Tampereen yliopisto

Sakke Purolainen

Korrespondenssianalyysi aineiston analysoinnin tukena

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Toukokuu 2022

Tiivistelmä

Sakke Purolainen: Korrespondenssianalyysi
Kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin kandidaattiohjelma
Toukokuu 2022

Korrespondenssianalyysi on monimuuttujamenetelmä, jota käytetään mallintamaan monimutkaista kategorista dataa. Tutkielman tarkoituksena on perehtyä korrespondenssianalyysin ominaisuuksiin, sekä havainnollistaa menetelmää käytännön tilanteissa.

Tutkielmassa esitellään ensin korrespondenssianalyysin idea. Sen jälkeen tarkastellaan menetelmän matemaattista taustaa vektoreiden ja matriisien avulla. Tarkastelussa keskitytään etenkin rivi- ja sarakeprofiileihin, riippumattomuuden testaamiseen ja kuvan muodostamiseen. Lopuksi sovelletaan sitä kahteen eri kyselytutkimusaineistoon R-ohjelmistoa apuna käyttäen sekä tehdään sen pohjalta tulkintoja. Tutkielmassa esitellään myös lyhyesti imputaatiomenetelmä, jolla pyritään paikkaamaan aineiston puuttuvia arvoja. Imputaatiota hyödynnetään ensimmäiseen kyselytutkimusaineistoon. Tutkielma osoittaa, että korrespondenssianalyysi on erittäin hyvä menetelmä mallintamaan kategorista dataa. Visuaalisuutensa ansiosta analyysin tulkitseminen on myös melko helppoa. Lisäksi tutkielmassa huomataan aineiston riittävyyden merkitys: menetelmän toimivuuden kannalta on tärkeää, että aineisto sisältää runsaasti havaintoja erityisesti silloin, jos kategorioita on useita.

Sisällys

1 Johdanto **4**

2 Menetelmä **5**

2.1 Rivi- ja sarakeprofiilit 5

2.2 Riippumattomuuden testaaminen 8

2.3 Kuvan muodostaminen 10

3 Soveltaminen aineistoihin **13**

3.1 Aineistot 13

3.2 Datan muokkaus 13

3.3 Puuttuvan tiedon paikkaaminen (Imputaatio) 14

3.4 Korrespondenssianalyysi pankkidataan 15

3.5 Korrespondenssianalyysi uraseurantadataan 16

4 Johtopäätökset **21**

Lähteet **22**

A Kuvaajien R-koodi **23**

1 Johdanto

Korrespondenssianalyysistä ei juuri suomenkielistä materiaalia internetistä löydy. Englannin kielellä materiaalia on jonkin verran, mutta monet englanninkielisistä artikkeleistakin ovat hyvin pintapuolisia. Vaikka korrespondenssianalyysi on hyvin lähellä pääkomponenttianalyysia, ei se jostain syystä ole läheskään yhtä yleisesti käytetty.

Le Roux ja Rouanet (2010) mainitsevat, että korrespondenssianalyysia on kehitetty usean henkilön toimesta, mukaan lukien R. A. Fisherin, joka oli yksi merkittävimmistä henkilöistä tilastotieteessä 1900-luvun alkupuolella. Ranskalainen tilastotieteilijä Jean-Paul Benzecri kuitenkin kehitti ja popularisoi menetelmän vasta 1960- ja 1970-luvuilla, ensin Ranskassa ja myöhemmin muualla Euroopassa. Hän myös antoi menetelmälle nykyään yleisesti käytetyn nimen analyse de correspondence (Le Roux .& Rouanet, 2010.)

Korrespondenssianalyysi on exploratiivinen menetelmä monimutkaiselle kategoriselle datalle (Glynn & Ribbson, 2014). Sillä pyritään esittämään tietoa kaksisuuntaisella ristitaulukolla. Taulukko sisältää kahden tai useamman kategorisen muuttujan frekvenssit. Korrespondenssianalyysilla saadaan rakennettua malli, joka kuvaa kahden muuttujan välistä vuorovaikutusta sekä rivien ja sarakkeiden yhteyttä toisiinsa (Rencher, 2002.)

Työn tavoite on esitellä korrespondenssianalyysin yleinen periaate sekä menetelmän matemaattinen tausta. Lopuksi menetelmää käytetään kahteen dataan ja tehdään niistä tulkinnat analyysin pohjalta.

2 Menetelmä

Tavanomaisen kaksisuuntaisen korrespondenssianalyysin lisäksi on olemassa korrespondenssianalyysin yleistys moniuloitteisiin tauluihin (Rencher, 2002). Käytännön esimerkit käsittelevät tässä tutkielmassa molempia tapauksia. Testatessa merkitsevyyttä voidaan käyttää joko χ^2-testiä tai lineaarista log-lineaarista mallia (Rencher, 2002). Molempien käyttö perustuu asymptoottisiin tuloksiin (Rencher, 2002).

Mikäli ristitaulukossa jokin solun frekvensseistä on hyvin pieni tai nolla, χ^2-approksimaatio ei ole kovinkaan tyydyttävä. Tässä tapauksessa joitakin muuttujia voi yhdistää keskenään, jotta saadaan nostettua solun frekvenssiä. Korrespondenssianalyysi voikin olla hyödyllinen työkalu tunnistamaan muuttujia, jotka ovat samankaltaisia, ja tästä syystä voitaisiin mahdollisesti yhdistää (Rencher, 2002.)

Menetelmä esittää tulokset kaksiulotteisena kuvaajana, joka visualisoi muuttujien yhteyttä intuitiivisesti (Glynn & Ribbson, 2014). Mikäli kahden rivin pisteet ovat lähellä toisiaan, niiden profiilit ovat rivien suhteen samankaltaisia. Yhtä lailla, mikäli sarakkeiden pisteet ovat lähellä toisiaan, ovat niiden profiilit samankaltaisia rivien suhteen (Rencher, 2002.) Menetelmän vahvuus onkin juuri visuaalisessa esityksessä.

2.1 Rivi- ja sarakeprofiilit

Rencher (2002) määrittelee rivi- ja sarakeprofiilit kirjassaan seuraavasti: Alla on ristitaulukko 2.1, jossa on a riviä ja b saraketta. Jokainen taulukon alkio n kuvaa frekvenssejä kaikille rivien ja sarakkeiden kombinaatioille. Reunajakaumien summat on esitetty summanotaatiolla: $n_{i.} = \sum_{j=1}^{b} n_{ij}$ ja $n_{.j} = \sum_{i=1}^{a} n_{ij}$. Kaikkien summa on merkitty yksinkertaisesti n (Rencher, 2002.)

