Sekaannusmatriisikorjaus ja sen vaikutus biologisiin indekseihin

Tommi Veistämö

26. kesäkuuta 2019

Tiivistelmä

Pohjaeläinten koneellinen tunnistaminen vähentäisi merkittävästi pohjaeläinten luokitteluun tarvittavaa työmäärää nykyiseen tilanteeseen verrattuna ja nopeuttaisi biologisten indeksien laskemista. Pohjaeläimistä laskettavat biologiset indeksit kertovat vesistöjen ekologisesta tilasta, joten niiden estimoiminen tarkasti on tärkeää. Koneellinen luokitteleminen aiheuttaa kuitenkin virheitä pohjaeläinten taksonomisten ryhmien tunnistamisessa, koska luokittelussa yksilö voi ominaisuuksiensa perusteella päätyä väärään taksonomiseen ryhmään. Näiden virheiden korjaamiseksi sovelletaan kolmea korjausmenetelmää: käyttäjän sekaannusmatriisikorjaus, tuottajan sekaannusmatriisikorjaus ja paras lineaarinen korjaus. Menetelmien toimivuutta tutkitaan 12 yleisen biologisen indeksin kohdalla. Tutkimuksen kohteena on luokitteluvirheen ja korjausmenetelmien vaikutus indeksien estimointiin.

Sekaannusmatriisista nähdään, mihin luokkiin pohjaeläimet kuuluvat ja mihin luokkiin ne on luokiteltu. Tätä tietoa voidaan käyttää luokittelun korjaamiseen aineistoissa, joissa oikeat luokat eivät ole tiedossa. Käyttäjän sekaannusmatriisikorjaus ja tuottajan sekaannusmatriisikorjaus ovat suhteellisen yksinkertaisia menetelmiä, kun taas paras lineaarinen korjaus on mutkikkaampi, lineaarista muunnosta optimoiva menetelmä. Menetelmät korjaavat luokittelun tuloksena saatavia pohjaeläinten taksonomisten ryhmien suhteellisia osuuksia. Näitä korjattuja osuuksia käyttäen voidaan laskea halutut biologiset indeksit.

Korjausten vaikutusta tutkitaan simulointikokeella, joka perustuu aikaisemmin toteutettuun pohjaeläinaineiston koneelliseen luokitteluun. Käyttäjän sekaannusmatriisikorjauksella saadaan lähes harhattomia arvoja pääosin kaikilla indekseillä. Menetelmä toimii myös, vaikka otoskoko olisi pieni ja luokittelija huonohko. Tuottajan sekaannusmatriisikorjaus ja paras lineaarinen korjaus vähentävät indeksien harhaa, mutta eivät niin hyvin kuin käyttäjän korjaus. Kaikki menetelmät eivät kuitenkaan toimi yhtä hyvin, jos sekaannusmatriisi on estimoitu erilaisesta populaatiosta kuin mihin korjausta käytetään. Tällöin paras lineaarinen korjaus on tarkin korjausmenetelmä, muttei kuitenkaan harhaton useimpien indeksien kohdalla. Käyttäjän sekaannusmatriisikorjaus on huonoin korjaus tällaisessa tapauksessa.

Käyttäjän sekaannusmatriisikorjaus on suositeltava vaihtoehto luokittelusta aiheutuvan harhan korjaamiseksi. Käyttäjän sekaannusmatriisikorjaus on lähes aina tutkituista menetelmistä paras harhan vähentämiseen ja poistaa parhaimmillaan kokonaan luokittelusta aiheutuvan harhan. Tosin käyttäjän sekaannusmatriisikorjaus toimii erinomaisesti vain, jos sekaannusmatriisi on estimoitu samanlaisesta populaatiosta kuin mitä tutkitaan. Muuten paras lineaarinen korjaus on suositeltava vaihtoehto luokittelusta aiheutuvan harhan korjaamiseksi.

Avainsanat: biologinen indeksi, koneellinen tunnistaminen, luokittelu, paras lineaarinen korjaus, pohjaeläin, sekaannusmatriisi, sekaannusmatriisikorjaus.

Sisältö

1 Johdanto

Vesistöjen ekologisen tilan tutkimiseksi vesistötutkijat seuraavat muun muassa pohjaeläinten laji- ja lukumääriä. Nämä jokien pohjassa elävät pieneliöt reagoivat vesistön olosuhteiden muutoksiin. Jonkin lajin lukumäärän suhteellinen muutos voi viitata vedenlaadun muutokseen ja esimerkiksi saasteiden lisääntymiseen vesistössä. Lajimäärien runsauden vuoksi on kehitetty biologisia indeksejä kuvaamaan pohjaeläinten ja niiden taksonomisten ryhmien lukumääriä, sekä niiden muutoksia. (Suomen ympäristökeskus, Aroviita, J. et al., 2012.)

Pohjaeläinten tunnistaminen näytteistä on hidasta ja aikaa vievää työtä asiantuntijaltakin. Tämän vuoksi koneellista tunnistamista on kehitetty ja tutkittu useissa töissä (Tirronen et al., 2009; Ärje et al., 2010; Kiranyaz et al., 2011; Ärje et al., 2013; Joutsijoki et al., 2014). Koneellisessa tunnistamisessa pohjaeläimestä mitataan muuttujia, joita käytetään päätteltäessä yksilön taksonomista ryhmää, eli luokiteltaessa pohjaeläintä. Koska koneen tekemä tunnistaminen on kuitenkin epätäydellistä, osa pohjaeläimistä luokitellaan väärään taksonomiseen ryhmään. Koneellisessa tunnistamisessa virheet ovat usein systemaattisia, ja näin ollen luokittelua voidaan parantaa tunnistamalla ja vähentämällä systemaattista virhettä. Parhaimmillaan tällainen korjaus parantaisi myös biologisten indeksien estimoinnin tarkkuutta, jota luokitteluvirhe voi heikentää.

Biologisen seurannan kannalta on tärkeää tietää, miten koneellinen luokittelu vaikuttaa biologisten indeksien arvoihin, sillä koneellista luokittelua ei kannata käyttää tilanteissa, joissa kiinnostavasta biologisesta indeksistä tulee liian harhainen. Ärje et al. (2017) havaitsivat joidenkin indeksien olevan herkempiä luokitteluvirheelle ja otoskoon vaihtelulle. He tutkivat erilaisten luokittelijoiden vaikutusta indeksien harhaan ja vaihteluun ilman korjausmenetelmiä. Muita merkittäviä sovellusaloja, joissa on tutkittu luokittelun lisäksi luokitteluvirheen vaikutusta indeksien laskemiseen, ovat esimerkiksi kaukokartoitus (Chen et al., 2010) ja tekstin tunnistus (Ciresan et al., 2011). Hess & Bay (1997) käyttivät kaukokartoituksessa bootstrap-menetelmää luottamusvälien laskemiseksi Simpsonin ja Shannonin monimuotoisuus -indekseille. Kuitenkaan useiden korjausmenetelmien vaikutusta erilaisiin indekseihin ei tietääksemme ole ennen tutkittu.

Tämän tutkimuksen tarkoituksena on selvittää, voidaanko pohjaeläinten luokittelua parantaa erilaisilla sekaannusmatriiseihin pohjautuvilla korjausmenetelmillä (Hay, 1998; Fortier, 1992; Card, 1982) ja arvioida minkälainen vaikutus tällä on pohjaeläimistä laskettuihin biologisiin indekseihin. Erityisesti kiinnostuksen kohteena on korjausmenetelmien vaikutus luokittelun jälkeisiin tunnuslukuihin, kuten biologisten indeksien harhaan ja vaihteluun.

Korjausmenetelmien vaikutus voi olla erilainen riippuen luokittelijan toimivuudesta, otos-

koosta tai jokityypistä. Tutkitaan, toimivatko korjausmenetelmät myös huonon luokittelijan kanssa vai tarvitseeko luokittelijan olla kohtuullisen hyvä, jotta korjausmenetelmästä olisi apua. Erilaisissa jokityypeissä taksonomisten ryhmien osuudet ovat erilaisia ja tämä voi vaikuttaa indeksien lisäksi myös korjausmenetelmien tarkkuuteen. Lisäksi tässä työssä tutkitaan otoskoon vaikutusta korjausmenetelmän toimintaan.

Tutkimuksessa käytettävä pohjaeläinaineisto on saatu Suomen ympäristökeskukselta. Aineisto koostuu 6585 kuvasta, joissa jokaisessa on yksi pohjaeläin. Erilaisia pohjaeläinten taksonomisia ryhmiä on aineistossa 32. Kolme asiantuntijaa ovat luokitelleet kuvissa olevat pohjaeläimet (Ärje et al., 2013), joten pohjaeläinten oikeat luokat ovat tiedossa. Aineistoa on käytetty myös aiemmissa tutkimuksissa (Ärje et al., 2013; Ärje et al., 2017).

Luokittelussa kone pyrkii tunnistamaan kuvasta pohjaeläimen lajin joidenkin piirteiden perusteella. Erilaiset geometriset piirteet (Duda et al., 2001), harmaasävyt (Trier et al., 1996) ja värisävyt (Drimbarean & Whelan, 2001) ovat selittäviä muuttujia. Jotta luokittelija tunnistaisi lajiluokkien ominaispiirteet, luokittelija koulutetaan aineiston avulla käyttäen jotain luokittelumenetelmää. Luokittelumenetelmiä on useita, kuten erilaiset päätöspuut, Bayes-luokittelijat ja lähimmän naapurin menetelmä. (Duda et al., 2001.) Tässä työssä käytetään kahta aikaisemmassa tutkimuksessa (Ärje et al., 2017) muodostettua sekaannusmatriisia. Nämä sekaannusmatriisit on estimoitu käyttäen kahta luokittelijaa, jotka ovat satunnainen metsä ja naiivi Bayes. Näistä ensimmäinen on kohtuullisen hyvä luokittelija ja jälkimmäinen huonohko.

Luokittelun onnistumista tutkitaan sekaannusmatriisien avulla. Siinä ennustettuja luokkia verrataan havaintojen todellisiin luokkiin. Saadusta matriisista nähdään oikein luokiteltujen osuus ja virheellisten luokittelujen osuus sekä laatu, eli mahdolliset luokittelun systemaattiset virheet. Sekaannusmatriisissa esitetään riveinä luokiteltu aineisto ja sarakkeina oikea tieto luokkiin kuulumisesta (Hess & Bay, 1997).

Sekaannusmatriisista voidaan laskea myös muita matriiseja, kuten käyttäjän ja tuottajan sekaannusmatriisit. Käyttäjän sekaannusmatriisi muodostetaan jakamalla sekaannusmatriisin solu rivisummallaan ja tuottajan sekaannusmatriisi saadaan käyttämällä sarakesummaa jakajana rivisumman sijasta. Näin ollen käyttäjän sekaannusmatriisista nähdään, mihin luokkaan havainnot oikeasti kuuluvat, kun ne on luokiteltu yhteen tiettyyn luokkaan. Tuottajan sekaannusmatriisista puolestaan nähdään, mihin luokkiin tietyn luokan havainnot luokittelualgoritmi luokittelee. (Hess & Bay, 1997.)

Koska sekaannusmatriiseissa on tarkka tieto luokittelun onnistumisesta, on houkuttelevaa käyttää tätä tietoa luokittelun parantamiseksi. Tässä tutkimuksessa selvitetään, saadaanko taksonomisten ryhmien osuudet estimoitua tarkemmin soveltamalla luokittelun tulok-

seen käyttäjän sekaannusmatriisikorjausta (Card, 1982) tai tuottajan sekaannusmatriisin käänteismatriisikorjausta (Hay, 1998). Kolmas tutkittava menetelmä on paras lineaarinen korjaus, joka minimoi muunnettujen estimaattien keskineliövirheen (Fortier, 1992).

Käytännössä tällaisia korjauksia varten tarvitaan sekaannusmatriisin estimaattori. Sekaannusmatriisin estimoimiseksi tarvitaan tieto sekä havaintoyksikön oikeasta luokasta että ennustetusta luokasta. Näin ollen käytettäessä yhtä aineistoa havaintoyksiköiden oikeat luokat olisivat tiedossa, eikä luokittelua tarvittaisi lainkaan. Tavoitteena onkin estimoida sekaannusmatriisi yhdestä aineistosta, jolloin sitä voidaan käyttää muista aineistoista tehtyjen luokittelujen korjaamiseen. Jos käytössä on vain yksi aineisto, voidaan sekaannusmatriisi estimoida osa-aineistosta, jolloin loput aineistosta voidaan luokitella ja luokittelu korjata sekaannusmatriisia käyttäen ilman tarvetta selvittää oikeita luokkia kaikista havainnoista (Hess & Bay, 1997).

Korjausmenetelmien vaikutus erilaisiin biologisiin indekseihin voi olla hyvinkin monimutkainen. Indeksien keskinäisten erojen vuoksi jokin korjausmenetelmä voi parantaa tietyn indeksin estimointia, mutta toiselle indeksille menetelmä voi lisätä harhaa. Näin ollen on järkevää tutkia kunkin korjausmenetelmän vaikutusta yhteen indeksiin kerrallaan. Jotkin indeksit ovat myös robusteja otoskoon vaikutukselle, kun taas joidenkin indeksien arvoihin otoskoko vaikuttaa suoraviivaisesti (Magurran, 2004). Tuloksia tulkittaessa on siten otettava huomioon otoskoko ja sen vaikutus.

Tässä työssä tutkittavat biologiset indeksit luokitellaan neljään eri tyyppiin: lajirikkautta, monimuotoisuutta, tasaisuutta/vallitsevuutta ja samankaltaisuutta mittaaviin (Magurran, 2004). Tähän tutkimukseen valitaan osittain samoja indeksejä, joita on käytetty työssä Ärje et al. (2017). Lajirikkautta mittaavia indeksejä ovat Margalefin monimuotoisuus, Chaon estimaattori lajimäärälle ja omana indeksinään lajien lukumäärä. Monimuotoisuutta mittaavia ovat Shannonin indeksi ja Simpsonin indeksi. Tasaisuuteen/vallitsevuuteen liittyy Berger-Parkerin indeksi. (Magurran, 2004.) Sørensenin samankaltaisuus, Canberran metriikka, euklidinen samankaltaisuus, Morisita-Hornin indeksi (Wolda, 1981), PMA-indeksi (Percent model affinity, Renkonen, 1938; Novad & Bode, 1992) ja Jaccardin samankaltaisuuskerroin (Jaccard, 1901) ovat puolestaan kahden näytteen samankaltaisuutta mittaavia indeksejä.

Tutkielman rakenne on seuraava: luvussa 2 käydään läpi koneellisen tunnistamisen vaiheet. Luvussa 3 tutustutaan sekaannusmatriisien teoriaan ja käydään läpi tuottajan ja käyttäjän sekaannusmatriisien väliset erot. Varsinaisten korjausmenetelmien teoriaa käydään lävitse luvussa 4. Näiden korjausten soveltamista pohjaeläinaineistoon käsitellään 5. luvussa. Samassa luvussa esitellään aineisto, biologiset indeksit, simulointiasetelma ja käsitellään

varsinaiset tulokset. Luvussa 6 tehdään johtopäätökset tuloksista ja pohditaan millaiseen jatkotutkimukseen olisi aihetta.

2 Kuvasta luokitteleminen - koneellisen tunnistamisen vaiheet

Koneellisessa tunnistamisessa jokin kohde tunnistetaan tiettyyn luokkaan kuuluvaksi. Tunnistamista tehdään koneellisesti monista syistä: usein tunnistaminen on nopeampaa koneellisesti kuin ihmisen tekemänä. Joissain luokittelutilanteissa kone pystyy tarkkuuteen, johon ihminen ei kykene. Työ voi myös olla raskasta, epämiellyttävää tai väsyttävää, jolloin ihmisten suorittamaan luokitteluun tulee herkästi virheitä. Luokittelu asiantuntijan tekemänä ei aina ole edes mahdollista käytännön syistä, esimerkiksi silloin kun matkapuhelimen halutaan reagoivan äänikomentoihin ilman viivettä.

Luokittelun kohteena voi olla lähtökohtaisesti minkälainen asia tahansa, kuten esine, olento, ääni, hiukkasen hajoamisen lopputuotteet ja niin edelleen. Luokiteltava asia täytyy ensin saada skannattua tietokoneelle ymmärrettävään muotoon. Usein luokiteltavat kohteet ovat fyysisiä kappaleita, jotka kuvataan kameralla. (Duda et al., 2001.) Tässä työssä keskitytään jatkossa kuvasta tehtävään luokitteluun.

Koneellisen kuvasta tunnistamisen vaiheita ovat kuvaus, segmentointi, piirteiden määritteleminen, luokittelu ja jälkikäsittely. Ensin kuvaamisessa ja segmentoinnissa haluttu kohde rajataan koneen tunnistamaan muotoon. Siitä kone määrittelee tietyille piirteille arvot ja käyttää näitä arvoja kohteen luokittelemiseksi johonkin luokkaan. Lopuksi arvioidaan luokittelun onnistuminen halutussa tutkimusongelmassa. (Duda et al., 2001.)

Kuvan täytyy olla tarpeeksi selkeä, jotta kohteen erityispiirteet näkyvät kuvassa. Valaistuksen täytyy olla kuvissa samankaltaista tai muuten vaarana on, että valaistus itsessään tai kohteiden varjot vaikuttavat kohteiden piirteiden tulkitsemiseen myöhemmissä vaiheissa. (Duda et al., 2001.)

Toinen vaihe on segmentointi, jolloin kuvissa olevat kiinnostavat kohteet erotellaan toisistaan omiksi kokonaisuuksikseen. Kohteiden erottelun hankaluus on siinä, että kohteita ei ole vielä millään tavalla luokiteltu, joten segmentoinnissa kohde pyritään tunnistamaan ja erottelemaan ilman tietoa kohteen laadusta tai ominaisuuksista. Usein kohteet voivat olla kuvissa päällekkäin tai lomittain, jolloin kohdetta ei edes voida täydellisesti erotella. (Duda et al., 2001.) Esimerkiksi pohjaeläinten tapauksessa eliöt voivat olla kiinni toisissaan, päällekkäin tai niiden raajat voivat näkyä kuvissa huonosti. Siitä huolimatta jokainen pohjaeläin pitäisi saada eriteltyä toisistaan kokonaisena yksilönä, jotta luokittelu onnistuu.

Pal & Pal (1993) mainitsee, että segmentointiin on olemassa jopa satoja menetelmiä, mutta suurin osa menetelmistä noudattaa samaa periaatetta. Kuva pyritään jakamaan mahdol-

lisimman homogeenisiin osiin ja usein vierekkäisten osien halutaan poikkeavan toisistaan merkittävästi, toisin sanoen osat erottaa toisistaan selkeä reuna. Saadut osat jaetaan tärkeisiin ja turhiin. Turhat hylätään ja tärkeät otetaan jatkokäsittelyyn.

Kun haluttu kohde on rajattu omaan kuvaansa, seuraavaksi halutaan mitata kohteen ominaispiirteet. Tätä kutsutaan piirteiden erottelemiseksi. Tarkoituksena on löytää kohteista piirteitä, jotka erottelisivat kohteet mahdollisimman hyvin toisistaan, eivätkä kärsisi kuvauksesta aiheutuneesta kohinasta. Mitä piirteitä käytetään ja miten nämä piirteet mitataan riippuu vahvasti tutkimusongelmasta. (Duda et al., 2001.) Piirteiden erottelemiseksi on kehitelty runsaasti erilaisia menetelmiä (esim. Trier et al., 1996).

Sovellusalasta ja ongelmasta riippumatta voidaan pohtia, millaisia piirteiden pitäisi olla. Trier et al. (1996) listaavat näiden piirteiden tärkeimmät ominaisuudet. Piirteiden pitäisi olla invariantteja. Tämä tarkoittaa sitä, että riippumatta millä tavalla edellä mainitut kuvaus ja segmentointi on tehty, piirteet pitäisi pystyä tunnistamaan ja mittaamaan tarkasti. Toisin sanoen ominaisuuksien täytyisi säilyä samana riippumatta siitä, missä kohdassa tai minkä kokoisena kohde on kuvassa. Kohde voi olla, tai se saattaa olla myöhemmin, skaalattu, venytetty, käännetty, vinoutettu tai käännetty täysin peilikuvakseen. Yleistä on myös, että kohde on pyörähtänyt jonkin verran ja on näin eri kulmassa yleiseen tapaukseen nähden. Näistä seikoista huolimatta piirteet pitäisi pystyä mittaamaan oikein. Mikäli emme voi luottaa piirteiden olevan invariantteja edellä mainittujen tapausten suhteen, kuvia tai kohteita voidaan standardoida edellä määriteltyjen ominaisuuksien, kuten koon tai asennon suhteen.

