Tilastotieteen pro gradu -tutkielma

Moniulotteinen korrespondenssianalyysi – sovelluksena jokien pohjaeläinaineisto

Marko Vikstedt

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
25. kesäkuuta 2015

Tiivistelmä

Vikstedt Marko: *Moniulotteinen korrespondenssianalyysi – sovelluksena jokien pohjaeläinaineisto.*
Tilastotieteen pro gradu -tutkielma, 30 s. + liitteet (23 s.). Matematiikan ja tilastotieteen laitos, Jyväskylän yliopisto, 25. kesäkuuta 2015.

Kerättävän tiedon määrä kasvaa jatkuvasti ja siten myös tutkittavien aineistojen koko kasvaa. Suurien ja monimutkaisten aineistojen tehokkaaseen analysointiin tarvitaan menetelmiä, joilla aineiston muuttujien väliset assosiaatiot voidaan tunnistaa. Tunnistamalla aineistosta mielenkiintoisimmat assosiaatiot voidaan jatkotutkimukset kohdentaa niihin, jolloin säästyy resursseja ja aikaa. Moniulotteinen korrespondenssianalyysi kuuluu näihin eksploratiivisiin menetelmiin. Tässä tutkielmassa menetelmää sovelletaan Suomen ympäristökeskuksen (SYKE) tuottamaan pohjaeläinaineistoon.

Aineisto koostuu mittauksista, jotka on tehty Suomen jokien koskialueilla. Koskialueita on 590 ja näistä on kerätty 5552 näytettä vuosina 2006 - 2012. Koskialueista otetuista näytteistä on tunnistettu ja laskettu niissä esiintyvät pohjaeläintaksonit, joita on 161 eri taksonia. Jokaiselle näytteelle tiedetään lisäksi koskialueen jokityyppi, pohjatyyppi, luonnontilaisuus sekä sijoittuminen Etelä- tai Pohjois-Suomeen. Tutkielman tavoitteena on selvittää pohjaeläinaineistoon liittyviä assosiaatioita, erityisesti liittyen hienoaines- ja kivipohjatyyppeihin. Jälkimmäinen jaotellaan vielä iso- ja pienkivityyppeihin.

Moniulotteisessa aineistossa olevaa informaatiota pyritään tiivistämään moniulotteisella korrespondenssianalyysilla helposti tulkittaviksi useimmin kaksiulotteisiksi kuviksi. Kuvia kutsutaan menetelmän yhteydessä kartoiksi, ja ne kuvaavat muuttujien välisiä assosiaatioita tiivistettynä kahden suurimman selitysosuuden omaavalle dimensiolle eli aliavaruudelle. Aliavaruudet määritetään muuttujien avulla. Tutkielmassa käytetään pohjaeläintaksoneita myös lisämuuttujina, jotka voidaan piirtää kartalle, mutta ne eivät vaikuta dimensioiden määrittämiseen.

Pohjatyyppien välillä voidaan tulkita moniulotteisen korrespondenssianalyysin perusteella olevan eroa hienoaines- ja kivipohjatyyppien suhteen. Hienoaineksen erottuminen näkyi kaikissa suoritetuissa analyyseissa. Kategorisia pohjaeläinmuuttujia käytettäessä havaittiin mäkärien (*Simuliidae*) olevan mahdollinen indikaattoritaksoni pohjatyyppien erottelussa hienoaines- ja kivipohjatyyppeihin.

Analyysitulosten perusteella voidaan myös todeta olevan eroa Pohjois- ja Etelä-Suomen välillä sekä luonnontilaisten ja ihmistoiminnan kuormittamien koskipaikkojen välillä. Pohjois- ja Etelä-Suomen väliselle jaolle voidaan tulosten perusteella esittää indikaattoritaksoneiksi kahta päiväkorentoa. Jokityypeistä savimaiden joet erottuvat muista jokityypeistä omana ryhmänään.

Tutkielman tulosten perusteella voidaan ehdottaa lisätutkimuksia siitä, mikä erottaa hienoaineksen kivipohjatyypeistä.

Avainsanoja: dimensio, eksploratiivinen menetelmä, kartta, moniulotteinen korrespondenssianalyysi, pohjaeläinaineisto, pohjatyyppi.

Sisältö

1 Johdanto

Kaksiulotteinen korrespondenssianalyysi (*(Simple) Correspondence Analysis – CA*) on monimuuttujamenetelmiin kuuluva, erityisesti kategorisia muuttujia sisältävien frekvenssiaineistojen analysointiin ja kuvailuun kehitetty työkalu. Teoreettisesti läheinen menetelmä on pääkomponenttianalyysi. Moniulotteinen korrespondenssianalyysi (*Multiple Correspondence Analysis – MCA*) on yleistetty versio kahden muuttujan korrespondenssianalyysista, jossa kahden muuttujan sijasta analysoidaan useampia muuttujia yhtä aikaa (Greenacre & Blasius, 2006). Tutkielmassa menetelmien teoriaa käsitellään niiltä osin, jotka ovat olleet sovelluksen kannalta keskeisiä analyysin toteuttamiseksi.

Kaksi- tai moniulotteisen korrespondenssianalyysin sijaan voitaisiin käyttää menetelmää nimeltä *Joint Correspondence analysis – JCA* (Greenacre, 1988). Tulokset JCA-menetelmässä saattavat kuvata paremmin aineiston assosiaatiorakenteita verrattuna moniulotteiseen korrespondenssianalyysiin. Menetelmää ei kuitenkaan käytetä tässä tutkielmassa, koska iteratiivisena menetelmänä se on laskennallisesti raskaampi.

Korrespondenssianalyysi eroaa pääkomponenttianalyysista siten, että kovarianssi- ja korrelaatiomatriisien tutkimisen sijaan analyysi suoritetaan frekvenssitaulukoiden avulla. Käytännössä korrespondenssianalyysi on kategoristen muuttujien ja pääkomponenttianalyysi jatkuvien muuttujien analysointimenetelmä. Tavoitteena korrespondenssianalyysissa on löytää maksimaalinen korrelaatiorakenne analysoitavan frekvenssitaulukon muuttujien välille ja esittää se helposti tulkittavassa muodossa.

Korrespondenssianalyysin tulosten tulkinnat suoritetaan kartan avulla. Kartta piirretään useimmiten kahden suurimman selitysosuuden saaneen dimension eli aliavaruuden perusteella. Jokaiselle muuttujan luokalle lasketaan koordinaatit, joiden avulla se voidaan sijoittaa kartalle. Kartalla esitettyjen muuttujien luokkien sijainnin perusteella voidaan tehdä tulkintoja muuttujien välisestä assosiaatiosta. Erityisesti analysoitavien muuttujien määrän ja aineiston koon kasvaessa suureksi, aineiston informatiivinen tiivistäminen kaksiulotteiseksi kartaksi helpottaa sen tulkintaa.

Korrespondenssianalyysi voidaan lukea tiedonlouhintamenetelmäksi (*data mining*). Tiedonlouhinta ja sen kehittäminen on noussut nyky-yhteiskunnassa yhdeksi tärkeimmistä data-analyysin ja dataperusteisen päätöksenteon apuvälineistä. Suurten aineistojen tehokkaan analysoinnin varmistamiseksi käytettävillä tiedonlouhintamenetelmillä aineistoista etsitään riippuvuusrakenteita. Näitä voidaan tämän jälkeen analysoida tarkemmin muilla menetelmillä: Esimerkiksi, jos moniulotteisen korrespondenssianalyysin kartasta havaitaan kahden muuttujan välinen mahdollinen assosiaatio, tätä voidaan tutkia χ^2-testillä näistä kahdesta muuttujasta muodostetusta uudesta ristiintaulukosta. Tällaiset tarkastelut rajataan työn ulkopuolelle.

Korrespondenssianalyysin on aiemmin todettu toimivan ”paikka×laji”-taulukoiden analyysimenetelmänä (ter Braak, 1985; Cadoret et al., 1995). Tyypillisesti suuria määriä nollahavaintoja sisältävien esiintyvyys- ja runsausdatojen analysoitavuus korrespondenssianalyysilla on ollut toimivampaa muihin kilpaileviin analyysimenetelmiin, esimerkiksi loglineaarisiin malleihin, verrattuna erityisesti laskennallisen helppouden ansiosta (ter Braak, 1985). Haluttaessa analysoida useampien muuttu-

jien välisiä rakenteita kaksiulotteinen korrespondenssianalyysi ei enää riitä. Työssä tutkitaan, toimiiko moniulotteinen korrespondenssianalyysi yhtä hyvin samankaltaisiin ekologisiin datoihin, joissa on paikan ja lajin lisäksi useita havaintopaikkaa kuvaavia muuttujia.

Sovellusaineistona käytetään Suomen ympäristökeskuksen (SYKE) tuottamaa pohjaeläinaineistoa. Pohjaeläinaineisto sisältää tietoa Suomen jokien koskipaikkojen pohjaeläinnäytteistä. Näytteistä on laskettu niissä olevat pohjaeläintaksonit ja näytepaikasta on kirjattu sen ekologiset piirteet, kuten pohjatyyppi ja jokityyppi. Pohjaeläinaineistoa käytetään yhtenä osana Suomen jokien ekologisen tilan luokittelussa (Suomen ympäristökeskus, 2012a).

Ekologisen tilan luokittelu on tärkeää vesiensuojelutyön ja sen resurssien kohdentamisen kannalta. Ekologisen tilan määritys tapahtuu vertaamalla taksonien jakaumia luonnontilaisissa (referenssi) ja ihmistoiminnan vaikutuksen alaisissa (impakti) joissa. Luokittelu perustuu Suomen ympäristökeskuksen ohjeeseen, jossa on esitetty aiempien tutkimustulosten avulla tehdyt ohjeet luokitteluun (Suomen ympäristökeskus, 2012b).

Tutkielman tavoitteena on selvittää moniulotteisen korrespondenssianalyysin avulla sovellusaineiston `pohjatyyppi`-muuttujan hienoaines-tason assosiaatiota muiden pohjatyyppien kanssa. Tarkoituksena on tutkia, pystytäänkö tällä menetelmällä havaitsemaan samoja tuloksia kuin mitä aiheesta aiemmin tiedetään (Suomen ympäristökeskus, 2012b; Meissner et al., 2013) sekä pystytäänkö mahdollisesti esittämään uusia aineistossa olevia assosiaatiorakenteita. Tavoitteena on myös tutkia muuttujan kategorisoinnin sekä lisämuuttujaksi valinnan vaikutusta sovelluksesta saataviin tuloksiin sekä indikaattoritaksonien havaitsemista moniulotteisella korrespondenssianalyysilla.

Tutkielmassa esitellään johdannon jälkeen tutkimusaineisto, käytettävät muuttujat muunnoksineen sekä tutkimusongelma (Luku 2). Seuraavana käsitellään korrespondenssianalyysin teoriaa aloittaen kaksiulotteisesta ja laajentaen moniulotteiseen versioon (Luvut 3 ja 4). Kaksiulotteisesta versiosta esitetään havainnollistava esimerkki. Luvussa 5 sovelletaan moniulotteista korrespondenssianalyysia pohjaeläinaineistoon. Viimeisenä tutkielmassa on esitetty yhteenveto ja liitteet.

2 Tutkimusaineisto ja tutkimusongelma

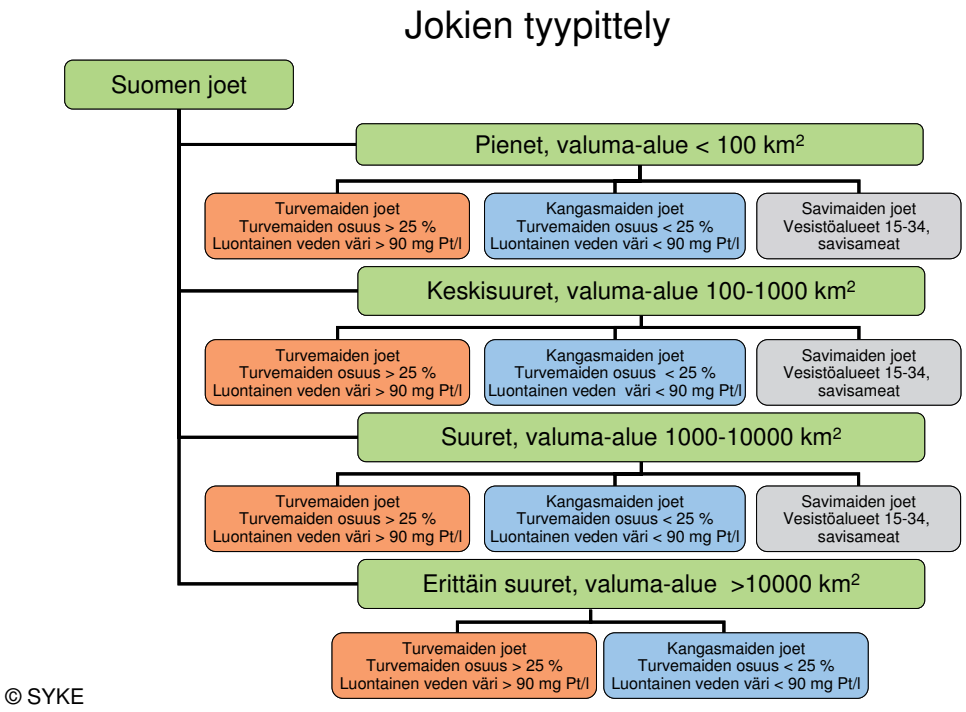
2.1 Tutkimusaineisto ja sen muokkaus

Tutkimusaineisto on Suomen ympäristökeskuksen (SYKE) Suomen jokien ekologista luokittelua varten tuottama pohjaeläinaineisto. Aineiston havainnot ovat vuosilta 2006-2012. Havaintoyksikkönä ja -paikkana on tiettyyn jokeen kuuluva koskialue, joita on 590 kappaletta. Aineistossa on yhteensä 5552 havaintoa jakautuen eri vuosille ja eri pohjatyyppeihin. Pohjatyyppejä aineistossa on kolme: hienoaines (h), pienkivi (pKi) sekä isokivi (iKi). Aineiston muuttujat on esitetty taulukossa 1 ja havaintomatriisia on havainnollistettu liitteessä A.1.

Jokainen koskialue on luokiteltu johonkin jokityyppiin (11 kpl). Jokityypit on jaettu valuma-alueen pinta-alan sekä sen pääasiallisen maaperän koostumuksen mukaan (Suomen ympäristökeskus, 2012a). Jokityyppejä ovat pienet, keskisuuret sekä suuret turve-, kangas- ja savimaiden joet (Pt, Pk, Psa, Kt, Kk, Ksa, St, Sk, Ssa). Erittäin suuret joet on jaettu kahteen tyyppiin: turve- ja kangasmaiden jokiin (ESt, ESk). Jokityyppien tarkemmat jakoperusteet on esitetty kuvassa 1. Sijainniltaan joet on jaettu pohjoisiin ja eteläisiin jokiin, jakautuen siten, että eteläisiin jokiin luetaan Oulunjoen vesistöalue sekä sitä eteläisemmät vesistöalueet. Luonnontilaisuudeltaan joet on jaettu vertailu- (referenssi) ja ihmistoiminnan alaisiin (impakti) jokiin. Luokittelu referenssi- ja impaktijokiin tehdään erilaisten fysikaalis-kemiallisten, biologisten sekä hydromorfologisten ominaisuuksien perusteella. Esimerkiksi fosforipitoisuuden, pohjaeläinten sekä vesirakenteiden (esim. padot) perusteella.

Taulukko 1: Analyysissa käytettävät muuttujat selityksineen.

Muuttuja	Arvot	Nimi aineistossa
joen ID numero	1 - 590	ID
näytevuosi	2006 - 2012	naytevuosi
pohjan koostumus	hienoaines (h)	pohjatyyppi
	pienkivi (pKi)	
	isokivi (iKi)	
jokityyppi	ks. kuva 1	jokityyppi
sijainti	pohjoinen – etelä	PE
luonnontilaisuus	kyllä (=1 eli referenssi)	ref
	ei (=0 eli impakti)	
taksonit	ks. liite A.4	



Kuva 1: Suomen jokien tyypittely valuma-alueen koon ja maaperän koostumuksen suhteen. Suomen ympäristökeskuksen luvalla. Lähde: Suomen ympäristökeskus (2013).

Aineiston keskeisintä osaa ovat eri pohjaeläintaksonien yksilömäärät näytteittäin. Aineistossa on 154 pohjaeläintaksonia. Osa taksoneista on tunnistettu suku- ja osa lajitasolle. Näytteiden taksonien yksilömäärät vaihtelevat yhdestä yksilöstä 359 775 yksilöön. Tarkasteltavat taksonit on valittu pääasiassa Suomen jokien ekologisen tilan luokittelussa käytetyn PMA-indeksin (*Percent Model Affinity*) laskemiseen käytettävien taksonilistojen mukaan (Suomen ympäristökeskus, 2012b, s. 110-114). Lisäksi joitakin taksoneita on valittu listan ulkopuolelta ja osa listalla olevista on poistettu asiantuntijaohjeiden mukaan. Lista analyysissa käytetyistä taksoneista on esitetty liitteessä A.4.

Alkuperäistä aineistoa muokattiin ennen analyysien soveltamista vielä siten, että aineistosta poistettiin puuttuvaa tietoa sisältävät havainnot, jos tietoa ei pystytty täydentämään ympäristöhallinnon OIVA-palvelusta saatavien tietojen eikä asiantuntijalausuntojen perusteella. Havaintoja, joissa oli puuttuvaa tietoa, oli 515, eli noin 8,5 % kaikista havainnoista. Poistamisen aiheuttaman harhan suuruus arvioitiin pieneksi: tietojen puuttumiselle ei löytynyt yhdistäviä tekijöitä aineistoa ja sen taustoja tutkittaessa. Puuttuvan tiedon lähempi tarkastelu rajataan kuitenkin työn ulkopuolelle.

2.2 Muuttujien muunnokset

Moniulotteinen korrespondenssianalyysi on tarkoitettu kategorisia muuttujia sisältävien taulukoiden analysointiin. Aineistossa taksonit ovat jatkuvia lukumäärämuuttujia, jotka kategorisoitiin moniulotteista korrespondenssianalyysia varten. Kategorisointi suoritettiin luomalla pohjaeläinmuuttujasta kaksiluokkainen muuttuja eli esiintyykö taksonia tietyllä havaintorivillä (1) vai ei (0). Tällaisella yksinkertaisella kategorisoinnilla pyrittiin laskennan nopeuttamiseen sekä indikaattoritaksonien löytämiseen. Indikaattoritaksonilla tai -lajilla tarkoitetaan tässä tutkielmassa pohjaeläintaksonia, jonka perusteella voidaan mahdollisesti tehdä päätelmiä siitä, mihin tarkasteltavan muuttujan luokkaan kyseinen koskipaikka kuuluisi.

Vaihtoehtona pohjaeläinmuuttujien kategorisoimiselle käytettiin niiden määrittelemistä lisä- tai täydentäviksi muuttujiksi (*supplementary variable*). Lisämuuttujia käytetään yleisimmin esittämään jotain aineistoon liittyvää mielenkiintoista osaa, joka kuitenkin saattaa olla jo esitettynä toisten muuttujien sisältämänä aineistossa. Esimerkkinä sovellusaineistosta voitaisiin mainita ”savimaa”-lisämuuttujan käyttämistä kuvaamaan savimaiden kaiken kokoisten jokien ”keskiarvoista” sijaintia.

Määriteltäessä muuttujat lisämuuttujiksi ne eivät osallistu kartan koordinaatiston määrittelyyn. Koordinaatisto ja dimensiot määritetään tässä tapauksessa muiden muuttujien avulla ja lisämuuttujat kuvataan pisteinä tähän ”valmiiseen” karttaan. Lisämuuttujien käyttöä on esitetty havainnollistavan esimerkin avulla kirjassa Greenacre & Blasius (2006, s. 70-74). Lisämuuttujiin liittyvää teoriaa on esitetty tarkemmin kirjassa Greenacre (1993, s. 95-102, s.149) sekä lyhyesti luvussa 4.2.3.

Näytteenottoaika alkuperäisessä aineistossa on esitetty päivän tarkkuudella, mikä sovellukseen nähden oli kuitenkin tarpeettoman tarkka. Näytteenottoaika muutettiin analyysia varten pelkäksi vuodeksi. Havaintopaikkojen välinen vaihtelu haluttiin ottaa huomioon analyysissa. Vaihtelun huomioimiseksi aineistoon luotiin `ID`-muuttuja, jonka tasoina oli jokainen koskipaikka. `ID`-muuttuja on luotu havaintopaikkojen nimien ja alkuperäisestä aineistosta löytyvän `Paikan id` -muuttujan perusteella.