Ristitaulukon frekvenssit n_{ij} voidaan muuttaa suhteellisiksi frekvensseiksi p_{ij} yksinkertaisesti jakamalla n: $p_{ij} = n_{ij}/n$. Suhteellisten frekvenssien matriisia kutsutaan korrespondenssimatriisiksi ja sitä merkitään kirjaimella $\boldsymbol{P}$ (Rencher, 2002.)

Taulukko 2.1. Suhteellisten frekvenssien ristitaulukko, jossa on a riviä ja b saraketta

	1	2	$\cdots$	b	Summa
1	p_{11}	p_{12}	$\cdots$	p_{1b}	$p_{1.}$
2	p_{21}	p_{22}	$\cdots$	p_{2b}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a	p_{a1}	p_{a2}	$\cdots$	p_{ab}	$p_{a.}$
Summa	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.b}$	1

Viimeinen sarake taulussa 2.1 sisältää rivisummat $p_{i.} = \sum_{j=1}^{b} p_{ij}$. Merkitään sarakevektoria vektorilla $\boldsymbol{r}$. $\boldsymbol{r}$ saadaan

$$\boldsymbol{r} = \boldsymbol{P}\boldsymbol{j} = (p_{1.}, p_{2.}, \cdots, p_{a.})' = (n_{1.}/n, n_{2.}/n \cdots, , n_{a.}/n)', \tag{2.1}$$

missä j on $a \times 1$ vektori numero ykkösiä. Samaan tapaan viimeinen rivi taulukossa 2.1 sisältää sarakesummat $p_{.j} = \sum_{i=1}^{a} p_{ij}$. Merkitään sarakevektoria vektorilla c ja se saadaan

$$\boldsymbol{c'} = \boldsymbol{j'}\boldsymbol{P} = (p_{.1}, p_{.2}, \cdots, p_{.b}) = (n_{.}1/n, n_{.2}/n, \cdots, n_{.b}/n), \tag{2.2}$$

missä j' on $1 \times a$ vektori numero ykkösiä. Korrespondenssimatriisi ja reunajakaumien summat taulukossa 2.1 voidaan esittää

$$\begin{pmatrix} \boldsymbol{P} & \boldsymbol{r} \\ \boldsymbol{c'} & 1 \end{pmatrix} = \left(\begin{array}{cccc|c} p_{11} & p_{12} & \cdots & p_{1b} & p_1 \\ p_{21} & p_{22} & \cdots & p_{2b} & p_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{a1} & p_{a2} & \cdots & p_{ab} & p_b \\ \hline p_{.1} & p_{.2} & \cdots & p_{.b} & 1 \end{array}\right)$$

Seuraavaksi jokainen $\boldsymbol{P}$:n rivi ja sarake muutetaan profiiliksi. Tällöin i:nnen rivin profiili r'_j, $i = 1, 2, \cdots, a$, esitetään jakamalla i:nnes rivi taulusta 2.1 sen reunajakauman summalla:

$$\boldsymbol{r'_i} = \left(\frac{p_{i1}}{p_{i.}}, \frac{p_{i2}}{p_{i.}}, \cdots, \frac{p_{ib}}{p_{i.}}\right) \tag{2.3}$$

Elementit jokaisessa $\boldsymbol{r_i'}$ ovat suhteellisia frekvenssejä ja siten niiden summa on 1:

$$\boldsymbol{r_i' j} = \sum_{j=1}^{b} \frac{n_{ij}}{n_{i.}} = \frac{n_{i.}}{n_{i.}} = 1 \tag{2.4}$$

Määritellään:

$$\boldsymbol{D_r} = diag(\boldsymbol{r}) = \begin{pmatrix} p_{1.} & 0 & \cdots & 0 \\ 0 & p_{2.} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{a.} \end{pmatrix} \tag{2.5}$$

Nyt riviprofiilien matriisi $\boldsymbol{R}$ voidaan esittää myös:

$$\boldsymbol{R} = \boldsymbol{D_r^{-1} P} = \begin{pmatrix} \boldsymbol{r_1'} \\ \boldsymbol{r_2'} \\ \vdots \\ \boldsymbol{r_a'} \end{pmatrix} = \begin{pmatrix} \frac{p_{11}}{p_{1.}} & \frac{p_{12}}{p_{1.}} & \cdots & \frac{p_{1b}}{p_{1.}} \\ \frac{p_{21}}{p_{2.}} & \frac{p_{22}}{p_{2.}} & \cdots & \frac{p_{2b}}{p_{2.}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_{a1}}{p_{a.}} & \frac{p_{a2}}{p_{a.}} & \cdots & \frac{p_{ab}}{p_{a.}} \end{pmatrix} \tag{2.6}$$

Samaan tapaan , j:nnes sarakeprofiili $\boldsymbol{c}_j$, $j = 1, 2, \cdots, , b$, määritellään jakamalla j:nnes sarake sen reunajakauman summalla:

$$\boldsymbol{c_j'} = \left(\frac{p_{j1}}{p_{j.}}, \frac{p_{j2}}{p_{j.}}, \cdots, \frac{p_{jb}}{p_{j.}} \right) \tag{2.7}$$

Elementit jokaisessa $\boldsymbol{c_i'}$ ovat suhteellisia frekvenssejä ja siten niiden summa on 1:

$$\boldsymbol{j' c}_j = \sum_{i=1}^{a} \frac{n_{ij}}{n_{.j}} = \frac{n_{.j}}{n_{.j}} = 1. \tag{2.8}$$

Määritellään

$$\boldsymbol{D_c} = diag(\boldsymbol{c}) = \begin{pmatrix} p_{.1} & 0 & \cdots & 0 \\ 0 & p_{.2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{.b} \end{pmatrix} \tag{2.9}$$

Nyt sarakeprofiilien matriisi $\boldsymbol{R}$ voidaan esittää myös:

$$\boldsymbol{R} = \boldsymbol{P D_c^{-1}} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_b) = \begin{pmatrix} \frac{p_{11}}{p_{.1}} & \frac{p_{12}}{p_{.2}} & \cdots & \frac{p_{1a}}{p_{.a}} \\ \frac{p_{21}}{p_{.1}} & \frac{p_{22}}{p_{.2}} & \cdots & \frac{p_{2a}}{p_{.a}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_{a1}}{p_{.1}} & \frac{p_{a2}}{p_{.2}} & \cdots & \frac{p_{ab}}{p_{.a}} \end{pmatrix} \tag{2.10}$$