Kuitenkaan kaikkia piirteitä ei aina voida mitata. Esimerkiksi kohde saattaa olla kuvassa sellaisessa kulmassa, että kohteen tiettyä osaa ei pystytä erottamaan. Kohde voi olla myös jollain tavalla epämuodostunut. Esimerkiksi seinäkello on kello, vaikka siitä puuttuisi sekuntiviisari. Kuitenkaan tämä kello ei enää täyttäisi piirrettä: “kellossa on kolme viisaria”. Vastaavasti jos piirteenä olisi sekuntiviisarin pituus, tätä ei pystyttäisi edellä mainitussa tapauksessa mittaamaan. Tämä ilmenee puuttuvana tietona piirteiden määrittelyssä (Duda et al., 2001).

Luokittelussa pyritään arvioimaan minkälaisen mallin mukaisesti aineisto on muodostunut. Luokittelija käyttää havaintoja tehdäkseen päätöksen luokkaan kuulumisesta. Prosessissa voidaan käyttää hyväksi myös prioritietoja. Ennakkotietona voi olla esimerkiksi, että yksi luokka on yleisempi kuin toinen tai että jonkin piirteen suhteen luokat poikkeavat tietyllä tavalla toisistaan. (Duda et al., 2001.)

Käytännössä havaitaan piirrevektori $\mathbf{x} = [x_1, x_2, \ldots, x_p]^T$, missä x_i kuvastaa tietyn tilastoyksikön piirrettä i, ja p on piirteiden lukumäärä. Periaatteessa mallissa voi olla kuinka

paljon piirteitä tahansa, mutta käytännössä jos aineiston koko on pieni verrattuna luokkien tai piirteiden määrään, suuridimensioisia malleja ei saada estimoitua. Varsinainen luokittelija pyrkii jakamaan $\mathbb{R}^p$-ulotteisessa piirreavaruudessa olevat havainnot mahdollisimman selvärajaisiin luokkiin. Mitä suurempaa hajonta on luokkien sisällä verrattuna luokkien väliseen hajontaan, sitä vaikeampaa on luokittelu. (Theodoridis et al., 2008.)

Luokittelijan opettamista varten aineisto jaetaan kahteen osaan: opetus- ja testiaineistoon. Opetusaineistolla rakennetaan luokittelijat ja käyttämällä näitä luokitteluja testiaineistoon saadaan selville mitkä tapaukset luokitellaan väärin testiaineistossa. (Theodoridis et al., 2008.) Väärinluokiteltujen osuus on yksinkertaisimpia testivirheen estimaatteja. Aineisto voidaan jakaa myös kolmeen osaan: opetus-, validointi- ja testiaineistoon. Tässä validointiaineistolla optimoidaan estimoitavat parametrit, jotta vältetään parametrien ylisovittuminen opetusaineistoon (Duda et al., 2001).

Luokitteluun on useita menetelmiä kuten Bayes-luokittelijat, päätöspuut ja lähimmän naapurin menetelmät. Usein menetelmien parametrit estimoidaan suurimman uskottavuuden menetelmällä. Tarkemmin menetelmistä voi lukea kirjallisuudesta (esimerkiksi Duda et al., 2001; Theodosidis et al., 2008).

Luokittelun tuloksena malli voi myös sopia liian hyvin aineistoon. Tällöin luokittelu ei juurikaan huomioi satunnaisvaihtelua aineiston synnyssä. Näin ollen mallin yleistäminen muihin aineistoihin on haastavaa, ja mallin parametrien arvot voivat vaihdella huomattavasti aineistosta riippuen. Yleisesti mallin monimutkaistuessa luokitteluvirhe pienenee tiettyyn pisteeseen asti, jonka jälkeen virhe alkaa taas kasvaa mallin sopiessa liian hyvin tiettyyn aineistoon. (Theodoridis et al., 2008.)

Jatkokäsittelyssä on yhden tai useamman luokittelijan tulokset, joita verrataan keskenään ja päätetään tulosten pohjalta parhaiten tutkimusongelmaan sopiva malli. Käytettäessä useita luokittelijoita pyritään arvioimaan paras mahdollinen luokittelija, joka minimoi luokitteluvirheen (Duda et al., 2001). Pelkän luokitteluvirheen lisäksi voidaan pohtia väärinluokittelusta aiheutuvaa riskiä. Joidenkin kohteiden luokitteleminen toiseen luokkaan on harmitonta, kun taas jokin väärinluokittelu voi aiheuttaa suuren tappion. Esimerkiksi pähkinälajin luokitteleminen väärään pussiin voi aiheuttaa kuluttujassa allergisen reaktion. Painotuksilla voidaan pyrkiä välttämään erityisen haitallisia väärinluokitteluja, mutta yleensä tällainen painottaminen heikentää luokittelun onnistumista kokonaisuudessaan (Duda et al., 2001).

Varsinaisen luokittelun ja jatkokäsittelyn välinen raja on häilyvä. Luokittelua voidaan pyrkiä parantamaan korjausmenetelmillä. Erilaisia luokittelumalleja ja korjauksia käytettäessä on tärkeintä muistaa, mikä on luokittelun varsinainen tavoite: halutaanko yksittäiset

kohteet luokitella mahdollisimman tarkasti, ovatko luokkien suhteelliset osuudet tärkeintä, vai käytetäänkö luokittelua jossain myöhemmässä laskutoimituksessa, esimerkiksi jonkin indeksin laskemisessa. Jälkikäsittelyssä ja mallien arvioinnissa on tärkeintä ymmärtää luokittelun vaikutus kyseessä olevaan ongelmaan ja pohtia luokittelun tarkkuutta kyseisessä kontekstissa. Joskus luokittelu saattaa olla laskennallisesti raskas, jolloin saattaa olla parempi käyttää laskennallisesti helpompaa mallia, vaikka luokittelu ei sillä onnistuisikaan yhtä tarkasti (Duda et al., 2001).

Jos jokin tunnistamisen vaihe epäonnistuu, sitä on myöhemmin vaikea korjata. Mikäli kuvaus ja segmentointi onnistuvat, luokittelua voidaan kokeilla eri malleilla käyttäen erilaisia piirteitä luokittelussa. Tutkimusongelmissa ei aina ole selkeää parasta mahdollista luokittelua, mutta tuloksia voidaan arvioida muun muassa luokitteluvirheen ja odotetun tappion perusteella ja huomioimalla luokitteluvirheen aiheuttama harha mahdollisissa jatkolaskennoissa. Arvioinnissa on oleellista tietää, minkä luokkien yksilöt luokitellaan vääriin luokkiin. Tämä tieto on sekaannusmatriisissa.

3 Sekaannusmatriisit

Luokittelun jälkeen tiedossa on kuhunkin luokkaan luokiteltujen tilastoyksiköiden osuudet. Jos lisäksi tiedetään kohteiden oikeat luokat, nämä tiedot voidaan koota sekaannusmatriisiksi. Siinä sarakkeet kuvaavat oikeita luokkia ja rivit luokittelun mukaisia luokkia. Matriisin solut kertovat, kuinka suuri osa havainnoista on luokiteltu tähän luokkaan, kun oikea luokka tiedetään.

Oletetaan, että populaatiossa tilastoyksiköt kuuluvat toisensa poissulkeviin luokkiin. Määritellään vektori $\mathbf{y} = (y_1, y_2, \ldots, y_k)$, missä k on luokkien lukumäärä. Tällöin y_i on luokkaan i kuuluvien tilastoyksiköiden lukumäärä otoksessa. Määritellään myös vektori $\mathbf{p} = (p_1, p_2, \ldots, p_k)$, jossa p_i on todennäköisyys, että satunnainen tilastoyksikkö kuuluu luokkaan i. Oletetaan, että luokkien lukumäärävektori $\mathbf{y}$ noudattaa multinomijakaumaa

$$(y_1, y_2, \ldots, y_k) \sim Multinom(N, p_1, p_2, \ldots, p_k),$$

missä N on otoksen kaikkien tilastoyksiköiden lukumäärä luokasta riippumatta. Tämä pätee kuitenkin vain, mikäli aineisto on hankittu satunnaisotannalla (Green, 1993).

Epätäydellisen luokittelun tuloksena saadaan vektori $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k)$, jossa $\tilde{y}_i$ on luokkaan i luokiteltujen havaintojen lukumäärä. Tämä vektori havaitaan luokittelun seurauksena, mutta luokittelun ollessa epätäydellistä sen alkiot eroavat vektorin $\mathbf{y}$ alkioista, jotka siis sisältävät oikeat lukumäärät. Luokittelun tuloksena saatavat luokkien lukumäärät noudattavat myös multinomijakaumaa

$$(\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k) \sim Multinom(N, p_1^*, p_2^*, \ldots, p_k^*),$$

jossa vektori $\mathbf{p}^* = (p_1^*, p_2^*, \ldots, p_k^*)$ sisältää todennäköisyydet p_i^* luokitella tilastoyksikkö luokkaan $i = 1, \ldots, k$, ja k on luokkien lukumäärä. Vektorien $\mathbf{p}$ ja $\mathbf{p}^*$ välistä yhteyttä käsitellään myöhemmin.

Luokittelu on harvoin virheetöntä. Kootaan seuraavaksi tieto virheiden todennäköisyyksistä sekaannusmatriisiin $\mathbf{P}$. Määritellään matriisi

$$\mathbf{P} = \begin{bmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ldots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{bmatrix},$$

missä $p_{ij} = P(\text{luokiteltu luokka on } i \ \& \text{ oikea luokka on } j)$. Sarakesumma $\Sigma_{i=1}^{k} p_{ij} = p_{.j}$ on todennäköisyys kuulua luokkaan j, $j = 1, \ldots, k$. Tätä merkittiin edellä symbolilla p_i.

Vastaavasti rivisumma $\Sigma_{j=1}^{k} p_{ij} = p_{i.}$ on luokitellun eli ennustetun luokan todennäköisyys p_i^*. Täydellisen luokittelun tapauksessa $\mathbf{P} = \mathbf{I}$, missä $\mathbf{I}$ on identiteettimatriisi.

Hay (1998) nostaa esille kolme tärkeintä sekaannusmatriisin käyttötapaa: yleisen tarkkuuden määrittäminen, ali- tai yliestimoinnin huomioiminen sekä tiettyjen virheiden (kuten myöhemmin esiteltävien tuottajan ja käyttäjän virheiden) määritteleminen. Sekaannusmatriisista nähdään, mitä luokkia määritellään väärin. Näin ollen sekaannusmatriisin diagonaalialkioiden summasta saadaan oikein luokiteltujen havaintojen osuus. Usein jotkin luokat menevät sekaisin keskenään tai johonkin luokkaan määritellään liian paljon ja vastaavasti johonkin luokkaan liian vähän havaintoja. Tällaisen systemaattisen virheen tapauksessa luokittelua voidaan korjata sekaannusmatriisin sisältämän tiedon avulla. Koska kokonaisvirhe ei anna tietoa luokitteluvirheen tyypistä, se voi olla epäinformatiivinen. Tilanteesta riippuen saattaakin olla mielekkäämpää tarkastella tuottajan tai käyttäjän sekaannusmatriisia, joihin paneudutaan seuraavaksi.

3.1 Tuottajan sekaannusmatriisi

Tuottajan sekaannusmatriisi saadaan tavallisesta sekaannusmatriisista $\mathbf{P}$. Tuottajan sekaannusmatriisin solun arvo kertoo todennäköisyyden, että luokkaan j kuuluva kohde luokitellaan luokkaan i. Toisin sanoen matriisin sarakkeesta voidaan lukea, kuinka monta prosenttia luokan havainnoista luokitellaan oikein ja mihin luokkiin havaintoja väärinluokitellaan. (Hess & Bay, 1997.)

Olkoon tuottajan sekaannusmatriisi $\mathbf{C} = [c_{ij}]$. Usein puhutaan myös tuottajan virheestä, joka määritellään $c_{ij} = P(\text{luokiteltu luokka} = i \mid \text{oikea luokka} = j)$, missä $i, j = 1, 2, \ldots, k$, ja k on luokkien lukumäärä (Green, 1993). Sekaannusmatriisin $\mathbf{P}$ avulla määriteltynä tämä virhe on

$$c_{ij} = \frac{p_{ij}}{p_{.j}}.$$

Näin määriteltynä tuottajan sekaannusmatriisi on

$$\mathbf{C} = \begin{bmatrix} c_{11} & \ldots & c_{1k} \\ \vdots & \ldots & \vdots \\ c_{k1} & \ldots & c_{kk} \end{bmatrix},$$

jossa jokaiselle luokalle j, todennäköisyys $c_{ij} \geq 0$ ja $\Sigma_i c_{ij} = 1$, joten sekaannusmatriisin $\mathbf{C}$ sarakkeen arvot summautuvat ykköseksi.

Aikaisemmin määriteltiin $\tilde{\mathbf{y}} \sim \text{Multinom}(N, p_1^*, \ldots, p_k^*)$. Määritellään nyt epätäydellisen

luokittelun todennäköisyydet $\mathbf{p}^*$ tuottajan sekaannusmatriisin avulla seuraavasti (Healy, 1981):

$$\mathbf{p}^* = \mathbf{C}\mathbf{p}. \tag{1}$$

Tuottajan sekaannusmatriisin todennäköisyydet ovat nimensä mukaisesti tuottajan tarvetta vastaavia. Tällä tarkoitetaan sitä, että usein luokittelijan tuottaja haluaa tietää, kuinka hyvin pystytään luokittelemaan halutut kohteet (Story & Congalton, 1986). Esimerkiksi luokiteltaessa puita vietäväksi sahalle luokittelun tekijää kiinnostaa, kuinka hyvin tietyt puulajit voidaan ylipäänsä luokitella.

3.2 Käyttäjän sekaannusmatriisi

Käyttäjän sekaannusmatriisi kuvastaa todennäköisyyttä, että luokkaan i luokiteltu kohde kuuluu oikeasti luokkaan j. Kun tiedetään luokiteltu luokka, millä todennäköisyydellä havainto kuuluu myös siihen luokkaan ja millä todennäköisyydellä johonkin muuhun luokkaan. Käyttäjän virhe määritellään vastaavasti kuin tuottajan virhe, mutta nyt kiinnostaa todennäköisyys $P(\text{oikea luokka} = j \mid \text{luokiteltu luokka} = i)$ (Green, 1993).

Olkoon käyttäjän sekaannusmatriisi $\mathbf{U} = [u_{ij}]$, missä $i, j = 1, 2, \ldots, k$, ja k on luokkien lukumäärä kuten aikaisemminkin. Tällöin matriisin

$$\mathbf{U} = \begin{bmatrix} u_{11} & \ldots & u_{1k} \\ \vdots & \ldots & \vdots \\ u_{k1} & \ldots & u_{kk} \end{bmatrix}$$

yksittäinen solu voidaan määritellä sekaannusmatriisin $\mathbf{P}$ avulla

$$u_{ij} = \frac{p_{ij}}{p_{i.}},$$

ja samaan tapaan kuin tuottajan sekaannusmatriisissa, myös käyttäjän sekaannusmatriisissa $u_{ij} \geq 0$, ja $\Sigma_j u_{ij} = 1$.

Havaintovektori $\tilde{\mathbf{y}}$ määriteltiin $\tilde{\mathbf{y}} \sim \text{Multinom}(N, p_1^*, \ldots, p_k^*)$. Luokittelutodennäköisyydet voidaan määritellä käyttäjän sekaannusmatriisilla seuraavasti (Hess & Bay, 1997):

$$\mathbf{p}^{*T} = \mathbf{p}^T \mathbf{U}^{-1}. \tag{2}$$

Käyttäjän sekaannusmatriisin todennäköisyyksien ajatellaan vastaavan käyttäjän tarvetta. Käyttäjä haluaa luokittelun vastaavan todellisuutta mahdollisimman tarkasti. Näin ollen

käyttäjän sekaannusmatriisi mittaa, kuinka hyvin luokittelun tuloksiin voidaan luottaa. (Story & Congalton, 1986.) Esimerkiksi sahan omistajan on pystyttävä luottamaan, että tilattu tavara vastaa luvattua, eli tiettyyn luokkaan luokiteltu puulaji on oikeasti kyseistä lajia.

Tuottajan sekaannusmatriisista voidaan havaita luokat, joiden tilastoyksiköt usein luokitellaan virheellisesti muihin luokkiin. Siinä mitataan siis eräänlaista puuttumista, kuinka moni havainto on virheellisesti luokiteltu muuhun luokkaan ja näin ollen puuttuu oikeasta luokasta. Käyttäjän sekaannusmatriisista puolestaan nähdään, mihin luokiteltuihin luokkiin tulee liian vähän tai liian paljon havaintoja muista luokista. Näin ollen virhettä syntyy, kun luokkaan kuulumattomia havaintoja luokitellaan ylimääräisenä tiettyyn luokkaan. (Story & Congalton, 1986.)

Käyttäjän ja tuottajan virheet liittyvät laskennallisesti toisiinsa. Nimittäin käyttäjän virhe voidaan laskea kaavalla

$$u_{ij} = \frac{p_i c_{ij}}{p_i^*}.$$

Tässä u_{ij} on käyttäjän virhe luokittelun ollessa i ja oikean luokan j, p_i on luokkaan i kuulumisen todennäköisyys, p_i^* on luokkaan i luokittelemisen todennäköisyys ja c_{ij} tuottajan virhe oikean luokan ollessa j ja luokittelun ollessa i. (Card, 1982.)

3.3 Sekaannusmatriisin estimointi

Käytännössä luokkiin kuulumisen todennäköisyydet estimoidaan aineistosta käyttäjän ja tuottajan sekaannusmatriisien avulla. Tätä estimointia varten määritellään, että varsinainen havaittu todennäköisyys luokitella havainto luokkaan i on

$$\tilde{p_i} = \frac{\tilde{y_i}}{N},$$

missä N on otoskoko (Fortier, 1992).

Edellä määriteltyä vektoria $\mathbf{p}^*$ ei havaita aineistosta, mutta se voidaan estimoida havaintojen avulla, sillä $E(\tilde{\mathbf{p}}|\mathbf{p}) = \mathbf{p}^*$. (Healy, 1981). Tätä tulosta voidaan hyödyntää kaavoja (1) ja (2) käytettäessä, jolloin vektoria $\mathbf{p}^*$ approksimoidaan vektorilla $\tilde{\mathbf{p}}$.

Määritellään sekaannusmatriisista aineiston avulla saatava estimaattori (taulukko 1) mukaillen artikkelia Prisley & Smith (1987). Olkoon havaittu sekaannusmatriisi $\mathbf{A} = [a_{ij}]$, missä $i, j = 1, 2, \ldots, k$, ja k on luokkien lukumäärä. Sekaannusmatriisin solu a_{ij} on luokkaan j kuuluvien kohteiden lukumäärä, jotka on luokiteltu luokkaan i. Näin ollen rivisumma kertoo, kuinka monta havaintoa luokiteltiin tiettyyn luokkaan i, ja sarakesummista nähdään,

Taulukko 1: Sekaannusmatriisi $\mathbf{A}$, jossa a_{ij} on luokkaan i luokiteltujen havaintojen lukumäärä, kun oikea luokka on j.

		Oikea luokka				Yhteensä
		$j = 1$	$j = 2$	...	$j = k$	
Luokittelun tulos	$i = 1$	a_{11}	a_{12}	...	a_{1k}	$N_{1.}$
	$i = 2$	a_{21}	a_{22}	...	a_{2k}	$N_{2.}$
	...	⋮	⋮	⋮	⋮	⋮
	$i = k$	a_{k1}	a_{k2}	...	a_{kk}	$N_{k.}$
	Yhteensä	$N_{.1}$	$N_{.2}$	...	$N_{.k}$	N

kuinka suuria luokat ovat oikeasti. Matriisin diagonaalilla ovat oikein luokiteltujen havaintojen lukumäärät ja ei-diagonaalilla olevat alkiot ovat väärinluokiteltuja tilastoyksiköitä.

