai sen approksimaatiosta. Jakauman generointimenetelmä valitaan ennen algoritmin suorittamista, joista vaihtoehtoisia tapoja ovat esimerkiksi bayesilaisittain normaalijakaumasta tai bootstrap-menetelmän avulla muodostamalla havaitusta aineistosta useita otoksia, joista estimoidaan imputointimalli jokaiselle muodostetulle otokselle. (van Buuren 2018, luku 3.2.)

Generointia suoritetaan niin kauan, kunnes imputoituja aineistoja on haluttu määrä sekä jokaiselle imputoiduille muuttujien arvoille on laskettu keskiarvo ja varianssi iterointikierroksittain. Lasketuista keskiarvoista ja variansseista voidaan muodostaa kuvaajat, joista tulisi havaita iterointien aikana tapahtuvaa konvergoitumista eli imputointikeskiarvojen ja -varianssien tulisi

jakautua samalla tavalla kaikissa imputoiduissa aineistoissa. Iterointien edetessä arvoissa ei pitäisi olla havaittavissa trendiä eli selkeästi havaittavissa olevaa kasvua tai laskua. (van Buuren 2018, luku 6.5.2.)

Muuttujille, joissa voidaan olettaa olevan selkeästi MNAR-puuttumisrakennetta, tehdään sensitiivisyysanalyysiä muokkaamalla imputointikeskiarvoja. Tällöin generoinnissa muodostuneille imputoitujen arvojen keskiarvoja muutetaan valitun vakioparametrin δ avulla: keskiarvoja voidaan joko laskea ($-\delta$) tai kasvattaa (δ), jolloin imputoinnista saatuja tuloksia muokataan vastaamaan todellisuutta. (van Buuren 2018, luku 3.8.) Tässä tutkielman aineistossa kysymykseen "Työtoverit puuttuvat asiaan, jos yhteisössä esiintyy työpaikkakiusaamista" saatetaan jättää vastaamatta, koska ei haluta antaa negatiivista vastausta. Tällöin havaittuun aineistoon perustuva imputointi saattaa antaa todellisuudesta poikkeavia arvoja, jolloin imputointikeskiarvoja muokataan negatiivisilla δ:n arvoilla.

Koulun hyvinvointiprofiilin aineiston mielipidekysymysten imputointiin käytetään PMM-menetelmää (*Predictive Mean Matching*). Menetelmässä käytettävä imputointiparametri generoidaan ensin bayesilaisittain normaalijakaumasta, jonka jälkeen imputointiin käytettäviä mallin selittävien muuttujien estimoituja arvoja parannetaan minimoimalla imputoitujen arvojen etäisyyksiä havaituista arvoista. Regressiokerroin $\dot{\beta}$ eli edellisessä algoritmissa esiintyvä imputointiparametri generoidaan bayesilaisittain normaalijakaumasta seuraavien vaiheiden mukaan (van Buuren 2018, luku 3.2.2):

1. Lasketaan aineiston ristitulomatriisi $S = X'_{obs}X_{obs}$, jossa X_{obs} tarkoittaa täydellisesti havaittua aineistoa matriisimuodossa ja X'_{obs} sen transpoosia.

2. Lasketaan $V = (S + diag(S)\kappa)^{-1}$.

3. Lasketaan regressiokertoimet $\hat{\beta} = VX'_{obs}y_{obs}$.

4. Muodostetaan satunnainen muuttuja $\dot{g}$, joka noudattaa χ^2 -jakaumaa vapausasteilla $n_1 - q$. Muuttuja n_1 tarkoittaa rivien eli havaintojen lukumäärää täydellisesti havaitussa aineistossa ja q tarkoittaa sarakkeiden eli muuttujien lukumäärää.

5. Lasketaan normaalijakautuneen posteriorijakauman varianssi $\dot{\sigma^2} = (y_{obs} - X_{obs}\hat{\beta})'(y_{obs} - X_{obs}\hat{\beta})/\dot{g}$, jossa y_{obs} kuvaa täydellisesti havaitun aineiston valittua vastemuuttujaa.

6. Arvotaan riippumattomaan virhemuuttujaan arvoja normaalijakaumasta $N(0,1)$ vektoriin $\dot{z_1}$.

7. Lasketaan $V^{1/2}$ Choleskyn hajotelmalla, jolloin saadaan diagonaalimatriisin V arvot neliöjuurella.

8. Lasketaan uusi estimoitu regressiokerroin $\dot{\beta} = \hat{\beta} + \dot{\sigma}\dot{z_1}V^{1/2}$.

Edeltäviä generoinnin vaiheita suoritetaan niin kauan kunnes kaikilla puuttuvia havaintoja sisältävillä jatkuvilla muuttujilla on imputointimalli eli $\dot{y} = \dot{\beta}X_{miss}$. Generointimenetelmän vaiheessa 2, kun muodostetaan matriisi V, valitaan myös positiivinen kerroin κ. Kertoimen avulla pyritään välttämään ongelmia singulaarimatriisin muodostumisessa havaintomatriisin kovarianssirakenteessa, joten kertoimen arvo on yleensä mahdollisimman lähellä nollaa. (van Buuren 2018, luku 6.3.2.) Tämän tutkielman moni-imputoinnissa κ on asetettu arvoksi 0.1.

Imputointimallien generoinnin jälkeen jatketaan mallien sovittamista aineistoon minimoimalla etäisyyksiä havaittujen arvojen ja mallien avulla muodostettujen imputoitujen arvojen välillä. Laskenta tapahtuu seuraavien vaiheiden mukaan (van Buuren 2018, luku 3.4):

1. Olkoon estimoidut havaitut arvot $\hat{y}_i$ ja estimoidut imputoidut havainnot $\hat{y}_j$ sekä näiden etäisyydelle valittu maksimiarvo η, jolloin saadaan kaava
$$|\hat{y}_i - \hat{y}_j| < \eta.$$

2. Valitaan lähin kandidaatti i, joka minimoi lausekkeen $|\hat{y}_i - \hat{y}_j|$.

3. Etsitään kandidaatit d minimoidun etäisyyden lausekkeen $|\hat{y}_i - \hat{y}_j|$ mukaan, ja valitaan yksi niistä havainnoksi. Valittuja d:tä on yleensä 3, 5 tai 10 kappaletta.

4. Imputoidaan havainto arpoen yksi arvo kandidaateista d siten, että kunkin kandidaatin poimintatodennäköisyys riippuu etäisyyksistä $|\hat{y}_i - \hat{y}_j|$.

Tässä tutkielmassa taustamuuttujien imputointiin on käytetty muuttujakohtaisesti sopivia menetelmiä: logistista regressiomallia kaksiarvoisille muuttujille, multinomiaalista logistista regressiomallia moniarvoisille nominaalisille muuttujille ja verrannollista logistista regressiomallia järjestysasteikollisille

muuttujille. PMM-menetelmä sopisi myös kategorisoiduille muuttujille, mutta tässä tutkielmassa menetelmää on käytetty ainoastaan suhdeasteikollisille mielipidekysymyksille.

Logistinen regressiomalli *logreg* on tarkoitettu kaksiarvoisille muuttujille, jolloin lasketaan todennäköisyyksien suhdetta eli voittamisen todennäköisyyden p suhdetta häviöön $1-p$ kaavalla $log(p/1-p)$, ja muodostetaan todennäköisyyksien suhteelle regressiomalli. Tällöin kaksiarvoisen muuttujan todennäköisyyksien suhde toimii imputointimallin vasteena, jota selitetään muilla aineiston muuttujilla: Muuttujille lasketaan estimaatit, jotka arvioivat jokaisen selittäjän todennäköisimmän arvon riippuen siitä, mikä on vasteen saamien arvojen todennäköisyyksien suhde. (Hilbe 2009, s. 297–313.)

Multinomiaalisen logistisen regressiomallin *polyreg* laskenta on lähes sama kuin logistisen regressiomallin, mutta vasteessa luokkia on enemmän kuin kaksi. Tällöin imputointimallin vasteena käytettäviä todennäköisyyksien suhteita ja siihen muodostettavia regressiomalleja tarvitaan $L-1$ kappaletta, jossa L tarkoittaa luokkien lukumäärää yhteensä. (Hilbe 2009, s. 385–399.)

Sen sijaan järjestysasteikollisille muuttujille sopii paremmin verrannollinen logistinen regressiomallinnus eli *polr* -menetelmä, jossa lasketaan todennäköisyyksien suhdetta kumulatiivisesti eli ensin lasketaan ensimmäisen kategorian todennäköisyys suhteessa muihin kategorian arvoihin. Tämän jälkeen lasketaan seuraavan kategorian ja ensimmäisen kategorian todennäköisyyksien summan suhde muihin muuttujan kategorioihin ja niin edelleen, kunnes kaikki muuttujan kategoriat on käyty läpi. Näiden kategorioiden muuttujien suhde toimii samalla tavalla vasteena, kuten edellä mainitussa logistisessa regressiomallissa. (Hilbe 2009, s. 353–376.)

Imputointien jälkeen saadaan aineistoja, joita voidaan ryhmitellä samalla tavalla kuin täydellisesti havaittua aineistoa. Ryhmiin jaotelluista imputoitujen aineistojen mielipidekyselyn muuttujista voidaan tutkia klusterien keskustojen keskiarvoja ja -hajontoja (Basagaña et al. 2013). Tällöin lasketaan imputoitujen aineistojen klusterien keskustojen perusteella keskiarvo $\overline{q}$ jokaiselle valitulle mielipidekyselyn muuttujalle seuraavalla kaavalla (van Buuren 2018, luku 2.3.2):

$$\overline{q} = \frac{1}{m} \sum_{l=1}^{m} \hat{q}_l,$$

jossa $\hat{q}_l$ on valitun muuttujan klusterin keskusta ryhmitellyssä imputoidussa aineistossa l ja m on imputoitujen aineistojen lukumäärä. Lisäksi voidaan

laskea keskiarvolle otoskeskihajonta seuraavalla kaavalla (van Buuren 2018):

$$B = \sqrt{\frac{1}{m-1}\sum_{l=1}^{m}(\hat{q}_l - \overline{q})^2},$$

jossa $\overline{q}$ on edellä laskettu keskiarvo sekä $\hat{q}_l$ on valitun muuttujan klusterin keskusta ryhmitellyssä imputoidussa aineistossa l. Lisäksi m on imputoitujen aineistojen lukumäärä. Tässä tutkielmassa klusterien keskustojen keskiarvo $\overline{q}$ ja varianssi B lasketaan ryhmittelymenetelmään valituille muuttujille. Tulokset on esitetty taulukossa 3.

4 Ryhmittelymenetelmät

Ryhmittelymenetelmät voidaan jakaa kovaan (*hard/crisp clustering*) ja sumeaan klusterointiin (*fuzzy/soft clustering*). Kovat ryhmittelymenetelmät ovat hyvin yleisesti käytettyjä: Niiden ehtona on, että yksi havainto voi sisältyä tasan yhteen klusteriin, jolloin havaintojen lukumäärä tietyssä klusterissa voidaan laskea. (Theodoridis & Koutroumbas 2008, s. 600 & 629.) Kovassa klusteroinnissa ryhmittely voi tapahtua havaintojen etäisyyksien perusteella, jolloin havainnoille ei lasketa ryhmittelytodennäköisyyksiä, kuten sumeissa ryhmittelymenetelmissä. (Hastie et al. 2009, s. 500.)