Vektori $\boldsymbol{r}$ on määritelty (2.1) $\boldsymbol{P}$:n rivisummien sarakevektoriksi. Se voidaan esittää sarakeprofiilien painotettuna keskiarvona:

$$\boldsymbol{r} = \sum_{j=1}^{b} p_{.j}\boldsymbol{c}_j \tag{2.11}$$

Samaan tapaan vektori $\boldsymbol{r}$ on määritelty (2.2) $\boldsymbol{P}$:n sarakesummien rivivektoriksi ja se voidaan esittää riviprofiilien painotettuna keskiarvona:

$$\boldsymbol{c}' = \sum_{i=1}^{a} p_{i.}\boldsymbol{r}_i' \tag{2.12}$$

Huomaa, että: $\sum_{j=1}^{b} p_{.j} = \sum_{i=1}^{a} p_{i.} = 1$, tai

$$\boldsymbol{j}'\boldsymbol{r} = \boldsymbol{c}'\boldsymbol{j} = 1, \tag{2.13}$$

missä ensimmäinen $\boldsymbol{j}$ on $a \times 1$ ja toinen on $b \times 1$. Tästä syystä $p_{.j}$:t ja $p_{i.}$:t ovat sopivia painoja painotetussa keskiarvossa (2.12) ja (2.13)(Rencher, 2002.)

2.2 Riippumattomuuden testaaminen

Kuten esittelyssä todettiin, ristitaulukon dataa voidaan käyttää selvittämään kahden kategorisen muuttujan välistä yhteyttä. Jos kahta muuttujaa merkitään x:llä ja y:llä, oletus riippumattomuudesta voidaan esittää todennäköisyyksinä samalla tavalla kuin Rencher (2002) kirjassaan seuraavasti:

$$P(x_i y_j) = P(x_i)P(y_j), \qquad i = 1, 2, \cdots, a; \quad j = 1, 2, \cdots, b, \tag{2.14}$$

missä x_i and y_j vastaa i:ttä riviä ja j:ttä saraketta ristitaulukossa. Käyttämällä taulukon (2.1) notaatiota voidaan estimoida (2.14)

$$p_{ij} = p_{i.}p_{.j}, \qquad i = 1, 2, \cdots, a; \quad j = 1, 2, \cdots, b. \tag{2.15}$$

χ^2-testisuure x:n ja y:n riippumattomuuden testaamiseksi on

$$\chi^2 = n\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}, \tag{2.16}$$

mikä on suurin piirtein (asymptoottisesti) χ^2-jakautunut vapausastein (a-1)(b-1). Kaksi vaihtoehtoista muotoa (2.16) ovat

$$\chi^2 = n \sum_{i=1}^{a} np_{i.} \sum_{j=1}^{b} \left[\left(\frac{p_{ij}}{p_{i.}} - p_{.j} \right)^2 / p_{.j} \right], \tag{2.17}$$

$$\chi^2 = n \sum_{j=1}^{b} np_{.j} \sum_{i=1}^{a} \left[\left(\frac{p_{ij}}{p_{.j}} - p_{i.} \right)^2 / p_{i.} \right], \tag{2.18}$$

vektori ja matriisi muotona (2.17) ja (2.18) voidaan kirjoittaa

$$\chi^2 = \sum_{i=1}^{a} np_{i.} (\boldsymbol{r}_i - \boldsymbol{c})' \boldsymbol{D}_c^{-1} (\boldsymbol{r}_i - \boldsymbol{c}), \tag{2.19}$$

$$\chi^2 = \sum_{j=1}^{b} np_{.j} (\boldsymbol{c}_j - \boldsymbol{r})' \boldsymbol{D}_r^{-1} (\boldsymbol{c}_j - \boldsymbol{r}), \tag{2.20}$$

missä $\boldsymbol{r}$, $\boldsymbol{c}$, $\boldsymbol{r}_i$, $\boldsymbol{c}_j$, $\boldsymbol{D}_r$ ja $\boldsymbol{D}_c$ on määritelty kohdissa (2.1), (2.2), (2.3), (2.7), (2.5) ja (2.9). Kohdassa (2.19) verrataan $\boldsymbol{r}_i$:tä $\boldsymbol{c}$:hen jokaisella i:llä ja $\boldsymbol{c}_j$:tä $\boldsymbol{r}$:ään jokaisella j:llä. Täten kumpikin näistä on yhtä pätevä testaamaan riippumattomuutta, kun verrataan p_{ij} ja $p_i p_i$ kaikilla i, j, koska kaikki χ^2 määritelmät (2.16)-(2.20) ovat yhtäpitäviä.

Täten, jos muuttujat x ja y ovat riippumattomia, voidaan olettaa että, ristitaulukon riveillä on samankaltaiset profiilit tai yhtä lailla, sarakkeilla on samankaltaiset profiilit, jos muuttujat ovat riippumattomia. Riviprofiileja voidaan verrata keskenään vertaamalla jokaista riviprofiilia $\boldsymbol{r}_i'$ riviprofiilien painotettuun keskiarvoon $\boldsymbol{c}'$ (2.2). Tämä vertailu on tehty kohdassa (2.19). Samalla tavalla voidaan verrata sarakeprofiileja (2.20).

χ^2-testisuure kohdassa (2.16) voidaan esittää vektori- ja matriisimuodossa

$$\chi^2 = n * tr[\boldsymbol{D}_r^{-1} (\boldsymbol{P} - \boldsymbol{r}\boldsymbol{c}') \boldsymbol{D}_c^{-1} (\boldsymbol{P} - \boldsymbol{r}\boldsymbol{c}')'] \tag{2.21}$$

$$= n \sum_{i=1}^{k} \lambda_i^2 \tag{2.22}$$

missä $\lambda_1^2, \lambda_2^2, \cdots, \lambda_k^2$ ovat nollasta poikkeavia $\boldsymbol{D}_r^{-1}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c^{-1}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')'$:n ominaisarvoja ja

$$k = rank[\boldsymbol{D}_r^{-1}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c^{-1}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')'] = rank(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}'). \tag{2.23}$$

$rank(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')$ on tavallisesti $k = min[(a-1),(b-1)]$. On selvää, että aste on vähemmän kun $min(a,b)$, sillä

$$(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')\boldsymbol{j} = \boldsymbol{P}\boldsymbol{j} - \boldsymbol{r}\boldsymbol{c}'\boldsymbol{j} = \boldsymbol{r} - \boldsymbol{r} = \boldsymbol{0} \tag{2.24}$$

(Rencher, 2002.)