2.3 Tutkimusongelma

Vuodesta 2014 alkaen jokien ekologisessa luokittelussa käytetään pohjatyypeistä ainoastaan iso- ja pienkivipohjatyyppejä. Hienoainesta ei käytetä, koska sen ei katsota tarjoavan merkittävää etua luokitteluun (Meissner et al., 2013). Tutkielman tavoitteena on selvittää, voidaanko moniulotteisella korrespondenssianalyysilla todeta hienoaineksen ja kahden muun pohjatyypin välillä eroa. Tavoitteena on myös selvittää, onko hienoaineksen ja muiden muuttujien välillä havaitsemattomia assosiaatioita ja näin mahdollisesti hyödyntää kerättyä informaatiota paremmin.

Menetelmälähtöisenä tavoitteena on vertailla lisämuuttujien sekä kategorisoitujen muuttujien käyttöä ja tuloksia pohjaeläinaineiston sovelluksessa. Tutkittavana kohteena on myös indikaattorilajien tunnistaminen moniulotteisen korrespondenssianalyysin avulla.

3 Kaksiulotteinen korrespondenssianalyysi

Tässä luvussa esitellään ensimmäisenä korrespondenssianalyysin kannalta keskeisiä historiallisia tuloksia. Tämän jälkeen esitellään keskeisiä käsitteitä, niiden merkintätapoja sekä suhteita muissa menetelmissä käytettäviin käsitteisiin ja merkintöihin. Luvun lopuksi esitetään esimerkki kaksiulotteisen korrespondenssianalyysin käytöstä. Tässä työssä käytetään vastaavia merkintöjä kuin kirjassa Greenacre & Blasius (2006).

3.1 Lyhyt katsaus historiaan

Ensimmäisiä askeleita korrespondenssianalyysin kehityksessä otettiin vuonna 1935, kun Hirschfeld (1935, myöh. Hartley) esitti kaavan kontingenssitaulun rivien ja sarakkeiden väliselle korrelaatiolle. 40-luvulla menetelmän kehittäjinä katsotaan olleen mm. Fisher (1940) sekä Guttman (1941). (Greenacre & Blasius, 2006). Menetelmän keksimisen jälkeen korrespondenssianalyysi oli useamman vuosikymmenen ajan lähes tuntematon menetelmä englanninkielisessä tutkimuksessa. Ranskassa menetelmä oli toisaalta hyvinkin suuressa suosiossa. Käännekohtana korrespondenssianalyysin yleiseen suosion kasvuun, etenkin englanninkielisessä kirjallisuudessa, voidaan katsoa olevan kirjan Benzécri et al. (1973, kirj. ranskaksi) ja erityisesti artikkelin Hill (1974) julkaisu (Clausen, 1998). Kiinnostusta menetelmää kohtaan lisäsivät myös teokset Lebart et al. (1984) sekä Greenacre (1984).

Korrespondenssianalyysi käsitteenä vakiintui 1960-luvulla Ranskassa: Aikaisemmin menetelmää tai matemaattiselta teorialtaan vastaavia menetelmiä kutsuttiin nimillä *"reciprocal averaging"* sekä *"dual (*tai *optimal) scaling"*. Näissä menetelmissä erona korrespondenssianalyysiin on se, että niiden tuloksia tarkastellaan numeerisessa muodossa. (Greenacre, 1984).

3.2 Peruskäsitteitä

Palautetaan ensin mieleen χ^2-testisuure, koska korrespondenssianalyysin yhteydessä esiintyvät käsitteet ja tulokset liittyvät läheisesti siihen. Olkoot ristiintaulukko $\mathbf{N}$, sen havaitut frekvenssit n_{ij}, $i = 1, \ldots, I$ ja $j = 1, \ldots, J$, sekä frekvenssien kokonaissumma n. χ^2-testisuuretta laskettaessa frekvenssitaulukon havaittuja frekvenssejä verrataan taulukon odotettuihin frekvensseihin, millä mitataan rivien ja sarakkeiden riippuvuutta. Korrespondenssianalyysin tarkoituksena on analysoida, millaista havaittu riippuvuusrakenne taulukossa on. Testisuure χ^2 määritellään

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - e_{ij})^2 / e_{ij}, \tag{1}$$

jossa n_{ij} ovat havaitut frekvenssit ja $e_{ij} = (n_{i\cdot} \times n_{\cdot j})/n$ odotetut frekvenssit. Kaavaan (1) liittyvät standardoidut jäännökset s'_{ij} ovat seuraavat

$$s'_{ij} = (n_{ij} - e_{ij})/\sqrt{e_{ij}}. \tag{2}$$

Korrespondenssianalyysin yhteydessä taulukon $\mathbf{N}$ suhteellisista frekvensseistä laskettuja marginaalijakaumia kutsutaan massoiksi (Greenacre, 2007) eli

$$\text{sarakemassat:} \quad c_j = \sum_{i=1}^{I} n_{ij}/n = n_{\cdot j}/n, \tag{3}$$

$$\text{rivimassat:} \quad r_i = \sum_{j=1}^{J} n_{ij}/n = n_{i\cdot}/n. \tag{4}$$

Vaihtoehtoisesti niitä voidaan nimittää myös suhteellisiksi sarake- tai rivisummiksi. Korrespondenssianalyysin suorittamiseen tarvitaan näiden massojen lisäksi korrespondenssimatriisi $\mathbf{P}$, joka sisältää massojen lisäksi suhteelliset osuudet $p_{ij} = n_{ij}/n$. Huomaa, että massat voidaan laskea myös korrespondenssimatriisin suhteellisten osuuksien avulla $c_j = \sum_i p_{ij}$ ja $r_i = \sum_j p_{ij}$.

Massoja käytetään korrespondenssimatriisin arvojen keskittämiseen ja normalisoimiseen eli standardointiin. Massojen avulla standardoituja arvoja nimitetään myös standardoiduiksi jäännöksiksi s_{ij} ja ne muodostavat standardoidun matriisin $\mathbf{S}$. Keskittäminen suoritetaan laskemalla erotus korrespondenssimatriisin solun p_{ij} ja sitä vastaavien sarake- ja rivimassojen tulon $r_i c_j$ välille. Normalisointi tehdään jakamalla saatu erotus sarake- ja rivimassojen tulon neliöjuurella seuraavasti

$$s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}. \tag{5}$$

Huomaa korrespondenssianalyysin standardoitujen jäännösten s_{ij} samankaltaisuus verrattuna χ^2–testisuureen standardoituihin jäännöksiin s'_{ij} kaavassa (2). Kaavan (5) sekä $n_{ij} = p_{ij} \cdot n$ ja $e_{ij} = r_i c_j \cdot n$ perusteella voidaan kirjoittaa

$$s'_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} = \frac{np_{ij} - nr_i c_j}{\sqrt{nr_i c_j}} = \frac{n(p_{ij} - r_i c_j)}{\sqrt{n}\sqrt{r_i c_j}} = \sqrt{n} \cdot s_{ij}. \tag{6}$$

Matriisimerkinnöin standardoitu $(I \times J)$-matriisi $\mathbf{S}$ saadaan seuraavasti

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-1/2}, \tag{7}$$

jossa $\mathbf{r}$ ja $\mathbf{c}$ ovat $(I \times 1)$- ja $(J \times 1)$-vektoreita, $\mathbf{D}_r = \text{diag}\{r_1, r_2, \ldots, r_I\}$ ja $\mathbf{D}_c = \text{diag}\{c_1, c_2, \ldots, c_J\}$ ovat $(I \times I)$- ja $(J \times J)$-matriiseja tässä järjestyksessä.

Standardoitujen jäännösten avulla voidaan laskea kokonaisinertia $\mathcal{I}$, joka kuvaa ristiintaulukon kokonaisvarianssia. Kokonaisinertia saadaan standardoitujen jäännösten neliöiden summana $\sum_i \sum_j s_{ij}^2$. Havaitaan, että kokonaisinertia saadaan myös χ^2–testisuureen avulla seuraavasti

$$\mathcal{I} = \chi^2/n, \tag{8}$$

ks. kaavat (1) ja (6). Taulukon riippuvuusrakenteen tarkastelu kokonaisinertian suhteen korrespondenssianalyysilla on mielekästä ainoastaan, jos χ^2–testin mukaan taulukossa on riippuvuutta. Korrespondenssianalyysin tavoitteena on esittää mahdollisimman suuri osuus kokonaisinertiasta alkuperäistä aineistoa vähäisemmillä dimensioilla.

3.3 Kokonaisinertian hajottaminen

Matriisin $\mathbf{S}$ assosiaatiorakenne saadaan esille singulaariarvohajotelman avulla

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{9}$$

jossa $\mathbf{\Sigma}$ on $(J \times J)$-diagonaalimatriisi siten, että diagonaalilla ovat singulaariarvot alenevassa järjestyksessä: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, missä r on matriisin $\mathbf{S}$ aste. Edellä $(I \times J)$-matriisin $\mathbf{U}$ sarakkeita kutsutaan vasemmanpuoleisiksi singulaarivektoreiksi ja $(J \times J)$-matriisin $\mathbf{V}$ sarakkeita vastaavasti oikeanpuoleisiksi singulaarivektoreiksi. Matriisit $\mathbf{U}$ ja $\mathbf{V}$ ovat ortogonaalisia eli $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.

Singulaariarvohajotelman todistus (Rao, 1973, s. 42-43):

Asetetaan, että $\mathbf{U}_i, i = 1, \ldots, r$, kuvaavat ortonormaaleja ominaisarvovektoreita, jotka vastaavat matriisin $\mathbf{S}\mathbf{S}^T$ nollasta eroavia ominaisarvoja $\sigma_i^2, i = 1, \ldots, r$. Edelleen asetetaan $\mathbf{V}_i = \sigma_i^{-1}\mathbf{S}^T\mathbf{U}_i$, jolloin $\mathbf{V}_i, i = 1, \ldots, r$ ovat myös ortonormaaleja. Oletetaan, että $\mathbf{U}_{r+1}, \ldots, \mathbf{U}_J$ ovat sellaisia vektoreita, että $\mathbf{U}_1, \ldots, \mathbf{U}_J$ on täysi ortonormaali joukko vektoreita eli $\mathbf{U}_1\mathbf{U}_1^T + \cdots + \mathbf{U}_J\mathbf{U}_J^T = \mathbf{I}$. Tällöin

$$\begin{aligned}
\mathbf{S} &= (\mathbf{U}_1\mathbf{U}_1^T + \cdots + \mathbf{U}_J\mathbf{U}_J^T)\mathbf{S} \\
&= (\mathbf{U}_1\mathbf{U}_1^T + \cdots + \mathbf{U}_r\mathbf{U}_r^T)\mathbf{S}, \ \textit{koska}\ \mathbf{U}_i^T\mathbf{S} = 0\ \textit{kaikille}\ i > r, \\
&= (\sigma_1\mathbf{U}_1\mathbf{V}_1^T + \cdots + \sigma_r\mathbf{U}_r\mathbf{V}_r^T) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,
\end{aligned}$$

jossa $\mathbf{U} = (\mathbf{U}_1 \vdots \cdots \vdots \mathbf{U}_r)$ ja $\mathbf{V} = (\mathbf{V}_1 \vdots \cdots \vdots \mathbf{V}_r)$.

Edellä olevan singulaariarvohajotelman (9) avulla voidaan kirjoittaa

$$\mathbf{S}^T\mathbf{S} = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \tag{10}$$

$$\mathbf{S}\mathbf{S}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T. \tag{11}$$

Huomaa, että oikeanpuoleiset singulaarivektorit matriisin $\mathbf{S}$ kaavassa (9) vastaavat $(J \times J)$-matriisin $\mathbf{S}^T\mathbf{S}$ ominaisarvovektoreita ja vastaavasti vasemmanpuoleiset singulaarivektorit vastaavat ominaisarvovektoreita $(I \times I)$-matriisissa $\mathbf{S}\mathbf{S}^T$. Neliöidyt singulaariarvot matriisissa $\mathbf{\Sigma}^2 = \text{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2, 0, \ldots, 0\}$ ovat samat kuin *ominaisarvot* $\lambda_1, \ldots, \lambda_r$ matriissa $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_r, 0, \ldots, 0\}$. Korrespondenssianalyysissa ominaisarvoja nimitetään pääinertioiksi ja ne summautuvat kokonaisinertiaksi (todistus liitteessä A.2)

$$\sum_i \sum_j s_{ij}^2 = \text{trace}(\mathbf{S}\mathbf{S}^T) = \text{trace}(\mathbf{S}^T\mathbf{S}) = \text{trace}(\mathbf{\Lambda}) = \sum_{i=1}^{r} \lambda_i. \tag{12}$$

Pääkomponenttianalyysin tapaan korrespondenssianalyysissa valitaan dimensioista tulkittavaksi ne, joiden pääinertiat (tai ominaisarvot) ovat suuremmat kuin keskimääräinen dimensioiden inertia. Toinen vaihtoehto on valita ne dimensiot, joille on sovelluskohteen mukaan järkevä tulkinta.

3.4 Graafinen esitys koordinaattimuuttujien avulla

Korrespondenssianalyysin tuloksia tulkitaan käyttäen karttaa, joka useimmiten kuvaa kaksi tärkeintä eli suurimman inertian (λ_1,λ_2) omaavaa dimensiota. Kartta piirretään laskemalla pistejoukolle, joka muodostuu sarake- tai rivivektoreista muodostetuista pisteistä, parhaat mahdolliset aliavaruudet. Nämä aliavaruudet kulkevat pistejoukon sentroidin kautta, jolla tarkoitetaan avaruuden geometrista keskipistettä. Parhaat eli optimaaliset aliavaruudet kulkevat aina pistejoukon sentroidin kautta (Greenacre, 1984, s. 44-45).

Sentroidin kautta kulkevan aliavaruuden määrittämiseen käytetään singulaariarvohajotelmaa, kaava (9), ja sen avulla saatuja ominaisarvoja, kaavat (10) ja (11). Suurimman ominaisvektorin suuntainen sentroidin kautta kulkeva suora määrää ensimmäisen pistejoukkoon liittyvän optimaalisen aliavaruuden. Kartan piirtämiseen tarvittavat koordinaatit ovat tälle aliavaruudelle projisoitujen pisteiden etäisyyksiä sentroidista. Seuraava optimaalinen aliavaruus määrittyy samalla tavalla kuin ensimmäinen, mutta se on kohtisuorassa ensimmäistä aliavaruutta kohden. Myös koordinaatit määrittyvät samalla tavalla. Edellisessä kappaleessa esitettyjen tulosten avulla voidaan laskea tarvittavat koordinaatit, joiden perusteella karttoja piirretään:

$$\begin{aligned}
&\text{rivien pääkoordinaatit:} && \mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Sigma}, && (13)\\
&\text{sarakkeiden pääkoordinaatit:} && \mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\boldsymbol{\Sigma}, && (14)\\
&\text{standardoidut rivikoordinaatit:} && \mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U}, && (15)\\
&\text{standardoidut sarakekoordinaatit:} && \mathbf{B} = \mathbf{D}_c^{-1/2}\mathbf{V}. && (16)
\end{aligned}$$

Edelliset tulokset on esitetty teoksessa Greenacre & Blasius (2006, s. 14) ja todistettu hieman eri merkinnöin kirjassa Greenacre (1984, s. 87-89). Kartan tekemiseksi matriisien $\mathbf{F}$, $\mathbf{G}$, $\mathbf{A}$ ja $\mathbf{B}$ avulla valitaan niiden ensimmäiset kaksi sarakevektoria siten, että käyttämällä: a) $\mathbf{F}$ ja $\mathbf{G}$ matriiseja saadaan muodostettua symmetrinen kartta, b) $\mathbf{A}$ ja $\mathbf{G}$ matriiseja saadaan epäsymmetrinen kartta sarakkeista ja c) $\mathbf{F}$ ja $\mathbf{B}$ matriiseja saadaan epäsymmetrinen kartta riveistä. Ensimmäiset kaksi sarakevektoria valitaan, koska ne edustavat kahta tärkeintä dimensiota. Kolmiulotteiseen kuvaan valittaisiin lisäksi kolmas sarakevektori. Seuraavan esimerkin symmetrisen kartan piirtäminen `R`-ohjelmistolla on esitetty vaiheittain liitteessä A.5.

3.5 Esimerkki

Johdatuksena korrespondenssianalyysiin esitän Greenacren ja Blasiuksen kirjassa esitetyn esimerkin, joka perustuu vuonna 2003 julkaistuun tutkimukseen (ISSP, 2003). Siinä verrattiin erilaisten tekijöiden vaikutusta kansalliseen identiteettiin. Esimerkissä tutkittiin viiden maan kansalaisten keskuudessa mielipidettä kysymykseen: "Jos maani menestyy kansainvälisessä urheilussa, tunnen ylpeyttä olla maani kansalainen." Vastausvaihtoehtoja oli viisi: 1) vahvasti samaa mieltä, 2) samaa mieltä, 3) ei kumpaakaan, 4) eri mieltä ja 5) vahvasti eri mieltä. Maita, joita esimerkissä verrattiin, oli viisi: Iso-Britannia (UK), Yhdysvallat (USA), Venäjä, Espanja sekä

Taulukko 2: Muuttujien maa ja ylpeys ristiintaulukko.

	Iso-Britannia	Yhdysvallat	Venäjä	Espanja	Ranska	Yhteensä
vahvasti samaa mieltä	230	400	**1010**	**201**	365	2206
samaa mieltä	329	471	530	639	478	2447
ei kumpaakaan	177	237	141	208	305	1068
eri mieltä	34	28	21	72	50	205
vahvasti eri mieltä	6	12	11	14	**97**	140
yhteensä	776	1148	1713	1134	1295	6066

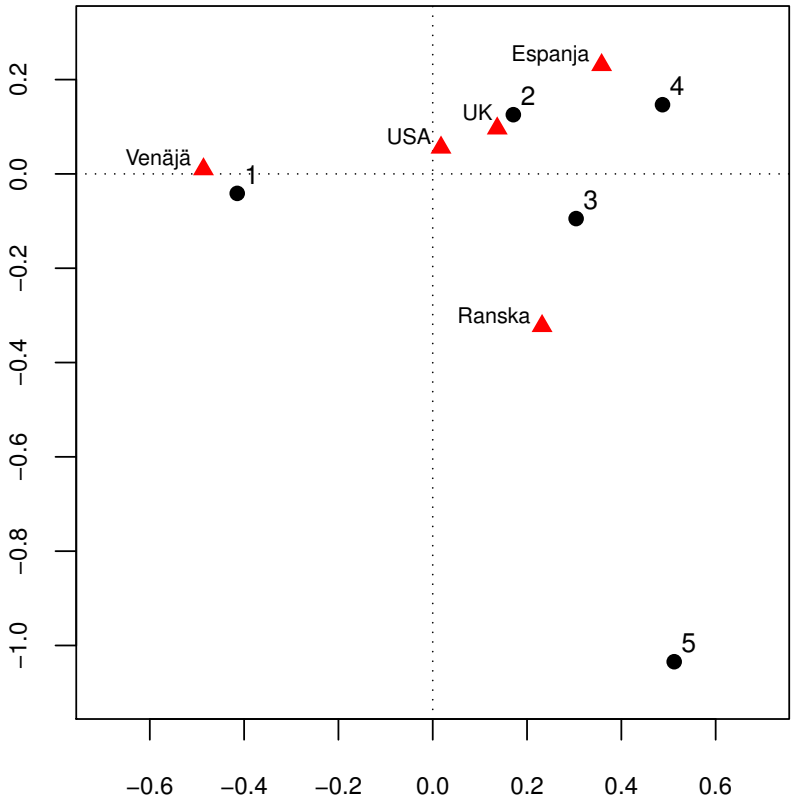
Ranska. Yksinkertaistuksen vuoksi aineistossa ei ole mukana henkilöitä, joiden vastauksista on puuttunut tietoa. Tässä luvussa esitetyt kuvat ja analyysi on toteutettu `R`-ohjelmiston `ca`-kirjaston `ca`-funktiolla (Nenadic & Greenacre, 2007). `R`-koodi on esitetty liitteessä A.5.