Seuraavaksi esitellään k:n keskiarvon ja k:n prototyypin ryhmittelyiden matemaattista taustaa. Nämä menetelmät ovat kovia ryvästysmenetelmiä, joissa lasketaan havaintopisteiden euklidisia etäisyyksiä toisistaan, ja pyritään minimoimaan klusterien sisäistä hajontaa siirtämällä havaintoja klusterista toiseen. K:n keskiarvon ja k:n prototyypin ryhmittelymenetelmien yhtenä vaatimuksena on, että ryhmien lukumäärä k on valittava ennen ryvästystä: Jos klustereiden määrä on huonosti valittu, algoritmi ei välttämättä ratkaise aineistoon sopivaa ensisijaista klusterirakennetta (Theodoridis & Koutroumbas 2008, s. 633–634). Lukumäärän valintaan voidaan käyttää apuna indeksejä, joista neljä esitellään luvussa 5.

4.1 *K*:n keskiarvon ryhmittelymenetelmä

K:n keskiarvon ryhmittelymenetelmää (*K-means clustering*) käytetään, kun muuttujien arvot ovat määrällisiä ja jatkuvia. Tällöin havaintojen etäisyydet voidaan laskea neliöidyn euklidisen etäisyyden $d(x_i, x_i') = ||x_i - x_i'||^2$ mukaan. Menetelmän algoritmin keskeisenä periaatteena on, että siinä muutetaan klustereiden keskipisteitä ja siirretään havaintojen sijainteja ryhmästä toiseen niin kauan, kunnes klustereiden havaintojen väliset varianssien summat ovat minimoitu. Tämä aloitetaan siten, että klusterien keskustoille annetaan satunnaiset alkuarvot, joista varianssien minimointiin tarkoitetut iterointikierrokset aloitetaan. Kierrokset tapahtuvat alla esitettyjen vaiheiden mukaan. (Hastie et al. 2009, s. 500–501.)

1. Aineiston alustava ryhmittely C muutetaan minimoimalla klusterien kokonaisvariansseja siten, että muutetaan klustereiden alkuperäisiä keskipisteitä, jotka ovat $m_1, ..., m_k$. Kokonaisvarianssi klusterille voidaan

laskea seuraavan kaavan avulla:

$$\sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2,$$

jossa $N_k = \sum_{i=1}^{N} I(C(i) = k)$ ja $I(C(i) = k)$ on indikaattorifunktio, joka saa arvon 0, jos havainto ei ole lähellä klusteria k ryhmittelyn C mukaan tai arvon 1, jos havainto on lähellä klusteria k ryhmittelyn C mukaan.

2. Muuttuneiden keskipisteiden $m_1, ..., m_k$ ja havaintojen $x_1, .., x_n$ välinen etäisyys minimoidaan ryhmittelemällä jokainen havainto siihen klusteriin k, jonka keskipiste on havaintoa lähimpänä. Tämä voidaan ilmaista kaavalla

$$C(i) = argmin_{1 \leq k \leq K} ||x_i - m_k||^2,$$

joka sijoitetaan edeltävään indikaattorifunktioon.

3. Vaiheita 1 ja 2 toistetaan niin kauan, kunnes havaintojen ryhmittely klusterista toiseen ei enää muutu.

K:n keskiarvon ryhmittelymenetelmä on hyvin yleisesti käytetty monissa sovelluksissa, koska sen laskenta on yksinkertaista ja nopeaa. Menetelmä on kuitenkin herkkä kohinalle ja poikkeaville havainnoille, mikä on hyvä huomioida tuloksia tulkittaessa (Theodoridis & Koutroumbas 2008, s. 633–634). Lisäksi muuttujien tulee olla jatkuvia, koska laskenta suoritetaan euklidisen etäisyyden mukaan. Tutkielmassa käytettävän aineiston taustamuuttujat eivät kuitenkaan ole jatkuvia, jolloin ne voidaan huomioida k:n keskiarvon ryhmittelyn muunnelman avulla eli k:n prototyypin ryhmittelymenetelmällä.

4.2 *K*:n prototyypin ryhmittelymenetelmä

K:n prototyypin ryhmittelymenetelmässä (*K-prototypes clustering*) muodostetaan valittu k määrä prototyyppejä eli mahdollisia klustereita, joita päivitetään niin kauan, että saadaan mahdollisimman pieni hajonta klusterin sisäisten havaintojen etäisyyksille. Menetelmän laskenta tapahtuu samalla tavalla kuin edellä mainitun k:n keskiarvon ryhmittelymenetelmän, mutta siinä huomioidaan numeeristen muuttujien lisäksi myös kategorisoidut muuttujat. Tällöin jatkuville muuttujille lasketaan ensin euklidisten etäisyyksien

mukaan etäisyysmatriisi, johon summataan kategoristen muuttujien vaikutus painokertoimen ja indikaattorifunktion avulla. Summausta ei tehdä, jos kategorisen muuttujan arvot ovat samoja: Tällöin indikaattorifunktio saa arvon nolla, jolloin myös havaintojen etäisyys asetetaan nollaksi. Arvojen ollessa erilaiset indikaattorifunktio saa arvon yksi, jolloin tehdään summaus lisäämällä euklidiseen etäisyysmatriisiin kategoriselle muuttujalle laskettu estimoitu etäisyys. (Huang 1997.)

Yhteenvetona k:n prototyypin ryhmittelymenetelmässä käytettävä erilaisuusmitta kategoristen muuttujien X_i ja numeeristen muuttujien Z_l välillä saadaan seuraavan kaavan avulla (Huang 1997):

$$d(X_i, Z_l) = \sum_{j=1}^{m_r} (x_{ij}^r - z_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} \delta(x_{ij}^c, z_{lj}^c),$$

jossa $\delta(x_{ij}^c, z_{lj}^c)$ on kategorisille muuttujille tehty indikaattorifunktio siten, että $\delta(x_{ij}^c, z_{lj}^c)$ saa arvon nolla, kun $x_{ij}^c = z_{lj}^c$ ja $\delta(x_{ij}^c, z_{lj}^c)$ saa arvon 1, kun $x_{ij}^c \neq z_{lj}^c$. Muuttujat x_{ij}^c ja z_{lj}^c ovat kategoristen muuttujien arvoja klusterissa l, kun taas muuttujat x_{ij}^r ja z_{lj}^r ovat numeerisia eli jatkuvien muuttujien arvoja klusterissa l. Kaavassa esiintyvä m_r tarkoittaa numeeristen muuttujien yhteenlaskettua määrää ja m_c kategoristen muuttujien lukumäärää. (Huang 1997.)

Estimoitu kategoristen muuttujien painokerroin γ_l riippuu numeeristen muuttujien keskihajonnoista klusterissa l. Painottamisen avulla voidaan kasvattaa kategoristen muuttujien saamien etäisyyksien arvoja, mutta jos kerroin saa arvon nolla, $\gamma_l = 0$, menetelmä vastaa tavallista k:n keskiarvon ryhmittelymenetelmää eikä kategoristen muuttujien erilaiset arvot vaikuta klusterointiin. (Huang, 1997.) Tässä tutkielmassa kategoristen muuttujien painokerroin γ_l estimoidaan automaattisesti mitta-asteikollisten mielipidemuuttujien keskihajontojen jakaumien mukaan, jolloin niiden merkitys huomioidaan painottaen, kun $x_{ij}^c \neq z_{lj}^c$.

Tässä tutkielmassa aineisto ryhmitellään k:n prototyypin ryvästysmenetelmän avulla, koska ryhmittelymenetelmä sopii tällaisille suhdeasteikollisille muuttujille (Hastie et al. 2009, s. 500). Ryhmittelyn toteutukseen on käytetty R -ohjelmiston version 3.4.4. `kproto` -funktiota, joka sijaitsee `clustMixType` -kirjastossa (Szepannek 2018).

4.3 Muuttujien valinta ryhmittelymenetelmissä

Muuttujan merkitystä ryhmittelymenetelmissä voidaan perustella luokitteluvirheen avulla siten, että jokaisen muuttujan arvoja sekoitetaan satunnaisesti useaan kertaan ja ryhmittelyt toistetaan: luokitteluvirheiden avulla voidaan verrata, miten sekoitettujen aineistojen ryhmittelyt poikkeavat alkuperäisen aineiston ryhmittelystä. (Fisher et al. 2019). Kyseinen luokitteluvirheisiin perustuva menetelmä, jolla muuttujien merkitystä ryhmittelyssä arvioidaan, muistuttaa Breimanin (2001) kehittämää luokittelumenetelmää nimeltä satunnainen metsä (*Random Forest*).

Tässä tutkielmassa käytetyssä muuttujien valinnassa tehdään k:n prototyypin ryhmittely koko aineistolle, joka sisältää kaikki aineiston muuttujat. Ryhmittelyn luotettavuuden lisäämiseksi valitaan sopiva ryhmien lukumäärä luvussa 5 esiteltävien kriteerien avulla. Klustereiden lukumäärän valinnan, ja sen perusteella tehdyn ryhmittelyn jälkeen, sekoitetaan aineiston muuttujien saamat arvot satunnaisesti 5 kertaa Pfaffelin (2019) kehittelemän R-ohjelmiston version 3.4.4. `FeatureImpCluster` -funktiolla, joka sijaitsee `FeatureImpCluster` -kirjastossa. Sekoitettuihin aineistoihin sovitetaan uudestaan ryhmittely alkuperäisen aineiston mukaan samalla ryhmien lukumäärällä, ja tutkitaan luokitteluvirheen keskiarvon avulla, miten alkuperäisen aineiston sekä sekoitettujen aineistojen klusterirakenteet poikkeavat toisistaan.

Luokitteluvirheen keskiarvo lasketaan siten, että summataan yhteen kaikkien sekoitettujen aineistojen saamat klusterien arvot, jotka eroavat alkuperäisestä, ja jaetaan havaintojen lukumäärällä: Jos luokitteluvirheiden saama keskiarvo on suuri, niin sekoitettujen aineistojen sovitetut ryhmittelyt poikkeavat alkuperäisen aineiston ryhmittelystä (taulukko 1). Tällaiset muuttujat sopivat hyvin ryhmittelymenetelmiin, koska muuttujien arvojen sekoittaminen muuttaa klusterirakennetta, ja siten arvot ovat toisistaan eroavia. (Pfaffel 2019.) Seuraavaksi esitellään taulukkomuodossa havainnollistava esimerkki muuttujien sekoittamisesta, ja sen vaikutuksesta klusterirakenteeseen. Taulukossa yhden muuttujan neljä arvoa sekoitetaan satunnaisesti, jolloin myös niiden saaman klusterin arvo siirtyy vastaamaan samaa riviä. Jos klusterien arvoissa k havaitaan eroja alkuperäisen ja sekoitetun aineiston välillä, niin ryhmämuuttujan saamat arvot poikkeavat toisistaan, ja luokitteluvirheen arvo kasvaa.

Taulukko 1: Yksinkertainen esimerkki aineiston muuttujien valinnasta. Ensimmäisessä sarakkeessa on aineiston alkuperäisen muuttujan neljä kuvitteellista arvoa asteikolla 0–100, kolmannessa ja viidennessä sarakkeessa nämä arvot on sekoitettu satunnaisesti. Toisessa, neljännessä ja kuudennessa sarakkeessa ovat klusterien arvot, joihin muuttujien arvot ovat jaoteltu. Kun verrataan sarakkeen k_0 arvoja sekoitettujen muuttujien (1. sekoitus ja 2. sekoitus) arvoihin k_1 ja k_2, saadaan luokitteluvirheen keskiarvoksi $\frac{(3+4)}{4} = \frac{7}{4} = 1,75$.