2.3 Kuvan muodostaminen

Rencher (2002) esittää kuvan muodostamisen kirjassaan seuraavasti: lasketaan rivi- ja sarakepisteet kaksiulotteiselle mallille datasta ristitaulukkoon. Aluksi skaalataan $\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}'$, jotta saadaan

$$\boldsymbol{Z} = \boldsymbol{D}_r^{-1/2}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c^{-1/2}, \tag{2.25}$$

jonka elementit ovat

$$z_{ij} = \frac{p_{ij} - p_{i.}p_{.j}}{\sqrt{p_{i.}p_{.j}}}, \tag{2.26}$$

ositetaan $\boldsymbol{Z}$ käyttämällä singulaariarvohajotelmaa

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}'. \tag{2.27}$$

$\boldsymbol{U}$ matriisin sarakkeet ovat $\boldsymbol{Z}\boldsymbol{Z}'$ matriisin skaalatut ominaisvektorit. Taas $\boldsymbol{V}$ matriisin sarakkeet ovat normalisoidut $\boldsymbol{Z}'\boldsymbol{Z}$ matriisin ominaisvektorit ja $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_k)$, missä $\lambda_1^2, \lambda_2^2, \cdots, \lambda_k^2$ ovat $\boldsymbol{Z}'\boldsymbol{Z}$:n ja $\boldsymbol{Z}'\boldsymbol{Z}$:n ominaisarvot. Ominaisvektorit $\boldsymbol{U}$ ja $\boldsymbol{V}$ vastaavat ominaisarvoja $\lambda_1^2, \lambda_2^2, \cdots, \lambda_k^2$. $\lambda_1, \lambda_2, \cdots, \lambda_k$ kutsutaan $\boldsymbol{U}$:n yksittäis arvoiksi.

Nyt voidaan laskea hajotelma $\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}'$:lle:

$$
\begin{aligned}
\boldsymbol{D}_r^{-1/2}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c^{-1/2} &= \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}' \\
\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}' &= \boldsymbol{D}_r^{-1/2}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}'\boldsymbol{D}_c^{-1/2} \\
&= \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{B}' = \sum_{i=1}^{k}\lambda_i \boldsymbol{a}_i\boldsymbol{b}_i'
\end{aligned} \tag{2.28}
$$

Missä $\boldsymbol{A} = \boldsymbol{D}_r^{-1/2}\boldsymbol{U}$, $\boldsymbol{B} = \boldsymbol{D}_c^{-1/2}\boldsymbol{V}$, $\boldsymbol{a}_i$ ja $\boldsymbol{b}_i$ ovat $\boldsymbol{A}$:n ja $\boldsymbol{B}$:n indeksiä i vastaavia sarakkeita.

Kohdassa (2.28) $\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}'$:n rivit on esitetty $\boldsymbol{B}'$:n rivien lineaarikombinaationa. Koordinaatit $\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}'$:n i:nnelle riville löytyy $\boldsymbol{A}\boldsymbol{\Lambda}$:n i:nneltä riviltä. Vastaavasti $\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}'$:n i:nnen sarakkeen koordinaatit löytyy $\boldsymbol{\Lambda}\boldsymbol{B}'$:n i:nneltä sarakkeelta.

Jotta saadaan koordinaatit rivienpoikkeamille $\boldsymbol{r}_i' - \boldsymbol{c}'$, $\boldsymbol{R}-\boldsymbol{j}\boldsymbol{c}'$:ssä ja sarakepoikkeamat $\boldsymbol{c}_j - \boldsymbol{r}$, $\boldsymbol{C}-\boldsymbol{r}\boldsymbol{j}'$:ssä, esitetään molemmat matriisit $\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}'$:n funktiona.

$$
\boldsymbol{R}-\boldsymbol{j}\boldsymbol{c}' = \boldsymbol{D}_r^{-1}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}'), \tag{2.29}
$$

$$
\boldsymbol{C}-\boldsymbol{r}\boldsymbol{j}' = \boldsymbol{D}_c^{-1}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}'). \tag{2.30}
$$

Näin rivipoikkeamien koordinaatit $\boldsymbol{R}-\boldsymbol{j}\boldsymbol{c}'$:ssä saadaan matriisioperaationa

$$
\boldsymbol{X} = \boldsymbol{D}_r^{-1}\boldsymbol{A}\boldsymbol{\Lambda} \tag{2.31}
$$

missä koordinaatit löytyvät $\boldsymbol{X}$:n sarakkeilta.

Samalla tavalla sarakepoikkeamien koordinaatit $\boldsymbol{C}-\boldsymbol{r}\boldsymbol{j}'$:ssä saadaan matriisioperaationa

$$
\boldsymbol{Y} = \boldsymbol{D}_c^{-1}\boldsymbol{B}\boldsymbol{\Lambda} \tag{2.32}
$$

missä koordinaatit löytyvät $\boldsymbol{Y}$:n sarakkeilta.

Nyt, jotta voidaan piirtää koordinaatit riviprofiilien poikkeamalle $\boldsymbol{r}_i' - \boldsymbol{c}'$, $i = 1, 2, \cdots, a$ kahdessa ulottuvuudessa, tällöin

$$
\boldsymbol{X}_1 = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{a1} & x_{a2} \end{pmatrix} \tag{2.33}
$$

Vastaavasti, jotta voidaan piirtää koordinaatit sarakeprofiilien poikkeamalle $\boldsymbol{c}'_j - \boldsymbol{r}'$, $j = 1, 2, \cdots, b$ kahdessa ulottuvuudessa, tällöin

$$\boldsymbol{Y}_1 = \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{b1} & y_{b2} \end{pmatrix} \tag{2.34}$$

Koska $\boldsymbol{A}$ ja $\boldsymbol{B}$ omaavat samat yksittäisarvot $\lambda_1, \lambda_2, \cdots, \lambda_k$, voidaan ne esittää samassa kuvassa. Eetäisyys kahden riviprofiilin välillä on

$$d_{ij}^2 = (\boldsymbol{r}_i - \boldsymbol{r}_j)' \boldsymbol{D}_c^{-1} (\boldsymbol{r}_i - \boldsymbol{r}_j) \tag{2.35}$$

Jos kaksi rivipistettä ovat lähellä toisiaan, nämä pisteet voidaan yhdistää yhdeksi kategoriaksi, mikäli on tarvetta kasvattaa χ^2 aproximaatiota. Rivi- ja sarakepisteen etäisyys ei ole niinkään merkitsevää, mutta niiden läheisyys on. Tällöin nämä kaksi kategoriaa esiintyvät yleisemmin yhdessä, kun voisi ajatella niiden esiintyvän sattumalta, mikäli ne olisivat riippumattomia (Rencher, 2002.)