Taulukossa 2 on esitetty eri vastausvaihtoehtojen frekvenssit maittain sekä maa- ja vastauskohtaiset frekvenssit eli marginaalijakaumat. Taulukosta huomataan, että eniten vastauksia on tullut kahteen ensimmäiseen luokkaan: ”vahvasti samaa mieltä” ja ”samaa mieltä”. Muodostamalla taulukko vastausten prosentuaalisista osuuksista maittain voidaan helpommin vertailla eri vastausvaihtoehtojen osuuksia eri maiden välillä. Prosentuaaliset osuudet on esitetty taulukossa 3. Taulukosta 3 huomataan, että Venäjällä vastaukset ovat painottuneet selkeästi enemmän ensimmäiseen luokkaan verrattuna muihin maihin, kun taas Ranskassa ja Espanjassa on enemmän kielteisiä vastauksia kahdessa viimeisessä luokassa verrattuna muihin. Tutkittaessa tilastollisesti muuttujien välisiä riippuvuuksia huomataan, että muuttujat ovat riippuvia ($\chi^2 = 879.3$, p-arvo < 0.01, $df = 16$).

Suoritettaessa korrespondenssianalyysi esimerkkiaineistolle saadaan graafinen kuvaus muuttujien välisestä assosiaatiosta. Kuvassa 2 on esitetty kaksiulotteinen symmetrinen kartta, joka on piirretty rivi- ja sarakepääkoordinaattien perusteella (kaavat 13 ja 14). Maiden ja vastausten välisen assosiaation täydelliseen kuvaamiseen tarvittaisiin neliulotteinen kuva (5×5)-ristiintaulukon tapauksessa. Tämä johtuu siitä, että marginaalien ollessa kiinnitettynä sama frekvenssitaulukko voidaan muodostaa, jos tunnetaan neljä viidestä luokan arvosta. Korrespondenssianalyysin ta-

Taulukko 3: Maan ja ylpeyden ristiintaulukon prosentuaaliset osuudet.

	Iso-Britannia	Yhdysvallat	Venäjä	Espanja	Ranska	maittain
vahvasti samaa mieltä	29.6	34.8	**59.0**	**17.7**	28.2	36.4
samaa mieltä	42.4	41.0	30.9	56.4	36.9	40.3
ei kumpaakaan	22.8	20.6	8.2	18.3	23.6	17.6
eri mieltä	4.4	2.4	1.2	6.5	3.9	3.4
vahvasti eri mieltä	0.8	1.1	0.6	1.2	**7.5**	2.3
yhteensä	100.0	100.0	100.0	100.0	100.0	100.0



Kuva 2: Symmetrinen kartta. Kartan piirtämiseen on käytetty taulukon 2 rivien ja sarakkeiden pääkoordinaatteja.

voitteena on esittää mahdollisimman suuri osa aineiston sisältämästä vaihtelusta, inertiasta, mahdollisimman vähäisillä dimensioilla, yleensä kaksiulotteisena.

Tämän aineiston tapauksessa korrespondenssianalyysin tavoite toteutuu hyvin, inertiasta 95.6 % voidaan esittää kaksiulotteisen kartan avulla. Kuten edellä mainittiin, tämän aineiston täydelliseen esittämiseen tarvittaisiin neliulotteinen kuva, lopun inertian esittäminen tapahtuisi lisäämällä kuvaan kaksi ulottuvuutta. Neliulotteisen kuvan tulkinta ja piirtäminen on kuitenkin käytännössä mahdotonta ja kolmiulotteisenkin vaikeaa, joten varsinkin nyt, kun suurin osa aineiston inertiasta voidaan esittää kahden dimension avulla, ei kolmiulotteisen kuvan piirtäminen ole mielekästä.

Tässä tapauksessa voidaan selittämättä jääneen inertian osuus tulkita olevan epäoleellista, koska osuus on pieni. Korrespondenssianalyysin tavoitteena on myös selkeyttää aineiston rakennetta ja esittää se mahdollisimman yksinkertaisesti. Korrespondenssianalyysissa kartan akselit ovat ortogonaalisia keskenään, ne selittävät yhdessä kartan selittämän osuuden kokonaisvaihtelusta. Tässä tapauksessa ensimmäinen, vaaka-akseli selittää 72.2 % ja toinen, pystyakseli 23.4 % vaihtelusta (Kuva 2). Kuvassa origo kuvaa keskiarvoista kuvitteellista ”maata”, mitä lähempänä maa on origoa sitä keskimääräisemmät vastaukset maalla on.

Kartassa vaakasuuntaisen akselin voidaan tulkita kuvaavan ylpeyttä, jota tietyn maan kansalaiset tuntevat maansa menestyessä kansainvälisessä urheilussa. Tätä tul-

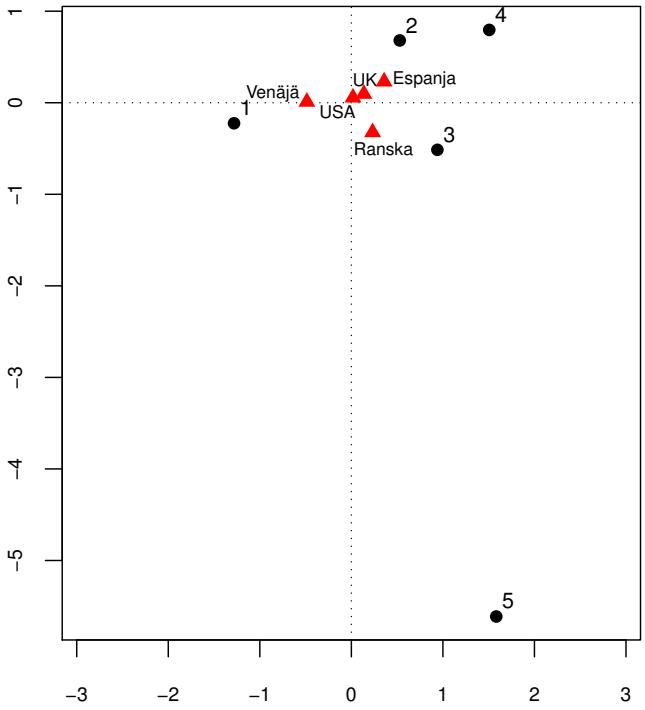
kintaa voidaan perustella sillä, että kuvassa 2 vastauskategoriat ovat järjestyksessä vasemmalta alkaen ”vahvasti samaa mieltä” loppuen oikealle ”vahvasti eri mieltä”. Vertaamalla maiden sijaintia vaakasuuntaisella akselilla voidaan tehdä sama päätelmä kuin taulukoiden 2 ja 3 perusteella: Venäjällä ollaan keskimäärin ylpeämpiä maan menestymisestä kansainvälisessä urheilussa verrattuna muihin maihin, kun taas Ranskassa ja Espanjassa tunnetaan vähiten ylpeyttä menestyksestä.

Toinen dimensio kuvaa lähinnä ”vahvasti eri mieltä” -vastauksen erilaisuutta maiden välillä verrattuna muihin vastauksiin. Ranskassa on enemmän ”vahvasti eri mieltä” -vastauksia kuin muissa maissa. Se näkyy siten, että Ranska on lähimpänä tätä vastausvaihtoehtoa myös kartassa. Tämän jälkimmäisen akselin selitysosuuden suuruutta kokonaisvaihtelusta voidaan selittää sillä, että erot Ranskan (7.5 %) ja muiden maiden (n. 1 %) vastauksissa tähän kategoriaan ovat niin suuret (Taulukko 3).

Edellä tarkastellussa symmetrisessä kartassa ainoa tarkka piste on kartan origo. Tämän takia ei voida tehdä päätelmiä siitä, kuinka kaukana maat ovat toisistaan. Maiden järjestys voidaan kuitenkin todeta: Venäjällä ollaan ylpeämpiä kuin Yhdysvalloissa. Voidaan myös todeta mihin suuntaan maa eroaa keskiarvoisesta maasta.

Haluttaessa tehdä päätelmiä myös maiden välisistä eroista täytyy piirtää epäsymmetrinen kartta. Epäsymmetrinen kartta piirretään yleensä siten, että kuvailevan muuttujan arvoiksi valitaan standardoidut koordinaatit, tämän aineiston tapauksessa rivien standardoidut koordinaatit, jotka saadaan kaavasta (15). Korrespondenssianalyysi suoritetaan esimerkin tapauksessa standardoitujen rivikoordinaattien ja sarakkeiden pääkoordinaattien perusteella. Kuvassa 3 on esitetty epäsymmetrinen kartta, joka on laskettu samasta aineistosta. Kartan voi huomata olevan samankaltainen kuin symmetrisessäkin tapauksessa lukuunottamatta muutosta akseleiden skaalassa ja vastausvaihtoehtojen sijainnissa.

Kuvassa 3 esitetyssä kartassa vastausvaihtoehtojen esittämät pisteet kuvastavat nyt kuvitteellisia maita, joissa tietyn vastausvaihtoehdon on valinnut 100 % kaikista vastaajista. Tämän kartan perusteella voidaan nyt nähdä, kuinka kaukana tietty maa on tällaisesta kuvitteellisesta 100 %:n maasta. Korrespondenssianalyysin tuloksia esitettäessä suositaan yleensä kuitenkin symmetrisen kartan käyttöä, koska epäsymmetrisessä kartassa standardoitujen koordinaattipisteiden käyttäminen ”työntää” pääkoordinaattien muodostamat pisteet lähelle origoa skaalauksesta johtuen. Kuvassa 3 tämä ei vielä ole suuri ongelma, mutta jos kartassa on useampia koordinaattipisteitä kartan selkeä esittäminen hankaloituu.



Kuva 3: Epäsymmetrinen kartta. Kartan piirtämiseen on käytetty taulukon 2 rivien standardoituja koordinaatteja ja sarakkeiden pääkoordinaatteja.

4 Moniulotteinen korrespondenssianalyysi

Moniulotteiseen korrespondenssianalyysiin liittyvät keskeisesti käsitteet indikaattorimatriisi sekä Burtin matriisi. Molemmilla matriiseilla voidaan esittää kontingenssitaulun useiden muuttujien välinen informaatio yksinkertaisesti ja yksikäsitteisesti. Näiden matriisien teoriaa esitetään tämän luvun alussa, koska niitä tarvitaan moniulotteisessa korrespondenssianalyysissa.

Matriisien esittelyn jälkeen käsitellään moniulotteiseen korrespondenssianalyysiin liittyvää teoriaa. Teoriaosuudessa näytetään korrespondenssimatriisin laskeminen sekä kerrotaan siihen liittyvästä sarakegeometriasta. Rivigeometria on läheisesti yhteydessä sarakegeometriaan, joten se on rajattu työn ulkopuolelle. Rivigeometriaa käsitellään kirjassa Greenacre (1984, s. 133-136), johon myös muu tämän luvun teoriasta perustuu. Teoria esitetään suhteuttaen sitä kaksiulotteiseen korrespondenssianalyysiin, jotta yhteys kaksi- ja moniulotteisen korrespondenssianalyysin välillä nähdään paremmin.

4.1 Monimuuttujaisten taulukoiden esittäminen

4.1.1 Indikaattorimatriisi

Kaksiulotteinen indikaattorimatriisi voidaan muodostaa kontingenssitaulun perusteella siten, että jokainen havainto esitetään omalla rivillään ja kontingenssitaulun

kahden muuttujan tasot esitetään matriisin sarakkeina. Sarakkeiden järjestys on siten, että kontingenssitaulun rivimuuttujan tasot esitetään ensin ja niiden perään taulun sarakemuuttujan tasot. Indikaattorimatriisi

$$\mathbf{Z} \equiv [\,\mathbf{Z}_1 \quad \mathbf{Z}_2\,], \tag{17}$$

joka on $(n \times (I + J))$-matriisi. Edellä $(n \times I)$-matriisi $\mathbf{Z}_1$ ja $(n \times J)$-matriisi $\mathbf{Z}_2$ sisältävät ristiintaulukon $\mathbf{N}$ muuttujien arvot sarakkeina.

Edellisen luvun esimerkissä esitetty taulukko 2 voidaan esittää indikaattorimatriisina, jossa olisi N riviä (=6066) ja $I + J$ saraketta (=10). Jokaisella edellisen luvun esimerkin indikaattorimatriisin rivillä on kahdeksan nollaa ja kaksi ykköstä. Esimerkiksi rivi, joka kuvaa ”samaa mieltä” - sekä ”Venäjä”-havaintoa olisi: [0, 1, 0, 0, 0; 0, 0, 1, 0, 0]. Näitä rivejä indikaattorimatriisissa olisi 530 kappaletta vastaten jokaista yksittäistä vastauskombinaatioluokkaan kuuluvaa havaintoa. Muutettaessa indikaattorimatriisi kontingenssitauluksi menetetään informaatiota, mutta ainoastaan informaatio yksittäisen henkilön vastauksesta.

Moniulotteinen indikaattorimatriisi muodostetaan samalla tavalla kuin kaksiulotteinenkin matriisi, sarakkeiden määrä vain kasvaa vastaamaan muuttujien ja niiden tasojen lukumäärää. Se on muotoa

$$\mathbf{Z} \equiv [\,\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \cdots \quad \mathbf{Z}_Q\,], \tag{18}$$

missä $\mathbf{Z}_K$, $K = 1, \ldots, Q$, ovat moniulotteisen ristiintaulukon muuttujien arvot sarakkeina. Moniulotteisessa indikaattorimatriisissa saatetaan menettää joidenkin muuttujien välinen mahdollinen assosiaatio.

4.1.2 Burtin matriisi

Toinen vaihtoehto indikaattorimatriisin sijasta on käyttää Burtin matriisia (Burt, 1950). Burtin matriisin avulla voidaan ottaa huomioon muuttujien väliset mahdolliset assosiaatiot. Matriisi on rakenteeltaan lohkomatriisi, jolloin jokaisessa lohkossa on yksi kaikista kahden muuttujan välisistä assosiaatiorakenteista. Burtin matriisin esitys voidaan verrannollistaa indikaattorimatriisin esitykseen. Kaksiulotteisessa tapauksessa Burtin matriisi on ainoastaan:

$$\mathbf{Z}^T\mathbf{Z} \equiv \begin{pmatrix} \mathbf{Z}_1^T\mathbf{Z}_1 & \mathbf{Z}_1^T\mathbf{Z}_2 \\ \mathbf{Z}_2^T\mathbf{Z}_1 & \mathbf{Z}_2^T\mathbf{Z}_2 \end{pmatrix} \equiv \begin{pmatrix} \mathbf{ID}_r & \mathbf{N} \\ \mathbf{N} & \mathbf{ID}_c \end{pmatrix},$$

missä $\mathbf{Z}$ on aiemmin esitetty indikaattorimatriisi ja $\mathbf{Z}^T\mathbf{Z}$ on $((I + J) \times (I + J))$-kokoa oleva Burtin matriisi. Matriisit $\mathbf{Z}_1$ ja $\mathbf{Z}_2$ sisältävät ensimmäisen ja toisen muuttujan tiedot tässä järjestyksessä. Matriisit $\mathbf{D}_r$ ja $\mathbf{D}_c$ ovat kontingenssitaulun $\mathbf{N}$ rivi- ja sarakemassat sisältävät diagonaalimatriisit. Monimuuttujaisessa tapauksessa Burtin matriisi on muotoa

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1^T\mathbf{Z}_1 & \mathbf{Z}_1^T\mathbf{Z}_2 & \cdots & \mathbf{Z}_1^T\mathbf{Z}_Q \\ \mathbf{Z}_2^T\mathbf{Z}_1 & \mathbf{Z}_2^T\mathbf{Z}_2 & \cdots & \mathbf{Z}_2^T\mathbf{Z}_Q \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}_Q^T\mathbf{Z}_1 & \mathbf{Z}_Q^T\mathbf{Z}_2 & \cdots & \mathbf{Z}_Q^T\mathbf{Z}_Q \end{pmatrix},$$

missä $\mathbf{Z}$ on aiemmin esitetty indikaattorimatriisi ja $\mathbf{Z}^T\mathbf{Z}$ on $((I + J + \cdots + Q) \times (I + J + \cdots + Q))$-kokoa oleva Burtin matriisi. Lohkomatriisin $\mathbf{Z}_K^T\mathbf{Z}_{K'}$ matriisi $\mathbf{Z}_K$ ja $\mathbf{Z}_{K'}$, $K, K' = 1, \ldots, Q$, ovat indikaattorimatriisin muuttujien sarakkeita.

4.2 Moniulotteisen korrespondenssianalyysin teoriaa

4.2.1 Sarakegeometria indikaattorimatriisin korrespondenssianalyysissa

Olkoon $\mathbf{N}$ edelleen kahden muuttujan välinen $(I \times J)$-kontingenssitaulu ja $\mathbf{Z}$ siihen liittyvä indikaattorimatriisi, jossa on n riviä ja $I + J$ saraketta. Kaavan (17) perusteella voidaan todeta, että $\mathbf{N} = \mathbf{Z}_1^T\mathbf{Z}_2$. Lisäksi huomataan, että matriisin $\mathbf{Z}$ jokainen rivimassa on $1/n$. Sarakemassat saadaan jakamalla taulukon $\mathbf{N}$ rivi- ja sarakemassat kahdella. Matemaattisemmin rivi- ja sarakemassat ovat

$$\mathbf{r}^Z = \frac{1}{n} \cdot 1_{(n\times 1)}, \tag{19}$$

$$\mathbf{c}^Z = \frac{1}{2}\begin{bmatrix}\mathbf{r}\\ \mathbf{c}\end{bmatrix}, \tag{20}$$

missä $\mathbf{r}^Z$ sisältää indikaattorimatriisin $\mathbf{Z}$ rivimassat ja $\mathbf{c}^Z$ sarakemassat. Matriisin $\mathbf{Z}$ sarakemassojen laskemiseen käytetään taulukon $\mathbf{N}$ rivi- ja sarakemassoja eli vektoreita $\mathbf{r}$ sekä $\mathbf{c}$.

Tällöin indikaattorimatriisin korrespondenssimatriisi sekä rivi- ja sarakemassojen diagonaalimatriisit ovat:

$$\mathbf{P}^Z = \frac{1}{2n}\mathbf{Z}, \tag{21}$$

$$\mathbf{D}_r^Z = \frac{1}{n}\mathbf{I}, \tag{22}$$

$$\mathbf{D}_c^Z = \frac{1}{2}\begin{bmatrix}\mathbf{D_r} & 0\\ 0 & \mathbf{D_c}\end{bmatrix}, \tag{23}$$

missä $\mathbf{P}^Z$ on indikaattorimatriisin $\mathbf{Z}$ korrespondenssimatriisi, $\mathbf{D}_r^Z$ sekä $\mathbf{D}_c^Z$ rivi- ja sarakemassojen diagonaalimatriisit. Matriisin $\mathbf{Z}$ sarakemassojen diagonaalimatriisin laskemiseen käytetään taulukon $\mathbf{N}$ rivi- ja sarakemassojen diagonaalimatriiseja $\mathbf{D}_r$ sekä $\mathbf{D}_c$.

Korrespondenssianalyysin yhteydessä indikaattorimatriisin standardoidut koordinaatit saadaan ratkaisemalla matriisit $\mathbf{\Gamma}_1^Z$ sekä $\mathbf{\Gamma}_2^Z$ seuraavista ominaisarvoyhtälöistä (Greenacre, 1984, s. 131),

$$\mathbf{D}_c^{-1}\mathbf{P}^T\mathbf{D}_r^{-1}\mathbf{P}\mathbf{\Gamma}_2^Z = \mathbf{\Gamma}_2^Z(2\mathbf{D}_\lambda^Z - \mathbf{I})(2\mathbf{D}_\lambda^Z - \mathbf{I}), \tag{24}$$

$$\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}\mathbf{P}^T\mathbf{\Gamma}_1^Z = \mathbf{\Gamma}_1^Z(2\mathbf{D}_\lambda^Z - \mathbf{I})(2\mathbf{D}_\lambda^Z - \mathbf{I}). \tag{25}$$

Edellä esitetyissä yhtälöissä $\mathbf{D_c}$, $\mathbf{D_r}$ ja $\mathbf{P}$ ovat samoja matriiseja kuin luvussa 3 esitetyt matriisit. Matriisit $\mathbf{\Gamma}_1^Z$ sekä $\mathbf{\Gamma}_2^Z$ ovat indikaattorimatriisin $\mathbf{Z}$ standardoitujen koordinaattien matriiseja. $\mathbf{D}_\lambda^Z$ on matriisin $\mathbf{Z}$ ominaisarvojen diagonaalimatriisi.

Kontingenssitaulun $\mathbf{N}$ ja indikaattorimatriisin $\mathbf{Z}$ analyyseissa saatavien ominaisarvojen yhteys on seuraava:

$$\lambda = (2\lambda^I - 1)^2, \tag{26}$$

missä λ on kontingenssitaulun ja λ^I indikaattorimatriisin analyysista saatava ominaisarvo. Kaavan (26) mukaisesta ominaisarvojen skaalautumisesta johtuen indikaattorimatriisia analysoitaessa ollaan kiinnostuneita niistä dimensioista, joiden ominaisarvot ovat suurempia kuin 1/2. Valittaessa dimensiot edellä mainitulla tavalla saadaan täsmälleen samat ominaisarvot kuin ristiintaulukon $\mathbf{N}$ korrespondenssianalyysissa kaksiulotteisessa tilanteessa.