Arvo	k_0	1. sekoitus	k_1	2. sekoitus	k_2
50	2	71	2	1	1
1	1	90	3	50	2
71	2	1	1	90	3
90	3	50	2	71	2

5 Sisäiset ja ulkoiset validointikriteerit ryhmittelymenetelmissä

Validointikriteerien indeksien avulla voidaan määritellä, millainen ryhmittely sopii parhaiten aineistoon. Kriteerien indeksit jaetaan yleensä kahteen: ulkoisiin ja sisäisiin kriteereihin (Aggarwal & Reddy 2014). Ulkoisessa validoinnissa oletetaan jo aikaisempaa tietoa aineistosta, ja siihen tehdyistä menetelmistä, mutta sisäinen validointi sopii aineistolähtöisiin ongelmiin, koska siinä ei tehdä oletuksia etukäteen aineiston ominaisuuksista. (Rendón et al. 2011).

Ulkoisen Rand-indeksin saamien arvojen avulla voidaan vertailla erilaisia klusterirakenteita. Tällöin voidaan esimerkiksi tutkia kahden eri ryhmittelymenetelmän tai kahteen eri aineistoon tehtyjen ryhmittelyiden klusterirakenteiden eroavaisuuksia. (Rand 1971.) Tässä tutkielmassa Rand-indeksiä käytetään täydellisesti havaitun aineiston ja moni-imputoitujen aineistojen klusterirakenteiden vertailemiseen, kun on käytetty k:n prototyypin ryhmittelymenetelmää.

Sisäisen validoinnin indekseistä voidaan erotella vielä relatiiviset indeksit, joita voidaan käyttää esimerkiksi klustereiden lukumäärän valitsemiseen. Tällöin aineiston ryhmittely suoritetaan yhden ryhmittelymenetelmän algoritmin mukaan useaan kertaan, joissa jokaisessa ryhmien lukumäärä on erilainen: Menetelmän tuloksista saatujen indeksien arvoja voidaan vertailla ja va-

lita niiden perusteella paras ryhmien lukumäärä aineistoon. (Theodoridis & Koutroumbas 2008, s. 747.) Tässä tutkielmassa sisäisiä validointikriteerien indeksejä on valittu useampi, jotta saadaan varmistettua paremmin aineistoon sopiva klustereiden lukumäärä. Kahdesta validointikriteeristä eli indeksistä tutkitaan niiden saamaa maksimiarvoa (Dunn-indeksi ja Calinski–Harabasz-indeksi) ja vastaavasti kahdesta muusta indeksistä tutkitaan minimiarvoa (Davies–Bouldin-indeksi ja C-indeksi). Indeksit lasketaan k:n prototyypin ryhmittelyille, joissa klustereiden lukumäärä asetetaan 2–10 ryhmäksi. Indeksien avulla voidaan tutkia, millä ryhmittelyllä indeksit saavat minimi- tai maksimiarvonsa.

Seuraavaksi esitellään yhden ulkoisen validointikriteerin Rand-indeksin ja neljän valitun sisäisten validointikriteerien laskennallista teoriaa, joiden taustalla olevat matemaattiset kaavat perustuvat yksittäisten havaintojen tiheyteen ja klustereiden erilaisuusmittaan. Tiheyttä kuvataan yleensä variansilla, jonka avulla voidaan mitata havaintojen hajontaa aineistossa. Klustereiden erilaisuusmitta sen sijaan osoittaa, kuinka kaukana kaksi eri klusteria sijaitsevat toisistaan. (Rendón et al. 2011.) Indeksien laskentaan on käytetty version 3.4.4. sisäisten validointikriteerien laskentaan `intCriteria` -funktiota ja ulkoisen Rand-indeksin laskentaan `extCriteria` -funktioita, jotka sijaitsevat `clusterCrit` -kirjastossa (Desgraupes 2017).

5.1 Rand-indeksi

Rand-indeksi on Randin (1971) kehittämä ulkoinen validointikriteeri, jonka tarkoituksena on mitata kahden eri ryhmittelyn samanlaisuutta. Niiden avulla voidaan tulkita aineiston arvojen jakautuneisuutta eri klustereihin, mutta niiden avulla ei voi tulkita ryhmittelymenetelmän tuottaman ryvästymisen sopivuutta. Laskennassa tarvitaan kaksi eri ryhmiteltyä aineistoa X ja Y, joista laskettujen havaintoparien lukumäärien yhtäläisyyksistä saadaan neljä mahdollista arvoa a, b, c ja d:

- a tarkoittaa niiden havaintoparien lukumäärää, jotka ovat samassa klustereissa sekä X että Y ryhmitellyssä aineistossa.
- b tarkoittaa niiden havaintoparien lukumäärää, jotka ovat eri klusterissa ryhmitellyssä aineistossa X ja Y.
- c tarkoittaa niiden havaintoparien lukumäärää, jotka ovat samassa klusterissa ryhmitellyssä aineistossa X, mutta eri klusterissa aineistossa Y.

- d tarkoittaa niiden havaintoparien lukumäärää, jotka ovat samassa klusterissa ryhmitellyssä aineistossa Y, mutta eri klusterissa aineistossa X.

Tällöin voidaan muodostaa Rand-indeksin laskukaava:

$$R = \frac{a+b}{a+b+c+d},$$

jossa $a+b+c+d$ tarkoittaa havaintoparien kokonaismäärää, ja se voidaan ilmaista myös kaavalla $N(N-1)/2$, kun N on aineistojen kaikkien havaintojen lukumäärä. Indeksi saa arvoja nollan ja yhden väliltä: Mitä lähempänä indeksin arvo on ykköstä niin sitä samanlaisempia ovat tutkittavat klusterirakenteet X ja Y. Sen sijaan indeksin arvon lähestyessä nollaa, sitä enemmän rakenteet poikkeavat toisistaan. (Rand 1971.)

5.2 Dunn-indeksi

Dunn-indeksi on Dunnin (1974) kehittämä validointikriteeri koville klusterointimenetelmille. Indeksin laskemisessa tarvitaan kahden eri klusterin minimietäisyys ja klusterin sisällä olevien havaintojen maksimietäisyys eli klusterin läpimitta. Tämän jälkeen voidaan muodostaa indeksin laskemiseen tarvittava kaava:

$$D_m = \min_{i=1,..,m} \left[\min_{j=i+1,..,m} \left(\frac{d(C_i, C_j)}{\max_{k=1,...,m} diam(C_k)} \right) \right],$$

jossa

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

on minimietäisyys klustereiden C_i ja C_j välillä ja

$$diam(C_k) = \max_{x,y \in C_k} d(x, y)$$

on klusterin C_k läpimitta. Dunn-indeksin arvo kasvaa, jos klusterit ovat etäällä toisistaan ja havaintojen etäisyydet ovat klustereiden sisällä pieniä. Tällöin ryhmittely on sopiva aineistolle, koska se erottelee hyvin havainnot toisistaan. Huomioitavaa on, että luotettavien päätelmien tekemiseksi indeksin olisi hyvä olla yli yhden, koska tällöin aineistoon on muodostunut hyvin eroteltavissa oleva klusterirakenne. (Dunn 1974.)

5.3 Davies–Bouldin-indeksi

Daviesin ja Bouldinin (1979) kehittämässä validointikriteerissä eli DB-indeksissä valitaan matriisi R_{ij} mittaamaan klusteroinnin järjestäytymisen onnistumista. Kaavat s_i ja s_j kuvaavat klustereiden C_i ja C_j etäisyyksien keskiarvoa sekä klustereiden välistä etäisyyttä kuvataan määreen $d(C_i, C_j)$ avulla, joka on identtinen symmetrisen d_{ij} etäisyysmatriisin kanssa. Klusterin C_i havaintojen etäisyyksien keskiarvo lasketaan kaavalla

$$s_i = \left(\frac{1}{n_i} \sum_{x \in C_i} ||x_i - \overline{x}_{C_i}|| \right),$$

jossa n_i tarkoittaa laskettujen etäisyysvektoreiden lukumäärää klusterissa C_i sekä $||x_i - \overline{x}_{C_i}||$ havainnon x_i ja klusterin keskipisteen $\overline{x}_{C_i}$ välistä etäisyyttä. Klusterin C_j etäisyyksien keskiarvo s_j lasketaan samalla tavalla, mutta tällöin i on j. (Davies & Bouldin 1979.)

Järjestäytymistä kuvaavalle matriisille on asetettu muutamia ehtoja, kun R_{ij} on positiivinen ja symmetrinen. Ehdot ovat seuraavat:

1. $R_{ij} \geq 0$.
2. $R_{ij} = R_{ji}$.
3. Jos $s_i = 0$ ja $s_j = 0$, niin $R_{ij} = 0$.
4. Jos $s_j > s_k$ ja $d_{ij} = d_{ik}$, niin $R_{ij} > R_{ik}$.
5. Jos $s_j = s_k$ ja $d_{ij} < d_{ik}$, niin $R_{ij} > R_{ik}$.

Yksinkertaisin kaava R_{ij} muodostamiseksi, joka täyttää annetut ehdot, on seuraava:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}.$$

Lisäksi R_i eli matriisin R_{ij} maksimiarvo $R_i = \max_{j=1,...,m} R_{ij}$ on määritelty kaikille $i = 1, ..., m$, kun i ei ole sama kuin j. Tällöin voidaan muodostaa Davies–Bouldin-indeksi:

$$DB_m = \frac{1}{m} \sum_{i=1}^{m} R_i.$$

Indeksi on keskiarvo kaikkien klustereiden välisille samankaltaisuudelle. Kun klusterit ovat keskenään mahdollisimman vähän yhtäläisiä, voidaan päätellä ryhmittelyn olevan sopiva aineistoon: Tällöin Davies–Bouldin-indeksin arvo on oltava mahdollisimman pieni. (Davies & Bouldin 1979.)

5.4 C-indeksi

C-indeksi on esitetty ensimmäistä kertaa Hubertin ja Levinin (1976) tutkimusartikkelissa. Se on myöhemmin todistettu toimivaksi Milliganin (1981) artikkelissa. Indeksin laskemiseksi tutkitaan jokaisen klusterin sisällä olevien havaintoparien etäisyyksiä toisistaan. Tällöin ei lasketa kokonaisvarianssia klusterin sisällä olevien havaintojen välillä, vaan oletetaan, että jokainen klusteri C_k sisältää $n_k(n_k - 1)/2$ paria, joiden yhteenlaskettu summa on:

$$N_W = \sum_{k=1}^{K} \frac{n_k(n_k - 1)}{2}.$$

Summausta N_W klustereiden sisältämien parien määrästä käytetään kolmen eri muuttujan laskennassa:

1. S_W on summa jokaisen klusterin havaintoparien välisistä etäisyyksistä eli summataan yhteen kaikki aineistoon muodostuneiden klusterien havaintoparien N_W etäisyydet.

2. S_{min} on summa pienimmistä N_W parien välisistä etäisyyksistä, kun otetaan huomioon kaikki ryhmittelyssä käytettävät havainnot.