On näytetty, että rivi- ja sarakepisteiden metriikka on sama. Täten molemmat pisteparvet voidaan esittää samassa kuvassa. Glynn ja Robinson (2014) esittävät kirjassaan, että kuvan jokainen ulottuvuus kuvaa prosenttiosuutta varianssin rakenteesta. Jos kuvataan yhtä ulottuvuutta, saadaan yksinkertaisesti jana, joka kuvaa datapisteiden eroa etäisyytenä toisistaan. Yleensä tämä kuitenkin kuvaa melko huonosti datapisteiden yhteyttä toisiinsa. Jos lisätään toinen dimensio eli y-akseli, saadaan kaksiulotteinen pisteparvi, joka on erittäin yleinen esitystapa korrespondenssianalyysissa sekä monissa muissa spesifikaatioiden vähentämistekniikoissa. Yleensä kaksi ensimmäistä dimensiota kattavat suurimman osan datan varianssista, ja ovat täten riittävä. Mitä lähempänä pisteet ovat toisiaan, sitä enemmän voidaan ajatella niillä olevan riippuvuutta (Glynn & Robinson,2014.)

3 Soveltaminen aineistoihin

3.1 Aineistot

Tutkielmassa sovelletaan korrespondenssianalyysia pankkidataan. Pankkidatassa on suoritettu kysely 36:lle pankissa työskentelevälle henkilölle. Kysely on eräänlainen työpaikkatyytyväisyyskysely, jossa on yli 60 erilaista kysymystä. Osa kysymyksistä on henkilötietoja ja suurin osa työhyvinvointiin liittyviä kysymyksiä, joihin on vastattu antamalla numero 1-5.

Toinen aineisto on Aarresaari-verkoston laatima ja toteuttama uraseurantakysely vuonna 2015 valmistuneille. Siinä on tutkittu 5 vuotta aiemmin valmistuneiden ylemmän korkeakoulututkinnon ja ns. päättyvän alemman korkeakoulututkinnon suorittaneiden sijoittumista työelämään. Kyselyssä on kartoitettu valmistuneiden tyytyväisyyttä tutkintoonsa, työuran kokonaisuutta sekä tämänhetkistä työtilannetta. Data koostuu 88 erilaisesta kysymyksestä. Vastaajia kyselyyn on ollut 6830.

Tutkielmassa käytetään R-ohjelmistoa kuvien laatimiseen. R-koodin lähteenä on käytetty pääasiassa kirjan Modern Applied Statistics with S. Fourth Edition (Venables & Ripley, 2002) materiaalia.

3.2 Datan muokkaus

Pankkidatassa on 31 havaintoa ja 63 muuttujaa. Ensin data siistitään sellaiseen muotoon, että jäljellä ovat ainoastaan solut, joihin on kerätty dataa. Data sisältää useita puuttuvia arvoja. Koska havaintoja on niin vähän, ei ole järkevää käyttää ainoastaan niitä havaintoja, joissa ei esiinny puutuvaa dataa. Niinpä onkin mielekästä generoida jotkin arvot puuttuvan datan tilalle. Tässä tapauksessa käytetään imputaatiota ja vielä täsmällisemmin, usean arvon imputaatiota. Pankkidatassa muuttujia on runsaasti, vaikkakin kategorisia muuttujia on hieman vähänlaisesti. Esimerkkiin on valittu kaksi muuttujaa: koulutusaste sekä ”Tärkein palkkauksesta, urakehityksestä”. Eri koulutusasteet ovat: peruskoulu, ammattikoulu, ylioppilas, keskiaste sekä korkeakoulu. Mahdolliset vastausvaihtoehdot kysymykseen ovat: Urakehitys on mahdollista työyhteisössäni, Palkkaukseni on oikeudenmukainen, Minulle tarjotaan riittävästi kouluttautumismahdollisuuksia, Ammattitaitoani hyödynnetään parhaalla mahdollisella tavalla, Uskon työskenteleväni pankissa vuonna 2003, Henkilöstö valitaan

objektiivisin perustein.

Uraseurantadata sisältää myös puuttuvia arvoja, mutta koska havaintoja on niin runsaasti, voidaan analyysissä käyttää niitä arvoja, joista ei puutu dataa. Datasta on valittu muuttujiksi koulutusala , työmarkkinatilanne ja työnantaja. Koulutusalamuuttuja koostuu luonnontieteistä, teknillisestä alasta, teologiasta, oikeustieteellisestä sekä taidealasta . Työmarkkinatilanne kuvaa sitä, millainen työsuhde valmistunella on viiden vuoden jälkeen. Se koostuu vaihtoehdoista: vakituinen kokopäivätyö, määräaikainen kokopäivätyö, osa-aikatyö, itsenäinen yrittäjä, useita rinnakaisia työsuhteita, työllistetty/harjoittelija, työtön työnhakija, työvoimakoulutus, päätoiminen opiskelu, perhevapaa (työsuhteessa), perhevapaa (ei työsuhteessa), työskentely apurahalla, työvoiman ulkopuolella, muu tilanne. Työnantajamuuttuja koostuu vaihtoehdoista: kunta tai kuntaryhmä, valtio, suuri yritys, pieni tai keskisuuri yritys, oma yritys, järjestö, säätiö tai seurakunta, yliopisto, ammattikorkeakoulu, muu työnantaja.

3.3 Puuttuvan tiedon paikkaaminen (Imputaatio)

Usean arvon imputaatiossa puuttuvat arvot paikataan joukolla M arvoja. Tyypillisesti ($5 \leq M \leq 10$) on sopiva määrä estimoimaan otosvaihtelua. Nummen (2019) mukaan imputaation vaiheet voidaan jakaa seuraavaan kolmeen vaiheeseen.

1. Luodaan M kappaletta imputaatioita.
2. Lasketaan M kappaletta parametriestimaatteja ja niille keskihajonnat.
3. Saadut estimaatit yhdistetään, jotta saadaaan yksittäinen estimaatti parametrille. (Tässä tapauksessa valitaan jokin sopiva imputaatio ja käytetään sitä. Jätetään yhdistämisvaihe tekemättä, sillä se ei ole tässä tapauksessa niin oleellista (Nummi, 2019.))