4.2.2 Keinotekoiset dimensiot ja dimensioiden selitysosuuksien korjaaminen

Useamman kuin kahden muuttujan tapauksessa indikaattorimatriisia analysoitaessa ominaisarvoista valitaan ne, jotka ovat suurempia kuin $1/K$, missä K on muuttujien lukumäärä. Dimensioiden määrä valitaan näin siksi, että indikaattorimatriisia muodostettaessa muuttujien tasojen muuntaminen sarakkeiksi synnyttää keinotekoisia dimensioita (*artificial dimensions*) (ks. luku 4.1.1). Käytännössä tämä tarkoittaa sitä, että yksi muuttuja ilmoitetaan useamman sarakkeen avulla, mikä luo uusia keinotekoisia dimensioita, jolloin tulosavaruus laajenee.

Tulosavaruuden laajetessa yksittäisten dimensioiden selitysosuus pienenee. Dimensioita kuvaavien ominaisarvojen selitysosuudet eivät kuvaa alkuperäiseen aineistoon liittyvää vaihtelua oikein verrattuna tilanteeseen, että keinotekoisia dimensioita ei olisi. Vaihtelun kuvaamisen parantamiseksi ominaisarvoja sekä dimensioiden selitysosuuksia korjataan lähemmäs todellista arvoa. Todellisina arvoina voidaan pitää JCA-menetelmällä saatuja arvoja (Greenacre, 1988). Ominaisarvojen korjaamiseen käytetään seuraavaa yhtälöä (Greenacre & Blasius, 2006, s. 67-68)

$$\lambda_I^c = \begin{cases} \left[\left(\frac{K}{K-1}\right)\left(\lambda_I - \frac{1}{K}\right)\right]^2, & \text{jos } \lambda_I > \frac{1}{K}, \\ 0, & \text{jos } \lambda_I \leq \frac{1}{K}, \end{cases}$$

missä λ_I^c on korjattu ominaisarvo, λ_I indikaattorimatriisin analyysista saatu ominaisarvo ja K muuttujien lukumäärä.

Selitysosuuden korjaamiseen on ehdotettu ainakin kahta tapaa. Ensimmäisessä tapauksessa dimensioon liittyvä ominaisarvo eli inertia jaetaan ominaisarvojen summalla (Benzécri, 1979). Tämä saattaa kuitenkin johtaa liian optimistiseen selitysosuuden arvioon. Paremman arvion selitysosuuden laskemiseksi on esittänyt Greenacre (1993, s. 144-145) ehdottamalla, että selitysosuus laskettaisiin vertaamalla korjattua inertiaa Burtin matriisin diagonaalin ulkopuolisten lohkojen keskiarvoiseen inertiaan. Tällä tavalla korjattuna selitysosuus on yleensä aliarvioitu verrattuna JCA-menetelmällä saatuun tulokseen, mutta on kuitenkin parempi kuin ilman korjausta. Keskiarvoinen inertia $\hat{I}$ lasketaan seuraavasti

$$\hat{I} = \frac{K}{K-1} \times \left(\sum_I \lambda_I^2 - \frac{L-K}{K^2} \right), \tag{27}$$

missä L on muuttujien tasojen yhteenlaskettu lukumäärä ja λ_I^2 Burtin matriisista saadut ominaisarvot (indikaattorimatriisin neliöidyt ominaisarvot). Tästä seuraa, että dimensioiden selitysosuudet τ_c ovat

$$\tau_c = \frac{\lambda^c}{\hat{I}} \tag{28}$$

4.2.3 Lisämuuttujat

Moniulotteisessa korrespondenssianalyysissa optimaaliset aliavaruudet eli dimensiot lasketaan aineiston muuttujien avulla. Jos aineistossa on muuttujia, joita ei haluta ottaa huomioon dimensioita määritettäessä, voidaan nämä muuttujat määritellä lisämuuttujiksi. Lisämuuttujat kuvataan analyysin tuloksena syntyvään karttaan, mutta ne eivät vaikuta kartan dimensioiden määrittämiseen.

Moniulotteisella korrespondenssianalyysilla ei pystytä analysoimaan jatkuvia muuttujia, ellei niitä kategorisoida jollain tavalla. Kategorisointi on usein kuitenkin vaikeaa, varsinkin jos muuttujien jakaumat ovat keskenään hyvin erilaisia kuten tässä aineistossa. Tällaisessa tapauksessa muuttujien määrittely lisämuuttujiksi mahdollistaa jatkuvien muuttujien käytön, ne eivät tosin osallistu kartan määrittämiseen, mutta ne voidaan kuitenkin esittää kartassa.

Kategorisille lisämuuttujille teoria ja laskukaavat on esitetty kirjassa Greenacre & Blasius (2006, s. 31-32; s. 70-74; s. 533-534). Jatkuvalle lisämuuttujalle pääkoordinaatit saadaan laskemalla dimensioiden ja lisämuuttujan väliset selitysasteet R^2 (*Squared correlation coefficient* tai *Coefficient of determination*) (Husson & Josse, 2014).

5 Aineiston analyysi

Aineistoon sovellettiin moniulotteista korrespondenssianalyysia siten, että eri taksonien yksilömääriä käytettiin ensin lisämuuttujina ja tämän jälkeen taksonien lukumäärät kategorisoitiin. Kategorisoinnin jälkeen analyysi suoritettiin ensin koko aineistolle ja tämän jälkeen jokaiselle aineiston keräämisvuodelle erikseen. Vuosikohtainen jako suoritettiin, koska haluttiin tutkia, pysyykö analyysin antama kuva aineistosta samanlaisena eri vuosina. Huomionarvoista on, että tulkinnat, joita tässä luvussa on esitetty, on tehty kahden suurimman selitysosuuden saaneen dimension perusteella ellei tekstissä toisin mainita.

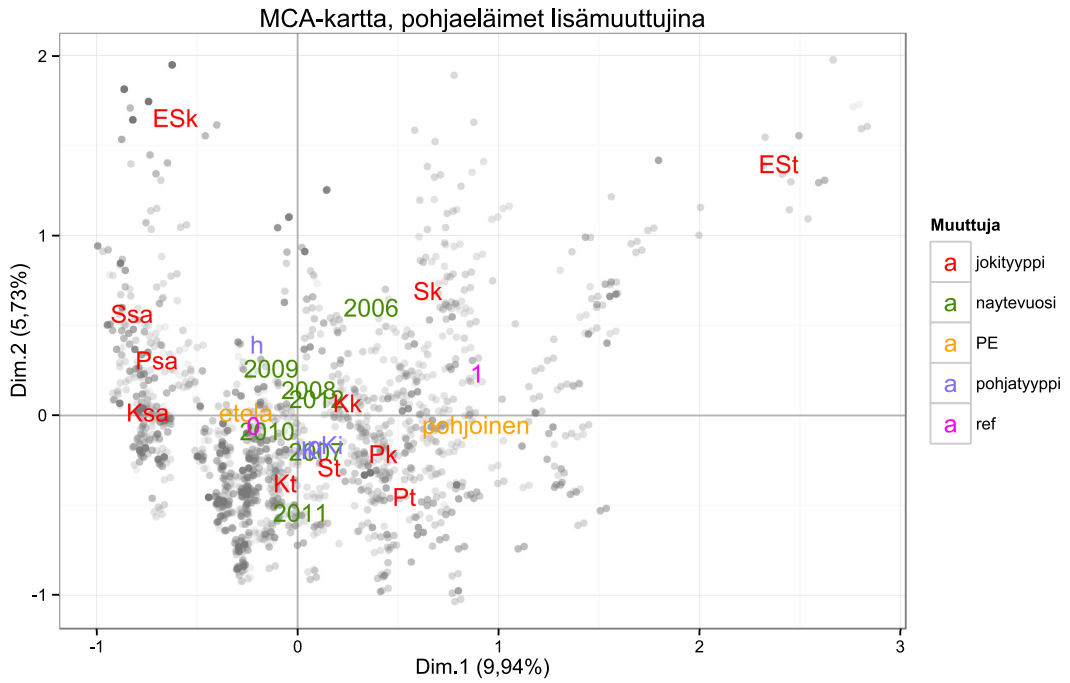
Analyysista saatavien karttojen tulkintaidea ei eroa aiemmin kaksiulotteisen korrespondenssianalyysin yhteydessä esitetystä symmetrisen kartan tulkinnasta (Luku 3.5). Kartan origo on ainoa tarkka piste kartalla, jolloin kartan avulla voidaan tulkita vain muuttujien suhteellisia sijoittumisia toisiinsa sekä dimensioakseleihin nähden. Etäisyydet eivät siis ole tarkkoja symmetrisessä kartassa. Kartoissa esitetään myös havaintojen jakauma käyttämällä harmaita pisteitä. Mitä tummempi piste on, sitä useampi havainto on samassa kohdassa. Havaintopiste kuvaa siis kyseiseen havaintoon liittyvää riviprofiilia.

Analyysi suoritettiin `R`-ohjelmistolla käyttäen `FactoMineR`-pakettia (Lê et al., 2008) ja sen funktiota `MCA`. Muita mahdollisia `R`-paketteja funktioineen ovat ainakin `MASS` (`mca`), `ade4` (`dudi.acm`), `ca` (`mjca`) sekä `homals` (`homals`). `FactoMineR`-pakettia käytettiin, koska se tarjosi laajan valikoiman erilaisia työkaluja aineiston analysointiin sekä selkeät käyttöohjeet. `R`-koodi aineiston analysoimiseksi on esitetty liitteessä A.6.

5.1 Taksonit lisämuuttujina

Tämän luvun analyysissa on mukana kaikki taulukossa 1 esitellyt muuttujat siten, että taksonit on asetettu lisämuuttujiksi. Taksonien lukumäärät haluttiin huomioida lisämuuttujiksi asettamisen kautta, koska niiden esiintymistä aineistossa olisi ollut vaikeaa kategorisoida siten, että kategorisointi ei olisi varmuudella ollut vaikuttamatta analyysin tulokseen. Aineistoon toteutettiin moniulotteinen korrespondenssianalyysi.

Kuvassa 4 on esitetty selitysosuudeltaan kahdesta suurimmasta dimensiosta muodostettu kartta. Siitä voidaan päätellä, että ensimmäinen dimensio, joka kuvaa 9,9 % kokonaisinertiasta, voisi kuvata havaintopaikan maantieteellistä sijaintia Suomessa pohjois–etelä-suunnassa (muuttuja `PE`) ja myös havaintopaikkojen kuulumista referenssi- tai impaktijokiin (`ref`). Edellisten tulkintojen tueksi voidaan vielä katsoa savimaiden jokityyppien sijoittuminen kartan ensimmäisen dimension vasempaan äärilaitaan, sillä savimaiden joet ovat kaikki Etelä-Suomessa. Tämän lisäksi kaikki niistä kuuluvat impaktijokiin.



Kuva 4: MCA-kartta, joka on piirretty ainoastaan jokimuuttujien avulla. Kartassa on esitetty jokimuuttujat tasoineen sekä havainnot harmailla pisteillä. Dimensiot ja niiden selitysosuudet on merkitty kuvaan.

Kartasta voidaan myös nähdä, kuinka `pohjatyyppi` jakautuu kahteen ryhmään, hienoainekseen (h) sekä iso- tai pienkivipohjaisiin jokiin (iKi, pKi). Tämä tukee aiempia päätelmiä siitä, että kivipohjaiset havainnot voidaan tutkia samoin perustein, mutta hienoainespohjaisille täytyy määrittää omat tutkimusparametrit (Meissner et al., 2013). Vuosien sijoittuminen ensimmäisellä dimensiolla lähelle dimension nollapistettä samoin kuin `PE`-muuttujan etelä-taso sekä `ref`-muuttujan impakti(0)-taso kertovat siitä, että maantieteellisesti ja referenssi-impakti-suhteeltaan jokien valinta on painottunut Etelä-Suomen impaktijokiin. Eniten vuosina 2009 ja 2010 ja vähiten vuonna 2006.

Kartassa 4 taustalla esitetty havaintojen (indikaattorimatriisin tapauksessa rivien) erikoinen jakautuminen ”pylväsmäisesti” johtuu erityisesti `jokityyppi`-muuttujasta. Havaintojen jakaumaa tarkasteltaessa pystyttiin toteamaan, että `jokityyppi`-muuttujaan liittyvät havainnot muodostivat pisteparven, johon `PE`-, `ref`- sekä `pohjatyyppi`-muuttujan tasot toivat hajontaa ensimmäisen dimension suunnassa ja `naytevuosi`-muuttujan havainnot jakautuivat sen sijaan toisen dimension suunnassa (Kuvat 7-11, Liite A.3).

Toisen ja kolmannen dimension korjatut selitysosuudet ovat lähes yhtäsuuret niiden ollessa 5,7 % ja 5,1 %. Viidelle suurimman selitysosuuden omaavalle dimensiolle laskettiin korjatut selitysosuudet, joita voi verrata muiden analyysien selitysosuuksiin (Taulukko 4).

Taulukko 4: Viiden ensimmäisen dimension korjatut selitysosuudet eri aineistoille moniulotteista korrespondenssianalyysia käytettäessä. Luvut on pyöristetty kahden desimaalin tarkkuuteen.

koko aineisto	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
lisämuuttujat	9.94	5.73	5.06	4.42	4.14
kategorisoidut muuttujat	**23.49**	**13.46**	6.36	3.95	2.90
vuosittain					
2006	**22.06**	**16.80**	6.38	4.12	3.19
2007	**18.30**	**11.30**	7.34	3.80	3.60
2008	**18.58**	**15.26**	6.94	3.91	3.69
2009	**28.72**	**13.21**	4.83	4.08	3.31
2010	**17.06**	**14.47**	7.74	4.37	2.98
2011	**33.05**	9.71	6.30	4.45	2.88
2012	**19.86**	**15.95**	7.23	5.86	4.49

Analyysissa etsittiin myös mahdollisia indikaattorilajeja jokimuuttujien eri tasoille. Indikaattoritaksonien etsiminen suoritettiin laskemalla pohjaeläintaksonin ja jokimuuttujan tason välinen euklidinen etäisyys viiden suurimman selitysosuuden saaneen dimension muodostamassa avaruudessa. Käsiteltäessä pohjaeläintaksoneita lisämuuttujina niiden esiintymisrunsaus otetaan huomioon laskettaessa niiden koordinaatteja kartassa. Etäisyydet oli helppo laskea `R`-ohjelmistolla, koska `MCA`-funktio on laskenut koordinaatit valmiiksi sekä tallentanut ne tuloksena saatuun objektiin. Taulukossa 5 on esitetty jokaiselle jokimuuttujaluokalle lähimmät viisi pohjaeläintaksonia. Tulosten perusteella savimaiden jokityypeille (Psa, Ksa, Ssa) löytyi kaksi indikaattoritaksonia (Luku 2.2): *Unio-* sekä *Calopteryx spp.* -taksonit. *Anodonta piscinalis* esiintyy pienissä ja keskisuurissa savimaiden joissa. Muita selkeitä indikaattoritaksoneita ei löytynyt viiden lähimmän taksonin joukosta.

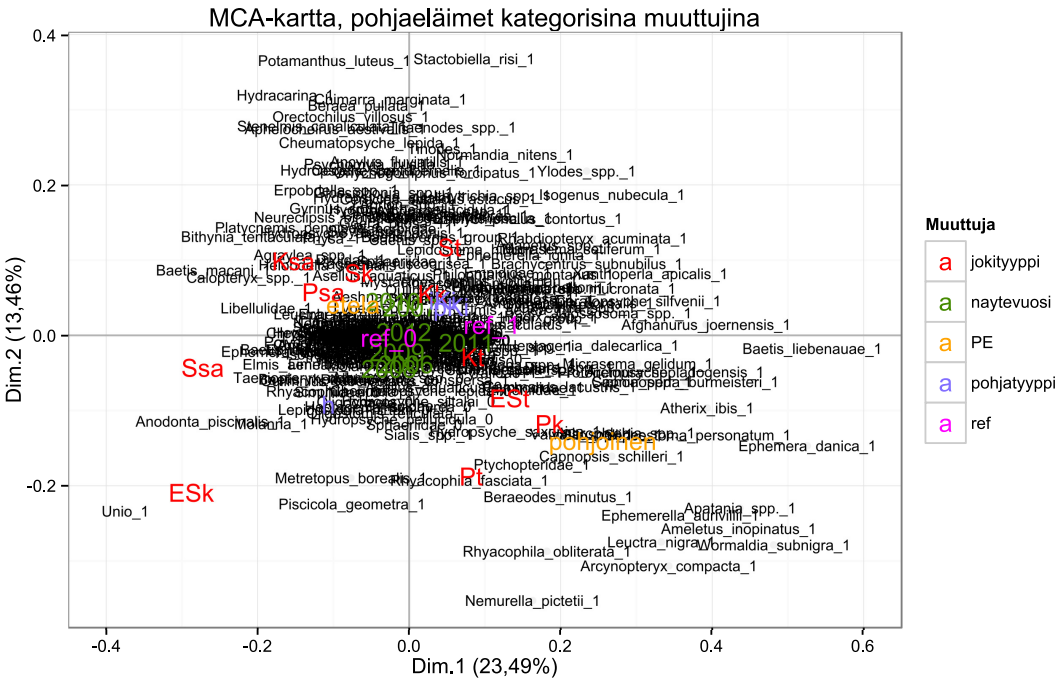
Taulukko 5: Moniulotteisen korrespondenssianalyysin antamien koordinaattien perusteella jokimuuttujien tasoja lähimpänä olevat pohjaeläintaksonit. Tekstissä tulkitut taksonit on korostettu lihavoimalla.(Joitakin taksonien nimiä on lyhennetty.)

Luokat					
ESk	*Kageronia fuscogrisea*	*Ephemera vulgata*	*Taeniopteryx nebulosa*	*Baetis vernus group*	*Oulimnius tubercula.*
ESt	*Ephemerella mucrona.*	*Arctopsyche ladogensis*	*Gyraulus spp.*	*Ceratopsyche silfvenii*	*Siphonoperla burmeis.*
Kk	*Afghanurus joernensis*	*Planorbiidae*	*Normandia nitens*	*Agraylea spp.*	*Philopotamus monta.*
Ksa	***Unio***	***Anodonta piscinalis***	*Molanna*	***Calopteryx spp.***	*Ancylus fluviatilis*
Kt	*Onychogomphus forci.*	*Rhyacophila nubila*	*Agapetus spp.*	*Bathyomphalus contor.*	*Isogenus nubecula*
Pk	*Baetis rhodani*	*Ephemerella aurivillii*	*Diura spp.*	*Limoniidae*	*Atherix ibis*
Psa	***Unio***	***Anodonta piscinalis***	***Calopteryx spp.***	*Gammarus lacustris*	*Ptychopteridae*
Pt	*Baetis rhodani*	*Elmis aenea*	*Oulimnius tubercula.*	*Rhyacophila nubila*	*Silo pallipes*
Sk	*Kageronia fuscogrisea*	*Ceratopogonidae*	*Cloeon spp.*	*Rhabdiopteryx acumi.*	*Aphelocheirus aestiva.*
Ssa	*Ephemera vulgata*	***Unio***	***Calopteryx spp.***	*Limnephilidae*	*Lype spp.*
St	*Beraea pullata*	*Arcynopteryx compac.*	*Afghanurus joernensis*	*Wormaldia subnigra*	*Aphelocheirus aestiva.*
Etelä	*Physa*	*Ancylus fluviatilis*	*Gammarus pulex*	*Dixiidae*	*Notidobia ciliaris*
Pohj.	*Diura spp.*	*Heptagenia dalecarlica*	*Ephemerella aurivillii*	*Arctopsyche ladogensis*	*Ameletus inopinatus*
h	*Lype spp.*	*Limnephilidae*	*Helobdella stagnalis*	*Ephemera vulgata*	*Limnius volckmari*
iKi	*Aphelocheirus aestiva.*	*Isogenus nubecula*	*Xanthoperla apicalis*	*Stactobiella risi*	*Glossosoma spp.*
pKi	*Normandia nitens*	*Habrophlebia spp.*	*Glossosoma spp.*	*Xanthoperla apicalis*	*Isogenus nubecula*
ref 0	*Platycnemis pennipes*	*Dytiscidae*	*Phryganea spp.*	*Libellulidae*	*Chrysomelidae*
ref 1	*Arcynopteryx compac.*	*Beraea pullata*	*Rhabdiopteryx acumi.*	*Glossiphonia spp.*	*Isoperla spp.*

5.2 Taksonit kategorisina muuttujina

Pohjaeläintaksonien ottaminen mukaan kategorisoituina muuttujina paransi selvästi kahden ensimmäisen dimension selitysosuuksia verrattuna luvun 5.1 analyysiin. Ensimmäisen dimension selitysosuus on 23,5 % ja toisen dimension 13,5 %. Seuraavien dimensioiden selitysasteet voi nähdä taulukosta 4. Täytyy huomioida, että analyysissa on nyt 154 muuttujaa enemmän, joten oletettavaa onkin, että selitysosuus nousee informaation määrän kasvaessa. Verrattuna kuvan 4 karttaan kuvan 5 kartan avulla voidaan siis tehokkaammin esittää aineiston riippuvuussuhteita ja aineistossa esiintyvää informaatiota tehtäessä tulkintaa ainoastaan selitysosuuksien perusteella. Tässä luvussa karttaan piirrettiin joki- ja pohjaeläinmuuttujat, mutta havaintojen jakauman kuvaaminen jätettiin pois. Tämä valinta tehtiin siksi, että erityisesti näillä kategorisoiduilla pohjaeläinmuuttujilla haluttiin etsiä mahdollisia indikaattoritaksoneita. Havaintojen jakauman sisältävä kartta (Kuva 12) on kuitenkin esitetty liitteessä A.3.