3. S_{max} on sen sijaan summa pisimmistä N_W parien välisistä etäisyyksistä otettaessa huomioon kaikki ryhmittelyn havainnot.

C-indeksin lopullinen laskentakaava näyttää seuraavalta:

$$C = \frac{S_W - S_{min}}{S_{max} - S_{min}}.$$

Tällöin klusterien sisällä olevien kaikkien parien etäisyyksistä vähennetään kaikkien havaintoparien pienimmät etäisyydet eli $S_W - S_{min}$: Jos erotus on mahdollisimman pieni, niin klusterin etäisyyksien ja lyhyimpien parien etäisyyksien välillä ei ole suurta eroa eli havainnot ovat ryhmittäytyneet tiettyyn paikkaan. Tämä vielä jaetaan erotuksella, jossa vähennetään kaikkien havaintoparien pisimmistä etäisyyksistä kaikkien havaintoparien lyhimmät etäisyydet eli $S_{max} - S_{min}$, jonka tulisi olla suurempi kuin $S_W - S_{min}$. Tällöin C-indeksin arvo on mahdollisimman pieni: ryhmittely sopii aineistoon, koska havaintoparit ovat klusterien sisällä lähekkäin, mutta kaikkien havaintoparien etäisyydet toisistaan ovat kauempana. (Hubert & Levin 1976.)

5.5 Calinski–Harabasz-indeksi

Calinskin ja Harabaszin (1974) esittelemän indeksin laskenta perustuu klustereiden sisällä olevien havaintojen välisten etäisyyksien summauksiin ja klustereiden sisäisiin variansseihin. Havainnot jaotellaan k:n prototyypin ryhmittelymenetelmän muodostamien klustereiden keskipisteiden ympärille. Keskipisteiden avulla voidaan laskea klusterien sisäisiä ja välisiä variansseja.

Sisäiset varianssit voidaan laskea seuraavan kaavan avulla:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}||x_i - x_{i'}||^2 = \sum_{k=1}^{K}N_k\sum_{C(i)=k}||x_i - \overline{x}_k||^2,$$

jossa $N_k = \sum_{i=1}^{N} I(C(i) = k)$ sisältää indikaattorifunktion $I(C(i) = k)$, jonka tulokseksi saadaan joko 0 eli havainto ei ole lähellä klusteria k ryhmittelyn C mukaan tai tulos 1 eli havainto on lähellä klusteria k ryhmittelyn C mukaan. Lisäksi jokaisesta klusterin havainnosta lasketaan etäisyys k:n prototyypin ryhmittelyssä klusterin k muodostuneeseen keskipisteen $\overline{x}_k$ arvoon. Jokaiselle klusterille saadut sisäiset varianssit summataan vielä yhteen. (Hastie et al. 2009, s. 509.) Kuva 4 havainnollistaa havaintojen etäisyyksien laskentaa keskipisteestä, ja etäisyyksien avulla voidaan laskea myös klusterin sisäinen varianssi.

Sisäisten varianssien lisäksi tarvitaan klustereiden välillä olevien etäisyyksien hajontaa, joka saadaan seuraavan kaavan avulla:

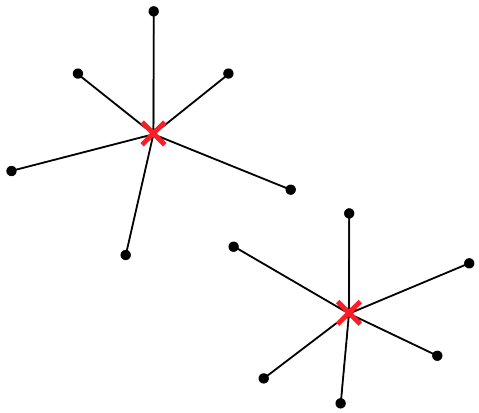
$$B(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i)\neq k} d_{ii'},$$

jossa $d_{ii'}$ tarkoittaa klusterissa sijaitsevien havaintojen keskiarvojen etäisyyttä keskiarvosta $\overline{x}$, joka sijaitsee klusterien välillä (Hastie et al. 2009, s. 508). Kuva 5 havainnollistaa klustereiden välisten etäisyyksien laskentaa, jolloin klustereiden keskipisteiden etäisyydet riippuvat niiden yhteisestä lasketusta keskiarvosta.

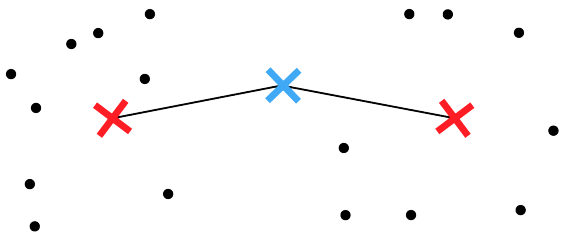
Näiden kahden kaavan avulla voidaan laskea Calinski–Harabasz-indeksi, jonka perusteella voidaan määrittää klusterien lukumäärän sopivuutta aineistoon. Indeksin kaava on seuraava:

$$CH = \frac{B(C)/(K-1)}{W(C)/(N-K)},$$

jossa K tarkoittaa klustereiden lukumäärää ja N havaintojen lukumäärää. Tällöin siis suhteutetaan klustereiden välillä olevaa etäisyyksien hajontaa klustereiden sisäiseen hajontaan. Jos indeksin arvo on suuri, niin ryhmien lukumäärä sopii aineistolle, koska tällöin klusterien välinen hajonta on mahdollisimman suuri, ja klustereiden sisällä oleva hajonta on mahdollisimman pieni. (Calinski & Harabasz 1974.)



Kuva 4: Havainnollistava kuva klustereiden sisäisten etäisyyksien laskennasta (Rezaei 2016). Kuvassa punaiset rastit ovat klustereille laskettuja keskiarvoja ja mustat pisteet ovat havaintoja. Piirretyt viivat kuvaavat laskettua etäisyyttä keskiarvosta.



Kuva 5: Havainnollistava kuva klustereiden välisten etäisyyksien laskennasta (Rezaei 2016). Kuvassa sininen rasti on klusterien välille laskettu keskiarvo sekä mustat pisteet ovat havaintoja ja punaiset rastit ovat klustereille lasketuja keskiarvoja. Piirretyt viivat kuvaavat laskettua etäisyyttä klustereiden sisäisten keskiarvojen etäisyydestä klustereiden väliseen keskiarvoon.

Ennen tuloksia esitellään vielä yhteenveto tutkielman työnvaiheista. Vaiheet ovat numeroitu allekkain algoritmiin, josta nähdään, miten k:n prototyypin ryhmittelymenetelmää ja moni-imputointia on sovellettu tutkielmassa käytettävään työhyvinvointiaineistoon. Lisäksi algoritmi selventää, missä vaiheissa sisäisiä indeksejä on laskettu klustereiden lukumäärän valitsemiseksi.

Yhteenveto Työjärjestys tutkielman ryhmittelyn ja moni-imputoinnin toteutuksesta.

1: Tehdään k:n prototyypin ryhmittely 2–10 ryhmälle täydellisesti havaitulle aineistolle, joka sisältää 76 muuttujaa ja 927 havaintoa.
   1.a: Valitaan aineistoon sopiva ryhmien lukumäärä Davies–Bouldin- ja Calinski–Harabasz-indeksien avulla.
2: Tehdään muuttujien valinta sopivalla ryhmien lukumäärällä perustuen sekoitettujen aineistojen avulla laskettuihin luokitteluvirheisiin.
   2.a: Ryhmitellään täydellisesti havaittu aineisto 2–10 ryhmällä, joka sisältää 24 valittua muuttujaa.
   2.b: Valitaan osa-aineistoon sopiva ryhmien lukumäärä Davies–Bouldin- ja Calinski–Harabasz-indeksien avulla.
   2.c: Tehdään taulukko täydellisesti havaitun aineiston 22 muuttujan k:n prototyypin ryhmittelystä muodostuneista klusterien keskustojen arvoista sekä lasketaan havaintojen lukumäärät klusterittain kahdelle taustamuuttujalle.
3: Muodostetaan moni-imputointimalli.
   3.a: Valitaan moni-imputointiin käytettävät selittävät muuttujat perustuen muuttujien hyödyllisyyteen imputoinnissa.
   3.b: Valitaan käytettävät imputointimallin menetelmät muuttujakohtaisesti.
   3.c: Tehdään sensitiivisyysanalyysi muuttujalle ”Työtoverit puuttuvat asiaan, jos yhteisössä esiintyy työpaikkakiusaamista”.
4: Imputoidaan MICE-algoritmilla viisi erilaista aineistoa 40 iteroinnilla.
   4.a: Ryhmitellään imputoidut aineistot 2–10 ryhmällä, joka sisältää aikaisemmin valitut 24 muuttujaa ja 1643 havaintoa.
   4.b: Valitaan aineistoon sopiva ryhmien lukumäärä Davies–Bouldin- ja Calinski–Harabasz-indeksien avulla.
5: Lasketaan Rand-indeksin arvot vertaillen moni-imputoituja aineistoja keskenään sekä täydellisesti havaittua aineistoa ja moni-imputoituja aineistoja.
6: Lasketaan viiden imputoidun aineiston klusterien keskustojen keskiarvot ja -hajonnat valitulle mielipidekyselyn muuttujalle sekä lukumäärien keskiarvot kahdelle taustamuuttujalle.

6 Tulokset

Sisäisten Davies–Bouldin- ja Calinski–Harabasz-indeksien perusteella sekä täydellisesti havaittuun että moni-imputoituun aineistoon sopii parhaiten 2 klusteria (liite D). Dunn-indeksin arvot ovat pienempiä kuin yksi, joten indeksin saamat arvot eivät ole luotettavasti tulkittavia (Dunn 1974). Lisäksi C-indeksin arvot ovat hyvin lähellä toisiaan, jolloin pienin indeksin saama arvo on haasteellista valita. C-indeksin arvoista voidaan kuitenkin todeta, että ryhmiä pitäisi valita enemmän kuin Davies–Bouldin- ja Calinski–Harabasz-indeksien perusteella. Tämä saattaa johtua laskennallisesta erosta, koska C-indeksissä muodostetaan havaintopareja aineistosta, ja tutkitaan niiden etäisyyksiä toisistaan: Indeksin laskennassa ei huomioida kaikkia yhden klusterin sisäisiä etäisyyksiä (Hubert et al. 1976).

Liitteen B taulukosta B1 voidaan todeta, että alle vuoden samassa koulussa työskennelleistä suurin osa on toisessa klusterissa, kun taas 6–10 vuotta työskennelleistä suurin osa on ensimmäisessä klusterissa. Liitteen B taulukosta B3 voidaan sen sijaan todeta, että täydellisesti havaitun aineiston ikämuuttujan ryhmistä 26–30-vuotiaat sekä 41–45-vuotiaat suurin osa havainnoista sijaitsee ensimmäisessä klusterissa. Sen sijaan alle 25- tai 25-vuotiaista sekä 51–55-vuotiaista suurin osa havainnoista ovat toisessa ns. positiivisesti ajattelevien klusterissa. Muissa ikäryhmissä vastaavia eroja ei ole havaittavissa, koska havaintojen lukumäärien erot klusterittain ovat 7 tai alle.