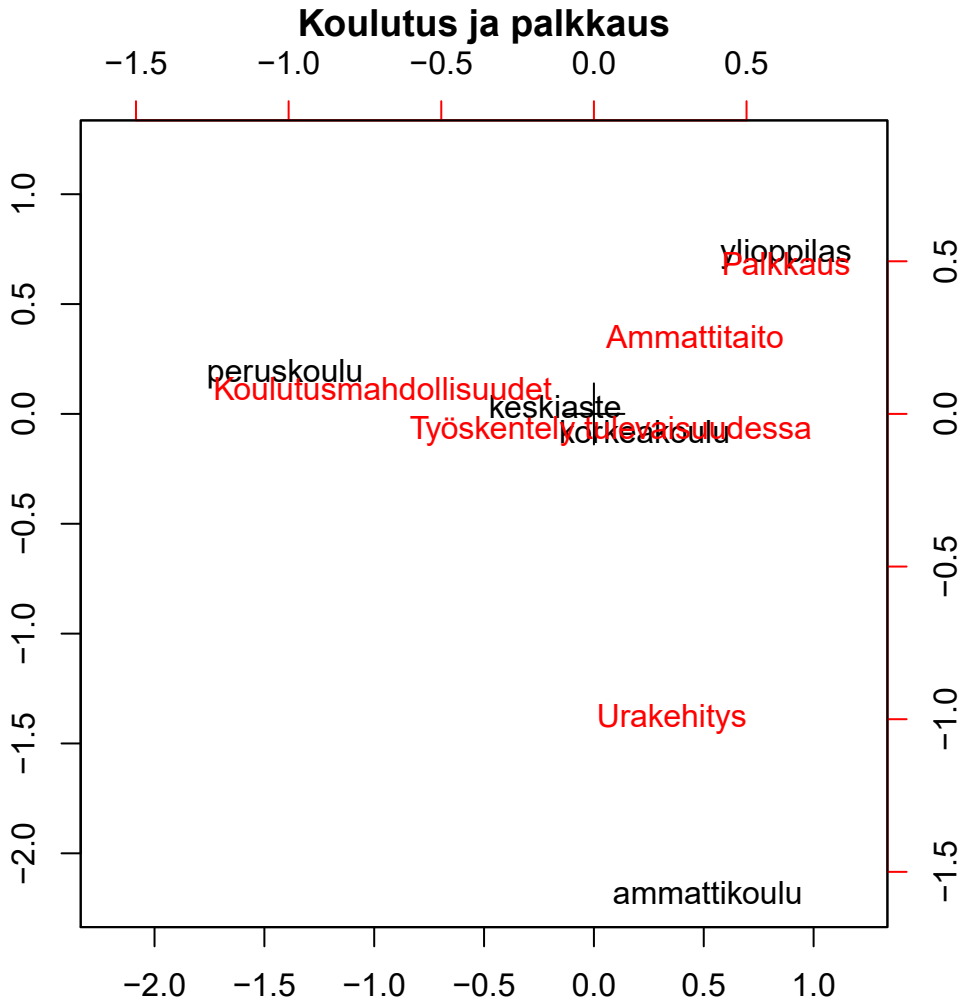
Työssä käytetty imputaation R-koodi on seuraavanlainen:

```
library(mice)
imp<-mice(pankkidata, m=5, maxit = 10, method="pmm"
, seed = 500, print=FALSE)
completePankkidata<-complete(imp, 2)
complete.cases(completePankkidata)
```

Näistä imputoiduista arvoista on valittu yksi paikkaamaan puuttuvaa tietoa.

3.4 Korrespondenssianalyysi pankkidataan

Sovelletaan korrespondenssianalyysia pankkidataan. Selvitetään pitävätkö eri koulutustaustoista tulevat työntekijät eri asioita tärkeimpänä heidän palkkauksessaan.



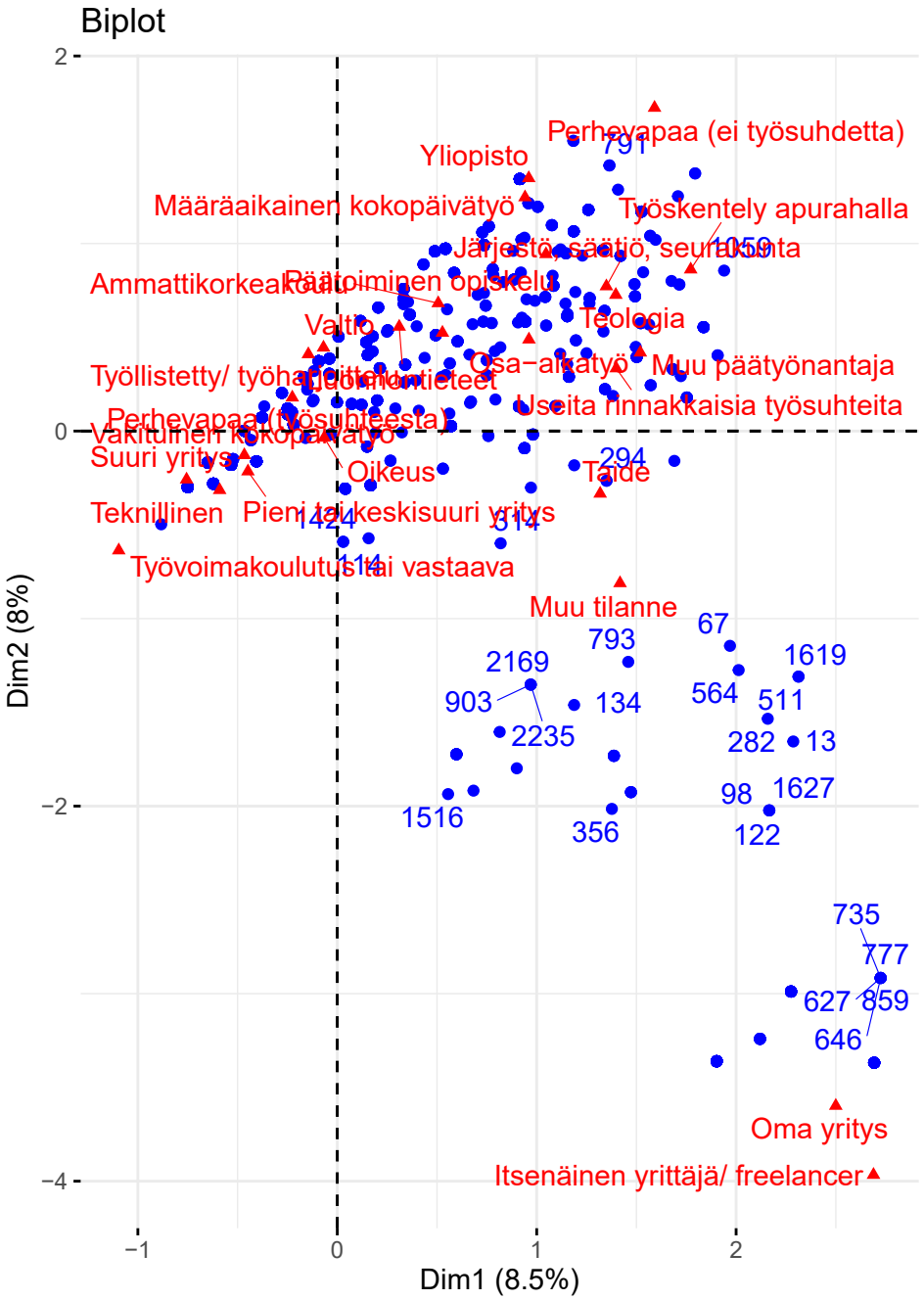
Kuva 3.1. Kuvassa on sovitettu molemmat muuttujat samaan kuvaan. Punaisella tekstillä näkyvät vastaukset ja mustalla tekstillä koulutustaustaa.