Muuttujista erityisesti vuosien väliset erot suhteessa toistensa sijainteihin ovat pienempiä kuin luvussa 5.1. Vuosimuuttujan tasot sijoittuvat muutenkin lähelle kartan origoa, minkä perusteella voidaan sanoa, että vuosien välillä ei ole suurta eroa. Maantieteellinen jako Pohjois- ja Etelä-Suomen välillä on mielekäs, kuten myös hienoaines- ja kivipohjatyyppien välillä. Referenssi- ja impaktijokien välillä voidaan myös nähdä ero ensimmäisen dimension suunnassa. Selvä ero on myös hienoaineksen sijainnilla verrattuna kivipohjatyyppeihin, kivipohjatyypit ovat lähes päällekkäin kartassa.



Kuva 5: MCA-kartta, jonka piirtämiseen on käytetty jokimuuttujien lisäksi pohjaeläinmuuttujien kategorisoituja muunnoksia. Kartassa on esitetty jokimuuttujat sekä taksonit tasoineen. Dimensiot ja niiden selitysosuudet on merkitty kuvaan.

Muuttujien välisten assosiaatioiden tulkinnat ovat kartassa järkeviä ja tukevat aikaisempia tuloksia ja oletuksia, kuten esimerkiksi Pohjois- ja Etelä-Suomen eroavaisuutta sekä pohjatyyppien jakoa hienoaines- ja kivipohjatyyppeihin (Suomen ympäristökeskus, 2012b). Savimaiden joet ovat lähellä hienoainespohjatyyppiä ja `PE`-muuttujan etelä-tasoa, kun taas referenssimuuttujan referenssi-taso on lähellä pohjois-tasoa. Jokityyppejä tutkiessa voidaan nähdä erottumista ensimmäisen dimension suunnassa savimaiden jokiin sekä turve- ja kangasmaiden jokiin. Erittäin suuret sekä suuret kangasmaiden joet ovat lähimpänä savimaiden jokia.

Ensimmäinen dimensio kuvaa tässäkin analyysissa jokien maantieteellistä sijaintia Pohjois- tai Etelä-Suomessa. Joki- ja pohjatyyppien sekä referenssi- ja impaktitasojen sijoittuminen kartalle tukee tätä tulkintaa. Toisen dimension tulkitseminen perustellusti pelkästään jokipaikkaa kuvaavien muuttujien avulla ei ole mahdollista. Toisin kuin kuvan 4 kartassa, tutkimalla havaintojen jakaumaa tässä tapauksessa ei pystytä avustamaan dimensioiden tulkinnassa.

Lisämuuttuja-analyysin mukaisesti pyrittiin kategorisoiduillakin pohjaeläintaksoneilla löytämään indikaattori-muuttujia. Selkeimmät tulkinnat saatiin maantieteelliseen jakoon Pohjois- ja Etelä-Suomeen sekä pohjatyyppien jakoon hienoaines- ja kivipohjatyyppeihin. Taulukossa 6 on esitetty pohjaeläintaksonien tasojen (esiintyy (1), ei esiinny (0)) viisi lähintä tasoa jokaiseen jokimuuttujan tasoon verrattuna. Tulokset tukevat päiväkorentojen *Ephemerella aurivillii* sekä *Heptagenia dalecarlica* pohjoiseen painottunutta esiintymistä sekä mäkärien (*Simuliidae*) esiintymisen vahvaa painottumista kivisiin pohjatyyppeihin. Tarkasteltaessa kuvan 5 karttaa voidaan havaita ainakin *Heptagenia dalecarlican* sijoittuminen samalle kohdalle pohjois-tason kanssa suhteessa ensimmäiseen dimensioakseliin, mikä vastaa taulukon 6 tulosta.

Kartan 5 mukaan vuodet ovat hyvin samankaltaisia, mutta toisaalta kartassa 4 vuosien välillä näyttäisi olevan eroa. Analysoimalla vuosia erikseen on mahdollista saada selitys tälle erolle. Samalla on mahdollista validoida tutkielmassa saatuja tuloksia, jos vuosittaisten karttojen perusteella voidaan esittää samansuuntaisia tuloksia kuin kartoissa, joissa on kaikki vuodet.

Taulukko 6: Moniulotteisella korrespondenssianalyysilla laskettujen koordinaattien perusteella jokimuuttujien tasoja lähimpänä olevat pohjaeläintaksonien tasot. Numerot 0 ja 1 taksonin nimen perässä kertovat, että kumpi ”ei esiinny” vai ”esiintyy” liittyy muuttujan tasoon. Tekstissä tulkitut taksonit on korostettu. (Joitakin taksonien nimiä on lyhennetty.)

Luokat					
ESk	*Unio_1*	*Anodonta piscinalis_1*	*Piscicola geometra_1*	*Elmis aenea_0*	*Simuliidae_0*
ESt	*Capnia spp._1*	*Xanthoperla apicalis_1*	*Philopotamus monta._1*	*Ceratopsyche newae_1*	*Arcynopteryx compa._1*
Kk	*Oligochaeta_1*	*Sphaeriidae_1*	*Simuliidae_1*	*Chironomidae_1*	*Oulimnius tubercu._1*
Ksa	*Physa_1*	*Gyrinus spp._1*	*Dixiidae_1*	*Gammarus pulex_1*	*Goera pilosa_1*
Kt	*Polycentropus flavoma._1*	*Nemoura spp._1*	*Leuctra spp._1*	*Cheumatopsyche lepi._0*	*Erpobdella spp._0*
Pk	*Limoniidae_1*	*Capnopsis schilleri_1*	*Leuctra spp._1*	*Habrophlebia spp._1*	*Siphlonurus spp._1*
Psa	*Dixiidae_1*	*Gammarus pulex_1*	*Physa_1*	*Elodes spp._1*	*Gyrinus spp._1*
Pt	*Hydropsyche saxonica_1*	*Rhyacophila fasciata_1*	*Brachycercus harrisella_1*	*Plectrocnemia cons._1*	*Metretopus borealis_1*
Sk	*Hydropsyche contuber._1*	*Nemoura spp._0*	*Ceratopsyche newae_1*	*Limnephilidae_0*	*Neureclipsis bimacu._1*
Ssa	*Anodonta piscinalis_1*	*Bithynia tentaculata_1*	*Taeniopteryx nebulosa_0*	*Elmis aenea_0*	*Leuctra spp._0*
St	*Heptagenia sulphurea_1*	*Philopotamus monta._1*	*Athripsodes spp._1*	*Baetis vernus group_1*	*Lepidostoma hirtum_1*
Etelä	***Ephemerella auriv._0***	***Heptagenia dalec._0***	*Diura spp._0*	*Ameletus inopinatus_0*	*Protonemura spp._0*
Pohj.	*Capnia spp._1*	***Heptagenia dalec._1***	***Ephemerella auriv._1***	*Arctopsyche ladogensis_1*	*Gammarus lacustris_1*
h	*Rhyacophila nubila_0*	*Baetis rhodani_0*	***Simuliidae_0***	*Taeniopteryx nebulosa_0*	*Isoperla spp._0*
iKi	*Rhyacophila nubila_1*	***Simuliidae_1***	*Baetis rhodani_1*	*Ephemera vulgata_0*	*Ceratopogonidae_0*
pKi	***Simuliidae_1***	*Oligochaeta_1*	*Taeniopteryx nebulosa_1*	*Ephemera vulgata_0*	*Chironomidae_1*
ref_0	*Arctopsyche ladogensis_0*	*Capnia spp._0*	*Ceratopsyche silfvenii_0*	*Micrasema setiferum_0*	*Micrasema gelidum_0*
ref_1	*Isoperla spp._1*	*Taeniopteryx nebulosa_1*	*Asellus aquaticus_0*	*Baetis rhodani_1*	*Philopotamus monta._1*

5.2.1 Vuosikohtaiset analyysit

Vuosittaisten aineistojen analyysit on tehty kategorisoiduilla pohjaeläinmuuttujilla. Kategorisoituja muuttujia käytettiin, koska haluttiin nähdä, erottuvatko eri pohjaeläin taksonit eri vuosina ja löytyykö pohjaeläimistä joitakin taksoneita joiden perusteella voisi tehdä päätelmiä joen tilasta. Eri vuosien aineistot ovat tulkinnoiltaan osittain samanlaisia, joten kaikkia vuosia ei käydä erikseen läpi. Vuosia vertaillaan karttojen perusteella ja kartoista pyritään esittämään mielenkiintoisimmat erot verrattuna toisten vuosien sekä lukujen 5.1 sekä 5.2 analyyseihin.

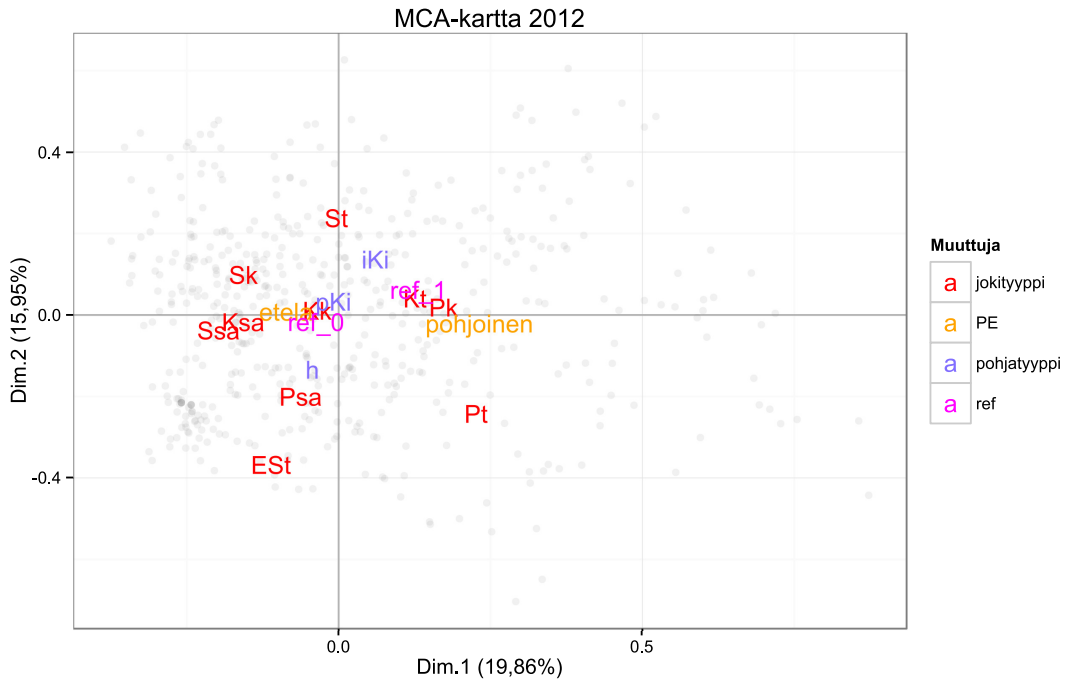
Koko aineiston havaintojen lukumäärä on 5552. Havaintojen määrän jakautuminen eri vuosille on esitetty taulukossa 7. Vuodet eroavat toisistaan havaintopaikkojen määrän ja havaintopaikkojen identiteetin suhteen. Vaikka jotkut havaintopaikat ovat olleet mukana jokaisena vuonna, osa on ollut mukana vain yhtenä vuotena. Vuoden 2011 aineistossa ei ole havaintoja erittäin suurista turvemaiden joista eikä suurista savimaiden joista. Erittäin suuria kangasmaiden jokia ei ole vuoden 2012 havaintojen joukossa.

Taulukko 7: Havaintomäärien jakautuminen eri vuosille ja puuttuvat muuttujien tasot.

Vuosi	Havaintojen määrä	Puuttuvat tasot
2006	470	-
2007	678	-
2008	641	-
2009	1149	-
2010	1206	-
2011	879	ESt, Ssa
2012	529	ESk
Yhteensä	5552	

Pohjatyyppien sijoittuminen kartalle eri vuosina on samanlaista kuin koko aineiston analyysissa 5.2. Vuosina 2007 ja 2008 sekä vuonna 2010 ei voi havaita karttaa lukemalla eroa kivipohjatyyppien välillä, mutta hienoainestyyppi eroaa näistä kahdesta selvästi (Kuvat 14, 15, 17; Liite A.3). Vuosina 2006, 2009 ja 2011 kivipohjatyyppien välillä voi havaita pientä eroa verrattuna edellä mainittuihin vuosiin, hienoaineksen erotessa kuitenkin jälleen selvästi kahdesta muusta tyypistä (Kuvat 13, 16, 18; Liite A.3). Erilaisin vuosi pohjatyyppien sijoittumista vertailtaessa on 2012 (Kuva 6), jolloin kaikki tyypit ovat selkeästi erillään ja pienkivipohjatyyppi kahden muun pohjatyypin puolivälissä.

Jokityyppien sijoittuminen kartalle on pieniä eroavaisuuksia lukuun ottamatta samanlaista jokaisena vuotena. Savimaiden joet ovat yhdessä kasassa tai linjamaisesti kartalla. Muutokset vuosien välillä saattavat johtua esimerkiksi dimensioiden ”asennosta” data-avaruudessa. Erittäin suuret kangasmaiden joet ovat lähellä suuria savimaiden jokia, nämä kaksi jokityyppiä muodostavat vuosina 2006-2009 selkeästi erottuvan ryhmän erillään muista jokityypeistä.



Kuva 6: MCA-kartta vuodelle 2012, jossa kartan piirtämiseen on käytetty jokimuuttujien lisäksi pohjaeläinmuuttujien kategorisoituja muunnoksia. Kartassa on selkeyden takia esitetty ainoastaan jokimuuttujat tasoineen sekä havainnot harmailla pisteillä. Dimensiot ja niiden selitysosuudet on merkitty kuvaan. Kuvassa erityistä on kivipohjatyyppien erottuminen.

Muuttujien `ref` sekä `PE` tasojen sijoittuminen kartalle noudattaa koko aineiston analyysissa nähtyä kaavaa, toisin sanoen etelä- ja impakti-tasot ovat lähellä toisiaan ja pohjois- ja referenssi-tasot taas lähellä toisiaan. Mielenkiintoisin ero analyysien välillä on vuoden 2006 kartassa, jossa `PE`-muuttuja näyttäisi muodostavan tulkinnan toiselle dimensiolle, kun taas referenssimuuttuja tulkinnan ensimmäiselle dimensiolle. Samana vuonna pienet savimaiden joet-taso on myös erittäin lähellä referenssi-tasoa, mikä on erikoista ottaen huomioon, että savimaiden joet ovat kaikki impaktijokia.

Karttojen dimensioiden tulkinnoissa mielenkiintoisia eroavaisuuksia on edellä esitetyn vuoden 2006 lisäksi vuosien 2007 ja vuoden 2010 kartoissa. Vuoden 2007 kartan ensimmäisellä dimensiolla voidaan huomata, että vasemmassa laidassa ovat savimaiden joet sekä erittäin suuret kangasmaiden joet. Keskellä kartassa ovat loput jokityypeistä lukuunottamatta erittäin suuria turvemaiden jokia, jotka sijaitsevat erillään kartan oikeassa reunassa. Erittäin mielenkiintoinen tulkinta voidaan esittää vuoden 2010 kartan toiselle dimensiolle, kokonsa puolesta ääripäihin sijoittuvat eli erittäin suuret sekä pienet joet ovat erillisenä ryhmänä verrattuna kooltaan ääripäiden väliin sijoittuviin jokiin. Poikkeuksena tästä ovat pienet savimaiden joet, jotka ovat samassa ryhmässä kartan keskelle sijoittuvien jokien kanssa. Savimaiden jokien on toisaalta jo todettu käyttäytyvän eri tavalla muihin jokityyppeihin verrattuna.

6 Yhteenveto

Tutkielmassa on esitetty kaksi- ja moniulotteisen korrespondenssianalyysin teoriaa ja niiden sovellusta käytännön esimerkkeihin. Pääpaino tutkielmassa on moniulotteisessa korrespondenssianalyysissa, jota on käytetty sovellusaineiston analyysiin. Tutkielman teoriaosuudessa esitetään ensin kaksiulotteisen korespondenssianalyysin teoriaa. Menetelmän teoriaan liittyvät tärkeimmät tulokset on todistettu tutkielmassa tai niistä on annettu viite kirjallisuuteen. Kaksiulotteisesta korrespondenssianalyysistä on lisäksi kuvattu kirjallisuudessa esitetty esimerkki.

Teoriaa laajennetaan moniulotteiseen korrespondenssianalyysiin tarvittavilta osilta. Tulosten tulkitseminen ja peruslaskentamenetelmät pysyvät samoina, suurimpien muutosten ollessa erot aineiston esittämisessä indikaattori- ja Burtin matriisin avulla sekä aineiston muokkauksesta näiksi matriiseiksi syntyvien keinotekoisten dimensioiden huomiointi. Lisäksi esitetään lyhyesti teoriaa lisämuuttujista, jotka laajentavat moniulotteisen korrespondenssianalyysin sovellusmahdollisuuksia antamalla esimerkiksi mahdollisuuden tutkia jatkuvia muuttujia.

Sovellusaineistona tutkielmassa on Suomen ympäristökeskuksen tuottama pohjaeläinaineisto. Aineiston muuttujina on erilaisia havaintopaikkaa kuvaavia muuttujia sekä pohjaeläintaksoneita. Moniulotteisen korrespondenssianalyysin avulla pyrittiin muodostamaan kuva aineiston mahdollisista assosiaatiorakenteista. Tarkoituksena oli erityisesti tutkia hienoaines- ja kivipohjatyyppien välillä aiemmin havaittua eroavaisuutta ja mahdollisesti löytää sille selitystä muiden muuttujien avulla.

Tärkeimpiä esille nousseita tuloksia olivat referenssi- ja impaktipaikkojen erottuminen, pohjatyyppien jakautuminen hienoaines- ja kivipohjatyyppeihin sekä Pohjois- ja Etelä-Suomen maantieteellinen erottuminen. Lisäksi analyysissa löydettiin joitakin mahdollisia indikaattorilajeja pohjatyypin, maantieteellisen pohjois–etelä-rajan sekä savimaiden jokien tunnistamisen apuvälineeksi.

Pohjatyyppien selkeä jakautuminen kartoissa hienoaines- ja kivipohjatyyppeihin vahvistaa aiempia käsityksiä eroavaisuudesta näiden välillä ja antaa aihetta tutkia asiaa jatkossakin eron syiden selvittämiseksi. Jatkotutkimuksissa voitaisiin ottaa kantaa myös savimaiden jokien eriytymiseen muista jokityypeistä sekä siihen miksi erittäin suurten kangasmaiden -jokityyppi esiintyy useasti lähellä savimaiden jokityyppejä. Indikaattoritaksonien havaitsemista moniulotteisen korrespondenssianalyysin avulla voitaisiin lisäksi kehittää.

Tutkielman avulla pystyttiin vastaamaan sille asetettuihin tutkimustavoitteisiin, tärkeimpänä pohjatyyppien välisen eroavaisuuden vahvistaminen. Analyysien tulokset herättävät kysymyksiä aineiston rakenteesta, mikä onkin tämän menetelmän lähtökohtaisena tarkoituksena: esittää aineiston assosiaatiorakenne sellaisessa helposti tulkittavassa muodossa, että sen avulla voidaan muotoilla uusia tutkimuskysymyksiä.