Aineisto jaetaan luokittelua varten opetus- ja testiaineistoksi, ja opetusaineisto voidaan erikseen jakaa vielä opetus- ja validointiaineistoon. Useimmiten sekaannusmatriisi estimoidaan koko testiaineistosta, esimerkiksi Schuldt et al. (2004) ja Csurka et al. (2004). Testiaineisto muodostetaan alkuperäisestä aineistosta erottamalla tietynkokoinen osa havainnoista, joita ei käytetä luokittelijan muodostamiseen. Testiaineistolla rakennetaan sekaannusmatriisi sekä tutkitaan luokittelun onnistumista. Luokittelua ja sekaannusmatriisin muodostamista varten jokaisen tilastoyksikön oikea luokka on tiedossa, joten luokittelua ei tähän aineistoon varsinaisesti tarvita. Sekaannusmatriisia voidaankin käyttää muista otoksista tehtyjen luokittelujen korjaamiseen, kunhan alkuperäisissä populaatioissa todennäköisyydet kuulua eri luokkiin ($\mathbf{p}$) ovat samat (Fortier, 1992).

Fielding & Bell (1997) listaavat useita tapoja aineiston jakamiseen. Yksinkertaisin menetelmä on muodostaa testiaineisto satunnaisotannalla koko aineistosta. Toinen tapa on käyttää ristiinvalidointia, jossa aineisto jaetaan l osaan ja vain yhtä osaa käytetään testiaineistona. Muut osat yhdistetään opetusaineistoksi. Luokittelu voidaan tehdä l kertaa käyttäen jokaisessa luokittelussa eri opetus- ja testiaineistoa. Näin varmistetaan, että luokittelija on koulutettu tarpeeksi suurella aineistolla, mutta myös se, että testiaineistoa on riittävästi luokittelutarkkuuden mittaamiseen. Esimerkiksi Ravi et al. (2005) jakoivat aineiston kymmeneen osaan tutkiessaan ihmisten aktiivisuutta kiihtyvyysmittarilla. Tällöin opetusaineiston koko oli 90 prosenttia koko aineistosta, luokittelu tehtiin kymmenen kertaa ja tuloksista laskettiin keskiarvot kymmenen luokittelun suhteen.

Välttämättä jakoa opetus- ja testiaineistoon ei tarvitse tehdä, vaan sekä luokittelu että

testaaminen voidaan tehdä täsmälleen samalla aineistolla. Tämä ei kuitenkaan johda hyviin tuloksiin, sillä sekaannusmatriisi on luotu samalla aineistolla kuin millä luokittelu on tehty ja näin ollen luokittelija on yleensä ylisovittunut testattavaan aineistoon. Tällöin luokittelutarkkuus vaikuttaa suuremmalta kuin mitä se oikeasti on. Edellistä tapaa voitaisiin parantaa prospektiivisellä otannalla. Tässä ensin käytettäisiin koko aineisto luokittelijan kouluttamiseen ja sen jälkeen kerättäisiin uusi aineisto, jota käytettäisiin testiaineistona. (Fielding & Bell, 1997.)

Aina ei ole välttämätöntä käyttää koko testiaineistoa sekaannusmatriisin luomiseen. Hess & Bay (1997) esittelevät kaukokartoituksessa käytetyn menetelmän, jossa luokitelluista havainnoista otetaan otos ja otokseen päätyneiden havaintojen oikeat luokat selvitetään. Toisin sanoen sekaannusmatriisi muodostetaan luokiteltujen havaintojen otoksesta. Otos on saatu joko satunnaisotannalla tai ositetulla otannalla siten, että rivisummat ovat kiinnitettyjä. Sekaannusmatriisin estimointi tällä tavalla on kannattavaa, jos käytössä on vain yksi aineisto, jonka oikeiden luokkien selvittäminen on hankalaa ja halutaan säästää resursseja.

Estimoidusta sekaannusmatriisista voidaan laskea tuottajan sekaannusmatriisi, kuten taulukossa 2. Healyn (1981) mukaisesti sekaannusmatriisin solu $\hat{c}_{ij}$ on todennäköisyys määritellä luokkaan j kuuluva kohde luokkaan i. Tuottajan sekaannusmatriisin solut $\hat{c}_{ij}$ saadaan alkuperäisestä sekaannusmatriisista jakamalla sekaannusmatriisin solut sarakesummillaan:

$$\hat{c}_{ij} = \frac{a_{ij}}{N_{.j}},$$

missä $N_{.j}$ on luokkaan j kuuluvien havaintojen summa.

Taulukko 2: Estimaattori tuottajan sekaannusmatriisille $\hat{\mathbf{C}}$ havaitun sekaannusmatriisin $\mathbf{A}$ avulla ilmaistuna.

		Oikea luokka				Yhteensä
		$j = 1$	$j = 2$	$\dots$	$j = k$	
Luokittelun tulos	$i = 1$	$\hat{c}_{11} = \frac{a_{11}}{N_{.1}}$	$\hat{c}_{12}$	$\dots$	$\hat{c}_{1k}$	$\Sigma_{j=1}^{k} \hat{c}_{1j}$
	$i = 2$	$\hat{c}_{21} = \frac{a_{21}}{N_{.1}}$	$\hat{c}_{22}$	$\dots$	$\hat{c}_{2k}$	$\Sigma_{j=1}^{k} \hat{c}_{2j}$
	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$i = k$	$\hat{c}_{k1} = \frac{a_{k1}}{N_{.1}}$	$\hat{c}_{k2}$	$\dots$	$\hat{c}_{kk}$	$\Sigma_{j=1}^{k} \hat{c}_{kj}$
	Yhteensä	$\Sigma_{i=1}^{k} \hat{c}_{i1} = 1$	1	$\dots$	1	

Aikaisemmin määriteltiin tuottajan sekaannusmatriisin lisäksi käyttäjän sekaannusmatriisi. Koska käyttäjän sekaannusmatriisi ei ole tiedossa, se estimoidaan aineistosta kuten

taulukossa 3. Käyttäjän sekaannusmatriisin solut $\hat{u}_{ij}$ ovat alkuperäisen sekaannusmatriisin soluja, jotka jaetaan rivisummillaan. Näin ollen kyseessä ovat osuudet

$$\hat{u}_{ij} = \frac{a_{ij}}{N_{i.}},$$

missä $N_{i.}$ on luokkaan i luokiteltujen havaintojen summa.

Taulukko 3: Estimaattori käyttäjän sekaannusmatriisille $\hat{\mathbf{U}}$ havaitun sekaannusmatriisin $\mathbf{A}$ avulla ilmaistuna.

		Oikea luokka				Yhteensä
		$j = 1$	$j = 2$	$\dots$	$j = k$	
Luokittelun tulos	$i = 1$	$\hat{u}_{11} = \frac{a_{11}}{N_{1.}}$	$\hat{u}_{12} = \frac{a_{12}}{N_{1.}}$	$\dots$	$\hat{u}_{1k} = \frac{a_{1k}}{N_{1.}}$	$\Sigma_{j=1}^{k} \hat{u}_{1j} = 1$
	$i = 2$	$\hat{u}_{21}$	$\hat{u}_{22}$	$\dots$	$\hat{u}_{2k}$	1
	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$i = k$	$\hat{u}_{k1}$	$\hat{u}_{k2}$	$\dots$	$\hat{u}_{kk}$	1
	Yhteensä	$\Sigma_{i=1}^{k} \hat{u}_{i1}$	$\Sigma_{i=1}^{k} \hat{u}_{i2}$	$\dots$	$\Sigma_{i=1}^{k} \hat{u}_{ik}$	

4 Korjausmenetelmiä

Jos selvitetään yhdestä otoksesta tai osa-aineistosta oikeat luokat ja tehdyn luokittelun tulos, saadaan sekaannusmatriisit. Sekaannusmatriiseilla voidaan korjata luokiteltuja osuuksia aineistossa, josta ei tiedetä havaintoyksiköiden oikeita luokkia. Toisin sanoen korjaus tehdään eri dataan, kuin mistä sekaannusmatriisi on estimoitu. Tässä luvussa esitetään tähän kolme menetelmää: korjaus tuottajan sekaannusmatriisilla, käyttäjän sekaannusmatriisilla ja paras lineaarinen korjaus, jossa käytetään myös tuottajan sekaannusmatriisia.

Korjauksiin liittyviä perusoletuksia on vain muutama: otanta on tehty satunnaisesti perusjoukosta ja jokaisen tilastoyksikön luokittelu tehdään riippumattomasti muista havainnoista. Joissain tutkimusongelmissa voidaan tarvita lisäoletuksia, kuten kaukokartoituksessa, jossa oletetaan, ettei virheellisesti luokiteltujen maastopisteiden välillä ole spatiaalista autokorrelaatiota. (Hess & Bay, 1997.)

4.1 Tuottajan sekaannusmatriisiin perustuva korjaus

Tuottajan sekaannusmatriisista tiedetään oikein- ja väärinluokiteltujen havaintojen osuudet. Tiedetään siis, mihin luokkiin tietyn luokan havaintoja luokitellaan, ja tätä tietoa voidaan käyttää luokittelun korjaamisessa.

Luokittelun korjaaminen perustuu määrittelyyn $\mathbf{p}^* = \mathbf{C}\mathbf{p}$. Yksinkertaisella matriisilaskulla voimme ratkaista

$$\mathbf{p} = \mathbf{C}^{-1}\mathbf{p}^*,$$

jos matriisi $\mathbf{C}$ on kääntyvä. (Fortier, 1992.) Oletetaan aluksi, että $\mathbf{C}$ on tunnettu, $\mathbf{p}^*$ estimoidaan ja estimaattoria merkittiin $\tilde{\mathbf{p}}$. Tällöin korjattu estimaattori on muotoa

$$\hat{\mathbf{p}}_c = \mathbf{C}^{-1}\tilde{\mathbf{p}},$$

ja estimaattorin odotusarvo on $\mathbf{E}[\hat{\mathbf{p}}_c|\mathbf{p}] = \mathbf{C}^{-1}\mathbf{E}[\tilde{\mathbf{p}}|\mathbf{p}] = \mathbf{C}^{-1}\mathbf{p}^* = \mathbf{p}$. Jos myös $\mathbf{C}$ on tuntematon, silloin

$$\hat{\mathbf{p}}_c = \hat{\mathbf{C}}^{-1}\tilde{\mathbf{p}}.$$

Tällöin korjauksen estimaatit ovat harhattomia suurimman uskottavuuden estimaatteja, jos havainnot on valittu yksinkertaisella satunnaisotannalla (Fortier, 1992) tai ositetulla otannalla niin, että rivisummat ovat kiinnitetyt (Buckland & Elston, 1994).

Sekaannusmatriisin estimaattori sisältää satunnaisvaihtelua. Tästä syystä korjaus ei yleensä kokonaan poista luokittelusta syntyvää luokitteluvirhettä, vaikka teoreettisesti mene-

telmä on harhaton. Jos luokittelija on huono, korjauksesta saatavat estimaatit voivat olla jopa virheellisempiä kuin suoraan aineistosta saatavat (Fortier, 1992).

Jos joidenkin luokkien otoskoot ovat pieniä, sekaannusmatriisi saattaa olla singulaarinen, jolloin käänteismatriisin laskeminen ei onnistu (Hay, 1998). Tässä työssä on matriisin kääntämiseksi käytetty Moore-Penrosen käänteismatriisia, joka esitellään myöhemmin. Toinen ongelma on, että korjausmenetelmällä voidaan saada populaatio-osuudelle negatiivisia arvoja. Tämä ongelma korjataan asettamalla vastaava osuus nollaksi (Fortier, 1992).

4.2 Käyttäjän sekaannusmatriisiin perustuva korjaus

Käyttäjän sekaannusmatriisista saadaan suhteellinen osuus, kuinka moni luokkaan i luokitelluista havainnoista kuuluu oikeasti muihin luokkiin. Tätä tietoa voidaan käyttää luokittelun korjaamiseen.

Aikaisemmin määriteltiin vektori $\mathbf{p} = (p_1, p_2, \ldots, p_k)$, jossa p_i on tilastoyksikön todennäköisyys kuulua luokkaan i, $i = 1, 2, \ldots, k$, missä k on luokkien lukumäärä. Nyt voidaan Cardia (1982) mukaillen johtaa harhaton suurimman uskottavuuden estimaatti korjatulle todennäköisyydelle $\hat{p}_{ij}$. Estimoidusta sekaannusmatriisista voimme laskea yhden solun frekvenssistä a_{ij} todennäköisyyden $\hat{p}_{ij}$, jolle pätee $\sum_{j=1}^{k} \hat{p}_{ij} = \tilde{p}_i$. Näiden solutodennäköisyyksien uskottavuusfunktioksi saadaan

$$L(\{\hat{p}_{ij}\}) = \prod_{i=1}^{k} \prod_{j=1}^{k} \hat{p}_{ij}^{a_{ij}}.$$

Muodostetaan Lagrangen kertoimet siten, että maksimoidaan logaritminen uskottavuusfunktio rajoitteen $\sum_i^k \sum_j^k \hat{p}_{ij} = 1$ avulla. Silloin log-uskottavuusfunktio

$$l(\{\hat{p}_{ij}\}) = \log(\hat{p}_{ij}) a_{ij} + \lambda\Big(1 - \sum_{i=1}^{k} \sum_{j=1}^{k} \hat{p}_{ij}\Big).$$

Derivoidaan edellä oleva funktio ja ratkaistaan se estimaatin $\hat{p}_{ij}$ suhteen

$$\begin{aligned} & \frac{\partial l}{\partial \hat{p}_{ij}} \Big[\log(\hat{p}_{ij}) a_{ij} + \lambda\Big(1 - \sum_{i=1}^{k} \sum_{j=1}^{k} \hat{p}_{ij}\Big)\Big] \\ \iff \quad & \frac{a_{ij}}{\hat{p}_{ij}} - \lambda = 0 \\ \iff \quad & \hat{p}_{ij} = \frac{a_{ij}}{\lambda}. \end{aligned} \tag{3}$$

Jos summaamme molemmat puolet sarakkeiden (eli j:n yli) huomaamme, että

$$\tilde{p}_i = \frac{N_{i.}}{\lambda} \iff \lambda = \frac{N_{i.}}{\tilde{p}_i}.$$

Sijoittamalla tämä lausekkeeseen (3) saadaan

$$\hat{p}_{ij} \quad = \quad \tilde{p_i}\frac{a_{ij}}{N_{i.}},$$

ja kun muistetaan käyttäjän sekaannusmatriisin määrittely, saadaan ratkaisuksi

$$\hat{p}_{ij} = \tilde{p_i}\hat{u}_{ij}.$$

Olkoon korjausmenetelmän avulla saatava korjattu todennäköisyys sarakevektorina $\hat{\mathbf{p}}_u$. Nyt samat laskelmat voidaan esittää matriisimuodossa:

$$\hat{\mathbf{p}}_u^T = \tilde{\mathbf{p}}^T\hat{\mathbf{U}}.$$

Toisin sanoen käyttäjän sekaannusmatriisilla voidaan suoraan korjata luokkien ennustettuja osuuksia. Tarkoituksena on, että korjattu estimaatti $\hat{\mathbf{p}}_u$ on mahdollisimman lähellä oikeita todennäköisyyksiä $\mathbf{p}$, mutta sekaannusmatriisin estimaattoriin $\hat{\mathbf{U}}$ sisältyy satunnaisvaihtelua, jonka vuoksi estimaattiin voi sisältyä virhettä.

4.3 Paras lineaarinen korjaus

Paras lineaarinen korjaus (Best-Linear-Corrector, BLC) on lineaarinen muunnos, joka minimoi estimaattien keskineliövirheen mielivaltaisille populaatio-osuuksille. Parhaassa lineaarisessa korjauksessa todennäköisyyksille p_i asetetaan priorijakauma, jota käytetään myöhemmin käsiteltävän neliösumman minimoimiseen. Menetelmän on esitellyt Fortier (1992), jonka esitystä tämä luku seuraa.

Tarkoituksena on määritellä lineaarinen muunnos $\mathbf{Q}\tilde{\mathbf{p}}$ vektorille $\tilde{\mathbf{p}}$. Toisin sanoen halutaan määritellä $k \times k$-matriisi $\mathbf{Q}$, joka minimoi lausekkeen

$$L_Q = E||\mathbf{Q}\tilde{\mathbf{p}} - \mathbf{p}||^2. \tag{4}$$

Matriisin $\mathbf{Q}$ ratkaisemista varten täytyy määritellä matriisi $\mathbf{M}$, joka on $k \times k$-dimensioinen, kuten muutkin matriisit tässä alaluvussa. Tämä matriisi lasketaan satunnaisvektorin $\mathbf{p}$ toisista momenteista, eli

$$\mathbf{M} = E(\mathbf{p}\mathbf{p}^T).$$

Oletetaan, että $\mathbf{p}$ noudattaa Dirichletin jakaumaa, jolloin

$$\mathbf{M}_{ii} = \frac{\alpha_i(\alpha_i + 1)}{\alpha_0(\alpha_0 + 1)} \text{ ja } \mathbf{M}_{ij} = \mathbf{M}_{ji} = \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0 + 1)}, i \neq j,$$

missä $\alpha_0 = \sum_{i=1}^{k} \alpha_i$, $\alpha_i = \pi_i \alpha_0$, kun $i = 1, \ldots, k$. Edellä π_i on vektorin $\mathbf{p}$ elementin p_i prioritodennäköisyys ja α-parametrit ovat Dirichletin jakauman parametreja. Jakauman varianssi riippuu parametrista α_0, joka tutkijan täytyy itse määritellä.

Määritellään vielä matriisi

$$\mathbf{S} = \frac{1}{N}\text{diag}(\mathbf{C}\boldsymbol{\pi}) - \mathbf{C}\Big(\frac{1}{N}\text{diag}(\boldsymbol{\pi}) - \mathbf{M}\Big)\mathbf{C}^T.$$

Lausekkeen (4) minimoi

$$\mathbf{Q} = \mathbf{M}\mathbf{C}^T\mathbf{S}^{-1},$$

joka samalla minimoi keskineliövirheen ja on näin ollen paras lineaarinen muunnos aineistosta saataville populaatio-osuuksille. Nyt korjattu estimaattori oikeille populaatio-osuuksille $\mathbf{p}$ on siten

$$\hat{\mathbf{p}}_{blc} = \mathbf{Q}\tilde{\mathbf{p}}.$$

Todistetaan edeltävä väite. Todistus seuraa Fortierin (1992) todistusta. Ratkaistaan

$$L_{\mathbf{Q}} = E[E[||\mathbf{Q}\tilde{\mathbf{p}} - \mathbf{p}||^2|\mathbf{p}]].$$

Merkitään matriisin $\mathbf{Q}$ rivejä $\mathbf{Q}_i, i = 1, 2, \ldots, k$, ja yksittäistä solua q_{ij}. Määritellään myös ehdollinen kovarianssimatriisi

$$cov(\tilde{\mathbf{p}}|\mathbf{p}) = \frac{1}{N}[\text{diag}(\mathbf{C}\mathbf{p}) - \mathbf{C} \cdot \text{diag}(\mathbf{p})\mathbf{C}^T].$$

Käyttämällä tätä apuna voidaan laskea kovarianssi

$$cov(\mathbf{Q}\tilde{\mathbf{p}}|\mathbf{p}) = \mathbf{Q}cov(\tilde{\mathbf{p}}|\mathbf{p})\mathbf{Q}^T = \frac{1}{N}[\mathbf{Q} \cdot \text{diag}(\mathbf{C}\mathbf{p})\mathbf{Q}^T - \mathbf{Q}\mathbf{C} \cdot \text{diag}(\mathbf{p})\mathbf{C}^T\mathbf{Q}^T],$$

josta saadaan

$$var(\mathbf{Q}_i\tilde{\mathbf{p}}|\mathbf{p}) = \frac{1}{N}[\mathbf{Q}_i \cdot \text{diag}(\mathbf{C}\mathbf{p})\mathbf{Q}_i^T - \mathbf{Q}_i\mathbf{C} \cdot \text{diag}(\mathbf{p})\mathbf{C}^T\mathbf{Q}_i^T]. \quad (5)$$

Aikaisemmin määriteltiin $E(\tilde{\mathbf{p}}|\mathbf{p}) = \mathbf{p}^* = \mathbf{C}\mathbf{p}$. Käytetään tätä tietoa ja määritellään harha B sekä sen neliö. Nyt harha on

$$\begin{aligned} B &= E(\mathbf{Q}_i\tilde{\mathbf{p}}|\mathbf{p}) - p_i \\ B^2 &= E(\mathbf{Q}_i\tilde{\mathbf{p}}|\mathbf{p})^2 - 2E(\mathbf{Q}_i\tilde{\mathbf{p}}|\mathbf{p})p_i + p_i^2 \\ &= (\mathbf{Q}_i\mathbf{C}\mathbf{p})^2 - 2\mathbf{Q}_i\mathbf{C}\mathbf{p}p_i + p_i^2 \\ &= \mathbf{Q}_i\mathbf{C}\mathbf{p}\mathbf{p}^T\mathbf{C}^T\mathbf{Q}_i^T - 2\mathbf{Q}_i\mathbf{C}\mathbf{p}p_i + p_i^2. \end{aligned} \quad (6)$$