Kuvasta on havaittavissa, että keskiaste ja korkeakoulu ovat vastanneet melko tasaisesti kaikkia vaihtoehtoja. Pelkän peruskoulun suorittaneet pitävät koulutusmahdollisuuksia tärkeimpänä. Ylioppilaat puolestaan pitävät ammattitaidon hyödyntämistä ja palkkauksen oikeudenmukaisuutta tärkeimpänä. Ammattikoulun käyneet taas näkevät urakehitys mahdollisuuksia tärkeimpänä. Kuva näyttää erittäin selkeitä tuloksia, mutta tulkintoja tehdessä täytyy ottaa erityisesti huomioon vastanneiden

henkilöiden vähyys. Esimerkiksi vain ammattikoulun käyneitä vastaajia on ainoastaan yksi. Tästä syystä tuloksille ei voida antaa kovinkaan suurta painoarvoa.

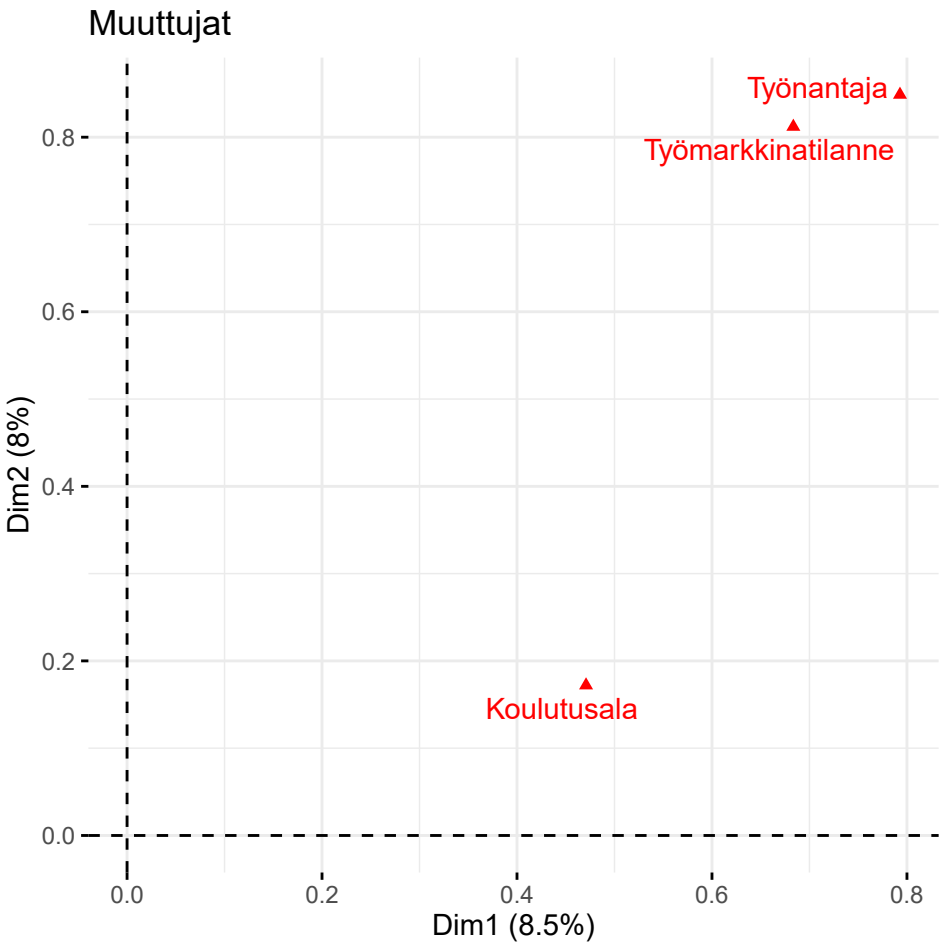
3.5 Korrespondenssianalyysi uraseurantadataan

Toisessa esimerkissä sovelletaan menetelmää uraseurantadataan. Pyritään selvittämään, millaisiin työsuhteisiin eri koulutusaloilta on päädytty.



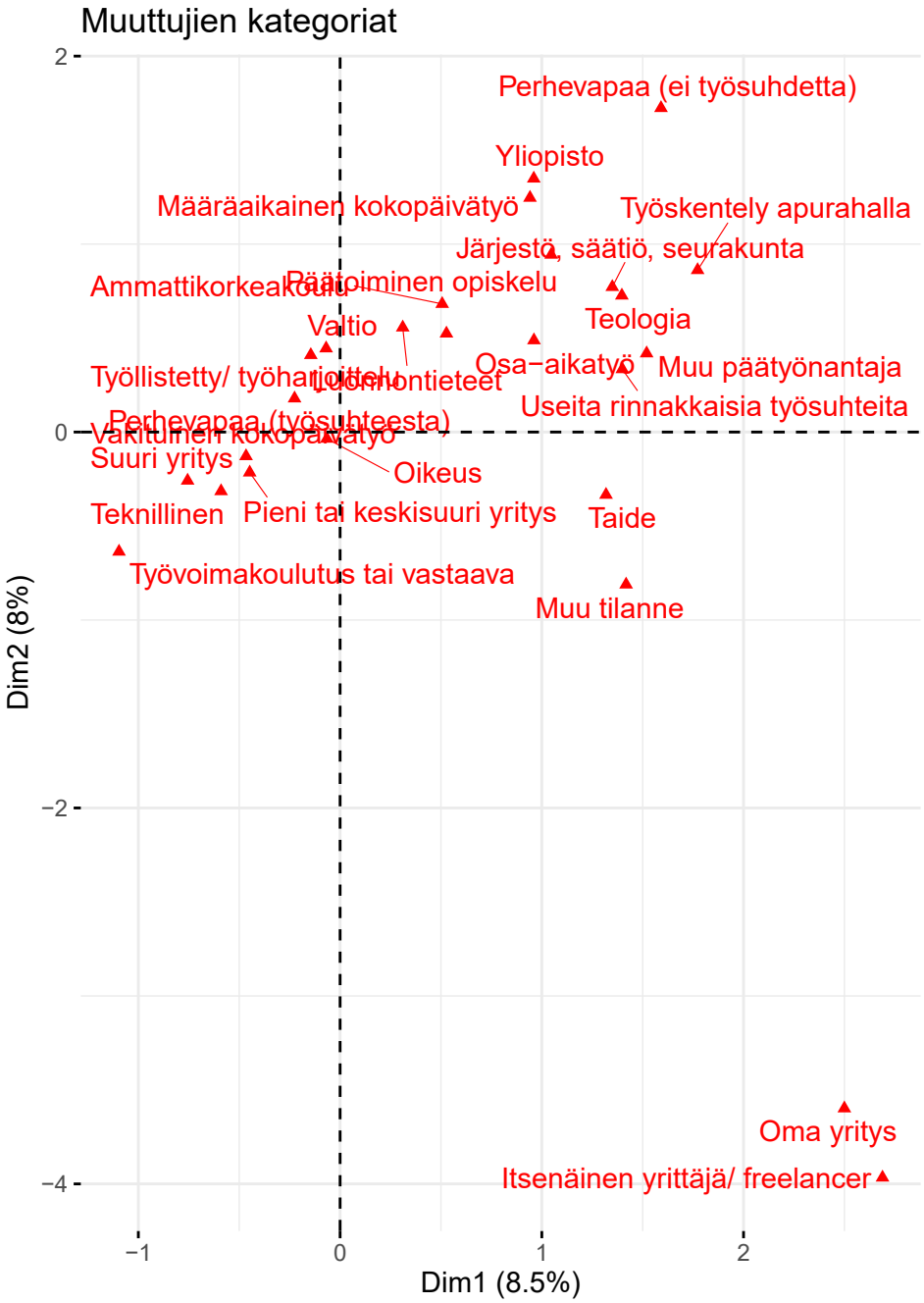
Kuva 3.2. Kuvassa on sovitettu kategoriat ja havainnot samaan kuvaan. Sinisellä värillä on merkitty havaintojen pisteet sekä monesko havainto on kyseessä. Punaisella taas on merkitty kategorioiden pisteet ja niiden nimet.