Kiitokset

Haluan kiittää tutkielmani ohjaajaa FT Salme Kärkkäistä tuesta, kommenteista ja kannustuksesta työn jokaisessa vaiheessa. Pohjaeläinaineiston saamisesta sovellusaineistoksi sekä asiantuntijalausunnoista aineistoon liittyen haluan esittää kiitokseni Suomen ympäristökeskuksen erikoistutkija Kristian Meissnerille. Avusta aineiston muokkauksessa haluan kiittää FM Johanna Ärjeä. Työn loppuvaiheen kommenteista haluan kiittää dos. Sara Taskista.
Perheelleni haluan esittää kiitoksen kaikesta mahdollisesta tuesta tutkielman tekemisen aikana.

Lähdeluettelo

Benzécri, J. P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [bin. mult.]. *Cahiers de l'analyse des données*, 4(3):377–378.

Benzécri, J.-P. et al. (1973). *L'Analyse des Données, volume II : L'Analyse des Correspondances*. Dunod, Paris.

Burt, C. (1950). The factorial analysis of qualitative data. *British Journal of Psychology (Statistical Section)*, 3:166–185.

Cadoret, L., Legendre, P., Adjeroud, M., & Galzin, R. (1995). Répartion spatiale des chaetodontidae dans différents secteurs récifaux de l'île de moorea, polynésie française. *Écoscience*, 2:129–140.

Clausen, S.-E. (1998). *Applied Correspondense Analysis: An introduction*. Series no. 07-121. Sage, Thousand Oaks, CA.

Fisher, R. A. (1940). The precisions of discriminant functions. *Annals of Eugenics*, 10:422–429.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Greenacre, M. J. (1988). Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, 75(3):457–467.

Greenacre, M. J. (1993). *Correspondence Analysis in Practice*. Academic Press, London.

Greenacre, M. J. (2007). *Correspondence Analysis in Practice*, 2nd edition. Chapman & Hall, London.

Greenacre, M. J. & Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods*. Statistics in the Social and Behavioral Sciences Series. Chapman & Hall, London.

Guttman, L. (1941). The quantification of a class of attributes: a theory and method of scale construction. In Horst, P. et al., editors, *The Prediction of Personal Adjustement*, p. 321–348. Social Science Recearch Council, New York.

Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Journal of the Royal Statistic society C (Applied Statistics)*, 23:340–354.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society*, 31:520–524.

Husson, F. & Josse, J. (2014). *Multiple Correspondence Analysis*, p. 165–184. Chapman and Hall/CRC.

ISSP (2003). International social survey program: National identity II. www.issp.org. Viitattu 27.5.2015.

Lebart, L., Morineau, A., & Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons, New York.

Lê, S., Josse, J., & Husson, F. (2008). Factominer: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

Meissner, K. et al. (2013). Jokien ja järvien biologinen seuranta – näytteenotosta tiedon tallentamiseen. http://www.ymparisto.fi/download/noname/%7BB948034F-7F9D-4EAB-A153-92FA2DDEDBBE%7D/29725. Viitattu 28.5.2015.

Nenadic, O. & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3):1–13.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd edition. John Wiley & Sons, New York.

Suomen ympäristökeskus (2012a). Jokien tyypittely. http://www.ymparisto.fi//fi-FI/Vesi/Pintavesien_tila/Pintavesien_tyypittely > Ohje pintaveden tyypin määrittämiseksi (pdf 850kb). Viitattu 22.2.2015.

Suomen ympäristökeskus (2012b). Ohje pintavesien ekologisen ja kemiallisen tilan luokitteluun vuosille 2012–2013 – päivitetyt arviointiperusteet ja niiden soveltaminen. http://hdl.handle.net/10138/41788 > OH_7_2012.pdf (pdf 6,198Mb). Viitattu 22.2.2015.

Suomen ympäristökeskus (2013). Jokien tyypittely. http://www.ymparisto.fi//fi-FI/Vesi/Pintavesien_tila/Pintavesien_tyypittely > Kaavio Jokien tyypittely Suomessa. Viitattu 22.2.2015.

ter Braak, C. J. F. (1985). Correspondence analysis of incidence and abundance data: properties in terms of unimodal response model. *Biometrics*, 41:859–873.

A Liitteet

A.1 Aineiston havaintomatriisin havainnollistus

Muuttujat järjestyksessä: Paikan nimi, ID, naytevuosi, jokityyppi, PE, pohjatyyppi, ref, Aeshna_spp., Afghanurus_joernensis, Agapetus_spp. ja Agraylea_spp. ...

```
Aittokoski_H   3 2006 Sk etela h   1 0 0 0 0 ...
Aittokoski_H   3 2006 Sk etela h   1 0 0 6 0 ...
Aittokoski_iKi 3 2006 Sk etela iKi 1 0 0 4 0 ...
Aittokoski_iKi 3 2006 Sk etela iKi 1 0 0 0 0 ...
Aittokoski_pKi 3 2006 Sk etela pKi 1 0 0 22 0 ...
Aittokoski_pKi 3 2006 Sk etela pKi 1 0 0 5 0 ...
```

A.2 Pääinertioiden summautuminen kokonaisinertiaksi

1. $\sum_i \sum_j s_{ij}^2 = \text{trace}(\mathbf{S}\mathbf{S}^T)$:

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1J} \\ s_{21} & s_{22} & \cdots & s_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ s_{I1} & s_{I2} & \cdots & s_{IJ} \end{pmatrix} \qquad \mathbf{S^T} = \begin{pmatrix} s_{11} & s_{21} & \cdots & s_{I1} \\ s_{12} & s_{22} & \cdots & s_{I2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1J} & s_{2J} & \cdots & s_{IJ} \end{pmatrix}.$$

Matriisi $\mathbf{S}$ on $(I \times J)$-matriisi ja $\mathbf{S^T}$ on $(J \times I)$-matriisi, joten matriisin $\mathbf{SS^T}$ dimensio on $(I \times I)$. Matriisin jälki määritellään neliömatriisin diagonaalialkioiden summana, joten lasketaan matriisista $\mathbf{SS^T}$ ainoastaan sen diagonaali:

$$\mathbf{SS^T} = \begin{pmatrix} \sum_{j=1}^{J} {s_{1j}}^2 & \cdots & \cdot \\ \vdots & \ddots & \vdots \\ \cdot & \cdots & \sum_{j=1}^{J} {s_{Ij}}^2 \end{pmatrix}.$$

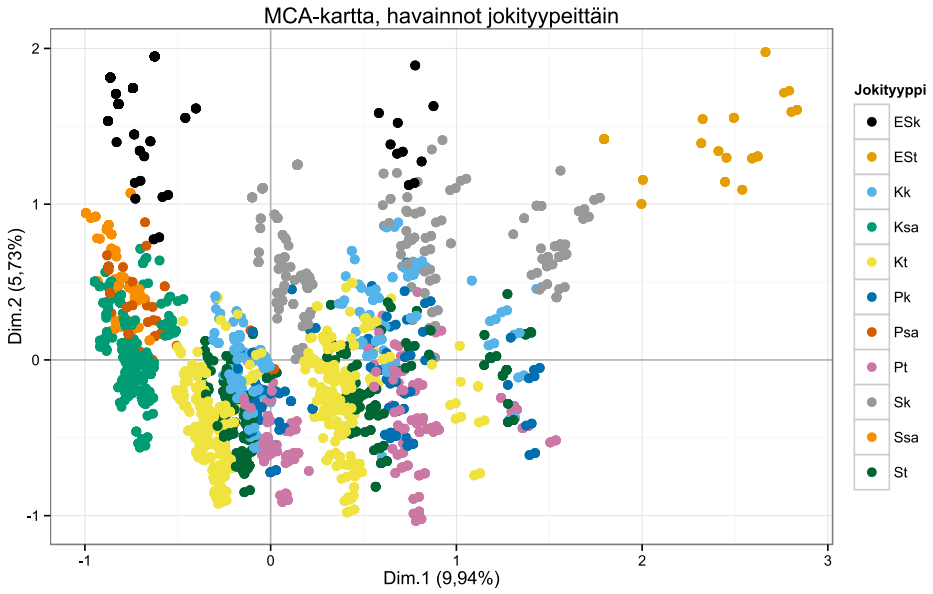
Koska $\mathbf{SS^T}$ on $(I \times I)$-matriisi, niin sen jälki $\text{trace}(\mathbf{SS^T}) = \sum_i \sum_j s_{ij}^2$.

2. $\text{trace}(\mathbf{S}\mathbf{S}^T) = \text{trace}(\mathbf{\Lambda})$:

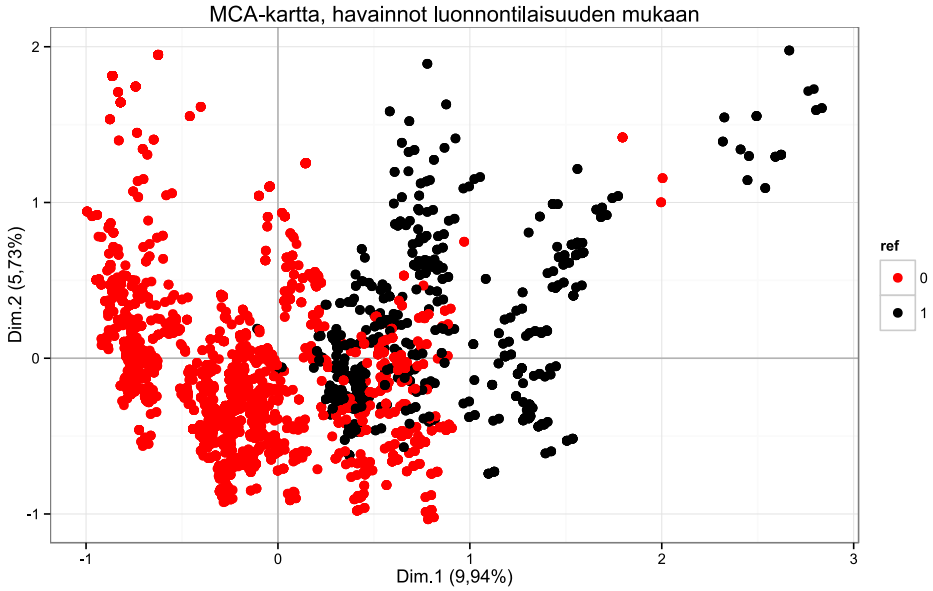
$$\begin{aligned} \text{trace}(\mathbf{S}\mathbf{S}^T) &= \text{trace}(\mathbf{U\Lambda U^T}) \\ &= \text{trace}(\mathbf{U^T U\Lambda}) \\ &= \text{trace}(\mathbf{I\Lambda}) \\ &= \text{trace}(\mathbf{\Lambda}). \end{aligned}$$

Ensimmäinen yhtäsuuruus seuraa kaavasta (11), toinen matriisin jäljen ominaisuuksista ja kolmas $\mathbf{U}$ matriisin ortogonaalisuudesta.

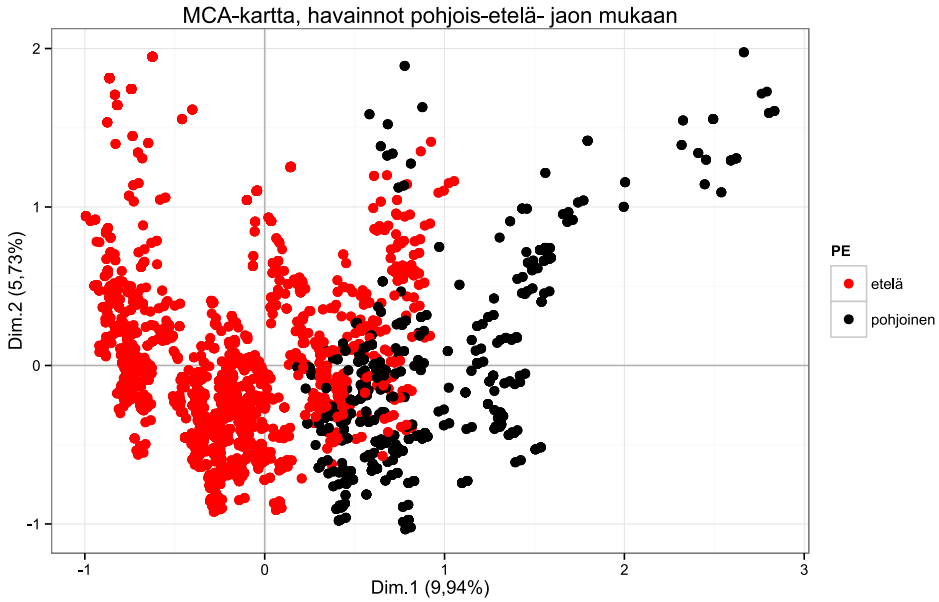
A.3 Kuvia moniulotteisista korrespondenssianalyyseista



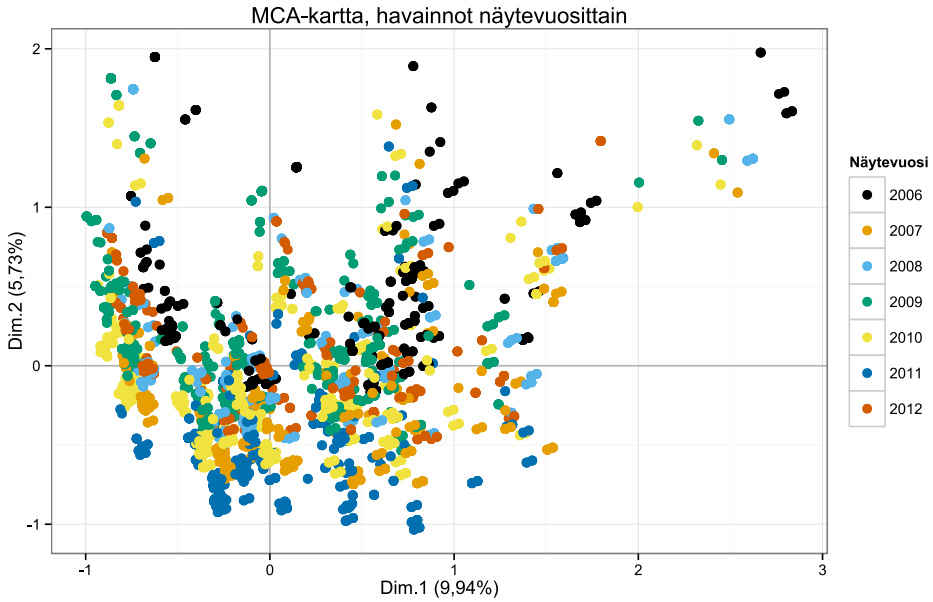
Kuva 7: MCA-kartta, jossa on esitetty havaintojen jakauma pohjaeläimet lisämuuttujina analyysista. Jokityypit on eroteltu toisistaan värein.



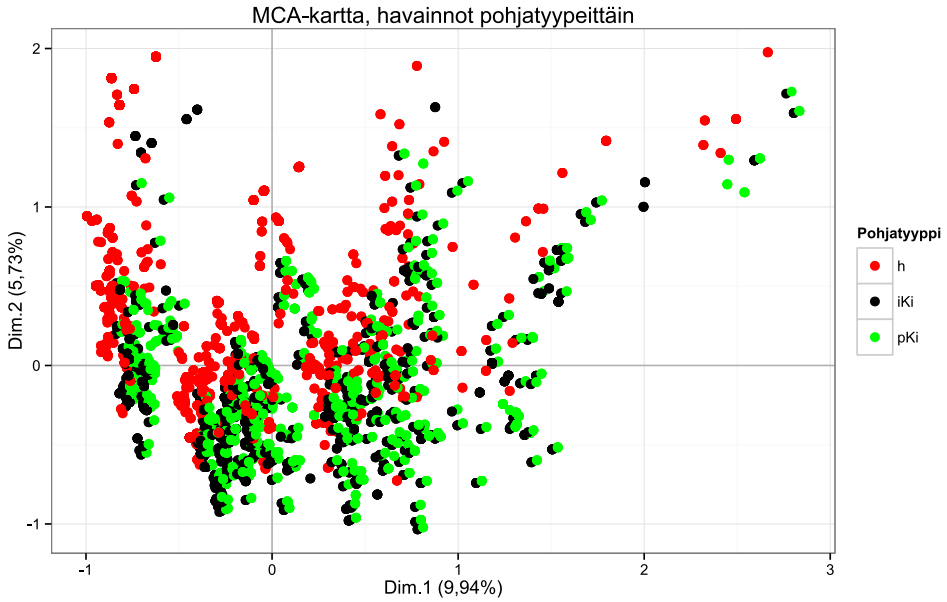
Kuva 8: MCA-kartta, jossa on esitetty havaintojen jakauma pohjaeläimet lisämuuttujina analyysista. Havainnot on eroteltu toisistaan värein siten, että punainen vastaa impakti- ja musta referenssijokia.



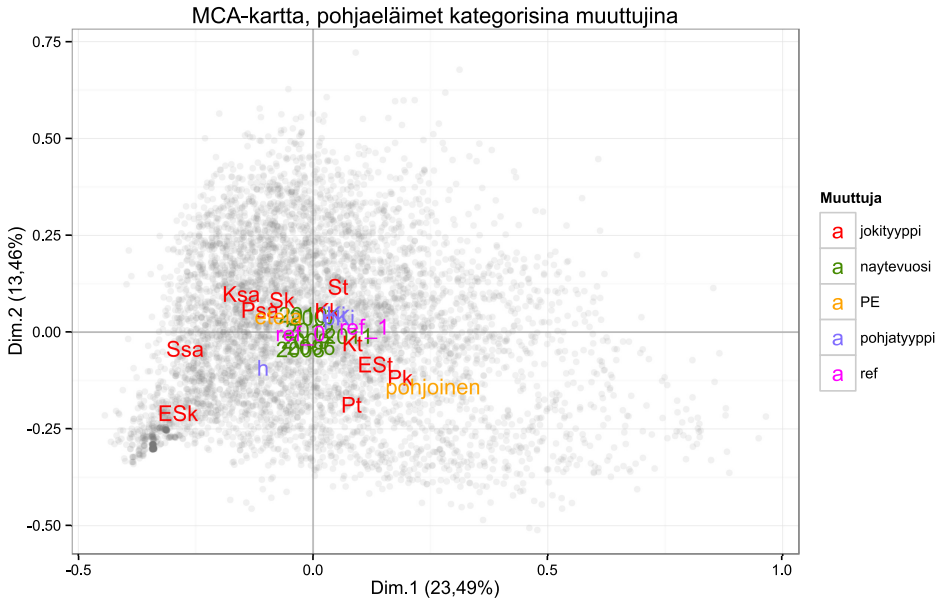
Kuva 9: MCA-kartta, jossa on esitetty havaintojen jakauma pohjaeläimet lisämuuttujina analyysista. Pohjois- ja etelä-tasot on eroteltu toisistaan värein.



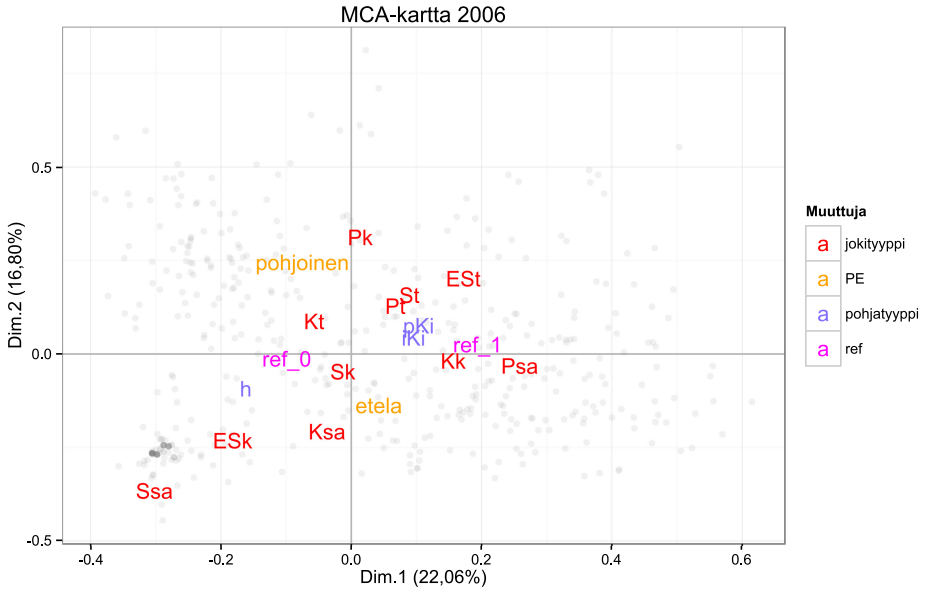
Kuva 10: MCA-kartta, jossa on esitetty havaintojen jakauma pohjaeläimet lisämuuttujina analyysista. Näytevuodet on eroteltu toisistaan värein.