Tarkastellaan odotusarvoa ja muokataan se toiseen muotoon

$$\begin{aligned} E(||\mathbf{Q}\tilde{\mathbf{p}} - \mathbf{p}||^2|\mathbf{p}) &= E\Big(\sum_i (\mathbf{Q}_i\tilde{\mathbf{p}} - p_i)^2|\mathbf{p}\Big) \\ &= \sum_i \Big(var(\mathbf{Q}_i\tilde{\mathbf{p}}|\mathbf{p}) + (E(\mathbf{Q}_i\tilde{\mathbf{p}}|\mathbf{p}) - p_i)^2\Big). \end{aligned}$$

Havaitaan, että sijoittamalla (5) ja (6) edelliseen lausekkeeseen, saadaan

$$E(||\mathbf{Q}\tilde{\mathbf{p}} - \mathbf{p}||^2|\mathbf{p}) = \sum_i [\mathbf{Q}_i\Big(\frac{1}{N}\text{diag}(\mathbf{Cp}) - \mathbf{C}\Big(\frac{1}{N}\text{diag}(\mathbf{p}) - \mathbf{pp}^T\Big)\mathbf{C}^T\Big)\mathbf{Q}_i^T - 2\mathbf{Q}_i\mathbf{Cp}p_i + p_i^2].$$

Otetaan odotusarvo vektorin $\mathbf{p}$ yli ja määrittelemällä matriisi $\mathbf{S}$ ja $\mathbf{M}$ saadaan

$$L_{\mathbf{Q}} = E||\mathbf{Q}\tilde{\mathbf{p}} - \mathbf{p}||^2 = \sum_i (\mathbf{Q}_i\mathbf{S}\mathbf{Q}_i^T - 2\mathbf{Q}_i\mathbf{C}\mathbf{M}_i^T + m_{ii}).$$

Lausekkeen $L_{\mathbf{Q}}$ minimoiminen vastaa osittaisderivaatan laskemista q_{ij} suhteen, kun derivaatta on asetettu nollaksi. Näin ollen

$$\frac{\partial L_{\mathbf{Q}}}{\partial q_{ij}} = 2\Big(\sum_l q_{il}s_{jl} - \mathbf{C}_j\mathbf{M}_i^T\Big) = 0,$$

missä s_{jl} on matriisin S solu. Matriisimuodossa sama voidaan esittää ja ratkaista matriisin $\mathbf{Q}$ suhteen

$$\begin{aligned} \mathbf{QS} &= \mathbf{MC}^T \\ \mathbf{Q} &= \mathbf{MC}^T\mathbf{S}^{-1}. \end{aligned}$$

□

Matriisin $\mathbf{M}$ estimointi on selkeästi ongelmallista, sillä vektorin $\mathbf{p}$ arvoja ei havaita, vaan korjauksen tarkoituksena on estimoida ne mahdollisimman tarkasti. Tämän vuoksi käytetään prioritodennäköisyyksiä $\boldsymbol{\pi}$. Fortier (1992) ehdottaa, että $\boldsymbol{\pi}$-vektoria estimoitaisiin tuottajan sekaannusmatriisikorjauksesta saatavilla estimaateilla $\hat{\mathbf{p}}_c$. Tuottajan sekaannusmatriisin estimaattoria $\hat{\mathbf{C}}$ tarvitaan myös matriisin $\mathbf{Q}$ laskemiseen.

Fortier (1992) toteaa, että kun $n \to \infty$, niin $\mathbf{Q} \to \hat{\mathbf{C}}^{-1}$ ja $L_Q \to 0$, joten otoskoon kasvaessa tuottajan sekaannusmatriisiin pohjautuva korjaus on samankaltainen parhaan lineaarisen korjauksen kanssa, ja näin ollen käänteismatriisikorjaus on riittävä tarkkojen estimaattien saamiseksi.

4.4 Käänteismatriisin laskeminen korjausmenetelmiä varten

Tuottajan sekaannusmatriisikorjausta ja parasta lineaarista korjausta varten joudutaan laskemaan käänteismatriisi. Tässä työssä käytettävässä, pohjaeläinaineistosta muodostetussa sekaannusmatriisissa on paljon soluja, joiden arvo on nolla. Pääosin tästä syystä sekaannusmatriisin käänteismatriisin laskeminen epäonnistuu. Matriisi on joko singulaarinen tai melkein singulaarinen, jolloin saadaan epäloogisia lajiosuuksia. Ongelma voidaan ratkaista laskemalla käänteismatriisi Moore-Penrosen käänteismatriisitekniikalla.

Moore-Penrosen käänteismatriisi (Penrose, 1955) määritellään seuraavasti: olkoon kääntyvä matriisi $\mathbf{X}$, joka yksikäsitteisesti toteuttaa seuraavat ehdot mille tahansa matriisille $\mathbf{A}$:

$$\begin{aligned} \mathbf{AXA} &= \mathbf{A} \\ \mathbf{XAX} &= \mathbf{X} \\ (\mathbf{AX})^* &= \mathbf{AX} \\ (\mathbf{XA})^* &= \mathbf{XA}, \end{aligned}$$

missä * viittaa konjugaattitranspoosiin. Näin määriteltynä matriisi $\mathbf{X}$ on kääntyvä ja mahdollisimman lähellä alkuperäistä kääntymätöntä matriisia $\mathbf{A}$. Matriisi $\mathbf{X}$ voidaan ratkaista tilanteesta riippuen jokaiselle mielivaltaiselle matriisille $\mathbf{A}$ seuraavilla tavoilla (Barata & Hussein, 2012):

$\mathbf{X} = \mathbf{A}^*(\mathbf{AA}^*)^{-1}$, jos $\mathbf{A}$:n rivit ovat lineaarisesti riippumattomia (AA^* on kääntyvä),

$\mathbf{X} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$, jos $\mathbf{A}$:n sarakkeet lineaarisesti riippumattomia (A^*A on kääntyvä).

Tätä matriisia käytetään singulaarisen matriisin $\mathbf{A}$ sijasta korjausmenetelmien laskemisessa.

5 Sovellus pohjaeläinaineistoon

Tässä tutkimuksessa sovelletaan korjausmenetelmiä pohjaeläinaineistoon ja tutkitaan menetelmien toimivuutta erilaisten biologisten indeksien suhteen. Työssä käytetään simulointiasetelmaa, jolla mallinnetaan tilannetta, jossa sekaannusmatriisi on estimoitu tietystä otoksesta ja sitä käytetään muista otoksista luokittelun tuloksena saatavien taksonomisten ryhmien osuuksien parantamiseen. Aineiston yksilöiden oikeat luokat ovat tiedossa, joten simulointien tuloksista laskettuja biologisia indeksejä verrataan oikeista luokista laskettuihin indekseihin.

Tässä tutkimuksessa korjausmenetelmien vaikutusta biologisiin indekseihin tutkitaan luokittelun, otoskoon ja jokityypin suhteen. Luokittelun suhteen kyseessä on luokittelijan toimivuus ennen korjausmenetelmiä, eli voidaanko myös huonon luokittelijan tapauksessa korjausmenetelmillä saada lähes harhattomia indeksien arvoja, vai onnistuvatko korjaukset kunnolla vain hyvälle luokittelijalle. Otoskoon vaikutusta tutkitaan sekä luokiteltavan näytteen että sekaannusmatriisin estimointiin käytetyn näytteen suhteen. Taksonomisten ryhmien osuudet vaihtelevat suuresti eri jokityyppien välillä ja monet indeksit mittaavat näitä eroja. Sekaannusmatriisin estimointi jokityyppikohtaisista otoksista voi olla vaikeaa, jos esimerkiksi jossain jokityypissä on paljon harvinaisia ryhmiä, joita ei aina havaita, joten myös jokityyppi voi vaikuttaa korjausmenetelmien onnistumiseen.

5.1 Aineiston kuvaus

Suomen ympäristökeskus on vuosina 2006-2013 kerännyt biologisen seurannan yhteydessä pohjaeläinaineiston, joka koostuu 6585 yksilöstä. Aineistossa on mukana 32 Suomen sisävesissä yleisesti esiintyvää taksonomista ryhmää, jotka on lueteltu liitteessä A. Aineiston keräystavasta johtuen pystytään erottamaan 24 erilaista jokityyppiä. Jokityypeistä on aineistossa edustettuina pääosin sekä luonnontilainen, että ihmisen vaikutuksesta kärsinyt joki.

Aineiston näytteet on kerätty seurannan yhteydessä ja asiantuntijat tunnistivat jokaisen yksilön taksonomisen ryhmän. Näytteet säilöttiin siten, että jokainen taksonominen ryhmä oli omassa alkoholiliuoksessaan. Vasta myöhemmin näytteet päätettiin skannata tietokoneelle, jolloin liuos kaadettiin petrimaljaan skannausta varten. Jokaisen liuoksen skannauksen jälkeen pohjaeläimet eroteltiin omiksi kuvikseen eli segmentointiin. Ennen segmentointia huomioitiin taustan värin ja kirkkauden vaikutus, jotteivat ne vaikuttaisi pohjaeläimestä mitattaviin piirteisiin.

Jokaisesta pohjaeläimestä laskettiin 64 piirrettä, joiden arvo mitattiin ja näin saatiin muo-

dostettua taksakohtaisesti jokaisen piirteen arvojen jakauma. Näiden tulosten avulla jokainen pohjaeläin luokiteltiin piirteidensä mukaisiin luokkiin, eli taksonomisiin ryhmiin. Kuvien käsittely ja piirteiden mittaus tehtiin käyttäen ImageJ-ohjelmaa (Rasband, 2010).

Piirteet voidaan jakaa geometrisiin piirteisiin ja värisävyihin liittyviin piirteisiin. Geometriset piirteet kuvastavat pohjaeläimen muotoa ja kokoa. Värisävyjä mitattiin sekä mustavalko- että väriskaalalla. Sininen, punainen ja vihreä ovat väriskaalassa mitatut värit ja jokainen väreihin liittyvä piirre piirre mitattiin jokaiselle värille erikseen. Sävyjä mitattaessa pohjaeläin eroteltiin maskiksi muusta kuvasta ja mittaukset tehtiin maskin rajaamasta alueesta, jotteivat taustavärit vaikuttaisi tulokseen. Tavoitteena on havaita tummemmat ja vaaleammat kohdat, joiden perusteella voidaan määrittää pohjaeläimen muotoa, rakennetta ja kokoa aivan kuten geometristen piirteiden osalta. Taulukossa 4 on listattu piirteet, joita on käyetty aikaisemmin toteutetussa luokittelussa. Esimerkiksi harmaasävyjen summa on kaikkien maskin pikseleiden arvojen summa ja Feretin halkaisija on pisin mahdollinen suora pohjaeläimen laidasta toiseen laitaan.

Taulukko 4: Luokittelussa käytettyjä piirteitä.

Kuvista mitattavat piirteet	
Harmaa- ja värisävyt:	Geometriset piirteet:
Keskiarvo	Pinta-ala
Keskihajonta	Ympärysmitta
Moodi	Keskipiste
Mediaani	Pohjaeläimen tasareunaisuus
Pikseleiden arvojen summa	Feretin halkaisija, koordinaatit ja kulma
Huipukkuus	Ympyrämäisyysarvot
Vinous	Ellipsin kulma ja halkaisijat
Minimi	Pienimmän halkaisijan pituus
Maksimi	Pohjaeläimen ympärille sovitetun neli-
Massakeskipiste	kulmion leveys, korkeus ja koordinaatit

Pohjaeläimet voivat olla epämuodostuneita, niistä voi puuttua raajoja tai ne voivat olla asennoissa, joista ne on hankala tunnistaa, jolloin kyseinen pohjaeläin ei enää täytä taksonomisen ryhmänsä ominaisuuksia. Tätä on pyritty kontrolloimaan piirteiden suurehkolla lukumäärällä. Epäsuhtaisuus joidenkin piirteiden kohdalla kompensoituu muiden piirteiden avulla. Kuvantaminen ja käytetyt piirteet on käsitelty tarkemmin artikkelissa Ärje et al. (2013).

5.2 Biologiset indeksit

Indeksit voidaan jakaa lajirikkautta, tasaisuutta/vallitsevuutta, monimuotoisuutta ja samankaltaisuutta mittaaviin indekseihin (Magurran, 2004). Lajirikkaus kertoo yksinkertaisesti lajien lukumäärän otoksessa tai estimoituna koko eliöyhteisössä. Tasaisuudella mitataan, ovatko ryhmät kooltaan yhtäsuuria vai ovatko jotkut ryhmät harvinaisempia kuin toiset. Vallitsevuus tutkii samaa ilmiötä vastakkaisesta näkökulmasta: tarkoitus on tutkia, onko eliöyhteisössä joitakin lajeja, joiden yksilöiden lukumäärät ovat paljon muiden lajien lukumääriä suurempia, jolloin nämä lajit dominoivat muita lajeja. Monimuotoisuutta mittaavat indeksit mittaavat sekä lajirikkautta että tasaisuutta ja ovat näin ollen usein yhdistelmiä niiden indekseistä. (Magurran, 2004.) Kahden näytteen samankaltaisuuden mittaamiseksi on kehitetty useita indeksejä. Nämä indeksit mittaavat pääosin sitä, ovatko näytteet lajimäärältään ja lajiosuuksiltaan samankaltaisia.

Tässä työssä tutkittavat indeksit ovat Margalefin monimuotoisuus, Chaon estimaattori lajimäärälle, lajien lukumäärä, Simpsonin indeksi, Shannonin indeksi, Berger-Parkerin indeksi, Sørensenin samankaltaisuus, Canberran metriikka, euklidinen samankaltaisuus, Morisita-Hornin indeksi, PMA-indeksi ja Jaccardin samankaltaisuuskerroin. Vaikka Shannonin tasaisuutta ja Simpsonin tasaisuutta ei suoranaisesti tutkita, ne ovat muunnoksia Shannonin ja Simpsonin indekseistä, joten tasaisuuteen liittyvät indeksit käyttäytyisivät tutkimuksessa samalla tavalla kuin alkuperäiset indeksit.

Yksinkertaisin lajirikkautta mittaava indeksi on taksonomisten ryhmien lukumäärä näytteessä. Mitä enemmän näytteessä on eri ryhmiä, sitä monipuolisempaa eliöstö on. Kuitenkin otoskoko vaikuttaa todennäköisyyteen havaita harvinaisempia lajeja. Margalefin monimuotoisuus pyrkii korjaamaan tätä ottamalla huomioon otoskoon. Merkitään otoskokoa N ja lajien lukumäärää S. Tällöin Margalefin monimuotoisuus on

$$D_{mg} = \frac{S-1}{\log N}.$$

Toinen vaihtoehto on käyttää Chaon estimaattoria. Chaon estimaattori arvioi lajien vähimmäislukumäärää huomioimalla harvinaiset lajit. Mikäli harvinaisia lajeja on paljon näytteessä, niin loogisesti harvinaisia lajeja on paljon myös havaitsematta. Merkitään F_1 lajien lukumäärää, joista on havaittu vain yksi yksilö ja F_2 lajien lukumäärää, joista on havaittu kaksi yksilöä. Chaon estimaattori tulee muotoon

$$S_{Chao} = S + \frac{F_1^2}{2F_2}.$$

Estimaattori voidaan myös yleistää tapauksiin, joissa tutkitaan useita näytteitä. (Magurran, 2004.)

Olkoon q_j suhteellinen osuus yksilöistä, jotka aineistossa kuuluvat luokkaan j. Silloin Shan-

nonin indeksi on

$$H' = -\sum_{j=1}^{k} q_j \log q_j.$$

Shannonin indeksi olettaa, että yksilöt on poimittu satunnaisesti äärettömän suuresta eliöyhteisöstä ja että kaikki lajit ovat edustettuina näytteessä. Siksi Shannonin indeksi toimii huonosti pienillä otoskoilla. Tästä syystä Shannonin indeksiä ei suositella käytettäväksi, mutta se on silti varsin yleisesti käytössä. (Magurran, 2004.)

Toisin kuin Shannonin indeksi, Simpsonin indeksi on hyvä monimuotoisuuden estimaatti pienilläkin otoskoilla. Indeksi mittaa todennäköisyyttä, että kaksi satunnaista yksilöä äärettömän suuresta perusjoukosta kuuluu samaan lajiin. Indeksi määritellään

$$D = \sum_{j=1}^{k} q_j^2.$$

Pieni indeksin arvo vastaa suurta monimuotoisuutta. (Magurran, 2004.)

Berger-Parkerin indeksi määritellään yksilömäärältään suurimman lajin yksilöiden lukumäärän N_{max} ja otoskoon N suhteena

$$d = \frac{N_{max}}{N}.$$

Indeksi mittaa vallitsevuutta, mutta otoksen lajimäärä vaikuttaa indeksin arvoon. Lajimäärän ollessa pieni indeksin arvo on yleensä suuri ja lajimäärän kasvaessa indeksin arvo pienenee. Tämä vaikutus katoaa vasta lajimäärän ollessa yli sata. Huolimatta tästä indeksi on hyvä vallitsevuuden mittari. (Magurran, 2004.)

Seuraavana esiteltävät indeksit mittaavat kahden näytteen samankaltaisuutta. Merkitään näitä näytteitä alaindekseillä a ja b. Toinen näyte voi olla vertailunäyte esimerkiksi luonnontilaisesta joesta, johon näytteitä verrataan. Täydellisen samankaltaisuuden tapauksessa jokaisen indeksin arvo on yksi. Mitä pienempi indeksin arvo on, sitä enemmän näytteet eroavat toisistaan. Indeksien pienin mahdollinen arvo on nolla, paitsi euklidisen samankaltaisuuden, jonka pienin arvo on -1.

Sørensenin samankaltaisuus määritellään

$$QS = \frac{2S_{ab}}{S_a + S_b},$$

missä S_a on näytteen a lajien määrä, S_b on näytteen b lajien määrä ja S_{ab} on lajien lukumäärä, jotka ovat molemmissa näytteissä. Sørensenin samankaltaisuus toimii useimmiten

ongelmitta. (Wolda, 1981.)

Canberran metriikka määritellään

$$1 - CM = 1 - \frac{1}{S_a + S_b - S_{ab}} \sum_{j=1}^{k} \frac{|n_{ja} - n_{jb}|}{n_{ja} + n_{jb}},$$

missä termi n_{ja} on lajin j yksilöiden lukumäärä aineistossa a ja vastaavasti n_{jb} aineistolle b. Indeksi ei ole ongelmaton, sillä Canberran metriikan arvo kasvaa epälineaarisesti, minkä lisäksi otoskoko vaikuttaa indeksin arvoihin. (Wolda, 1981.)

Kun N_a ja N_b ovat otosten a ja b otoskoot, Morisita-Hornin indeksi on

$$C_\lambda = \frac{2 \sum_{j=1}^{k} n_{ja} n_{jb}}{D_a D_b N_a N_b},$$

missä D on Simpsonin indeksi. Morisita-Hornin indeksi toimii useimmiten ilman ongelmia. (Wolda, 1981.)

Määritellään q_{ja} luokan j suhteelliseksi osuudeksi ensimmäisestä näytteestä ja q_{jb} on saman luokan osuus toisesta näytteestä. Euklidinen samankaltaisuus saadaan neliöidystä euklidisesta etäisyydestä muotoiltuna

$$1 - D_{euk}^2 = 1 - \sum_{j=1}^{k} (q_{ja} - q_{jb})^2.$$

Euklidinen samankaltaisuus on epälineaarinen ja näin ollen toimii huonosti joissain tilanteissa (Wolda, 1981).

PMA-indeksi (Percent model affinity index) mittaa kahden näytteen lajiosuuksien absoluuttista eroa. Tämä erotus vähennetään numerosta yksi, jolloin suurempi PMA-indeksin arvo viittaa näytteiden samankaltaisuuteen. Tällöin

$$PMA = 1 - \frac{1}{2} \sum_{j=1}^{k} |q_{ja} - q_{jb}|,$$

missä q_{ja} ja q_{jb} ovat näytteiden osuudet. (Renkonen, 1938 sekä Novak & Bode, 1992.) PMA-indeksin ominaisuuksia on käsitelty tarkemmin artikkelissa Ärje et al. (2016).