Biplot-kuva antaa osviittaa siitä, miten eri havainnot ovat sijoittuneet eri kategorioiden välillä. Voidaan huomata, että ”oma yritys” kategoriaan päin on joitakin arvoja kallellaan, mutta suurimmaksi osaksi arvot ovat keskittyneet lähelle origoa.



Kuva 3.3. Kuvassa on esitetty muuttujien yhteys toisiinsa.

Kuvasta huomataan, että työnantajan ja työmarkkinatilanteen välillä on korrelaatiota. Kuitenkaan näitä ei ole tarpeellista yhdistää.



Kuva 3.4. Kuvassa kategorioiden yhteys toisiinsa.

Kuva 3.4 on muuten sama kuin kuva 3.2, mutta kuvassa 3.4 on eritelty ainoastaan kategoriat. Kuvasta huomataan, että teknilliseltä alalta työllistytään yleisimmin erikokoisiin yrityksiin. Myös vakituinen kokopäivätyö on erittäin yleinen tällä alalla. Oikeustieteellinen on melko keskellä, eli se on liitettävissä jossakin määrin kaikkiin kategoriohin, joskin siihen liittyy valtiollinen sektori vahvasti, mikä käy hyvin järkeen. Teologian alalla työllistytään luonnollisesti seurakuntaan, mikä selittää pisteiden lähes päällekkäisyyden. Taideala on melko irrallaan muista kategorioista. Ai-

noastaan ”muu tilanne” on sen läheisyydessä. Taideala on myös hieman kallellaan ”itsenäinen yrittäjä” suuntaan verrattuna muihin aloihin.

4 Johtopäätökset

Kuten tutkielmassa on havaittu, korrespondenssianalyysi on erittäin hyvä menetelmä visualisoimaan kategorista dataa. Vaikka menetelmän matemaattinen esitys on pitkähkö, on sen avulla kuitenkin verrattain helppo saada yksinkertainen ja havainnollistava kuvaus datan käyttäytymisestä. Yksinkertaisen korrespondenssianalyysin tekeminen R-ohjelmalla on suhteellisen helppoa, eikä kuvankaan tulkitseminen ole kovinkaan hankalaa. Kuvasta saattaa tosin muodostua melko sekava, jos muuttujia on useampia. Toisaalta korrespondenssianalyysille löytyy vaihtoehtoisiakin menetelmiä, eikä se tunnu olevan yleisesti erityisen käytetty menetelmä.

Menetelmän ongelma kuitenkin on siinä, että menetelmää voidaan hyödyntää ainoastaan kategorisiin muuttujiin. Työssä huomattiin myös, että havaintoja pitää olla melko paljon, etenkin multiple korrespondenssianalyysia tehtäessä, jotta analyysilla saadaan aikaiseksi hyviä tuloksia. Tilanteiden joukko, mihin menetelmä soveltuu hyvin, tuntuu olevan verrattain kapea. Se saattaa osaltaan selittää sen, miksi menetelmä ei ole kovin yleinen.

Tutkimusta voisi jatkaa esimerkiksi vertailemalla korrespondenssianalyysia toisiin monimuuttujamenetelmiin, esimerkiksi pääkomponenttianalyysiin.

Lähteet

[1] Glynn, D. & Robinson, J.(2014) Corpus Methods for Semantics *Quantative studies in polysemy and synonymy*. University of Paris VIII, University of Sussex.

[2] Nummi, T.(2019) *Statistical analysis with missing data*. Tampere: University of Tampere Faculty of Natural Sciences.

[3] Rencher, A. *Methods of Multivariate Analysis. Second Edition*. Brigham Young University, 2002.

[4] Venables, V.N. & Ripley, B.D.(2002) *Modern Applied Statistics with S. Fourth Edition*.

[5] Le Roux, B. & Rouanet, H.(2010) *Multiple Correspondense Analysis*. Los Angeles: SAGE.

A Kuvaajien R-koodi

```
con4<-table(completePankkidata\$C60,completePankkidata\$C51)
con4 <- as.data.frame.matrix(con4)
corresp(con4)
par("mar")
par(mar=c(2,2,2,2))
biplot(corresp(con4, nf = 2), xlim=c(-2.2, 1.2), ylim=c(-2.2, 1.2));
title("Koulutus ja palkkaus", line = 2.2)


res.mca <- MCA(dfura5, graph = TRUE)
print(res.mca)

fviz_mca_var(res.mca, choice = "mca.cor",
             title = "Muuttujat",
             repel = TRUE,
             ggtheme = theme_minimal())


fviz_mca_biplot(res.mca,
             title = ("Biplot"),
             repel = TRUE,
             ggtheme = theme_minimal())

fviz_mca_var(res.mca,
                title = ("Muuttujien kategoriat"),
                repel = TRUE,
                ggtheme = theme_minimal())
```