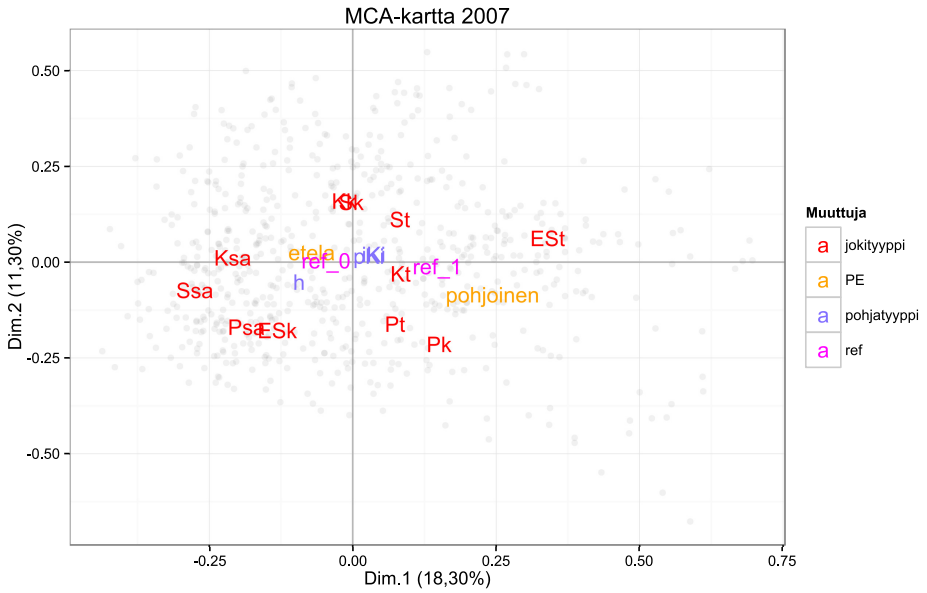
Kuva 11: MCA-kartta, jossa on esitetty havaintojen jakauma pohjaeläimet lisämuuttujina analyysista. Pohjatyypit on eroteltu toisistaan värein.



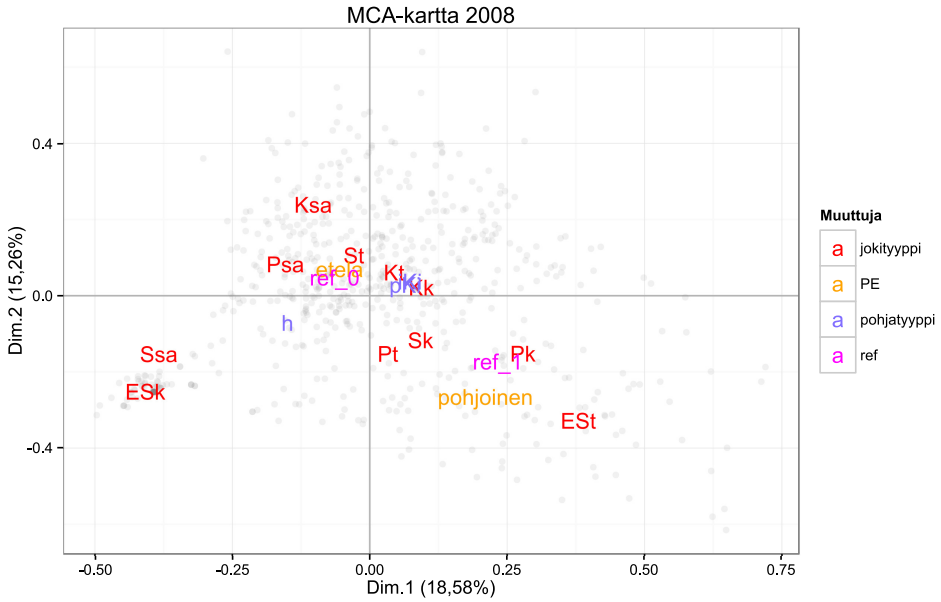
Kuva 12: MCA-kartta, jossa kartan piirtämiseen on käytetty jokimuuttujien lisäksi pohjaeläinmuuttujien kategorisoituja muunnoksia. Kartassa on selkeyden takia esitetty ainoastaan jokimuuttujat tasoineen sekä havainnot harmailla pisteillä. Dimensiot ja niiden selitysosuudet on merkitty kuvaan.



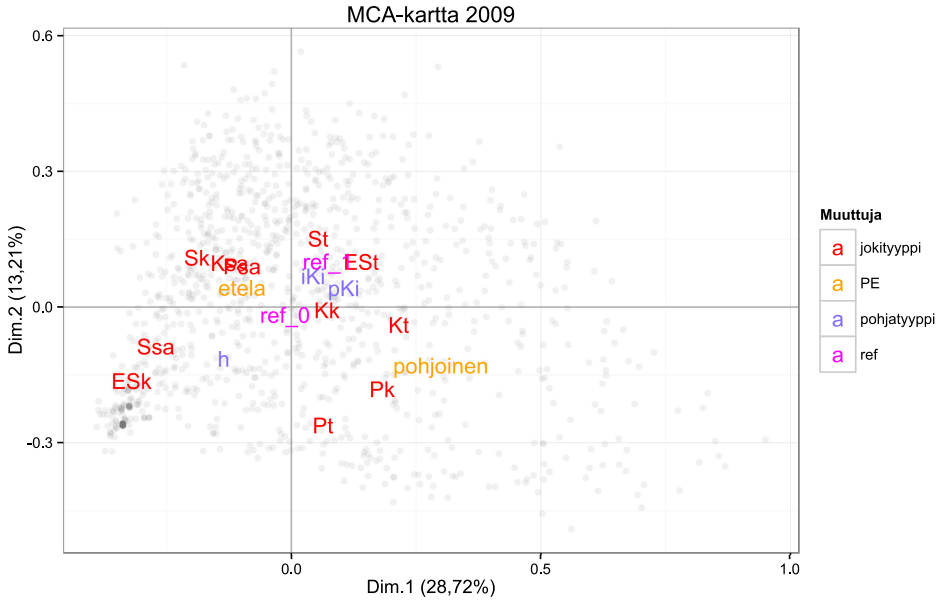
Kuva 13: MCA-kartta vuodelle 2006, jossa kartan piirtämiseen on käytetty jokimuuttujien lisäksi pohjaeläinmuuttujien kategorisoituja muunnoksia. Kartassa on selkeyden takia esitetty ainoastaan jokimuuttujat tasoineen sekä havainnot harmailla pisteillä. Dimensiot ja niiden selitysosuudet on merkitty kuvaan.



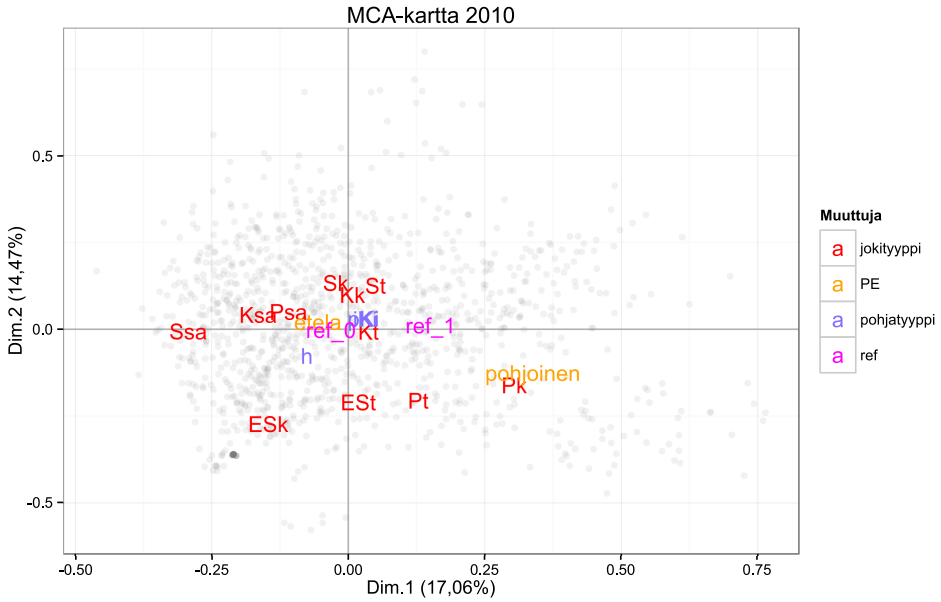
Kuva 14: MCA-kartta vuodelle 2007, jossa kartan piirtämiseen on käytetty jokimuuttujien lisäksi pohjaeläinmuuttujien kategorisoituja muunnoksia. Kartassa on selkeyden takia esitetty ainoastaan jokimuuttujat tasoineen sekä havainnot harmailla pisteillä. Dimensiot ja niiden selitysosuudet on merkitty kuvaan.



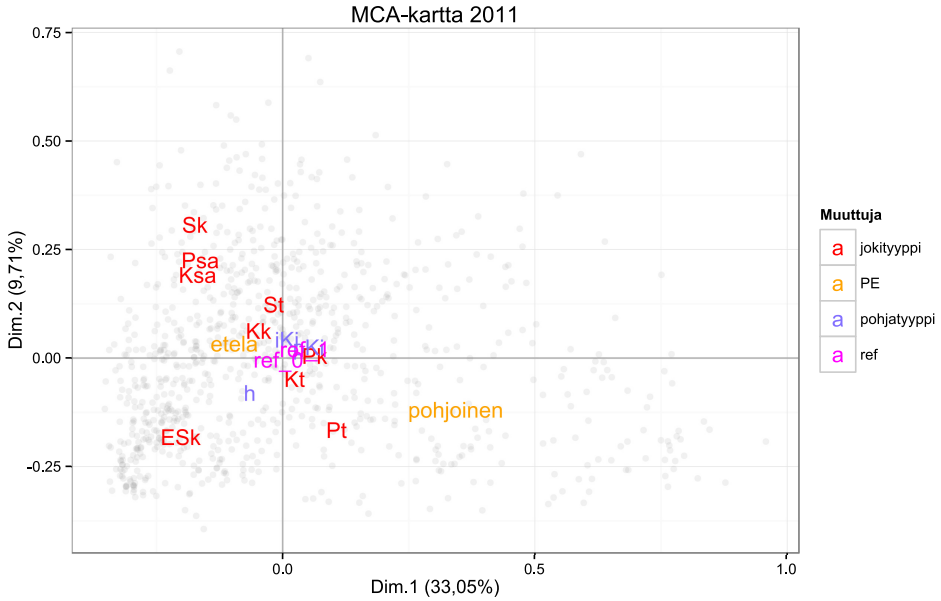
Kuva 15: MCA-kartta vuodelle 2008, jossa kartan piirtämiseen on käytetty jokimuuttujien lisäksi pohjaeläinmuuttujien kategorisoituja muunnoksia. Kartassa on selkeyden takia esitetty ainoastaan jokimuuttujat tasoineen sekä havainnot harmailla pisteillä. Dimensiot ja niiden selitysosuudet on merkitty kuvaan.



Kuva 16: MCA-kartta vuodelle 2009, jossa kartan piirtämiseen on käytetty jokimuuttujien lisäksi pohjaeläinmuuttujien kategorisoituja muunnoksia. Kartassa on selkeyden takia esitetty ainoastaan jokimuuttujat tasoineen sekä havainnot harmailla pisteillä. Dimensiot ja niiden selitysosuudet on merkitty kuvaan.



Kuva 17: MCA-kartta vuodelle 2010, jossa kartan piirtämiseen on käytetty jokimuuttujien lisäksi pohjaeläinmuuttujien kategorisoituja muunnoksia. Kartassa on selkeyden takia esitetty ainoastaan jokimuuttujat tasoineen sekä havainnot harmailla pisteillä. Dimensiot ja niiden selitysosuudet on merkitty kuvaan.



Kuva 18: MCA-kartta vuodelle 2011, jossa kartan piirtämiseen on käytetty jokimuuttujien lisäksi pohjaeläinmuuttujien kategorisoituja muunnoksia. Kartassa on selkeyden takia esitetty ainoastaan jokimuuttujat tasoineen sekä havainnot harmailla pisteillä. Dimensiot ja niiden selitysosuudet on merkitty kuvaan.

A.4 Tutkimuksessa käytetyt taksonit

Taulukko 8: Aineiston taksonit sekä vertailuna PMA-indeksin laskemiseen käytetyt taksonit.

Aineistossa käytetyt	Mukana PMA-indeksissä
Aeshna spp.	Kyllä
Afghanurus joernensis	Kyllä
Agapetus spp.	Kyllä
Agraylea spp.	Kyllä
Agrion spp.	Kyllä
Agrypnia spp.	Kyllä
Ameletus inopinatus	Kyllä
Amphinemura borealis	Kyllä
Amphinemura spp.	Kyllä
Ancylus fluviatilis	Kyllä
Anodonta piscinalis	Kyllä
Apatania spp.	Kyllä
Aphelocheirus aestivalis	Kyllä
Arctopsyche ladogensis	Kyllä
Arcynopteryx compacta	Kyllä
Asellus aquaticus	Kyllä
Astacus astacus	Ei
Atherix ibis	Kyllä
Athripsodes spp.	Kyllä
Baetis liebenauae	Kyllä
Baetis macani	Kyllä
Baetis niger group	Kyllä
Baetis rhodani	Kyllä
Baetis vernus group	Kyllä
Bathyomphalus contortus	Kyllä
Beraea pullata	Kyllä
Beraeodes minutus	Kyllä
Bithynia tentaculata	Kyllä
Brachycentrus subnubilus	Kyllä
Brachycercus harrisella	Ei
Caenis spp.	Kyllä
Calopteryx spp.	Kyllä
Capnia spp.	Kyllä
Capnopsis schilleri	Kyllä
Centroptilum luteolum	Kyllä
Ceraclea spp.	Kyllä
Ceratopogonidae	Kyllä
Ceratopsyche newae	Kyllä

Ceratopsyche silfvenii	Kyllä
Cheumatopsyche lepida	Kyllä
Chimarra marginata	Kyllä
Chironomidae	Ei
Chrysomelidae	Kyllä
Cloeon spp.	Kyllä
Cordulecaster boltoni	Kyllä
Corixidae	Kyllä
Diura spp.	Kyllä
Dixiidae	Kyllä
Dytiscidae	Kyllä
Elmis aenea	Kyllä
Elodes spp.	Kyllä
Empididae	Kyllä
Ephemera danica	Kyllä
Ephemera vulgata	Kyllä
Ephemerella aurivillii	Kyllä
Ephemerella ignita	Kyllä
Ephemerella mucronata	Kyllä
Erpobdella spp.	Kyllä
Gammarus lacustris	Kyllä
Gammarus pulex	Kyllä
Glossiphonia spp.	Kyllä
Glossosoma spp.	Kyllä
Goera pilosa	Kyllä
Gyraulus spp.	Kyllä
Gyrinus spp.	Kyllä
Habrophlebia spp.	Kyllä
Haliplidae	Kyllä
Helobdella stagnalis	Kyllä
Heptagenia dalecarlica	Kyllä
Heptagenia sulphurea	Kyllä
Hydracarina	Ei
Hydraena spp.	Kyllä
Hydrophilidae	Kyllä
Hydropsyche angustipennis	Kyllä
Hydropsyche contubernalis	Kyllä
Hydropsyche pellucidula	Kyllä
Hydropsyche saxonica	Kyllä
Hydropsyche siltalai	Kyllä
Hydroptila spp.	Kyllä
Isogenus nubecula	Kyllä
Isoperla spp.	Kyllä
Ithytrichia spp.	Kyllä
Kageronia fuscogrisea	Kyllä
Lepidostoma hirtum	Kyllä
Leptophlebiidae	Kyllä

Leuctra nigra	Kyllä
Leuctra spp.	Kyllä
Libellulidae	Ei
Limnephilidae	Kyllä
Limnius volckmari	Kyllä
Limoniidae	Kyllä
Lype spp.	Kyllä
Metretopus borealis	Ei
Micrasema gelidum	Kyllä
Micrasema setiferum	Kyllä
Molanna	Ei
Molannodes tinctus	Kyllä
Muscidae	Kyllä
Mystacides spp.	Kyllä
Nemoura spp.	Kyllä
Nemurella pictetii	Kyllä
Neureclipsis bimaculata	Kyllä
Normandia nitens	Kyllä
Notidobia ciliaris	Kyllä
Oecetis spp.	Kyllä
Oligochaeta	Ei
Oligostomis reticulata	Kyllä
Onychogomphus forcipatus	Kyllä
Ophiogomphus cecilia	Kyllä
Orectochilus villosus	Kyllä
Oulimnius tuberculatus	Kyllä
Oxyethira spp.	Kyllä
Philopotamus montanus	Kyllä
Phryganea spp.	Kyllä
Physa	Ei
Piscicola geometra	Kyllä
Planorbiidae	Kyllä
Platycnemis pennipes	Kyllä
Plectrocnemia conspersa	Kyllä
Polycentropus flavomaculatus	Kyllä
Polycentropus irroratus	Kyllä
Potamanthus luteus	Kyllä
Protonemura spp.	Kyllä
Psychodidae	Kyllä
Psychomyia pusilla	Kyllä
Ptychopteridae	Kyllä
Pyralidae	Kyllä
Radix spp.	Kyllä
Rhabdiopteryx acuminata	Kyllä
Rhyacophila fasciata	Kyllä
Rhyacophila nubila	Kyllä
Rhyacophila obliterata	Kyllä

Sciomyzidae	Kyllä
Semblis spp.	Kyllä
Sericostoma personatum	Kyllä
Sialis spp.	Kyllä
Silo pallipes	Kyllä
Simuliidae	Kyllä
Siphlonurus spp.	Kyllä
Siphonoperla burmeisteri	Kyllä
Sisyra spp.	Kyllä
Somatochlora metallica	Kyllä
Sphaeriidae	Kyllä
Stactobiella risi	Kyllä
Stenelmis canaliculata	Kyllä
Tabanidae	Kyllä
Taeniopteryx nebulosa	Kyllä
Tinodes	Ei
Tipulidae	Kyllä
Triaenodes spp.	Kyllä
Unio	Ei
Valvata spp.	Kyllä
Wormaldia subnigra	Kyllä
Xanthoperla apicalis	Kyllä
Ylodes spp.	Kyllä
	Acentrella lapponica
	Dinocras cephalotes
	Erotesis baltica
	Metretopus borealis