Jaccardin (1901) kehittelemä samankaltaisuuskerroin mittaa matemaattisesti kahden joukon leikkauksen suhdetta niiden yhdisteeseen. Pohjaeläinten tapauksessa verrataan molemmissa näytteissä esiintyvien lajien lukumäärää S_{ab} yhteensä havaittujen lajien lukumäärään, eli

$$J = \frac{S_{ab}}{S_a + S_b - S_{ab}}.$$

Näitä kaikkia indeksejä käytetään yleisesti biologisissa sovelluksissa ja siksi ne ovat mukana myös tässä tutkimuksessa. Indeksien omien ominaisuuksien lisäksi luokittelulla on oma vaikutuksensa indeksien mittaamiseen, jolloin jotkin indeksit voivat toimia paremmin kuin toiset. Tavoitteena on tutkia luokitteluvirheen vaikutusta indeksien mittaamiseen. Käytännössä halutaan biologisilta ominaisuuksiltaan hyviä indeksejä, jotka pystytään mittaamaan mahdollisimman harhattomasti.

5.3 Simulaatiomalli

Tutkitaan korjausmenetelmien ominaisuuksia simulointimallin avulla. Oletetaan, että käytössä on valmis sekaannusmatriisi, jonka ominaisuuksia ei tässä tarkastella lähemmin (sekaannusmatriisi muodostettu luokittelun tuloksena artikkelissa Ärje et al., 2017). Muodostetaan useita otoksia, jotka luokitellaan ja luokittelua korjataan korjausmenetelmin. Simulointimallin perustana on luvussa 5.1 esitelty pohjaeläinaineisto ja analysointi toteutetaan R-ohjelmistolla (R Core Team, 2017). Kuvien piirtämisessä on käytetty R-pakettia `ggplot2` (Wickham, 2016) ja Moore-Penrosen käänteismatriisi on laskettu paketin `MASS` (Venables & Ripley, 2002) `ginv`-funktiolla.

Pohjaeläinaineisto on luokiteltu kahdella menetelmällä: satunnaisella metsällä ja naiivi Bayes -menetelmällä. Satunnainen metsä on hyvä luokittelija kyseiseen aineistoon, sen luokitteluvirhe on alkuperäisessä testiaineistossa 20.5%. Naiivi Bayes sen sijaan on huono luokittelija, sen luokitteluvirhe on peräti 48.7 %. Käytännössä näin huonon luokittelijan käyttäminen ei ole kannattavaa, mutta mikäli korjausmenetelmät toimivat tehokkaasti, huononkin luokittelijan käyttäminen tulisi mahdolliseksi. Simuloinnit tehdään käyttäen molempia luokitteluista laskettuja sekaannusmatriiseja.

Aineiston perusteella on tiedossa tyypilliset taksonomisten ryhmien osuudet eri jokityypeissä eli vektorit $\mathbf{p}$. Valitaan yhden tyypillisen jokityypin osuudet $\mathbf{p}$. Multinomijakaumasta voidaan nyt näitä osuuksia käyttämällä simuloida $M + 1$ kappaletta otoksia, joiden koko on N. Pieneksi otoskooksi on valittu 300 yksilöä, joka vastaa keskimäärin noin kymmentä yksilöä ryhmää kohti, koska taksonomisten ryhmien lukumäärä on 32. Suureksi otoskooksi valittiin tuhat yksilöä eli hieman yli kolmekymmentä yksilöä taksaa kohden. Sekaannusmatriisi on muodostettu yhtä suurella otoskoolla kuin otokset. Simulointien määräksi on valittu 1001, joten $M = 1000$.

Otokset simuloidaan multinomijakaumasta $(y_1, y_2, \ldots, y_k) \sim Multinom(N, p_1, p_2, \ldots, p_k)$. Käytetään aiemmin muodostettua tuottajan sekaannusmatriisin estimaattia $\hat{\mathbf{C}}$. Luokitel-

lut lukumäärävektorit estimoidaan käyttämällä tietyn otoksen tietyn ryhmän yksilöiden lukumäärää y_i otoskokona. Kun sekaannusmatriisin vastaavan sarakkeen arvoja käytetään multinomijakauman todennäköisyyksinä, voidaan simuloida epätäydellistä luokittelua vastaava tulos $\tilde{y}_i$. Se saadaan multinomijakaumista $Multinom(y_i, \hat{c}_{1i}, \hat{c}_{2i}, \ldots, \hat{c}_{ki})$. Laskemalla jakauman alkiot jokaisen $i = 1, 2, \ldots, k$ suhteen, saadaan k määrä jakaumia. Summaamalla jokaisen jakauman i:net alkiot, saadaan kyseiseen luokkaan luokiteltujen alkioiden lukumäärä. Kun tämä toistetaan kaikille ryhmille i, saadaan ryhmien luokitellut lukumäärät, joita luvussa 3 merkittiin siis vektorilla $\tilde{\mathbf{y}}$. Toistamalla tämä kaikille otoksille, saadaan M kappaletta luokiteltuja lukumääriä $\tilde{\mathbf{y}}$ ja jokaista $\tilde{\mathbf{y}}$ vektoria vastaa alkuperäinen $\mathbf{y}$ vektori. Näin ollen jokaisesta luokittelusta on tiedossa oikeat ja luokitellut lukumäärät ryhmittäin. Viimeistä otosta käytetään korjausmenetelmissä tarvittavan sekaannusmatriisin estimaattina. Tämä vastaisi sitä, että yhdestä otoksesta on selvitetty oikeat luokat ja näin siitä on pystytty luomaan sekaannusmatriisi, jota käytetään muiden otosten korjaamiseen.

Nyt saatua sekaannusmatriisia ja siitä muodostettua käyttäjän ja tuottajan sekaannusmatriisia käytetään korjausmenetelmissä. Käyttäjän sekaannusmatriisikorjauksessa matriisi ei usein ole kääntyvä ja siksi käytetään Moore-Penrosen käänteismatriisia. Parasta lineaarista korjausta varten on asetettu $\alpha_0 = 100$. Dirichletin jakauman mielessä α_0 arvo on kohtuullisen pieni, jolloin varianssi on suuri. Kokeilujen perusteella tulokset eivät juuri muutu, kun $0.1 < \alpha_0 < 3000$. Sopiva arvo on löydetty simuloimalla erilaisia tilanteita. Myös parhaassa lineaarisessa korjauksessa tulee ongelmia matriisin kääntämisen kanssa, sillä matriisin $\mathbf{Q}$ määrittelyssä käytetään käänteismatriisia $\mathbf{S}^{-1}$, joten sovelletaan Moore-Penrosen käänteismatriisia kuten tuottajan sekaannusmatriisin tapauksessa.

Jokaisen indeksin arvo lasketaan otos kerrallaan oikeilla arvoilla, luokittelun mukaisilla arvoilla ja kaikkien korjausten tuloksilla. Joidenkin indeksien laskemiseen tarvitaan yksilöiden lukumäärät lajeittain. Koska sekaannusmatriisikorjaukset tehdään osuuksille, pitää saadut korjatut osuudet kertoa otoskoolla N, jotta saadaan korjattu lukumäärävektori. Aikaisemmin määriteltiin, että mikäli korjauksen seurauksena lajiosuus on negatiivinen, osuus asetetaan nollaksi. Tämän vuoksi ennen lukumäärävektoreiden laskemista lajiosuudet on skaalattu niin, että niiden kokonaissumma on 100 %.

Toisin sanoen oikea indeksin arvo saadaan, kun tiedetään kuinka monta yksilöä on otoksessa missäkin taksonomisessa ryhmässä. Luokitellusta otoksesta estimoituja osuuksia kutsutaan raakaestimaateiksi. Kolmella korjausmenetelmällä saadaan kolme luokittelusta korjattua estimaattia lajiosuuksille. Osuuksista laskettuja viittä lukumäärävektoria käytetään indeksien laskemisessa. Koska samankaltaisuusindeksit vertaavat kahta näytettä toisiinsa, tässä oikeaa lukumäärävektoria käytetään referenssituloksena, johon raakaestimaatteja ja

korjauksia verrataan.

Samankaltaisuusindeksin tapauksessa mitä lähempänä arvo on numeroa yksi, sitä lähempänä luokitellut arvot ovat oikeita arvoja. Muiden indeksien tapauksessa luokitellusta indeksin arvosta vähennetään oikea indeksin arvo. Näin ollen saadaan raakaestimaateilla ja kaikille kolmella korjausmenetelmillä lasketuille indekseille jakaumat, joiden odotusarvo olisi nolla, jos ne vastaisivat oikeiden indeksien jakaumia. Positiivisen arvon tapauksessa korjatun indeksin arvo on suurempi kuin sen pitäisi olla ja erotuksen ollessa nolla korjaus on odotusarvoisesti harhaton.

Tulokset esitetään viiksilaatikoina jokainen luokittelija ja otoskoko kerrallaan. Sen lisäksi esitetään taulukko indeksien tunnusluvuista, jotta saadaan parempi käsitys indeksien tavallisista arvoista, sekä siitä kuinka merkittäviä viiksilaatikoissa näkyvät erot ovat.

5.4 Tulokset

Tutkitaan sekaannusmatriisikorjauksen vaikutusta biologisiin indekseihin eri luokittelijoilla ja erilaisilla otoskoilla. Tutkitaan ensin tavallisia tunnuslukuja indeksittäin. Taulukossa 5 on listattu tunnuslukuja indeksien oikeilla arvoilla, raakaestimaateilla lasketuilla arvoilla, sekä kaikilla kolmella korjausmenetelmällä lasketuilla arvoilla. Taulukko on laskettu tuhannella otoksella, joiden jokaisen otoskoko on tuhat ja luokittelijana on satunnainen metsä. Vaikka tutkimuksessa on mukana 32 taksonomista ryhmää, keskimäärin yhdessä otoksessa on vain 23.81 ryhmää.

Oikeiden arvojen vaihtelu on suurempaa kuin korjauksien tai raakaestimaattien. Esimerkiksi Chaon estimaattori on suurimmillaan 47, kun korjausmenetelmillä se on suurimmillaan 27. Korjausmenetelmien tuloksena indeksien arvot keskittyvät lähemmäs keskiarvoaan, kun otoksen perusteella arvojen pitäisi poiketa enemmän. Kaikkein pienintä vaihtelu on käyttäjän sekaannusmatriisikorjauksella saaduilla indekseillä. Pienen vaihtelun seurauksena joissain epätyypillisissä otoksissa indeksien arvot voivat poiketa paljonkin oikeasta, mutta pieni vaihtelu voi olla myös eduksi. Indeksien oikeat arvot perustuvat otoksiin, jotka voivat poiketa joen tavanomaisesta populaatiosta. Esimerkiksi johonkin otokseen voi päätyä huomattavan vähän lajeja, mutta korjausmenetelmä korjaisi lajimäärän suuremmaksi. Tällöin korjattu lajimäärä olisi lähempänä oikeaa joessa olevien lajien määrää. Toisin sanoen näytteen poikkeavuus voi johtua sattumasta sen sijaan, että tutkittava joki poikkeaisi tavanomaisesta. Tällaisessa tapauksessa korjausmenetelmä tasoittaisi otoksesta aiheutuvaa vaihtelua.

Taulukko 5: Oikeilla arvoilla, raakaestimaateilla ja kolmella korjausmenetelmällä lasketut keskiarvot, keskihajonnat, minimit ja maksimit jokaiselle indeksille. Arvot on laskettu tuhannesta otoksesta otoskoon ollessa tuhat ja käytetty luokittelija on satunnainen metsä. Jokainen indeksi on omalla rivillään. Ylhäältä alaspäin katsottuna indeksit ovat: lajimäärä, Chaon estimaattori, Margalefin monimuotoisuus, Shannonin indeksi, Simpsonin indeksi, Berger-Parkerin indeksi, Sørensenin samankaltaisuus, PMA, Canberran metriikka, Euklidinen samankaltaisuus, Morisita-Hornin indeksi ja Jaccardin samankaltaisuuskerroin. Huomioi, että suurin osa indeksien arvoista on pyöristetty kahden desimaalin tarkkuudella.

	Oikeat indeksien arvot				Raakaestimaattien arvot				Käyttäjän korjauksen tulokset				Tuottajan korjauksen tulokset				Parhaan lineaarisen korjauksen tulokset			
Indeksi:	$\overline{\text{Ind}}$	sd	min	max	$\overline{\text{Ind}}$	sd	min	max	$\overline{\text{Ind}}$	sd	min	max	$\overline{\text{Ind}}$	sd	min	max	$\overline{\text{Ind}}$	sd	min	max
S	23.81	1.17	21.0	27.0	28.89	0.87	26.0	31.0	22.49	0.52	21.0	23.0	25.41	1.01	22.0	27.0	25.41	1.01	22.0	27.0
S_{Chao}	25.06	2.71	21.0	47.0	29.46	1.60	26.0	40.0	22.63	0.75	21.0	26.0	25.41	1.01	22.0	27.0	25.41	1.01	22.0	27.0
D_{mg}	3.30	0.17	2.90	3.76	4.04	0.13	3.62	4.34	3.11	0.08	2.90	3.19	3.54	0.15	3.04	3.77	3.54	0.15	3.04	3.77
H'	2.55	0.03	2.46	2.62	2.81	0.03	2.72	2.90	2.52	0.02	2.44	2.58	2.53	0.04	2.37	2.65	2.54	0.04	2.38	2.65
D	0.11	0.00	0.10	0.12	0.09	0.00	0.08	0.10	0.11	0.00	0.10	0.12	0.11	0.01	0.09	0.13	0.11	0.01	0.09	0.13
d	0.19	0.01	0.16	0.23	0.19	0.01	0.15	0.23	0.19	0.01	0.16	0.23	0.19	0.01	0.16	0.23	0.19	0.01	0.16	0.23
QS	1.00	0.00	1.00	1.00	0.90	0.02	0.84	0.96	0.95	0.02	0.85	1.00	0.91	0.03	0.81	0.98	0.91	0.03	0.81	0.98
PMA	1.00	0.00	1.00	1.00	0.87	0.01	0.84	0.90	0.96	0.01	0.94	0.98	0.95	0.01	0.91	0.98	0.95	0.01	0.91	0.98
$1 - CM$	1.00	0.00	1.00	1.00	0.67	0.02	0.58	0.76	0.83	0.04	0.68	0.94	0.70	0.03	0.59	0.79	0.70	0.03	0.59	0.78
$1 - D^2_{euk}$	1.00	0.00	1.00	1.00	1.00	0.00	0.99	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00
C_λ	1.00	0.00	1.00	1.00	0.97	0.01	0.95	0.99	1.00	0.00	0.99	1.00	1.00	0.00	0.98	1.00	1.00	0.00	0.98	1.00
J	1.00	0.00	1.00	1.00	0.74	0.04	0.66	0.84	0.69	0.03	0.59	0.72	0.70	0.04	0.56	0.81	0.70	0.04	0.56	0.81

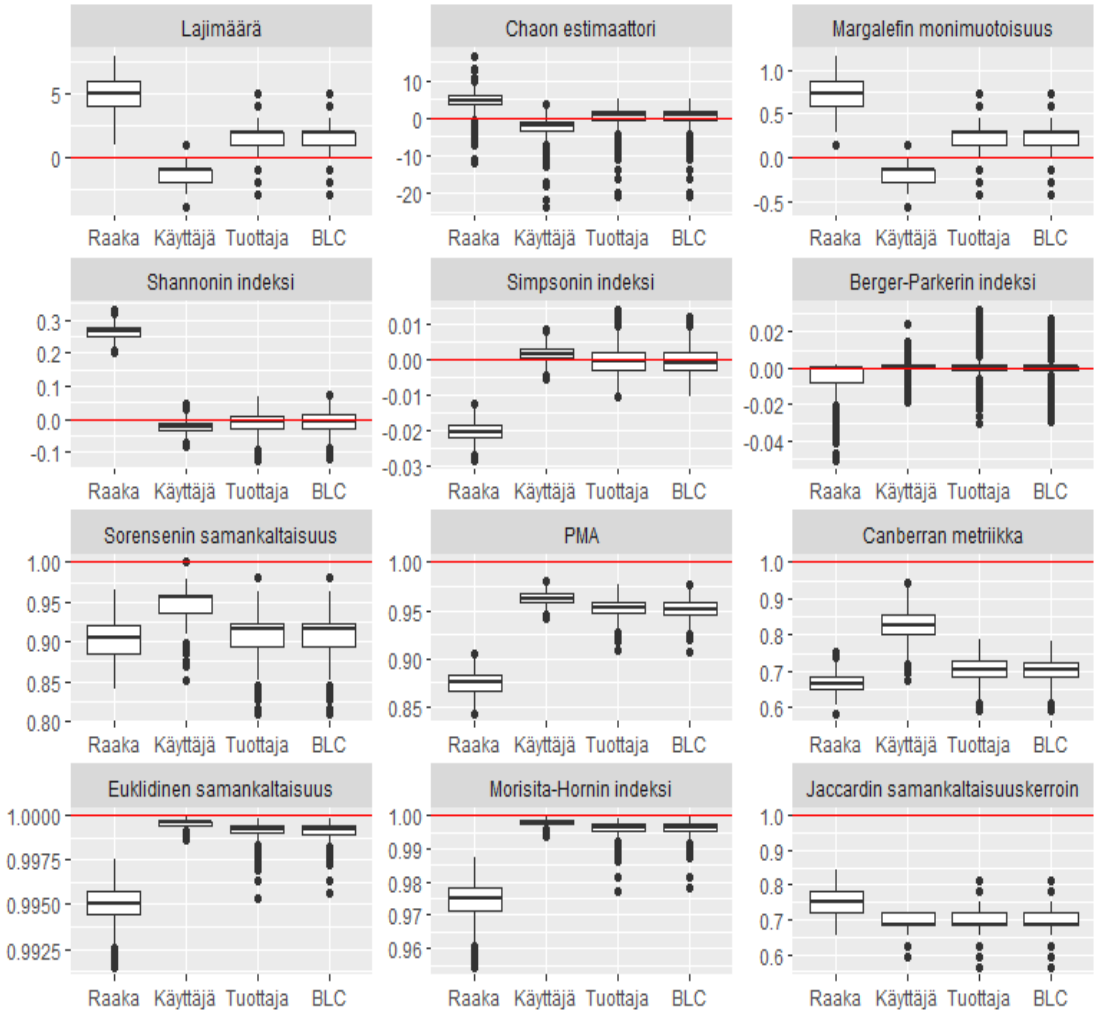
Myöhemmin korjausmenetelmillä ja raakaestimaateilla saadut indeksien arvot esitetään poikkeamina oikeasta arvosta. Indeksien mittakaava on tärkeää huomioida tuloksia tulkittaessa. Ero toisessa desimaalissa ei ole kovin merkittävä Chaon estimaattorin kohdalla (keskiarvo on 25.06), mutta on merkittävä Simpsonin indeksiä tutkittaessa (keskiarvo on 0.11).

Tutkitaan seuraavaksi viiksilaatikkokuvien avulla korjausmenetelmillä laskettujen indeksien poikkeamaa oikeista indeksien arvoista. Samankaltaisuusindekseissä verrataan oikeaa ja korjattua indeksiä. Täydellisen samankaltaisuuden tapauksessa indeksi saa arvon yksi, muilla indekseillä korjatusta indeksin arvosta vähennetään oikea indeksi. Näin ollen mitä lähempänä korjattu indeksi on nollaa, sitä tarkempi korjaus.

Käytetään edellä määriteltyä simulointimallia ja tarkastellaan raakaindeksejä ja korjattuja indeksejä. Kuvassa 1 on hyvällä luokittelijalla ja tuhannen yksilön otoskoolla saadut tulokset. Paras lineaarinen korjaus on kuvaan lyhennetty termillä BLC. Vaikka korjausmenetelmät toimivat jokseenkin hyvin ja usein korjaus on parempi kuin pelkkään luokitteluun perustuva raakakerroin, silti korjattuihin indekseihin jää harhaa. Kuvassa indeksien asteikot eroavat, joten poikkeavuutta indeksin oikeasta arvosta täytyy arvioida huolellisesti. Pääosin kaikki korjatut indeksit keskittyvät nollan tuntumaan lajimäärää, tasaisuutta ja vallitsevuutta mittaavissa indekseissä ja samankaltaisuusindekseissä lähelle ykköstä. Näin ollen korjatut indeksit ovat parhaassa tapauksessa liki harhattomia ja useimmiten vain vähän harhaisia.