Taulukosta 2 voidaan todeta, että täydellisesti havaitussa aineistossa ensimmäinen klusteri edustaa negatiivisesti koulutyöhyvinvoinnistaan ajattelevia eli kysymyksiin vastataan pienempiä arvoja, kun taas toisessa klusterissa on positiivisesti työhyvinvointiinsa suhtautuvia eli kysymyksiin vastataan suurempia arvoja. Lisäksi klusterissa 1 vastaajat kokevat stressiä ja väsymystä useammin kuin klusterissa 2 olevat henkilöt.

Taulukko 2: Valituista mielipidekysymyksistä täydellisesti havaitun aineiston klusterien 1 ja 2 keskustat. Taulukon sarakkeiden nimistä K_1 tarkoittaa klusterin 1 keskustaa, K_2 tarkoittaa klusterin 2 keskustaa, jotka ovat muodostuneet luvussa 4.2 esiteltävän k:n prototyypin laskennassa.

Kysymykset	K_1	K_2
Koulun opetustilat ovat hyvät	41.86	64.94
Työtä on koulussa sopivasti	43.63	73.15
Työskentely koulussa ei tunnu liian kiireiseltä	33.06	64.09
Työjärjestys on hyvä	56.32	79.43
Tarvittaessa saan apua työterveyshuollosta	62.26	78.66
Koulumme työntekijät viihtyvät hyvin yhdessä	63.68	84.74
Yhteistyö sujuu hyvin koulussamme	64.11	86.63
Työyhteisön jäseniä kohdellaan tasapuolisesti	54.45	85.25
Työtoverit puuttuvat asiaan jos yhteisössä esiintyy työpaikkakiusaamista	56.26	78.27
Kouluyhteisössä rohkaistaan ilmaisemaan oma mielipiteeni	59.55	81.80
Minulta ei odoteta liikaa työssäni	53.99	80.26
Työskentelytahti on minulle sopiva	59.04	84.54
Saan kiitosta jos olen suoriutunut hyvin työstäni	57.33	80.64
Ruokailu on rentouttava tauko	29.22	51.82
Meidän koulussamme ei ole työpaikkakiusaamista	61.68	86.73
Tarvittaessa saan työnohjausta	48.87	71.89
Ruokailutila on viihtyisä	47.38	68.09
Koulun piha on viihtyisä	47.91	68.20
Sääntöjen rikkomisia käsitellään oikeudenmukaisesti	63.18	81.53
Mielipiteeni otetaan huomioon koulun kehittämisessä	59.40	80.78
Olen väsynyt	63.80	47.22
Työni koulussa aiheuttaa minulle stressiä	59.08	36.04

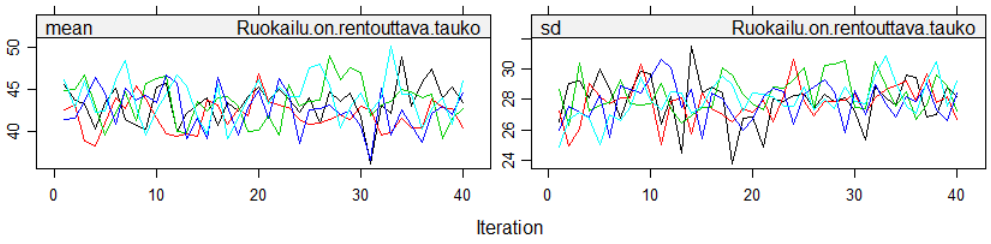
Ennen aineiston moni-imputointia valitaan jokaiselle imputoitavalle muuttujalle selittävät muuttujat. Tässä tutkielmassa imputointimallin selittäväksi muuttujiksi otetaan mukaan liitteen A taulukossa esitetyt valitut muuttujat, koska niiden avulla tehdään ryhmittely myös moni-imputoituihin aineistoihin. Tutkittaessa muiden aineiston muuttujien hyödyllisyyttä selittävänä muuttujana voidaan laskea jokaiselle muuttujalle luvun 3.1 toisessa kappaleessa esitelty `outflux`-kerroin. Aineistossa on 13 muuttujaa, joiden `outflux`-kerroin on alle 0.5: Muuttujissa on liikaa puuttuvuutta toimiakseen imputointimallin selittävänä muuttujana. Nämä kaikki muuttujat ovat terveydentilaa kuvaavia muuttujia, joista kaksi ryhmittelyn kannalta tärkeää muuttujaa säilytetään: ”Olen väsynyt” ja ”Työni koulussa aiheuttaa minulle stressiä”. Sen sijaan loput 11 terveydentilan muuttujaa poistetaan alkuperäisestä aineistosta, jolloin imputoitavaan aineistoon jää 65 muuttujaa. Muuttujat poistetaan, koska luvussa 4.3 esiteltyjen luokitteluvirheiden perusteella muuttujilla ei ole perusteltua syytä pitää ryhmittelymenetelmien analyyseissä ja koska `outflux`-kertoimen perusteella muuttujista ei ole hyötyä imputoinnissa. Lisäksi asetetaan `quickpred`-funktion avulla käytettävissä olevien tilastoyksiköiden määräksi vähintään 50 prosenttia sekä imputoitavien muuttujien ja selittävien muuttujan välinen korrelaatio tulee olla vähintään 0.2, koska tällöin suurin osa puuttuvista havainnoista voidaan imputoida 15–25 selittävän muuttujan avulla: tämä on riittävä määrä selittäjiä imputoitaville muuttujille (van Buuren 2018, luku 9.1.6).

Tässä tutkielmassa tehdään viiden aineiston imputointi 40 iteroinnilla. Imputoidut aineistot voidaan ryhmitellä ja valita sopiva ryhmien lukumäärä Davies–Bouldin- ja Calinski– Harabasz-indeksien avulla. Ryhmiteltyjä moni-imputoituja aineistoja vertaillessa klusterirakenteet ovat hyvin samanlaisia: Rand-indeksien arvot ovat yli 0.9.

Liitteen C taulukossa C1 on esitetty Rand-indeksit, joiden avulla verrataan täydellisesti havaitun aineiston ($N = 927$) klusterirakennetta täyteen moni-imputoituun aineistoon ($N = 1643$). Taulukossa C2 on sen sijaan arvottu moni-imputoiduista aineistoissa saman verran havaintoja kuin täydellisesti havaitussa aineistossa on ($N = 927$). Taulukoista voidaan tarkastella Davies–Bouldin- ja Calinski–Harabasz-indeksien perusteella kahden klusterin saamia Rand-indeksin tuloksia. Tällöin noin puolet moni-imputoitujen aineistojen havainnoista ovat samassa klusterissa kuin täydellisesti havaitussa aineistossa. Rand-indeksin arvosta ei voida kuitenkaan määrittää esimerkiksi sitä, onko osa täydellisesti havaitun aineiston vastauksista eri klusterissa kuin moni-imputoitujen aineistojen ryhmittelyssä, mikä kertoisi parem-

min klusterirakenteen muuttumisesta alkuperäisestä täydellisesti havaitusta aineistosta.

Moni-imputoinnin konvergenssitarkasteluiden perusteella voidaan päätellä imputoinnin onnistuvan hyvin valituilla selittävillä muuttujilla. Kuvassa 6 esitellään esimerkkinä yhden imputoidun muuttujan konvergointitulokset. Kuvasta voidaan havaita, että imputoitujen arvojen keskiarvoissa ja -hajonnoissa ei tapahdu 40 iteroinnin aikana suuria muutoksia. Muutamat yksittäiset poikkeamat tasaantuvat iterointien edetessä, eikä imputaatiokeskiarvot ja -hajonnat kasva tai laske iterointien aikana, joten trendiä ei ole havaittavissa. Muiden muuttujien imputoitujen arvojen konvergointitulokset ovat samankaltaisia, joten niitä ei esitellä tarkemmin tässä työssä. Konvergoitumista ei olla kuitenkaan tutkittu tieteellisesti riittävän systemaattisesti moni-imputoinnin yhteydessä, joten kuvaajiin ei voida sokeasti luottaa (van Buuren 2018).



Kuva 6: Yhden muuttujan konvergointituloksista havainnollistava esimerkki, jossa muuttujan ”Ruokailu on rentouttava tauko” keskiarvot (vasemmalla) ja keskihajonnat (oikealla) jokaiselle imputoinnille, kun imputoitujen aineistojen määrä on 5 ja iterointien määrä 40. Konvergoituminen ei ole täydellistä, mutta suuria poikkeamia ei ole havaittavissa.

Moni-imputoitujen aineistojen klustereiden keskustojen keskiarvoista voidaan todeta samaa kuin täydellisesti havaitusta aineistosta (taulukko 3): Klusteri 1 edustaa negatiivisesti koulutyöhyvinvoinnistaan ajattelevia, kun taas klusterissa 2 positiivisesti työhyvinvointiinsa suhtautuvia. Lisäksi ensimmäisessä klusterissa olevat kokevat enemmän stressiä ja väsymystä kuin toisessa klusterissa olevat.

Taulukko 3: Valituista mielipidekysymyksistä imputoitujen aineistojen klusterien 1 ja 2 keskustojen keskiarvot sekä keskihajonnat, joiden laskentaa on esitelty luvussa 3.2. Taulukon sarakkeiden nimistä $\overline{q}_1$ tarkoittaa klusterin 1 keskustojen keskiarvoa, $\overline{q}_2$ tarkoittaa klusterin 2 keskustojen keskiarvoa, B_1 tarkoittaa klusterin 1 keskihajontaa ja B_2 tarkoittaa klusterin 2 keskihajontaa.

Kysymykset	$\overline{q}_1$	$\overline{q}_2$	B_1	B_2
Koulun opetustilat ovat hyvät	43.98	62.25	8.95	9.10
Työtä koulussa on sopivasti	40.43	69.99	14.76	14.79
Työskentely koulussa ei tunnu liian kiireiseltä	30.27	61.46	15.48	15.55
Työjärjestys on hyvä	52.66	76.65	12.01	12.03
Tarvittaessa saan apua työterveyshuollosta	60.16	77.53	8.45	8.78
Koulumme työntekijät viihtyvät hyvin yhdessä	60.51	81.37	10.32	10.42
Yhteistyö sujuu hyvin koulussamme	60.46	82.78	11.17	11.11
Työyhteisön jäseniä kohdellaan tasapuolisesti	48.54	81.62	16.51	16.51
Työtoverit puuttuvat asiaan jos yhteisössä esiintyy työpaikkakiusaamista	51.82	76.80	12.42	12.48
Kouluyhteisössä rohkaistaan ilmaisemaan oma mielipiteeni	56.03	78.90	11.44	11.46
Minulta ei odoteta liikaa työssäni	51.19	76.80	12.79	12.72
Työskentelytahti on minulle sopiva	54.55	81.71	13.54	13.49
Saan kiitosta jos olen suoriutunut hyvin työssäni	54.99	78.72	11.93	11.70
Ruokailu on rentouttava tauko	30.44	50.13	9.73	9.83
Meidän koulussamme ei ole työpaikkakiusaamista	57.79	84.10	13.06	13.07
Tarvittaessa saan työnohjausta	46.45	70.51	12.04	12.07
Ruokailutila on viihtyisä	49.80	67.55	9.01	8.85
Koulun piha on viihtyisä	45.96	64.91	9.56	9.44
Sääntöjen rikkomisia käsitellään oikeudenmukaisesti	58.22	78.17	10.07	9.92
Mielipiteeni otetaan huomioon koulun kehittämisessä	54.39	78.13	12.11	11.86
Olen väsynyt	64.31	47.98	8.81	8.07
Työni koulussa aiheuttaa minulle stressiä	59.23	36.97	11.16	11.13

Tarkastellessa moni-imputoitujen aineistojen taustamuuttujia työvuosista saadaan vastaavia tuloksia kuin täydellisesti havaitusta aineistosta: ”alle 1 vuotta” -kategoriassa suurin osa on positiivisessa klusterissa 2, kun taas kauemmin työskennelleet eli kategoriassa ”6–10 vuotta” olevista suurin osa on klusterissa 1 (liite B, taulukko B2). Ikäryhmiin liittyvät ryhmittelytulokset ovat lähes samoja kuin täydellisesti havaitun aineiston tulokset. Imputoiduissa aineistoissa eniten klustereiden välisiä eroja on ikäryhmissä alle 25- tai 25-, 41–45- ja 51–55-vuotiaat siten, että alle 25- tai 25-vuotiaista ja 51–55-vuotiaista suurin osa havainnoista ovat toisessa ns. positiivisesti ajattelevien klusterissa sekä suurin osa 41–45-vuotiaista on ensimmäisessä ns. negatiivisesti ajattelevien klusterissa. Ainoastaan 26–30-vuotiaiden ikäryhmässä ei ole havaittavissa samanlaista selkeää eroa klustereiden välillä kuin täydellisesti havaitussa aineistossa. (liite B, taulukko B4.)