A.5 R-koodi korrespondenssianalyysiesimerkin kuville ja analyysille

```
install.packages("ca")
library(ca)

UK <- c(230,329,177,34,6) # Datan luonti
US <- c(400,471,237,28,12)
Rus <- c(1010,530,141,21,11)
Spn <- c(201,639,208,72,14)
Fra <- c(365,478,305,50,97)
Cnam <- c("UK","USA","Venäjä","Espanja","Ranska")
Rnam <- c("1","2","3","4","5")
intsport <- c(UK, US, Rus, Spn, Fra)
M <- matrix(intsport, byrow = F, ncol = 5) # Data matriisiksi
rownames(M) <- Rnam # Rivien nimeäminen
colnames(M) <- Cnam # Sarakkeiden nimeäminen

M <- as.table(M) # Analyysi suoritetaan "table"-luokan objektille
prop.table(M, 1) # Riviprosentit
prop.table(M, 2) # Sarakeprosentit
fit <- ca(M) # Analyysin suoritus
print(fit) # Perustulokset
summary(fit) # Laajemmat tulokset

## Seuraavassa piirretään symmetrinen kartta, jota on hieman
## paranneltu selkeyden takia. Pelkkä plot(fit) riittäisi.
par(pin =c(4.7244,4.7244) mai = c(0.5,0.5,0.1,0.1), cex = 1.3,
cex.axis = 0.7)

# Symmetrinen kartta
plot(fit, ylim = c(-1.1,0.3), xlim = c(-0.7,0.7), labels = c(2,0))
text(x = c(0.095,-0.05,-0.57,0.26,0.14), y = c(0.13,0.08,0.03,0.25,-0.3),
labels = Cnam, cex = 0.65)

## Seuraavassa piirretään epäsymmetrinen kartta.
## Riittävä olisi plot(fit, map = "colprincipal")
plot(fit, map = "colprincipal", xlim = c(-2,2), labels = c(2,0))
text(x = c(0.15,-0.15,-0.85,0.8,0.35), y = c(0.25,-0.1,0.1,0.25,-0.5),
labels = Cnam, cex = 0.65)
```

```
## Menetelmä ja SVD vaiheittain ##

library(base)

P <- M/sum(M) # Korrespondenssimatriisin teko
cm <- apply(P, 2, sum) # Sarakemassat
rm <- apply(P, 1, sum) # Rivimassat
S <- (P - (rm %*% t(cm))) / sqrt(rm %*% t(cm)) # Standardoitumatriisi
SV <- svd(S,LINPACK = T) # Singulaariarvot
SIGMA <- diag(SV$d) # Singulaariarvojen diag.matriisi
U <- SV$u # Vasemmat singulaarivektorit
V <- SV$v # Oikeat singulaarivektorit

MCOL <- diag(1/sqrt(cm)) %*% V # Sarakkeiden standardoidut koordinaatit
MROW <- diag(1/sqrt(rm)) %*% U # Rivien standardoidut koordinaatit
MSCOL <- MCOL %*% SIGMA # Sarakkeiden pääkoordinaatit
MSROW <- MROW %*% SIGMA # Rivien pääkoordinaatit
xr <- MSROW[,1] # Rivien pääkoordinaatit ensimmäiselle dimensiolle
yr <- MSROW[,2] # Rivien pääkoordinaatit toiselle dimensiolle
xc <- MSCOL[,1] # Sarakkeiden pääkoordinaatit ensimmäiselle dimensiolle
yc <- MSCOL[,2] # Sarakkeiden pääkoordinaatit toiselle dimensiolle

plot(fit, ylim = c(-1.1,0.3), xlim = c(-0.7,0.7), labels = c(2,0))
text(x = c(0.095,-0.05,-0.57,0.26,0.14), y = c(0.13,0.08,0.03,0.25,-0.3),
labels = Cnam, cex = 0.65) # Kuvan piirto ca-funktion avulla
points(xc, yc, pch = 3, col = 1) # Itse lasketut sarakepääkoordinaatit
points(xr, yr, pch = 3, col = 3) # Itse lasketut rivipääkoordinaatit
```

A.6 R-koodi MCA:n kuville ja analyysille

```
#### Kategorisoitujen muuttujien luonti
presens <- data3
for(i in 12:length(presens[1,])){
  for(j in 1:length(presens[,1])){
    if(presens[j,i]!=0){
      presens[j,i]<-1
    }
  }
}

names(presens[1:16])
pres <- presens[-c(1,2,4,6,7)]

#### Faktorointi
attach(pres)
ID <- as.factor(ID)
naytevuosi <- as.factor(naytevuosi)
jokityyppi <- as.factor(jokityyppi)
PE <- as.factor(PE)
pohjatyyppi <- as.factor(pohjatyyppi)
ref <- as.factor(ref)

pres[,c(1:161)] <- data.frame(apply(pres[c(1:161)], 2, as.factor))
levels(pres[,c(161)])

#### Vuosittaiset taulukot
i <- NULL
v6 <- NULL
v7 <- NULL
v8 <- NULL
v9 <- NULL
v10 <- NULL
v11 <- NULL
v12 <- NULL
for(i in 1:length(pres[,2])){
  if(naytevuosi[i]=="2006"){
    v6 <- c(v6,i)
  }
  if(naytevuosi[i]=="2007"){
    v7 <- c(v7,i)
  }
  if(naytevuosi[i]=="2008"){
    v8 <- c(v8,i)
  }
```

```
  if(naytevuosi[i]=="2009"){
    v9 <- c(v9,i)
  }
  if(naytevuosi[i]=="2010"){
    v10 <- c(v10,i)
  }
  if(naytevuosi[i]=="2011"){
    v11 <- c(v11,i)
  }
  if(naytevuosi[i]=="2012"){
    v12 <- c(v12,i)
  }
}
pres06 <- pres[v6,-2]
pres06[,c(1:160)] <- data.frame(apply(pres06[c(1:160)], 2, as.factor))
pres07 <- pres[v7,-2]
pres07[,c(1:160)] <- data.frame(apply(pres07[c(1:160)], 2, as.factor))
pres08 <- pres[v8,-2]
pres08[,c(1:160)] <- data.frame(apply(pres08[c(1:160)], 2, as.factor))
pres09 <- pres[v9,-2]
pres09[,c(1:160)] <- data.frame(apply(pres09[c(1:160)], 2, as.factor))
pres10 <- pres[v10,-2]
pres10[,c(1:160)] <- data.frame(apply(pres10[c(1:160)], 2, as.factor))
pres11 <- pres[v11,-2]
pres11[,c(1:160)] <- data.frame(apply(pres11[c(1:160)], 2, as.factor))
pres12 <- pres[v12,-2]
pres12[,c(1:160)] <- data.frame(apply(pres12[c(1:160)], 2, as.factor))

#### Kategorioiden määrä
cats = apply(pres, 2, function(x) nlevels(as.factor(x)))
cats
cats06 = apply(pres06, 2, function(x) nlevels(as.factor(x)))
cats06
cats07 = apply(pres07, 2, function(x) nlevels(as.factor(x)))
cats07
cats08 = apply(pres08, 2, function(x) nlevels(as.factor(x)))
cats08
cats09 = apply(pres09, 2, function(x) nlevels(as.factor(x)))
cats09
cats10 = apply(pres10, 2, function(x) nlevels(as.factor(x)))
cats10
cats11 = apply(pres11, 2, function(x) nlevels(as.factor(x)))
cats11
cats12 = apply(pres12, 2, function(x) nlevels(as.factor(x)))
cats12
```

```
############### MCA ################

#library(MASS)
#?mca
install.packages("FactoMineR")
install.packages("ca")
install.packages("ggplot2")
library("FactoMineR")
library("ca")
library("ggplot2")

#### MCA frekvenssidatalle

dataMCA<-data3[-c(1,2,4,6,7)]
names(dataMCA[1:10])
dataMCA[,c(1:6)] <- data.frame(apply(dataMCA[c(1:6)], 2, as.factor))
catsFREQ = apply(dataMCA[1:6], 2, function(x) nlevels(as.factor(x)))
mca_FREQ <- MCA(dataMCA, graph=F, quanti.sup = c(7:161), method="Burt")
summary(mca_FREQ)

#### Datakehikko muuttujien koordinaateista
mca_FREQ_vars_df = data.frame(mca_FREQ$var$coord, Variable =
c(rep(names(catsFREQ), catsFREQ),names(dataMCA[,7:161])))
mca_FREQ$var$coord[591:615,1:5]
piirtoNimetFREQ <- c(rep(names(catsFREQ), catsFREQ),names(dataMCA[,7:161]))
jokidataVarFREQ = data.frame(mca_FREQ$var$coord[591:615,1:5], Variable =
piirtoNimetFREQ[591:615])

#### Datakehikko havaintojen koordinaateista
mca_FREQ_obs_df = data.frame(mca_FREQ$ind$coord)
jokidataObsFREQ = data.frame(mca_FREQ$ind$coord)

#### Kartan piirtäminen muuttujien tasoista
ggplot(data = jokidataObsFREQ, aes(x = Dim.1, y = Dim.2)) +
  theme_bw() +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(colour = "gray50", alpha = 0.1) +
  geom_text(data = jokidataVarFREQ,
            aes(x = Dim.1, y = Dim.2,
                label = rownames(jokidataVarFREQ), colour = Variable)) +
  labs(x = "Dim.1 (9,94%)", y = "Dim.2 (5,73%)") +
  ggtitle("MCA-kartta, pohjaeläimet lisämuuttujina") +
   scale_colour_manual(name = "Muuttuja", values =
   c("red","chartreuse4","orange","lightslateblue","magenta"))
```

```
#### Korjatut inertiat ja dimensioiden selitysasteet
levelsFREQ <- NULL
for(i in 1:length(dataMCA[1,])){
  levelsFREQ <- c(levelsFREQ,levels(dataMCA[,i]))
}
inert_FREQ <- NULL
for(i in 1:20){
  #inert_pres12[i] <- (mca_pres$eig[i,1]-(1/160))
  inert_FREQ[i] <- ((6/(6-1))^2)*((sqrt(mca_FREQ$eig[i,1])-(1/6))^2)
}
average.inertiaFREQ <- (6/(6-1))*(sum(mca_FREQ$eig[,1])-
((length(levelsFREQ)-6)/6^2))
percentage_inert_FREQ <- NULL
for(i in 1:20){
  percentage_inert_FREQ[i] <- inert_FREQ[i]/average.inertiaFREQ*100
}
sum(percentage_inert_FREQ)

#### MCA kategorisoidut muuttujat
mca_pres <- MCA(pres, graph=F, method="Burt")
summary(mca_pres)

#### Datakehikko muuttujien koordinaateista
mca_pres_vars_df = data.frame(mca_pres$var$coord, Variable =
rep(names(cats), cats))
mca_pres$var$coord[591:615,1:5]
piirtoNimet <- rep(names(cats), cats)
jokidataVar = data.frame(mca_pres$var$coord[591:615,1:5], Variable =
piirtoNimet[591:615])
#### Datakehikko havaintojen koordinaateista
mca_pres_obs_df = data.frame(mca_pres$ind$coord)
jokidataObs = data.frame(mca_pres$ind$coord)

#### Kartan piirtäminen muuttujien tasoista
ggplot(data = jokidataObs, aes(x = Dim.1, y = Dim.2)) +
  theme_bw() +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(colour = "gray50",alpha = 0.1) +
  geom_text(data = jokidataVar,
            aes(x = Dim.1, y = Dim.2,
                label = rownames(jokidataVar), colour = Variable)) +
  ggtitle("MCA-kartta, pohjaeläimet kategorisoituina muuttujina") +
  labs(x = "Dim.1 (23,49%)", y = "Dim.2 (13,46%)") +
   scale_colour_manual(name = "Muuttuja", values =
   c("red","chartreuse4","orange","lightslateblue","magenta"))
```

```
#### Korjatut inertiat ja dimensioiden selitysasteet
inert_pres <- NULL
for(i in 1:175){
  #inert_pres[i] <- (mca_pres$eig[i,1]-(1/161))
  inert_pres[i] <- ((161/(161-1))^2)*((sqrt(mca_pres$eig[i,1])-(1/161))^2)
}
average.inertia <- (161/(161-1))*(sum(mca_pres$eig[,1])-((925-161)/(161^2)))
for(i in 1:175){
  percentage_inert_pres[i] <- inert_pres[i]/average.inertia*100
}
sum(percentage_inert_pres)

#### Vuosittain

#### 2006
mca_pres06 <- MCA(pres06, graph=F, method="Burt")
mca_pres06$eig

#### Datakehikko muuttujien koordinaateista
mca_pres06_vars_df = data.frame(mca_pres06$var$coord, Variable =
 rep(names(cats06), cats06))
piirtoNimet06 <- rep(names(cats06), cats06)
jokidataVar06 = data.frame(mca_pres06$var$coord[83:100,1:5], Variable =
 piirtoNimet06[83:100])

#### Datakehikko havaintojen koordinaateista
mca_pres06_obs_df = data.frame(mca_pres06$ind$coord)
jokidataObs06 = data.frame(mca_pres06$ind$coord)

#### Kartan piirtäminen muuttujien tasoista
ggplot(data = jokidataObs06, aes(x = Dim.1, y = Dim.2)) +
  theme_bw() +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(colour = "gray50", alpha = 0.1) +
  geom_text(data = jokidataVar06,
            aes(x = Dim.1, y = Dim.2,
                label = rownames(jokidataVar06), colour = Variable)) +
  ggtitle("MCA-kartta 2006") +
  labs(x = "Dim.1 (22,06%)", y = "Dim.2 (16,80%)") +
   scale_colour_manual(name = "Muuttuja", values =
   c("red","orange","lightslateblue","magenta"))
```

```
#### Korjatut inertiat ja dimensioiden selitysasteet
levels06 <- NULL
for(i in 1:length(pres06[1,])){
  levels06 <- c(levels06,levels(pres06[,i]))
}
inert_pres06 <- NULL
for(i in 1:87){
  #inert_pres06[i] <- (mca_pres06$eig[i,1]-(1/160))
  inert_pres06[i] <- ((160/(160-1))^2)*((sqrt(mca_pres06$eig[i,1])-(1/160))^2)
}
average.inertia06 <- (160/(160-1))*(sum(mca_pres06$eig[,1])-
((length(levels06)-160)/160^2))
percentage_inert_pres06 <- NULL
for(i in 1:87){
  percentage_inert_pres06[i] <- inert_pres06[i]/average.inertia06*100
}
```

⋮

Tästä poistettu koodit vuosien 2007-2012 analyyseista tilan säästämiseksi.

⋮

```
#### Inertia ja selitysasteet

#### Analyyseista saadut ominaisarvot/pääinertiat:
mca_FREQ$eig[1:10,1]
mca_pres$eig[1:10,1]
mca_pres06$eig[1:10,1]
mca_pres07$eig[1:10,1]
mca_pres08$eig[1:10,1]
mca_pres09$eig[1:10,1]
mca_pres10$eig[1:10,1]
mca_pres11$eig[1:10,1]
mca_pres12$eig[1:10,1]

#### Korjatut pääinertiat
inert_FREQ[1:10]
inert_pres[1:10]
inert_pres06[1:10]
inert_pres07[1:10]
inert_pres08[1:10]
inert_pres09[1:10]
inert_pres10[1:10]
inert_pres11[1:10]
inert_pres12[1:10]
```

```
#### Korjattujen pääinertioiden selitysosuudet (dimensioiden selitysosuudet)
percentage_inert_FREQ[1:10]
percentage_inert_pres[1:10]
percentage_inert_pres06[1:10]
percentage_inert_pres07[1:10]
percentage_inert_pres08[1:10]
percentage_inert_pres09[1:10]
percentage_inert_pres10[1:10]
percentage_inert_pres11[1:10]
percentage_inert_pres12[1:10]

#### Indikaattorilajit
#### Funktio, joka laskee euklidisenetäisyyden viiden dimension "tarkkuudella"
etaisyys <- function(piste1,piste2){
  x1<-piste1[1]
  y1<-piste1[2]
  z1<-piste1[3]
  u1<-piste1[4]
  k1<-piste1[5]
  x2<-piste2[1]
  y2<-piste2[2]
  z2<-piste2[3]
  u2<-piste2[4]
  k2<-piste2[5]
  euk<- sqrt((x1-x2)^2+(y1-y2)^2+(z1-z2)^2+(u1-u2)^2+(k1-k2)^2)
}

#### Kategorisoitujen muuttujien indikaattorilajit
etaisyydet <- matrix(nrow=18,ncol=310)
t <- 1
p <- 1
for(j in 598:615){
  for(k in 616:925){
    etaisyydet[t,p] <- etaisyys(as.vector(mca_pres_vars_df[j,c(1:5)],mode =
     "numeric"), as.vector(mca_pres_vars_df[k,c(1:5)],mode = "numeric"))
  p <- p+1
  print(p-1)
  }
  p <- 1
  t <- t+1
}
indikaattorilajit <- matrix(nrow = 18, ncol = 5, dimnames =
list(c("ESk","ESt","Kk","Ksa","Kt","Pk","Psa","Pt","Sk","Ssa",
"St","Etelä","Pohjoinen","h","iKi","pKi","ref_0","ref_1")))
indikaattorietaisyydet <- matrix(nrow = 18, ncol = 5, dimnames =
  list(c("ESk","ESt","Kk","Ksa","Kt","Pk","Psa","Pt","Sk","Ssa",
  "St","Etelä","Pohjoinen","h","iKi","pKi","ref_0","ref_1")))
```

```
#### Lisämuuttujien indikaattorilajit
mca_FREQ$quanti.sup$coord[,1]
etaisyydet_FREQ <- matrix(nrow=18,ncol=155)
t <- 1
p <- 1
for(j in 598:615){
  for(k in 1:155){
    etaisyydet_FREQ[t,p] <- etaisyys(as.vector(mca_pres_vars_df[j,c(1:5)],
    mode = "numeric"),as.vector(mca_FREQ$quanti.sup$coord[k,],
    mode = "numeric"))
    p <- p+1
    print(p-1)
  }
  p <- 1
  t <- t+1
}
indikaattorilajit_FREQ <- matrix(nrow = 18, ncol = 5, dimnames =
  list(c("ESk","ESt","Kk","Ksa","Kt","Pk","Psa","Pt","Sk","Ssa",
  "St","Etelä","Pohjoinen","h","iKi","pKi","ref_0","ref_1")))
indikaattorietaisyydet_FREQ <- matrix(nrow = 18, ncol = 5, dimnames =
 list(c("ESk","ESt","Kk","Ksa","Kt","Pk","Psa","Pt","Sk","Ssa",
 "St","Etelä","Pohjoinen","h","iKi","pKi","ref_0","ref_1")))

#### Havaintojen tarkastelu lisämuuttuja-analyysissa
havjt <- data.frame(jokidataObsFREQ,jokityyppi)
havpt <- data.frame(jokidataObsFREQ,pohjatyyppi)
havnv <- data.frame(jokidataObsFREQ,naytevuosi)
havref <- data.frame(jokidataObsFREQ,ref)
havpe <- data.frame(jokidataObsFREQ,PE)

varit <- c("#000000", "#E69F00", "#56B4E9", "#009E73",
"#F0E442", "#0072B2", "#D55E00", "#CC79A7",
"#999999", "#F99000", "#006633")

#### Jokityyppi
ggplot(data = havjt, aes(x = Dim.1, y = Dim.2)) +
  theme_bw() +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(aes(colour = jokityyppi),cex=3) +
  labs(x = "Dim.1 (9,94%)", y = "Dim.2 (5,73%)") +
  ggtitle("MCA-kartta, havainnot jokityypeittäin") +
  scale_colour_manual(name = "Jokityyppi", values = varit)
```

```
#### Pohjatyyppi
ggplot(data = havpt, aes(x = Dim.1, y = Dim.2)) +
  theme_bw() +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(aes(colour = pohjatyyppi),cex=3) +
  labs(x = "Dim.1 (9,94%)", y = "Dim.2 (5,73%)") +
  ggtitle("MCA-kartta, havainnot pohjatyypeittäin") +
  scale_colour_manual(name = "Pohjatyyppi", values =
  c("red", "black", "green"))

#### Näytevuosi
ggplot(data = havnv, aes(x = Dim.1, y = Dim.2)) +
  theme_bw() +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(aes(colour = naytevuosi),cex=3) +
  labs(x = "Dim.1 (9,94%)", y = "Dim.2 (5,73%)") +
  ggtitle("MCA-kartta, havainnot näytevuosittain") +
  scale_colour_manual(name = "Näytevuosi", values = varit)

#### Luonnontilaisuus
ggplot(data = havref, aes(x = Dim.1, y = Dim.2)) +
  theme_bw() +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(aes(colour = ref),cex=3) +
  labs(x = "Dim.1 (9,94%)", y = "Dim.2 (5,73%)") +
  ggtitle("MCA-kartta, havainnot luonnontilaisuuden mukaan") +
  scale_colour_manual(name = "ref", values = c("red", "black"))

#### Pohjois-etelä- jako
ggplot(data = havpe, aes(x = Dim.1, y = Dim.2)) +
  theme_bw() +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(aes(colour = PE),cex=3) +
  labs(x = "Dim.1 (9,94%)", y = "Dim.2 (5,73%)") +
  ggtitle("MCA-kartta, havainnot pohjois-etelä- jaon mukaan") +
  scale_colour_manual(name = "PE", values = c("red", "black"))

#### Pohjaeläinkuva
pohel = data.frame(mca_pres$var$coord[616:length(mca_pres$var$coord[,1]),1:5],
 Variable = piirtoNimet[616:length(piirtoNimet)])
```

```
ggplot(data = pohel, aes(x = Dim.1, y = Dim.2)) +
  xlim(-0.4,0.6) +
  theme_bw() +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(colour = "gray50",alpha = 0.1) +
  geom_text(data = pohel,
            aes(x = Dim.1, y = Dim.2,
                label = rownames(pohel)), cex = 3) +
  geom_text(data = jokidataVar,
            aes(x = Dim.1, y = Dim.2,
                label = rownames(jokidataVar), colour = Variable)) +
  ggtitle("MCA-kartta, pohjaeläimet kategorisina muuttujina") +
  labs(x = "Dim.1 (23,49%)", y = "Dim.2 (13,46%)") +
  scale_colour_manual(name = "Muuttuja", values =
  c("red","chartreuse4","orange","lightslateblue","magenta"))
```