Jos tarkastellaan ensin lajirikkauteen, tasaisuuteen ja monimuotoisuuteen liittyviä indeksejä, huomataan sekaannusmatriisikorjauksen parantavan raakaestimaatteja oikeiden indeksin arvojen suuntaan. Lajirikkautta mittaavat indeksit ovat korjattuina yleensä tarkkoja ja parempia kuin korjaamattomat raakaestimaatit. Shannonin, Simpsonin ja Berger-Parkerin indekseissä korjaukset ovat tarkkoja. Yleisesti yksikään korjausmenetelmä ei ole parempi kuin toinen. Kaikkien menetelmien vaihtelu on vähäistä, eli indeksin arvot poikkeavat oikeasta indeksistä suunnilleen yhtä paljon joka otoksessa. Tosin käyttäjän sekaannusmatriisikorjauksen vaihtelu on vielä pienempää tasaisuutta ja monimuotoisuutta mittaavissa indekseissä muihin menetelmiin verrattuna. Vaihtelun vähäisyys on tärkeää, jotta tulokseen voidaan luottaa otoksesta ja sekaannusmatriisista riippumatta.

Korjausmenetelmien välillä on eroja samankaltaisuutta mittaavissa indekseissä. Nyt korjausmenetelmän toimivuus riippuu tutkittavasta indeksistä. Sørensenin samankaltaisuudessa käyttäjän sekaannusmatriisi tuottaa parhaan tuloksen ja muut korjausmenetelmät ovat keskimäärin samantasoisia kuin raakaestimaatit. PMA-indeksissä käyttäjän sekaannusmatriisikorjauksella saadaan tarkkoja arvoja. Muut korjaukset ovat huonompia, mutta



Kuva 1: Indeksien arvot raakaestimaateilla ja eri korjausmenetelmillä verrattuna oikeaan arvoon käytettäessä hyvää luokittelijaa (satunnainen metsä) ja otoskoon ollessa tuhat. Indeksien jakaumat ovat tuhannesta näytteestä. Harhattomassa tilanteessa kuuden ylimmän kuvaajan arvot olisivat nolla ja kuudessa alimmassa yksi. Tätä on merkitty punaisella poikkiviivalla. Huomioi, että indeksien asteikot eroavat (y-akseli). Paras lineaarinen korjaus on lyhennetty BLC.

parempia silti kuin raakaestimaatit. Euklidinen samankaltaisuus ja Morisita-Hornin indeksi käyttäytyvät samalla tavalla: käyttäjän korjaus on lähes harhaton, raakaestimaatitkin toimivat hyvin. Muut korjaukset ovat harhattomampia kuin raakaestimaatit, mutta eivät yhtä hyviä kuin korjaus käyttäjän sekaannusmatriisilla. Canberran metriikassa käyttäjän korjaus on paras, sen jälkeen raakaestimaatit, tuottajan korjaus ja paras lineaarinen korjaus yhtä hyviä. Jaccardin samankaltaisuuskerroin poikkeaa muista indekseistä, sillä raakakerroin antaa parhaimman indeksin arvon. Korjatut indeksit ovat harhaisia.

Käyttäjän sekaannusmatriisikorjaus on yleensä paras valinta. Tarkasteltaessa kaikkia in-

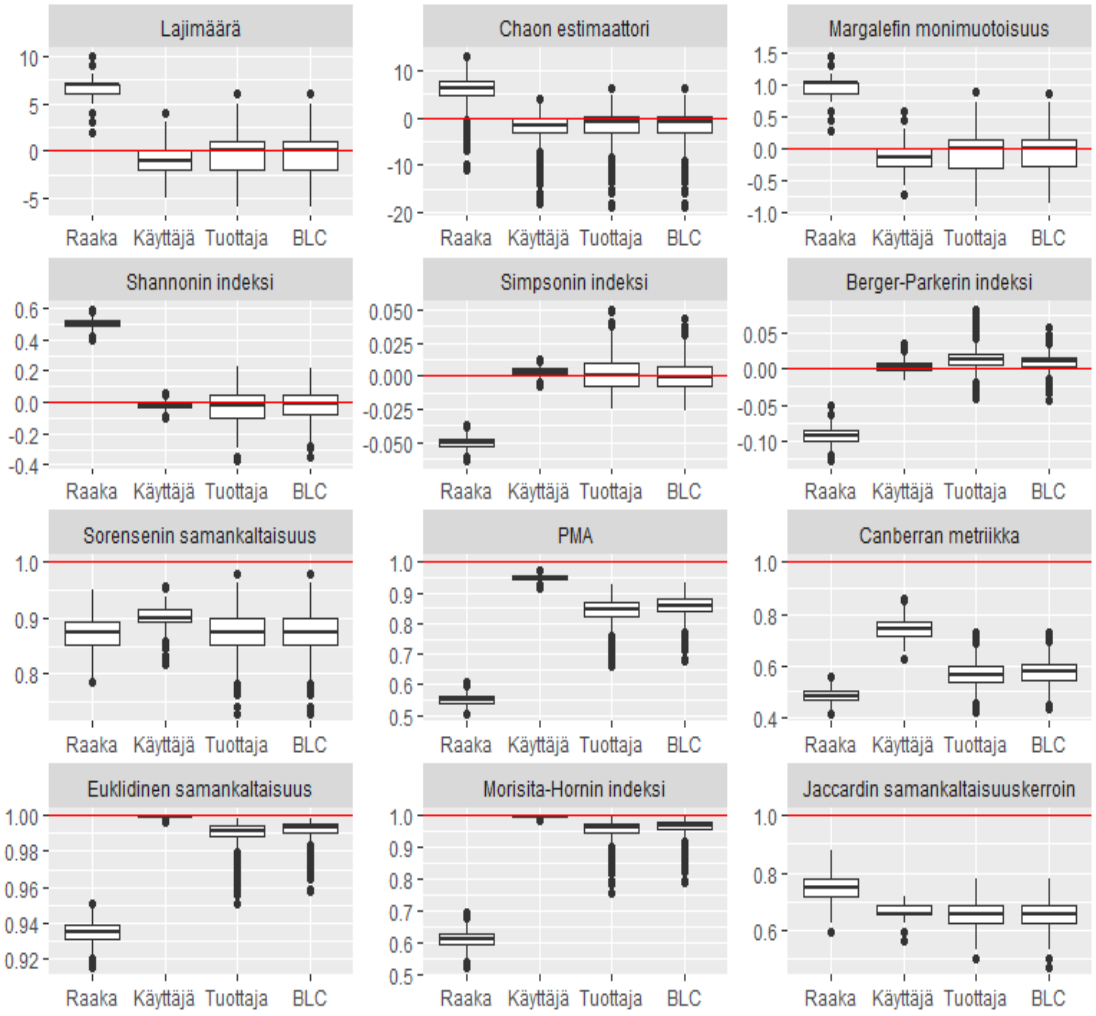
deksejä kokonaisuutena huomataan käyttäjän sekaannusmatriisikorjauksen toimivan lähes aina hyvin. Samankaltaisuusindekseissä se on aina paras korjaus ja Jaccardin samankaltaisuuskerrointa lukuunottamatta parempi kuin raakaestimaatit. Muiden indeksien kohdalla käyttäjän korjaus on lähellä oikeita arvoja. Viiksilaatikoiden perusteella vaihtelu on pientä, eli yksittäisissä otoksissa korjauksen epäonnistuminen on epätodennäköistä. Sen sijaan muiden korjausten kohdalla havaitaan, että joidenkin otosten kohdalla korjaus on epäonnistunut. Viiksilaatikoissa on huomattavan paljon poikkeavia arvoja. Näin ollen tuottajan sekaannusmatriisikorjauksella ja parhaalla lineaarisella korjauksella indeksien arvo voi tietyssä otoksessa poiketa paljonkin oikeasta, eivätkä indeksien arvot ole jokaisessa tilanteessa erityisen luotettavia. Yksi syy tähän on näiden menetelmien käänteismatriisien singulaarisuus, jonka vuoksi matriiseja joudutaan muokkaamaan kääntyvään muotoon. Suuren vaihtelun lisäksi tuottajan sekaannusmatriisikorjaus ja paras lineaarinen korjaus ovat selkeän harhaisia Canberran metriikan ja Sørensenin samankaltaisuuden osalta. Suurimmassa osassa indeksejä kaikki korjaukset ovat kuitenkin parhaimmillaan lähes harhattomia.

Canberran metriikka ja Jaccardin samankaltaisuuskerroin ovat vaikeimmin korjattavissa. Molemmat indeksit pohjautuvat lajimääriin, jolloin lajimäärässä oleva harha kertautuu näitä indeksejä laskettaessa. Muiden indeksien kohdalla on ainakin yksi korjaus, jolla saadaan hyviä tuloksia. Huolimatta erilaisesta laskutavasta paras lineaarinen korjaus ja tuottajan korjaus antavat usein samankaltaisen indeksin arvon. Näiden korjausmenetelmien viiksilaatikoissa on vain vähän eroja.

5.4.1 Luokittelijan vaikutus

Tarkastellaan seuraavaksi korjattujen indeksien arvoja, kun käytössä on huono luokittelija eli naiivi Bayes. Kuvassa 2 indeksit ja korjausmenetelmät ovat viiksilaatikoina, kuten aikaisemmin kuvassa 1. Lajimäärää, tasaisuutta ja monimuotoisuutta mittaavat indeksit ovat edelleen tarkkoja ilman suurta vaihtelua tarkkuudessa. Poikkeaviksi laskettujen arvojen määrä on hieman lisääntynyt, mutta korjaus on silti useimmiten tarkka. Margalefin monimuotoisuudessa tuottajan korjauksen ja parhaan lineaariseen korjauksen vaihtelu on suurempaa kuin käyttäjän sekaannusmatriisikorjauksessa.

Samankaltaisuusindekseissä korjaukset toimivat pääosin yhä hyvin, mutta indeksit ovat harhaisempia kuin hyvän luokittelijan tapauksessa. Esimerkiksi hyvällä luokittelijalla PMA-indeksin mediaani oli korjausmenetelmästä riippuen 0.95-0.97, kun huonolla luokittelijalla ne ovat 0.85-0.95 välissä. Kuitenkin raakaestimaatteihin verrattuna korjaukset toimivat hyvin. Käyttäjän sekaannusmatriisi on paras korjausmenetelmä. Paras lineaarinen korjaus



Kuva 2: Tuhannesta näytteestä, huonolla luokittelijalla (naiivi Bayes) ja tuhannen yksilön otoskoolla saadut indeksien arvot verrattuina oikeisiin arvoihin. Harhattomassa tilanteessa korjausten arvot olisivat nolla tai yksi indeksistä riippuen. Tätä on merkitty punaisella poikkiviivalla.

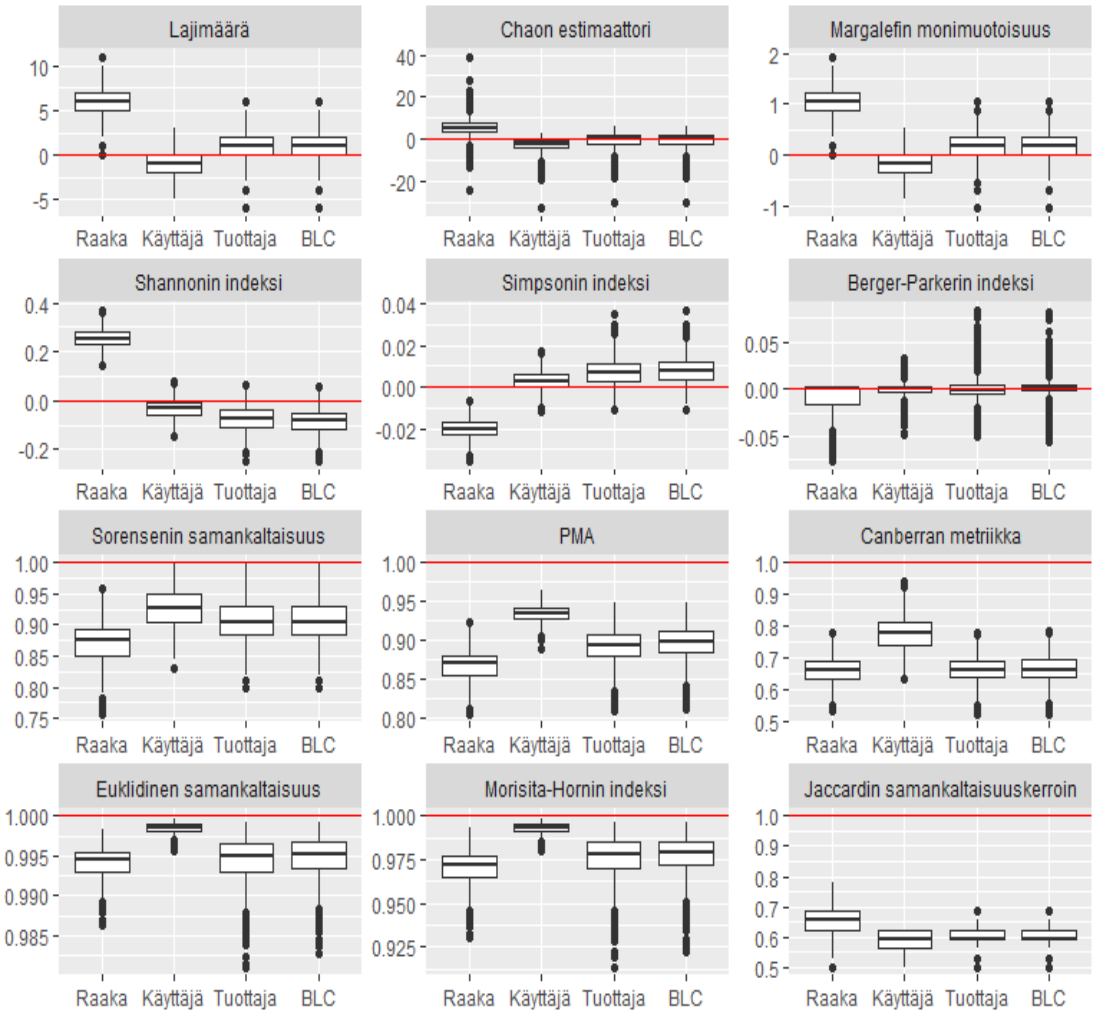
on hieman parempi kuin tuottajan korjaus ja sen vaihtelu on pienempää.

Korjausmenetelmät korjaavat selkeästi harhaa, vaikka alkuperäinen luokittelija luokitteli puolet pohjaeläimistä väärin. Ainoastaan Jaccardin samankaltaisuuskerroin ei ole korjattavissa. Käyttäjän sekaannusmatriisikorjaus on yleensä paras korjausmenetelmä, joskin erot näkyvät käytännössä vain samankaltaisuusindekseissä.

5.4.2 Otoskoon vaikutus

Tutkitaan otoskoon vaikutusta erikseen hyvällä ja huonolla luokittelijalla. Kuvassa 3 on hyvällä luokittelijalla lasketut indeksit, kun otoskoko on 300 ja kuvassa 4 on samalla otos-

koolla saadut indeksit huonoa luokittelijaa käyttäen. Hyvällä luokittelijalla otoskoolla ei ole merkittävää vaikutusta indeksien harhaan. Lajimäärään liittyvissä indekseissä käyttäjän sekaannusmatriisi aliarvioi lajimäärää, mutta keskimäärin vain muutaman lajin verran. Samankaltaisuusindekseissä harha kasvaa jonkin verran ja poikkeavien arvojen määrä lisääntyy. Esimerkiksi suurella otoskoolla euklidisessa samankaltaisuudessa yli 75 % kaikkien korjausmenetelmien korjatuista arvoista oli 0.998 tai enemmän. Pienellä otoskoolla vastaava arvo on 0.992.



Kuva 3: Indeksien arvojen poikkeamat oikeista indekseistä, kun otoskoko on 300, mutta luokittelija hyvä (satunnainen metsä). Simulointi on tehty tuhat kertaa. Punainen poikkiviiva vastaa harhatonta tilannetta.

Huonolla luokittelijalla pienempi otoskoko aiheuttaa harhaa tuottajan korjauksessa ja parhaassa lineaarisessa korjauksessa. Shannonin ja Simpsonin indeksit ja kaikki samankaltaisuusindeksit ovat harhaisempia näiden korjausten suhteen. Sørensenin samankaltaisuut-

ta lukuunottamatta käyttäjän sekaannusmatriisilla korjatut samankaltaisuusindeksit ovat harhaisempia kuin suurella otoskoolla, mutta ero ei ole erityisen suuri. Tasaisuus- ja monimuotoisuusindekseissä käyttäjän korjaus toimii hyvin, lajimäärää korjaus aliestimoi.

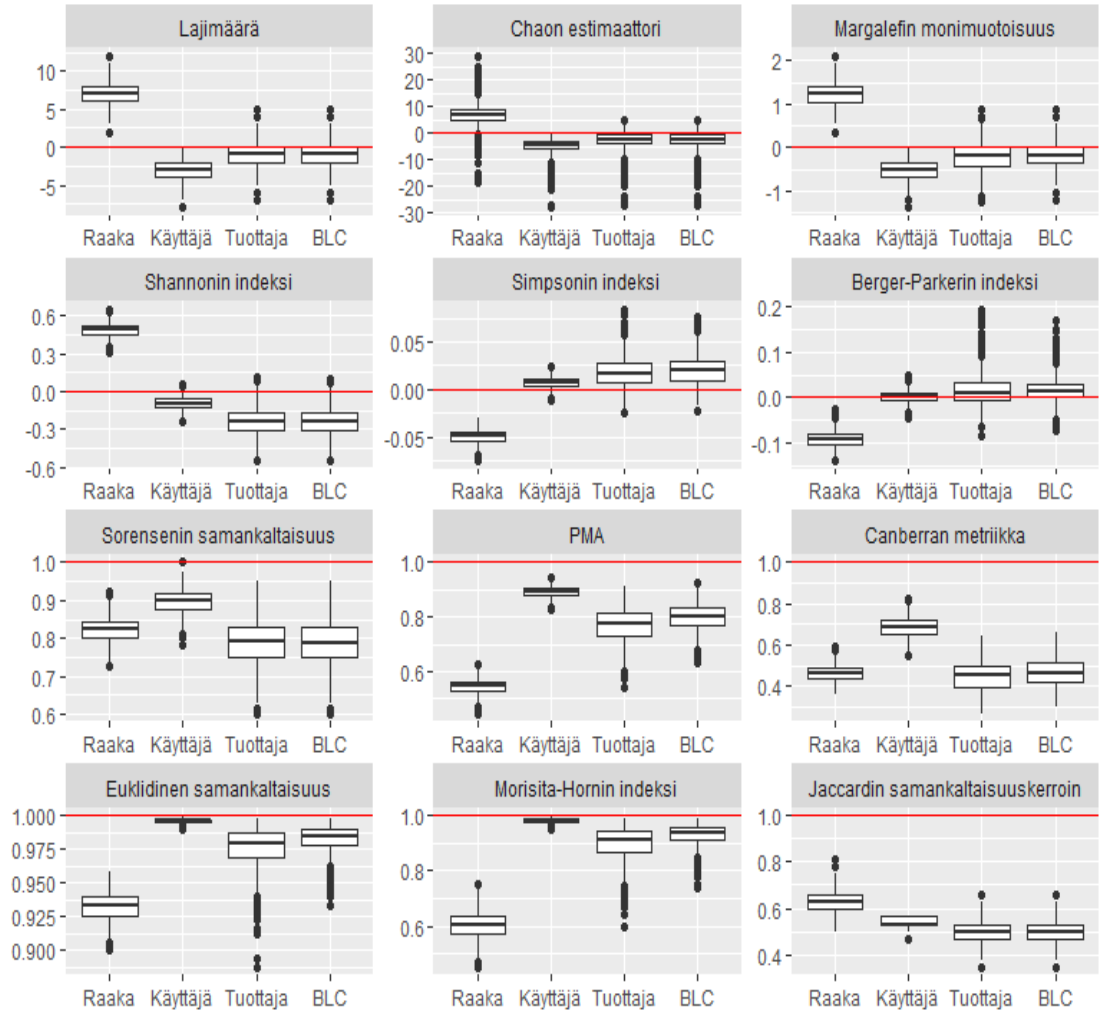
Luokittelijasta riippumatta pienemmällä otoskoolla ei saada yhtä tarkkoja tuloksia kuin suuremmalla otoskoolla, mutta esimerkiksi käyttäjän korjauksella saadut indeksit ovat parhaimmillaan likimain harhattomia. Tämä riippumatta siitä, että luokittelija on huono ja otoskoko pieni. Indeksit eivät kuitenkaan ole harhattomia, esimerkiksi käyttäjän korjauksella Simpsonin indeksi poikkeaa noin 0.01 oikeasta arvosta, joka on huomionarvoinen poikkeama, kun taulukon 5 perusteella Simpsonin indeksin keskiarvo on 0.11. Kuitenkin otoskoko on pieni ja luokittelija erittäin huono, joten siihen nähden korjausmenetelmillä saadaan todella hyviä tuloksia.