Kuvasta 3 voidaan todeta kysymysten ”Työtoverit puuttuvat asiaan, jos yhteisössä esiintyy työpaikkakiusaamista” ja ”Tarvittaessa saan työnohjausta” olevan neljän eniten puuttuvia havaintoja sisältävien muuttujien joukossa. Näistä kysymyksistä ”Työtoverit puuttuvat asiaan, jos yhteisössä esiintyy työpaikkakiusaamista” saattaa puuttua havaintoja sen takia, että kysymykseen ei haluta antaa negatiivista vastausta. Tällöin muuttujassa olisi MNAR-puuttuvuusrakenne, joten tämän kysymyksen kohdalla tehtiin sensitiivisyysanalyysiä vähentämällä jokaisesta generoidusta imputointikeskiarvosta -20 ja -40 yksikköä, mutta klusterien välisten keskustojen keskiarvojen erot eivät suuresti muuttuneet: vähentämällä 20 yksikköä keskiarvo pieneni ensimmäisessä klusterissa arvoon 48.33 ja toisessa arvoon 74.14 sekä vähentämällä 40 yksikköä ensimmäisen klusterin keskiarvo pieneni arvoon 45.27 ja toisen arvoon 71.75. Sensitiivisyysanalyysillä ei ole juurikaan vaikutuksia muiden muuttujien ryhmittelykeskiarvoihin.

7 Pohdinta

Tässä työssä päätavoitteena oli soveltaa k:n prototyypin ryhmittelymenetelmää ja moni-imputointia Koulun hyvinvointiprofiilin aineistoon. K:n prototyypin ryhmittelymenetelmä valittiin aineiston muuttujien erilaisten mittaasteikkojen vuoksi, koska tällöin voitiin huomioida jatkuvien muuttujien lisäksi kategoriset muuttujat. Toinen syy oli se, että luvussa 4.3 esitelty muuttujien valintaan käytetty algoritmi toimi hyvin tässä menetelmässä. Lisäksi menetelmä oli nopea, vaikka aineisto olikin sarakemäärältään suuri.

Ryhmiteltävien muuttujien valintaa tehtiin, koska tutkielmaa haluttiin rajata ja tulosten raportointia selkeyttää. Muuttujien valinta voisi myös perustua aineiston muuttujien sisältöä koskevaan tutkimuskysymykseen, mutta tässä työssä pyrittiin löytämään erityisesti ryhmittelymenetelmiin soveltuvia muuttujia. Tässä tutkielmassa muuttujien valinnan algoritmi suoritettiin nyt vain täydellisesti havaitulle aineistolle, ja samat muuttujat valittiin myös moni-imputoiduista aineistoista. Tällöin ryhmittelyyn soveltuvia muuttujia ei valittu enää moni-imputoiduista aineistoista erikseen, jolloin algoritmin avulla olisi voitu saada erilaisia tuloksia. Erot muuttujien luokitteluvirheiden välillä olivat kuitenkin pieniä, joten virheet olisivat todennäköisesti olleet pieniä myös moni-imputoiduissa aineistoissa (liite A, sarake 5).

Työssä käytetty moni-imputointi on todettu luotettavaksi menetelmäksi, koska siinä pystytään useamman aineiston avulla iteroimaan imputoitavia havaintoja. Imputointi antaa tällöin tarkemman vertailtavan tuloksen täydellisesti havaittuun aineistoon (van Buuren 2018, luku 2.1.2). Tutkielman tuloksissa esiteltiin imputoitujen aineistojen klusterien keskustojen keskiarvot ja -hajonnat, kuten Basagaña et al. (2013) artikkelissa. Lisäksi aineiston valittuja muuttujia saatiin konvergenssitarkastelun perusteella imputoitua hyvin.

Muutamia ainoastaan klusterointimenetelmille kehitettyjä imputointimenetelmiä on olemassa, mutta niitä on lähinnä sovellettu simuloiduille puuttuville havainnoille: esimerkiksi satunnaista imputointia (*Clustering-Based Random Imputation*) on sovellettu puuttuville havainnoille, mutta menetelmä soveltuu erityisesti tapauksiin, joissa puuttuminen on täysin satunnaista (MCAR) (Zhang et al. 2006). Tutkielmassa käytetyssä aineistossa joidenkin osioiden vastausten puuttuminen täysin ei vastaa satunnaiselle puuttumiselle asetettuja ehtoja, jotka esiteltiin luvussa 3: osion kysymyksen vastauksen puuttuminen on riippuvainen saman osion kysymyksistä.

Puuttuvan tiedon rakenne voi olla myös MNAR, jolloin luotettavuuden li-

säämiseksi voitaisiin tehdä sensitiivisyysanalyysejä. Puuttuvista arvoista olisi kuitenkin pitänyt olla tällöin enemmän tietoa, koska MNAR-rakenne ei ole täysin varma: Terveydentilaa kuvaaviin kysymyksiin on saatettu jättää vastaamatta myös esimerkiksi ajanpuutteen vuoksi, koska kysymykset olivat lomakkeessa viimeisenä. Sen lisäksi kysymys ”Tarvittaessa saan työnohjausta” sisälsi paljon puuttuvaa: työnohjausta ei useammassa peruskoulussa ole lainkaan saatavilla, ja siksi kysymykseen jätetään vastaamatta (Konu 2016-2019). Tällöin näiden kysymysten sensitiivisyysanalyyseihin ei olisi tarvetta.

Sensitiivisyysanalyysiä tehtiin kuitenkin muuttujalle ” Työtoverit puuttuvat asiaan jos yhteisössä esiintyy työpaikkakiusaamista.”, koska tämä muuttuja oli neljän eniten puuttuvia havaintoja sisältävien joukossa ja kysymykseen negatiivisesti suhtautuvat saattavat jättää helpommin vastaamatta (Konu 2016-2019). Lisäksi kysymyksen puuttuvien tietojen lukumäärä erottuu saman osion muista kysymyksistä (liite E). Tuloksissa ei kuitenkaan havaittu suuria muutoksia klustereiden keskustojen arvoissa. Tässä tutkielmassa klustereiden määrä oli vain kaksi, joka saattaa vaikuttaa siihen, ettei sensitiivisyysanalyysissä tapahtunut suuria muutoksia. Tällöin havaintojen lukumäärä klustereissa on suuri, jolloin yhden muuttujan arvojen muutokset eivät vaikuta merkitsevästi klustereiden keskustojen muodostumiseen (Hastie et al. 2009, s. 500–501).

Tutkielmassa käytettiin sisäisiä indeksejä tulkitsemaan sopivaa klustereiden lukumäärää aineistossa. Indeksien laskenta perustuu usein klustereiden sisäisten ja ulkoisten varianssien laskentaan tai klustereiden erilaisuusmittaan (Rendón et al. 2011). Useamman eri indeksin avulla voitiin varmistaa oikea ryhmien lukumäärä, koska k:n prototyypin ryhmittelyssä klustereiden lukumäärä on valittava ennen menetelmän suorittamista. Sopiva klusterien lukumäärä voitaisiin myös valita mielivaltaisesti, mutta tässä työssä haluttiin varmistaa oikeanlainen klusterirakenne indeksien avulla.

Aikaisemmassa tutkimuksessa ryhmiteltiin oppilaiden kokemaa hyvinvointia, jossa ryhmien lukumäärä oli valittu mielivaltaisesti: ryhmien määräksi valittiin 10. (Kylväjä et al. 2019.) Tässä tutkielmassa lopullisiin tuloksiin valittu ryhmien lukumäärä perusteltiin Davies–Bouldin- ja Calinski–Harabasz-indeksien avulla, koska näistä saatiin yhtenevä ja luotettavasti tulkittava tulos.

Rand-indeksiä käytettiin vertailemaan moni-imputoitujen aineistojen klusterirakennetta täydellisesti havaittuun aineistoon. Vertailtavien aineistojen ei tarvitse silloin olla samankokoisia, jolloin voitiin tutkia havaintomäärältään suurempaa moni-imputoitua aineistoa pienempään täydellisesti havaittuun aineistoon: Esimerkiksi ulkoisen Hubertin Γ -indeksin laskennassa muodostetaan korrelaatiomatriisi kahden vertailtavan klusterirakenteen välille, jolloin aineistojen tulisi aina olla samankokoisia. (Theodoridis & Koutroumbas 2008, s. 740.)

Aikaisemmassa tutkimuksessa Rand-indeksiä on käytetty tutkimaan erilaisten imputointimenetelmien onnistumista ryhmiteltyihin aineistoihin, joissa puuttuminen on simuloitu (Somasundaram & Nedunchezhian 2011). Tällöin kokonaan havaittu aineisto on ollut käytettävissä, joka ei vastaa todellisten aineistojen puuttuvan tiedon ongelmaa. Tässä tutkimuksessa imputoinnin onnistumisen tarkastelu ei ollut mahdollista Rand-indeksin avulla, koska täyttä alkuperäistä aineistoa ei ole olemassa. Täydellisesti havaitun ja moni-imputoidun aineiston vertailulla ei siis pystytä todistamaan imputoinnin onnistumista, mutta pystytään tutkimaan aineistojen klusterirakenteiden yhtäläisyyksiä: Mitä samanlaisempia aineistot ovat, sitä todennäköisemmin täydellisesti havaitun aineiston arvot ovat samoissa klustereissa myös moni-imputoidussa aineistossa.

Tutkielmaa voitaisiin jatkaa sovittamalla myös muita ryhmittelymenetelmiä, jolloin voitaisiin paremmin arvioida k:n prototyypin ryhmittelyn sopivuutta: esimerkiksi K-medoids -menetelmän avulla voidaan myös huomioida kategoriset ja jatkuvat muuttujat. Menetelmä ei olisi yhtä herkkä poikkeaville havainnoille kuin k:n prototyypin ryhmittely, koska klusterien keskipisteitä ei valita satunnaisesti vaan asetetaan jokaiselle klusterille haluttu keskusta. Lisäksi eri ryhmittelyiden klusterirakenteiden sopivuutta aineistoon voisi vertailla ulkoisten validointikriteerien indeksien avulla. (Theodoridis & Koutroumbas 2008, s. 635 & 738.)

Kovissa ryhmittelymenetelmissä jokainen havainto on sijoitettava johonkin klusteriin, joka ei välttämättä ole mielekästä kaikkien havaintojen tapauksessa (Hastie et al. 2009, s. 510–512). Kaikkien muuttujien havaintoja ei tarvitse sijoittaa vain yhteen klusteriin, jos käytettäisiin sumeita ryhmittelymenetelmiä. Sumean ryhmittelymenetelmän EM-algoritmin soveltamisessa aineistoon oli kuitenkin omat haasteensa, kun muuttujia oli paljon: Muuttujien valinta osoittautui päiviä vieväksi tulosten odottamiseksi, koska aineiston muuttujien lukumäärä oli suuri.

Koulun hyvinvointiprofiilin aineistoon voitaisiin sovittaa myös jokin muu

tilastollinen malli, kuten yleistetty lineaarinen regressiomalli, ja tehdä jatkotutkimuksia halutusta vasteesta. Tähän työhön valitut muuttujat, ja niistä saadut tulokset, voivat antaa jo viitteitä siitä, millaisia työhyvinvointia selittäviä muuttujia mallinnukseen kannattaa sisällyttää.

Lähteet

1. Aggarwal, C.C. & Reddy, C.K. 2014: *Data Clustering: Algorithms and Applications.* Chapman and Hall/CRC. Data Mining and Knowledge Discovery Ser 31 (1),23, 572–602.

2. Basagaña, X., Barrera-Gómez, J., Benet, M., Antó, J.M. & Garcia-Aymerich, J., 2013. A framework for multiple imputation in cluster analysis. *American Journal of Epidemiology*, 177(7), 718-725.

3. van Buuren, S. & Groothuis-Oudshoorn, K. 2011: Mice: multivariate imputation by chained equations in R. *Journal of Statistical Software* 45 (3), 1–67.

4. van Buuren, S. 2018: *Flexible Imputation of Missing Data.* 2. painos. USA: Chapman & Hall/CRC.

5. Breiman, L. 2001. Random forests. *Machine Learning* 45(1), 5–32.

6. Calinski, T. & Harabasz, J. 1974: A dendrite method for cluster analysis. *Communications in Statistics– Theory and Methods*, 3, 1–27.

7. Davies, D. & Bouldin, D. 1979: A cluster separation measure.*IEEE Transactions on Pattern Analysis and Machine Intelligence.* PAMI-1 (2), 224–227.

8. Desgraupes B. 2017: Clustering indices. `clusterCrit` R -manual. France: University Paris Ouest.

9. Dunn J.C. 1974: Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics* 4, 95–104.

10. Fisher, A., Rudin, C. & Dominici, F. 2019: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20, 1–81.

11. Hastie, T., Tibshirani, R. & Friedman, J. 2009: *The Elements of Statistical learning: Data mining, Inference, and Prediction*, 2. painos. New York: Springer.

12. Hilbe, J. 2009: *Logistic Regression Models.* USA: Chapman & Hall/CRC.

13. Huang, Z. 1997: Clustering large data sets with mixed numeric and categorical values. *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 21–34.

14. Hubert, L. J. & Levin, J. R. 1976: A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin* 83, 1072–1080.

15. Konu A. 2002: Oppilaiden hyvinvointi koulussa. *Acta Universitatis Tamperensis* 887. Tampereen yliopisto. http://acta.uta.fi/teos.php?id=7159.

16. Konu, A. 2010: Koululaisten hyvinvoinnin arviointi ja alakoulujen hyvinvointi 2000-luvulla. *Tampere University Press: Tunne- ja sosiaalisten taitojen vahvistaminen kouluyhteisössä*, 13–32.

17. Konu, A., Viitanen, E. & Lintonen, T. 2010: Teachers' well-being and perceptions on leadership practices. *Journal of Workplace Health Management* 3 (1), 44–57.

18. Konu, A. 2016-2019: Koulun hyvinvointiprofiili 2016-2019. Käytetty aineisto. Tampere: Tampereen yliopisto. https://koulunhyvinvointiprofiili.fi

19. Kylväjä, M., Kumpulainen, P. & Konu, A. 2019: Application of data clustering for automated feedback generation about student well-being. *Tampere University; ACM.*

20. Little, R. & Rubin, D. 2002: *Statistical Analysis with Missing Data.* John Wiley & Sons. Incorporated, Newy York. Available from: ProQuest Ebook Central.

21. Milligan, G. W. 1981: A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46, 187–199.

22. Pfaffel, O. 2019: Feature importance in k-means clustering. R-koodi saatavilla osoitteesta: https://github.com/o1iv3r/FeatureImpCluster

23. Rand, W.M. 1971: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66 (336), 846–850.

24. Rendón, E., Abundez, I., Arizmendi, A. & Quiroz, Elvia M. 2011: Internal versus external cluster validation indexes. *International Journal of Computers and Communications* 5(1), 27–34.

25. Rezaei, M. 2016: Clustering validation. *The University of Eastern Finland Dissertations in Forestry and Natural Sciences* No 225. Joensuu, Finland: Grano Oy.

26. Szepannek, G. 2018: `clustMixType`: User-Friendly Clustering of Mixed-Type Data in R.*The R Journal* Vol. 10/2, 200–208.

27. Somasundaram, R.S. & Nedunchezhian, R., 2011. Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. *International Journal of Computer Applications*, 21(10), 14-19.

28. Theodoridis, S. & Koutroumbas, K. 2008: *Pattern recognition*, 4. painos. USA: Elsevier.

29. Zhang, C., Qin, Y., Zhu, X., Zhang, J. & Zhang, S. 2006: Clustering-based missing value imputation for data preprocessing. *IEEE International Conference on Industrial Informatics*, 4th, 1081–1086

Liite A

Taulukko A1: Valituista mielipidekysymyksistä tärkeimpiä tunnuslukuja, kun vastaukset saavat arvoja asteikolla 0–100. Taulukon sarakkeiden nimistä $\overline{x}$ tarkoittaa vastausten keskiarvoa, Md mediaania ja NA tarkoittaa yhteensä laskettujen puuttuvien lukumäärää kysymyksittäin. Lisäksi $Error_{\overline{x}}$ tarkoittaa luokitteluvirheen keskiarvoa, jonka laskennasta on kerrottu tarkemmin luvussa 4.3.

Kysymykset	$\overline{x}$	Md	NA	$Error_{\overline{x}}$
Koulun opetustilat ovat hyvät	53.51	57.0	96	0.02
Työtä koulussa on sopivasti	55.85	61.0	113	0.03
Työskentely koulussa ei tunnu liian kiireiseltä	46.20	41.0	118	0.03
Työjärjestys on hyvä	65.04	66.0	133	0.02
Tarvittaessa saan apua työterveyshuollosta	69.45	71.0	133	0.02
Koulumme työntekijät viihtyvät hyvin yhdessä	71.33	72.0	85	0.02
Yhteistyö sujuu hyvin koulussamme	72.06	73.0	77	0.02
Työyhteisön jäseniä kohdellaan tasapuolisesti	65.80	70.0	105	0.03
Työtoverit puuttuvat asiaan jos yhteisössä esiintyy työpaikkakiusaamista	64.93	64.0	198	0.02
Kouluyhteisössä rohkaistaan ilmaisemaan oma mielipiteeni	68.18	68.0	137	0.02
Minulta ei odoteta liikaa työssäni	64.44	65.0	131	0.02
Työskentelytahti koulussa on minulle sopiva	68.63	72.0	113	0.02
Saan kiitosta jos olen suoriutunut hyvin työssäni	67.31	68.0	142	0.02
Ruokailu on rentouttava tauko	40.32	38.0	126	0.02
Meidän koulussamme ei ole työpaikkakiusaamista	71.52	76.0	105	0.02
Tarvittaessa saan työnohjausta	58.73	62.0	200	0.02
Ruokailutila on viihtyisä	59.21	64.0	118	0.02
Koulun piha on viihtyisä	55.77	63.0	104	0.02
Sääntöjen rikkomisia käsitellään oikeudenmukaisesti	68.93	70.0	138	0.02
Mielipiteeni otetaan huomioon koulun kehittämisessä	66.86	67.0	127	0.02
Olen väsynyt	54.91	59.5	391	0.02
Työni koulussa aiheuttaa minulle stressiä	46.54	49.0	398	0.02

Liite B

Taulukko B1: Täydellisesti havaitun aineiston muuttujan ”Olen työskennellyt tässä koulussa” kategorioissa havaintojen lasketut lukumäärät. Tällöin ensimmäisessä sarakkeessa on klusterien arvot k, johon havainto on ryhmitelty sekä muissa sarakkeissa havaintojen lukumäärät kategorioittain.

k	alle 1 vuotta	1–5 vuotta	6–10 vuotta	yli 10 vuotta
1	63	137	95	171
2	93	131	75	162
Yht.	156	268	170	333

Taulukko B2: Viiden imputoituidun aineiston havaintojen keskiarvot muuttujassa ”Olen työskennellyt tässä koulussa”. Kategorioiden keskiarvot ovat laskettu jakamalla imputoitujen aineistojen havaintojen lukumäärien summa viidellä.

k	alle 1 vuotta	1–5 vuotta	6–10 vuotta	yli 10 vuotta
1	122	240	157	280
2	181	256	129	277
Yht.	304	496	286	557

Taulukko B3: Täydellisesti havaitun aineiston klusterien k ikämuuttujan havaintojen lukumäärät. Tällöin ensimmäisessä sarakkeessa on klusteri k, johon havainto on luokiteltu ja ensimmäisellä rivillä 10 eri ikäryhmää.

k	≤ 25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65	$66 \geq$
1	21	55	50	64	84	72	63	42	13	2
2	39	37	46	65	58	68	96	42	8	2
Yht.	60	92	96	129	142	140	159	84	21	4

Taulukko B4: Keskiarvot ikämuuttujan 10 ryhmälle viidestä imputoiduista aineistoista molemmille klustereille k. Kategorioiden keskiarvot ovat laskettu jakamalla imputoitujen aineistojen havaintojen lukumäärien summa viidellä.

k	≤ 25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65	$66 \geq$
1	57	62	79	105	142	121	113	85	29	5
2	84	71	79	112	125	117	145	86	22	3
Yht.	142	133	157	217	268	238	258	171	51	8

Liite C

Taulukko C1: Rand-indeksien arvot klusterin lukumäärille k, kun verrataan ryhmiteltyjä sovitettuja moni-imputoituja aineistoja ($N = 1643$) täydellisesti havaittuun aineistoon ($N = 927$): Mitä lähempänä indeksin arvo on ykköstä niin sitä enemmän aineistojen klusterirakenteet muistuttavat toisiaan. Sarakkeet *Imp. data 1,...,Imp. data 5* tarkoittavat imputoituja aineistoja.

k	*Imp. data 1*	*Imp. data 2*	*Imp. data 3*	*Imp. data 4*	*Imp. data 5*
2	0.5046	0.5037	0.5032	0.5044	0.5026
3	0.6077	0.6106	0.6048	0.6149	0.6085
4	0.6922	0.6923	0.6970	0.7010	0.6928
5	0.7417	0.7403	0.7326	0.7185	0.7361
6	0.7699	0.7621	0.7780	0.7816	0.7784
7	0.7984	0.8002	0.7912	0.7956	0.8028
8	0.8145	0.82144	0.8237	0.8236	0.8098
9	0.8388	0.8337	0.8417	0.8264	0.8389
10	0.8469	0.8375	0.8435	0.8439	0.8502