Satunnaisvaihtelu indeksien arvoissa lisääntyy hieman pienellä otoskoolla. Tässä sekä sekaannusmatriisi, että näytteet on saatu pienellä otoskoolla. Liitteessä B havaitaan, että sekaannusmatriisin laskemisessa käytetty otoskoko on merkittävämpi tulosten kannalta. Pieni näytteen koko vaikuttaa lajimäärän estimointiin, mutta muuten sen vaikutus ei ole suuri. Jos sekaannusmatriisi on laskettu pienellä otoskoolla, kaikkien indeksien tarkkuus heikkenee.

Jos halutaan verrata hyvää luokittelijaa pienellä otoskoolla (kuva 3) ja huonoa luokittelijaa suurella otoskoolla (kuva 2), huomataan korjausmenetelmien välillä eroja. Käyttäjän sekaannusmatriisi antaa jopa harhattomampia tuloksia suuren otoskoon huonolla luokittelijalla kuin pienen otoskoon hyvällä luokittelijalla. Tuottajan sekaannusmatriisikorjaus ja paras lineaarinen korjaus sen sijaan ovat harhattomampia pienellä otoskoolla ja hyvällä luokittelijalla kuin suuren otoskoon huonolla luokittelijalla. Näin ollen käyttäjän korjauksessa otoskoon riittävyys on tärkeintä, menetelmä toimii hyvin myös huonolla luokittelijalla. Muiden menetelmien kohdalla luokittelun onnistuminen on tärkeämpää kuin otoskoko. Käyttäjän sekaannusmatriisikorjauksen käyttäminen on erityisen hyödyllistä silloin, kun käytössä on huono luokittelija, mutta otoskoko on riittävä. Tuottajan sekaannusmatriisikorjaus ja paras lineaarinen korjaus ovat joidenkin indeksien, erityisesti lajimäärää ja monimuotoisuutta mittaavissa indekseissä varteenotettava vaihtoehto, mikäli otoskoko on pieni.

5.4.3 Sekaannusmatriisin estimointiin käytettävän jokityypin vaikutus

Edellisissä simuloinneissa ryhmien osuuksia $\mathbf{p}$ estimoitiin ennalta asetetusta jokityypistä. Tämä jokityyppi on yleinen, mutta jokityypeissä on eroja. Joissakin jokityypeissä on vain

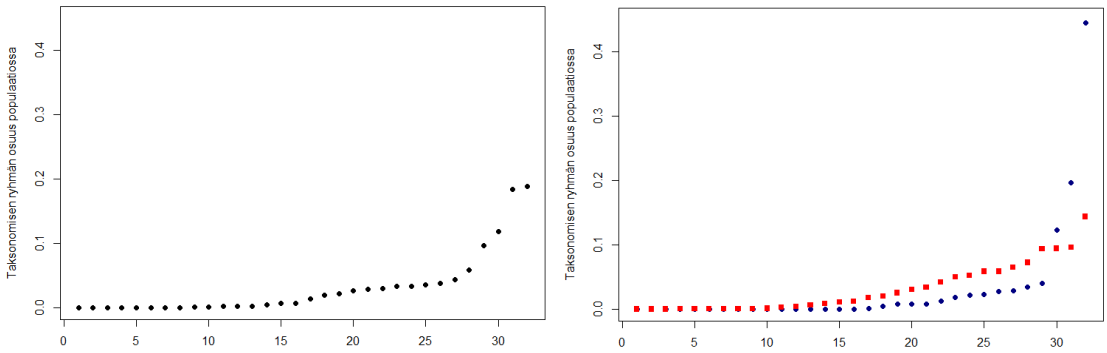


Kuva 4: Huonolla luokittelijalla (naiivi Bayes) ja 300 pohjaeläimen otoskoolla saatujen indeksien ero oikeaan indeksiin. Jos arvo on punaisen poikkiviivan kohdalla, se on täsmälleen oikea indeksin arvo. Indeksien jakaumat on saatu simuloimalla tuhat otosta.

muutamia dominoivia ryhmiä ja joissain tyypeissä on paljon jokseenkin yhtä suuria ryhmiä. Jossakin jokityypissä yleinen taksonominen ryhmä voi olla harvinainen toisessa. Jos simuloinnit tehdään jokityypin pohjalta, joka edustaa toista ääripäätä, korjaukset onnistuvat pääosin samalla tavalla kuin keskiverto-jokityypissä. Tuottajan sekaannusmatriisikorjaus ja paras lineaarinen korjaus yliestimoivat lajimäärän, kun lajeja jokityypissä on erityisen vähän tai erityisen paljon. Käyttäjän sekaannusmatriisikorjaus aliestimoi lajimäärän, kuten aikaisemminkin, mutta yleensä kyse on muutaman lajin erosta oikeaan lajimäärään.

Kuvassa 5 on esitetty käytettyjen jokityyppien taksonomisten ryhmien suhteelliset osuudet, joista ilmenee jokityyppien erot. Vasemmanpuoleisessa kuvassa on tyypillinen jokityyppi, jota on käytetty edellä olleissa analyyseissä. Oikeanpuoleisessa kuvassa on joki, jonka poh-

jaeläimistä suurin osa kuuluu kahteen taksonomiseen ryhmään ja joki, jossa pohjaeläimet ovat jakaantuneet tasaisesti eri taksonomisiin ryhmiin. Kuvissa taksonomiset ryhmät ovat suhteellisen osuuden mukaan suuruusjärjestyksessä jokainen jokityyppi kerrallaan.

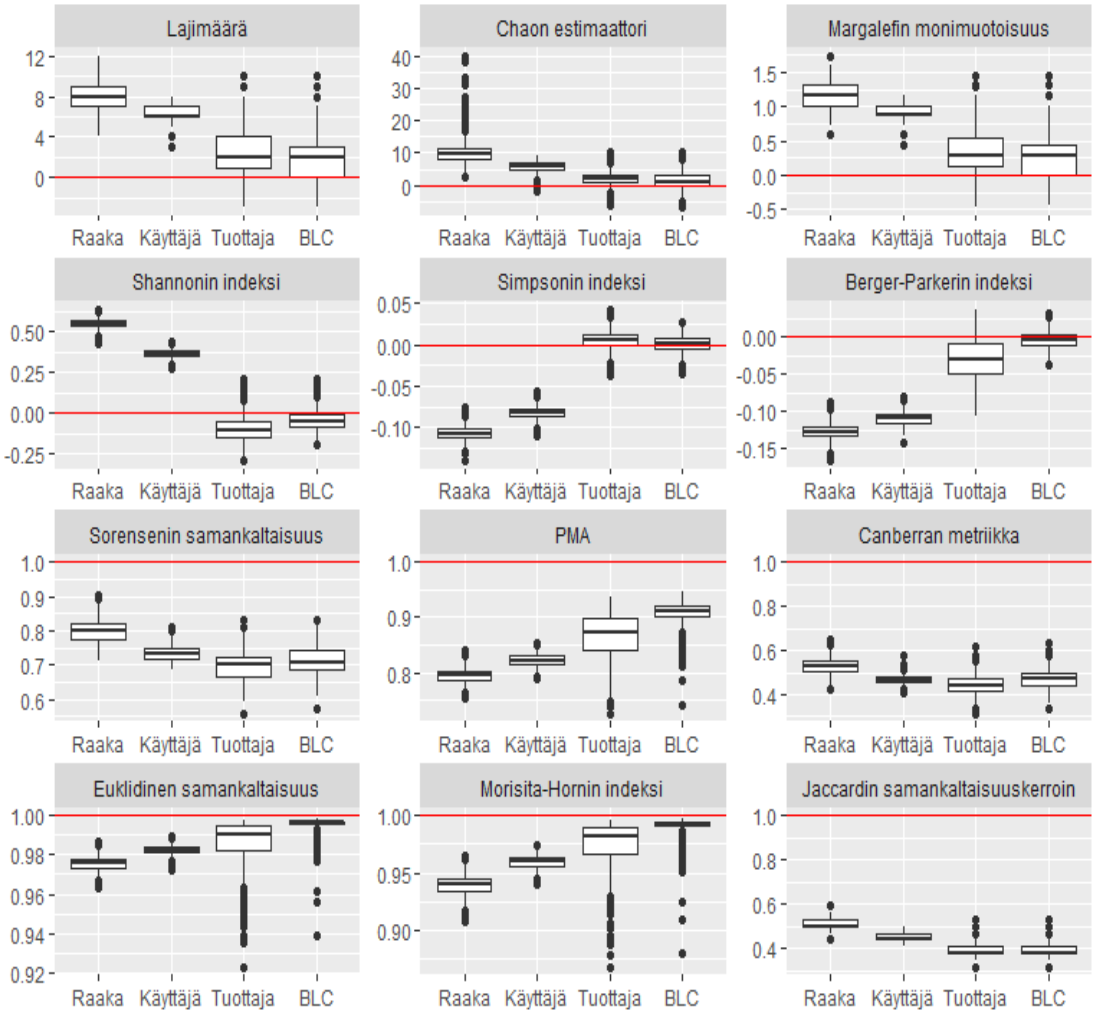


Kuva 5: Taksonomisten ryhmien suhteelliset osuudet eri jokityypeissä. Vasemmanpuoleisessa kuvassa tutkimuksessa pääosin käytetty jokityyppi. Oikeanpuoleisessa kuvassa on kaksi hyvin erilaista jokityyppiä, joita on käytetty havainnollistamaan jokityypin vaikutusta. Punaisilla neliöillä on esitetty jokityyppi, jossa taksonomisten ryhmien runsaudet ovat jokseenkin samat. Sinisillä ympyröillä on esitetty jokityyppi, jossa muutama laji dominoi koko populaatiota. Huomioi, että taksonomiset ryhmät on esitetty järjestyksessä pienimmästä osuudesta suurimpaan.

Jokityypillä ei siis ole merkittävää vaikutusta, kunhan sekaannusmatriisi on muodostettu samasta jokityypistä kuin mistä otos otetaan. Kuitenkin käytännön tilanteessa on mahdollista, että sekaannusmatriisi on aikaisemmin muodostettu jostain tietystä jokityypistä, jota sitten käytetään myös muista jokityypeistä hankittujen otosten korjaamiseen. Koska taksonomisten ryhmien osuudet voivat erota huomattavan paljon jokityypeittäin, voi edellä mainittu menettely johtaa harhaisiin indekseihin.

Väärästä jokityypistä muodostetun sekaannusmatriisin käyttäminen voi pahimmillaan johtaa hyvin harhaisiin indekseihin. Kuvassa 6 on hyvällä luokittelijalla ja suurella otoskoolla lasketut indeksien arvot korjauksille ja raakaestimaateille. Tässä sekaannusmatriisi estimoitiin mahdollisimman monipuolisesta jokityypistä, jossa on paljon taksonomisia ryhmiä. Otokset simuloitiin jokityypistä, jossa kolme ryhmää dominoi populaatiota ja muut taksonomiset ryhmät ovat hyvin pieniä (kuva 5, oikeanpuoleiset kuvaajat).

Lajimäärään liittyvät estimaattorit ovat kaikki ylöspäin harhaisia, sillä sekaannusmatriisin estimaattori on muodostettu lajirikkaasta jokityypistä. Käyttäjän korjaus on muita korjauksia harhaisempi, muuten korjausmenetelmien välillä ei ole eroja. Tasaisuutta ja monimuotoisuutta mittaavissa indekseissä käyttäjän sekaannusmatriisikorjaus on vain vähän raakaestimaatteja parempi. Tuottajan sekaannusmatriisikorjaus on kohtuullisen hyvä es-



Kuva 6: Jos sekaannusmatriisi on estimoitu erilaisesta jokityypistä, niin hyvästä luokittelijasta (satunnainen metsä) ja suuresta otoskoosta huolimatta korjaukset eivät onnistu hyvin. Punainen poikkiviiva vastaa indeksin oikeaa arvoa, johon raakaestimaatteja ja korjausmenetelmiä verrataan. Jakaumat on saatu tuhannesta näytteestä otoskoon ollessa jokaisessa tuhat pohjaeläintä.

timaattori. Paras lineaarinen korjaus on näissä indekseissä parhaimmillaan harhaton ja korjausmenetelmistä paras.

Samankaltaisuusindekseissä jokainen korjausmenetelmä on harhainen. Kuitenkin PMA-indeksin, euklidisen samankaltaisuuden ja Morisita-Hornin indeksin korjaaminen parhaalla lineaarisella korjauksella onnistuu hyvin. Tuottajan korjaus onnistuu vähän huonommin ja käyttäjän selkeästi huonommin. Sørensenin samankaltaisuus, Canberran metriikka ja Jaccardin samankaltaisuuskerroin ovat tarkimmillaan kun käytetään raakaestimaatteja.

Jos sekaannusmatriisi on estimoitu erilaisesta populaatiosta kuin mihin sitä käytetään, paras lineaarinen korjaus on paras korjausmenetelmä. Käyttäjän sekaannusmatriisikorjaus on

kaikkein huonoin menetelmä. Ainoastaan luokittelijan ollessa huono käyttäjän korjaus voi olla joidenkin indeksien kohdalla paras vaihtoehto. Tämä tulos on ristiriidassa aiemmin saman jokityypin sekaannusmatriisilla saatujen tulosten kanssa. Vaikuttaakin siltä, että korjausmenetelmän valinta riippuu sekaannusmatriisin muodostamisesta, eikä niinkään luokittelijasta tai otoskoosta. Liitteessä C on viiksilaatikot pienen otoskoon ja huonon luokittelijan tapauksissa. Hyvällä luokittelijalla ja pienellä otoskoolla käyttäjän sekaannusmatriisi saattaa olla paras korjausmenetelmä lajimäärää estimoitaessa.

Jaccardin samankaltaisuuskerrointa lukuunottamatta jokaisen indeksin arvoja pystytään korjaamaan jollakin korjausmenetelmällä. Sekaannusmatriisin ollessa samanlaisesta populaatiosta (sama $\mathbf{p}$) kuin mistä otos otetaan, käyttäjän sekaannusmatriisikorjaus on aina paras tai yhtä hyvä valinta kuin muut korjaukset. Erityisesti huonolla luokittelijalla käyttäjän korjaus on selkeästi paras. Lisäksi korjatun ja oikean indeksin erotuksilla on pieni vaihtelu. Näin ollen yksittäisen otoksen korjaus poikkeaa harvoin merkittävän paljon oikeasta indeksistä. Pienellä otoskoolla muut korjaukset eivät eroa yhtä selkeästi käyttäjän korjauksesta, mutta se on silti paras. Sen sijaan jos sekaannusmatriisi on estimoitu erilaisesta populaatiosta, paras lineaarinen korjaus on käyttökelpoisin korjausmenetelmä. Erilaisesta populaatiosta estimoiminen on vastoin korjausmenetelmien oletusta siitä, että otanta tehdään satunnaisesti perusjoukosta, joten tällainen tilanne on lähtökohtaisesti ongelmallinen.

6 Yhteenveto

Tutkimuksen pääpainopiste oli selvittää sekaannusmatriisikorjausten toimivuutta estimoitaessa biologisia indeksejä pohjaeläinnäytteistä. Tutkitut kolme menetelmää onnistuivat useimmiten korjaamaan luokittelua ja vähentämään luokittelusta johtuvaa indeksien harhaa, jopa luokittelun onnistuessa huonosti. Otoskoolla ei myöskään ollut suurta vaikutusta indeksien harhaan. Kuitenkaan korjaukset eivät toimineet jokaisessa tapauksessa. Oleellista on, että onko sekaannusmatriisi estimoitu samasta tai samankaltaisesta populaatiosta kuin mistä näyte on saatu.

Jos näyte ja sekaannusmatriisin estimaatti ovat samasta populaatiosta, niin käyttäjän sekaannusmatriisikorjaus on erinomainen valinta. Lähes kaikissa indekseissä korjaus on lähestulkoon harhaton, sen vaihtelu on vähäistä ja se toimii hyvin myös huonolla luokittelijalla ja pienellä otoskoolla. Menetelmä on myös hyvin yksinkertainen toteuttaa, joten se on myös käytännöllisesti katsoen hyödyllinen.

Jos tiedetään, etteivät sekaannusmatriisi ja näyte ole samanlaisesta populaatiosta, silloin paras lineaarinen korjaus on suositeltava korjausmenetelmä. Sillä korjaus onnistuu useimmiten hyvin, mutta joidenkin indeksien kohdalla korjaus ei onnistu kunnolla. Menetelmän toimivuus johtunee siitä, että se on optimoitu yli kaikkien mahdollisten osuusvektoreiden $\mathbf{p}$ (Fortier, 1992). Vaikka lähtökohtaisesti sekaannusmatriisia ei pitäisi estimoida erilaisesta populaatiosta kuin mihin sitä käytetään, käytännössä tämä ei ole aina mahdollista. Esimerkiksi olisi työlästä muodostaa oma sekaannusmatriisi jokaiselle tässä työssä käytetyille 24 eri jokityypille.

Miksi indeksit ovat harhaisia? Yksi tekijä on otoskoko. Harhattomuus toteutuu otoskoon ollessa hyvin suuri: lisätutkimuksessa huomattiin, että käyttäjän sekaannusmatriisikorjauksella lasketut indeksit ovat liki harhattomia käytettäessä otoskokona sataa tuhatta yksilöä, jolloin keskimäärin luokassa olisi kolme tuhatta pohjaeläintä. Tällaiset otoskoot ovat kuitenkin käytännössä epärealistisia. Edes tällaisella otoskoolla tuottajan sekaannusmatriisikorjaus ja paras lineaarinen korjaus eivät ole harhattomia. Syynä tähän lienee matriisien singulaarisuusongelma. Sekaannusmatriisin muokkaaminen kääntyväksi lisännee harhaa indekseihin. Tutkimusta voisi laajentaa käyttämällä erilaisia tapoja estimoida singulaarisia käänteismatriiseja ja tutkimalla näiden vaikutusta luokittelusta aiheutuvaan harhaan indekseissä.

Korjausmenetelmiä käytettiin tiettyjen biologisten indeksien laskemiseen pohjaeläinaineistosta. Saatuja tuloksia ei voida suoraviivaisesti yleistää koskemaan muuntyyppisiä aineistoja ja indeksejä, mutta tuloksia voidaan pitää suuntaa antavina. Vaikka esimerkiksi suurin

osa tutkituista indekseistä saatiin liki harhattomaksi, kaikki indeksit eivät olleet korjattavissa. Ongelmallisin indeksi on Jaccardin samankaltaisuuskerroin, mutta myös Canberran metriikan ja Sørensenin samankaltaisuuden estimaatit olivat harhaisia. Näille indekseille voitaisiin kehittää indeksikohtaisia korjausmenetelmiä, joilla indeksien estimointia saataisiin tarkemmaksi. Yleisesti osuuksien korjausmenetelmien harhattomuus ei näytä takaavan korjattujen indeksien harhattomuutta luokittelun suhteen eikä harhattomuutta muutoinkaan. Indeksien varsinaisten ominaisuuksien tutkiminen ei kuitenkaan ollut tämän työn tavoitteita.

Korjausmenetelmillä voidaan parantaa indeksien estimointia, jolloin saadaan entistä tarkempaa tietoa vesistöjen kunnosta. Nopeutensa ansiosta koneellisen tunnistamisen myötä voitaisiin myös tutkia entistä useampia näytteitä vuosittain, jolloin saataisiin nykyistä laajemmin tietoa vesistöjen kunnosta. Tutkituilla korjauksilla indeksit voidaan estimoida tarkasti ja näin koneellinen tunnistaminen ja luokittelu ovat entistä houkuttelevampia vaihtoehtoja perinteiselle manuaaliselle tunnistamiselle.

Viitteet

Barata, J. C. A. & Hussein, M. S. (2012). The Moore–Penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1-2):146–165.

Buckland, S. & Elston, D. (1994). Use of groundtruth data to correct land cover area estimates from remotely sensed data. *Remote Sensing*, 15(6):1273–1282.

Card, D. H. (1982). Using known ap category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, 48(3):431–439.

Chen, X. H., Yamaguch, i. Y., & Chen, J. (2010). A new measure of classification error: Designed for landscape pattern index. *International Archives of Photogrammetry and Remote Sensing and Spatial Information Sciences*, 38(8):759–762.

Ciresan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2011). Convolutional neural network committees for handwritten character classification. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on Document Analysis and Recognition*, pages 1135–1139. IEEE.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision*, volume 1, pages 1–2. Prague.