Taulukko C2: Rand-indeksin arvot klusterien lukumäärille k, kun verrataan ryhmiteltyjä moni-imputoituja aineistoja täydellisesti havaittuun aineistoon. Imputoiduista aineistoista (*Imp. data 1,...,Imp. data 5*) on nyt satunnaisesti arvottu lukuja samanverran kuin täydellisesti havaitussa aineistossa on ($N = 927$). Havaitaan, että tulokset ovat hyvin samoja kuin edeltävässä taulukossa, mutta pienempiä.

k	*Imp. data 1*	*Imp. data 2*	*Imp. data 3*	*Imp. data 4*	*Imp. data 5*
2	0.4995	0.4995	0.4995	0.4995	0.4995
3	0.5453	0.5436	0.5438	0.5480	0.5444
4	0.6241	0.6243	0.6223	0.6258	0.6223
5	0.6727	0.6713	0.6693	0.6694	0.6676
6	0.7142	0.7098	0.7109	0.7174	0.7143
7	0.7482	0.7475	0.7464	0.7501	0.7516
8	0.7523	0.7703	0.7624	0.7681	0.7730
9	0.7957	0.7971	0.7926	0.7967	0.7931
10	0.8024	0.8075	0.8039	0.7988	0.8010

Liite D

Taulukko D1: Sisäisten indeksien arvot, kun tarkastellaan täydellisesti havaittua aineistoa kaikilla 76 muuttujalla. Havaitaan, että Calinski–Harabasz- ja Davies–Bouldin-indeksin perusteella aineistoon voidaan valita kaksi klusteria. Tällöin Calinski–Harabasz-indeksin arvo on maksimissa ja Davies–Bouldin-indeksin arvo minimissä verrattaessa muihin klusterien määriin. Klusterien lukumäärän valinta päädyttiin tekemään näiden kahden indeksien perusteella, mistä kerrotaan tarkemmin luvussa 6.

k	C-indeksi	Davies–Bouldin	Dunn	Calinski–Harabasz
2	0.25	2.11	0.23	169.48
3	0.24	2.70	0.23	115.63
4	0.23	3.05	0.23	87.79
5	0.22	3.11	0.21	72.97
6	0.23	3.13	0.22	61.67
7	0.23	2.80	0.22	51.68
8	0.23	3.29	0.21	48.80
9	0.20	3.21	0.19	44.21
10	0.19	2.95	0.21	39.86

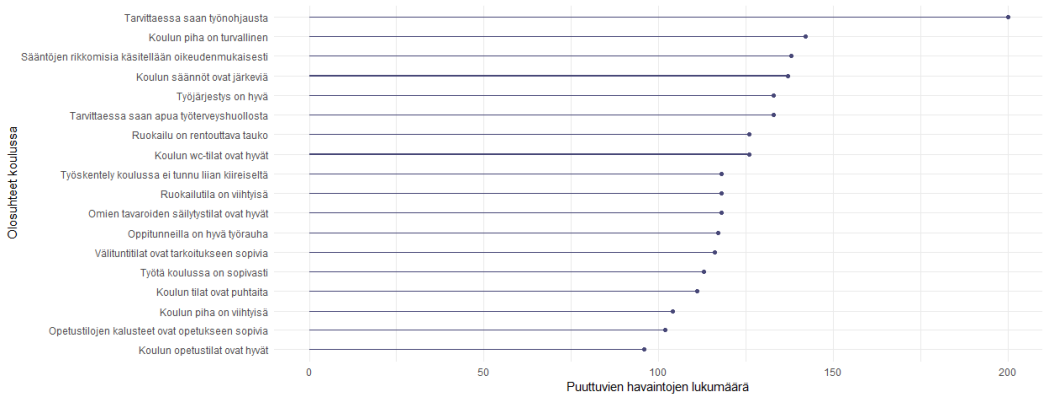
Taulukko D2: Sisäisten indeksien arvot, kun tarkastellaan täydellisesti havaittua aineistoa tuloksiin valituilla muuttujilla, jotka ovat esitelty liitteessä A. Havaitaan, että Calinski–Harabasz- ja Davies–Bouldin-indeksin perusteella aineistoon voidaan valita kaksi klusteria. Tällöin Calinski–Harabasz-indeksin arvo on maksimissa ja Davies–Bouldin-indeksin arvo minimissä verrattaessa muihin klusterien määriin.

k	C-indeksi	Davies–Bouldin	Dunn	Calinski–Harabasz
2	0.23	1.75	0.15	279.88
3	0.22	2.27	0.15	192.68
4	0.19	2.29	0.14	155.00
5	0.20	2.36	0.16	129.96
6	0.20	2.50	0.14	111.36
7	0.18	2.41	0.16	101.31
8	0.19	2.73	0.14	88.23
9	0.19	2.67	0.13	80.01
10	0.18	2.55	0.15	75.55

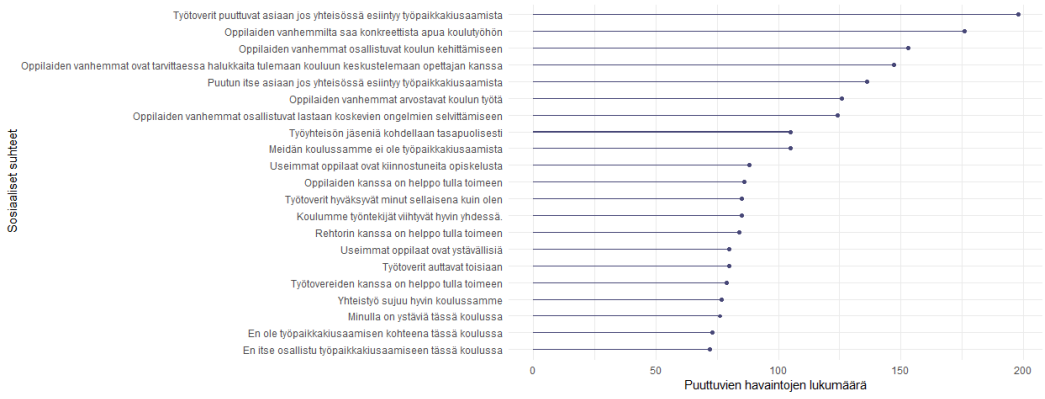
Taulukko D3: Sisäisten indeksien arvot, kun tarkastellaan ensimmäistä imputoitua aineistoa. Havaitaan, että Calinski–Harabasz- ja Davies–Bouldin-indeksin perusteella aineistoon voidaan valita kaksi klusteria. Tällöin Calinski–Harabasz-indeksin arvo on maksimissa ja Davies–Bouldin-indeksin arvo minimissä verrattaessa muihin klusterien määriin. Muiden imputoitujen aineistojen sisäisten indeksien arvot ovat samankaltaisia, joten ne sivuutetaan.

k	C-indeksi	Davies–Bouldin	Dunn	Calinski–Harabasz
2	0.23	1.78	0.12	482.64
3	0.21	2.27	0.14	337.13
4	0.20	2.25	0.14	274.19
5	0.19	2.40	0.15	228.37
6	0.18	2.39	0.15	199.89
7	0.18	2.33	0.14	178.97
8	0.18	2.58	0.13	157.49
9	0.18	2.56	0.14	146.74
10	0.18	2.53	0.13	133.29

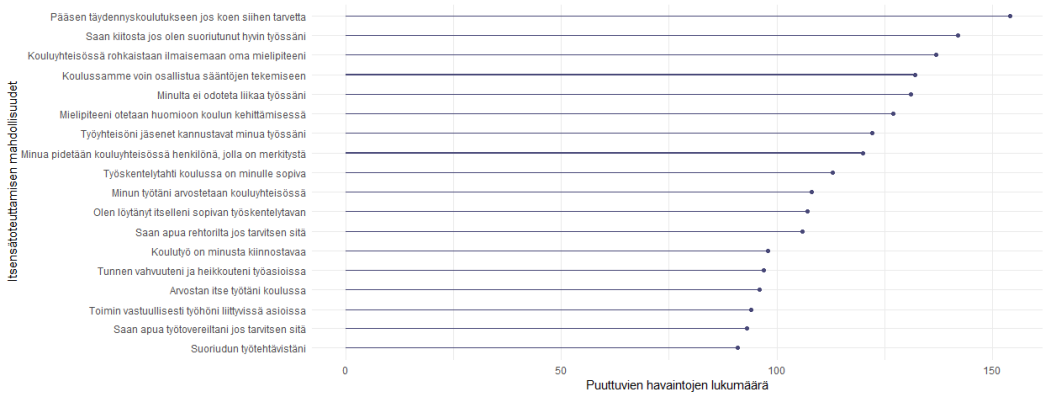
Liite E



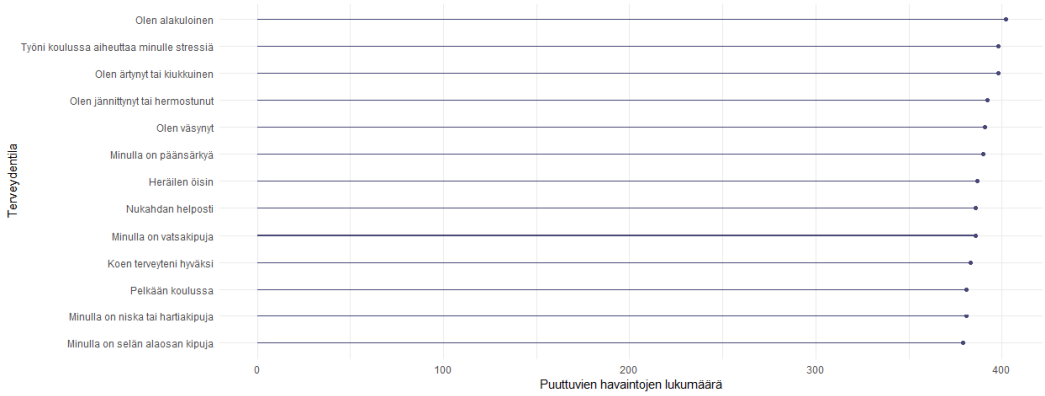
Kuva 7: Puuttuvien tietojen lukumääriä kysymyksittäin osiosta ”koulun olosuhteet” siten, että pystyakselilla on kysymykset ja vaaka-akselilla puuttuvien havaintojen yhteenlaskettu määrä. Kysymyksessä ”Tarvittaessa saan työnohjausta” on eniten puuttuvaa tässä osiossa.



Kuva 8: Puuttuvien tietojen lukumääriä kysymyksittäin osiosta ”sosiaaliset suhteet” siten, että pystyakselilla on kysymykset ja vaaka-akselilla puuttuvien havaintojen yhteenlaskettu määrä.



Kuva 9: Puuttuvien tietojen lukumääriä kysymyksittäin osiosta ”itsensätoteuttamisen mahdollisuudet” siten, että pystyakselilla on kysymykset ja vaaka-akselilla puuttuvien havaintojen yhteenlaskettu määrä.



Kuva 10: Puuttuvien tietojen lukumääriä kysymyksittäin osiosta ”terveydentila” siten, että pystyakselilla on kysymykset ja vaaka-akselilla puuttuvien havaintojen yhteenlaskettu määrä.