Drimbarean, A. & Whelan, P. F. (2001). Experiments in colour texture analysis. *Pattern Recognition Letters*, 22(10):1161–1167.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, Inc., New York, second edition.

Fielding, A. H. & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1):38–49.

Fortier, J. (1992). Best linear corrector of classification estimates of proportions of objects in several unknown classes. *The Canadian Journal of Statistics*, 20(1):23–33.

Green, E. (1993). Assessing classification probabilities for thematic maps. *Photogrammetric and Remote Sensing*, 59(5):635–639.

Hay, A. M. (1998). The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9(8):1395–1398.

Healy, J. (1981). The effects of misclassification error on the estimation of several population proportions. *Bell System Technical Journal*, 60(5):697–705.

Hess, G. R. & Bay, J. M. (1997). Generating confidence intervals for composition-based landscape indexes. *Landscape Ecology*, 12:309–320.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579.

Joutsijoki, H., Meissner, K., Gabbouj, M., Kiranyaz, S., Raitoharju, J., Ärje, J., Kärkkäinen, S., Tirronen, V., Turpeinen, T., & Juhola, M. (2014). Evaluating the performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates. *Ecological Informatics*, 20:1–12.

Kiranyaz, S., Ince, T., Pulkkinen, J., Gabbouj, M., Ärje, J., Kärkkäinen, S., Tirronen, V., Juhola, M., Turpeinen, T., & Meissner, K. (2011). Classification and retrieval on macroinvertebrate image databases. *Computers in Biology and Medicine*, 41(7):463–472.

Magurran, A. E. (2004). *Measuring Biological Diversity*. Blackwell Publishing.

Novak, M. & Bode, R. (1992). Percent model affinity: a new measure of macroinvertebrate community composition. *Journal of the North American Benthological Society*, 11(1):80–85.

Pal, N. R. & Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294.

Penrose, R. (1955). A generalized inverse for matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 406–413. Cambridge University Press.

Prisley, S. & Smith, J. (1987). Using classification error matrices to improve the accuracy of weighted land-cover models. *Photogrammetric Engineering and Remote Sensing (USA)*, 53(9):1259–1263.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rasband, W. (1997-2010). *ImageJ*. U.S. National Institutes of Health, Bethesda, Maryland, USA. http://rsb.info.nih.gov/ij/.

Rasband W. (1997). *ImageJ Manual*. http://rsbweb.nih.gov/ij/docs/menus/analyze.html.

Ravi, N., Dandekar, N., Mysore, P., & Littman, M. L. (2005). Activity recognition from accelerometer data. In *American Association for Artificial Intelligence*, volume 5, pages 1541–1546.

Renkonen, O. (1938). *Statistisch-ökologische Untersuchungen über die terrestrische Käferwelt der finnischen Bruchmoore.* PhD thesis, Societas zoologica-botanica Fennica Vanamo.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 32–36. IEEE.

Story, M. & Congalton, R. G. (1986). Accuracy assessment: a user's perspective. *Photogrammetric Engineering and Remote Sensing*, 52(3):397–399.

Suomen ympäristökeskus, Aroviita, J. et al. (2012). Ohje pintavesien ekologisen ja kemiallisen tilan luokitteluun vuosille 2012-2013 - päivitetyt arviointiperusteet ja niiden soveltaminen. https://helda.helsinki.fi/bitstream/handle/10138/41788/OH_7_2012.pdf?

Theodoridis, S., Koutroumbas, K., et al. (2008). Pattern recognition. *IEEE Transactions on Neural Networks*, 19(2):376.

Tirronen, V., Caponio, A., Haanpää, T., & Meissner, K. (2009). Multiple order gradient feature for macro-invertebrate identification using support vector machines. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 489–497. Springer.

Trier, O. D., Jain, A. K., Taxt, T., et al. (1996). Feature extraction methods for character recognition-a survey. *Pattern recognition*, 29(4):641–662.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S.* Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wolda, H. (1981). Similarity indices, sample size and diversity. *Oecologia*, 50(3):296–302.

Ärje, J., Kärkkäinen, S., Turpeinen, T., & Meissner, K. (2013). Breaking the curse of dimensionality in quadratic discriminant analysis models with a novel variant of a Bayes classifier enhances automated taxa identification of freshwater macroinvertebrates. *Environmetrics*, 24(4):248–259.

Ärje, J., Choi, K.-P., Divino, F., Meissner, K., & Kärkkäinen, S. (2016). Understanding the statistical properties of the percent model affinity index can improve biomonitoring related decision making. *Stochastic Environmental Research and Risk Assessment*, 30(7):1981–2008.

Ärje, J., Kärkkäinen, S., Meissner, K., Iosifidis, A., Ince, T., Gabbouj, M., & Kiranyaz, S. (2017). The effect of automated taxa identification errors on biological indices. *Expert Systems with Applications*, 72:108–120.

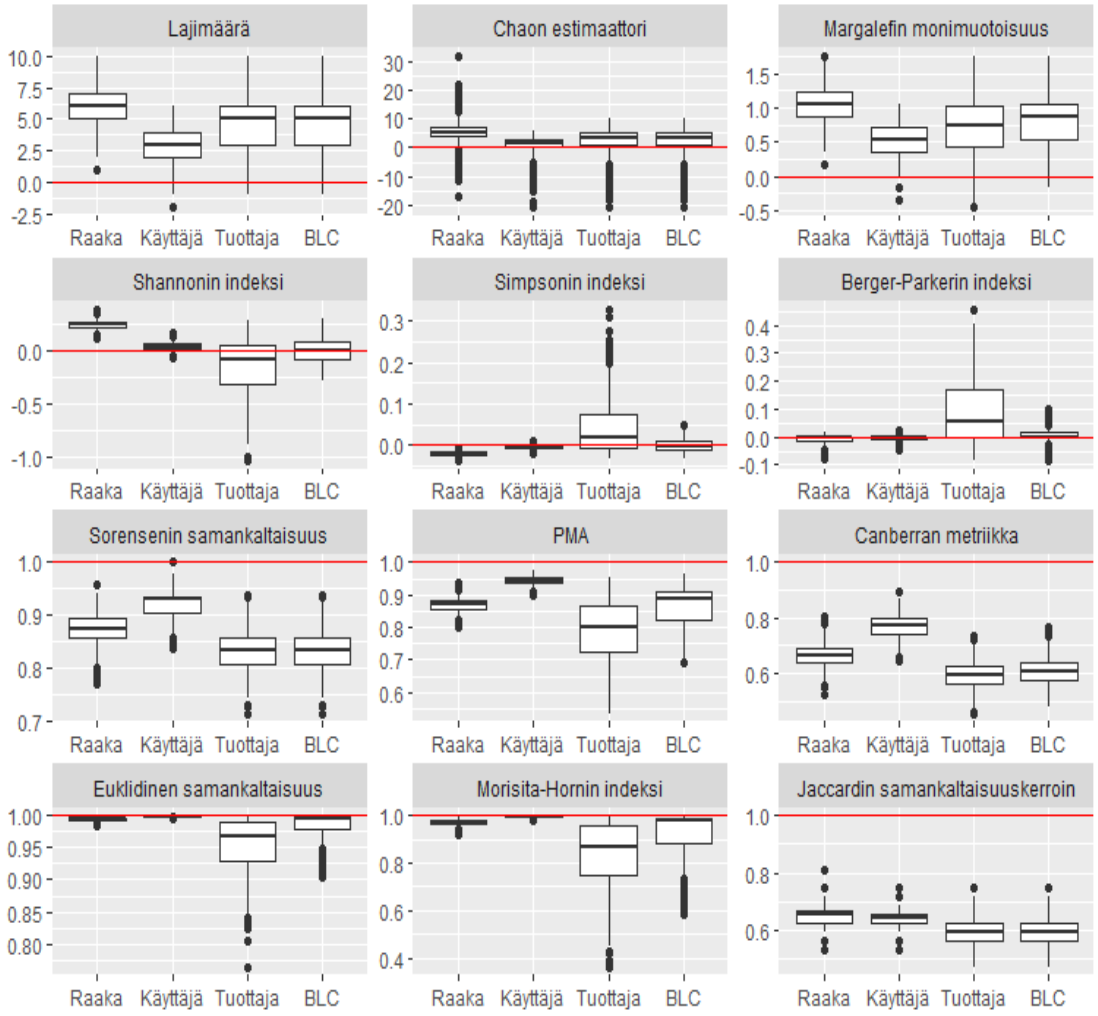
Liitteet

Liite A: Pohjaeläinten taksonomiset ryhmät

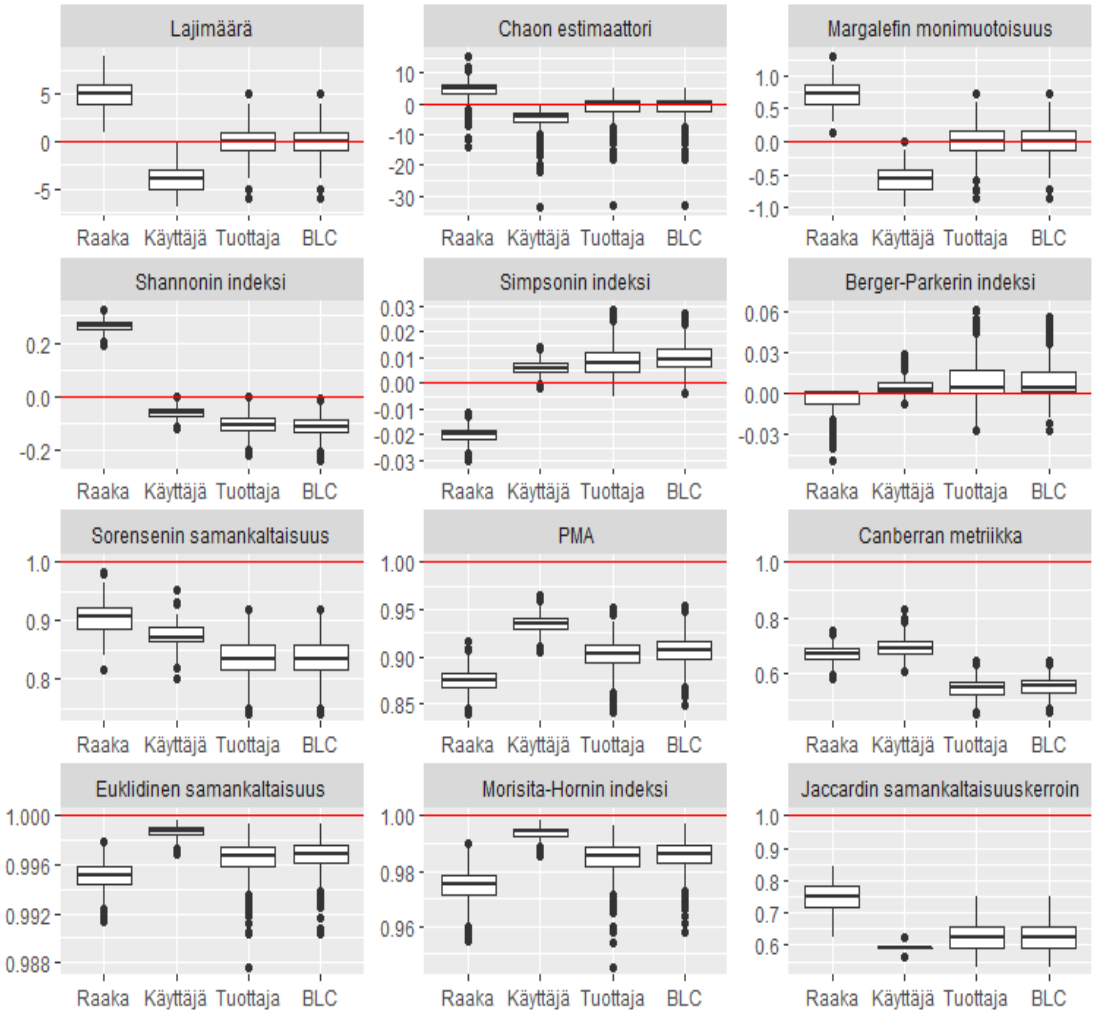
Liite A1: Aineiston pohjaeläinten taksonomiset ryhmät (k=32).

Pohjaeläinten taksonomiset ryhmät		
Ameletus inopinatus	Diura spp.	Isoperla spp.
Arctopsyche ladogensis	Elmis aenea	Leuctra spp.
Asellus aquaticus	Ephemerella aurivillii	Limnius volckmari
Baetis niger group	Ephemerella ignita	Micrasema gelidum
Baetis rhodani	Ephemerella mucronata	Micrasema setiferum
Bithynia tentaculata	Habrophlebia spp.	Nemoura spp.
Caenis spp.	Heptagenia dalecarlica	Sphaeriidae
Corixidae	Hydraena spp.	Protonemura spp.
Ceratopsyche silfvenii	Hydropsyche pellucidula	Rhyacophila nubila
Ceratopogonidae	Hydropsyche saxonica	Taeniopteryx nebulosa
Cheumatopsyche lepida	Hydropsyche siltalai	

Liite B: Otoksen koko sekaannusmatriisin estimoimisessa

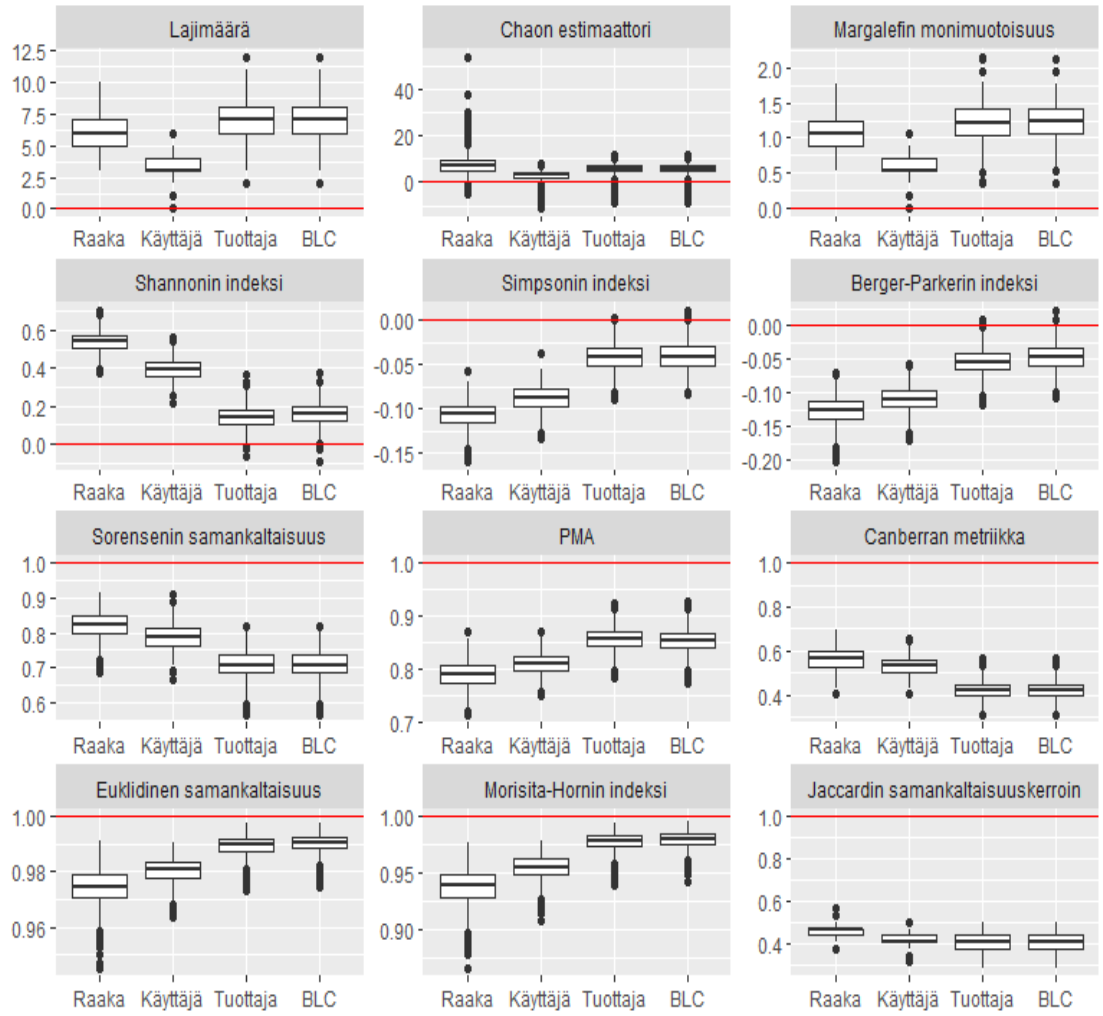


Liite B1: Indeksien poikkeamat oikeista arvoista eri korjausmenetelmillä, kun sekaannusmatriisin koko on 1000 ja otoksen 300. Punainen poikkiviiva vastaa harhatonta indeksin arvoa. Luokittelu tehtiin käyttäen satunnaista metsää.

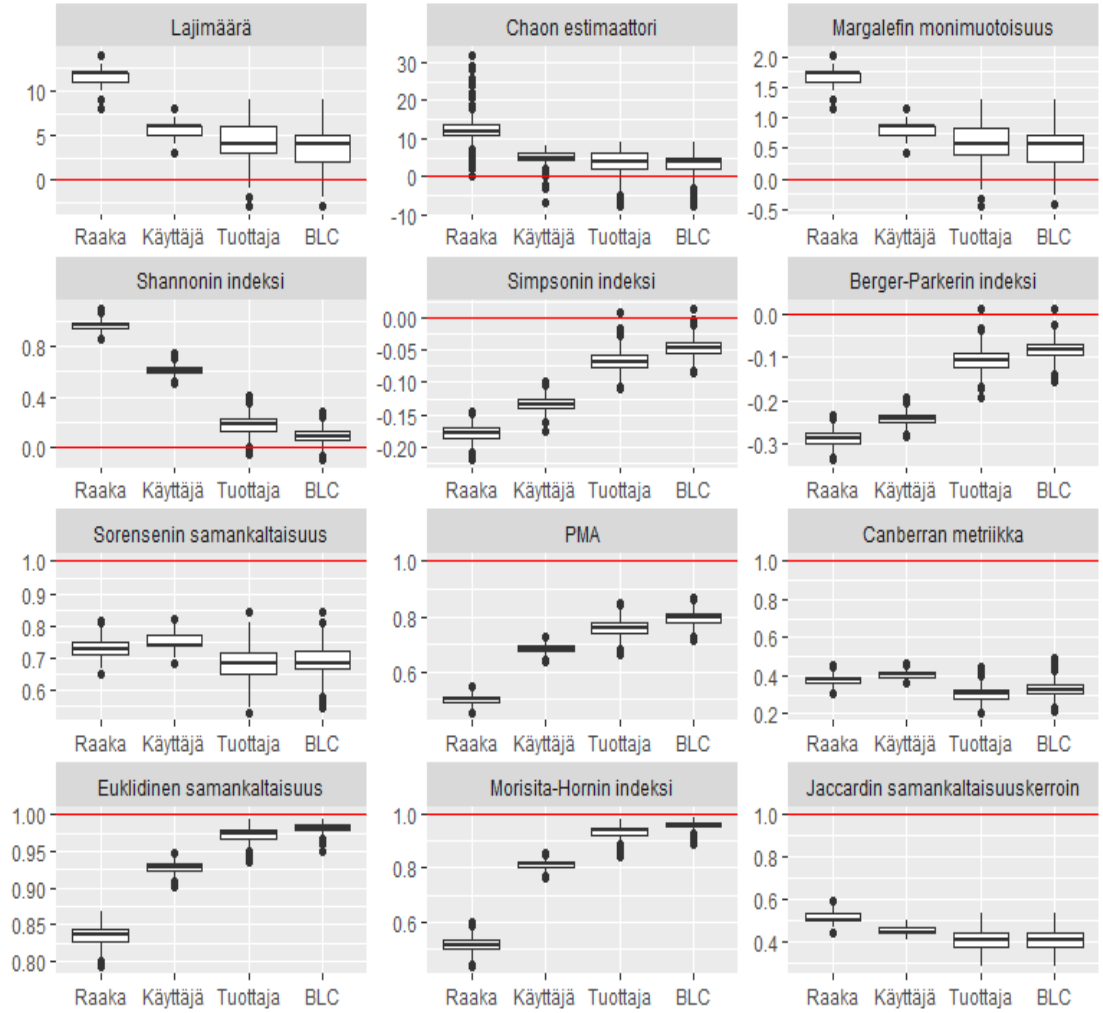


Liite B2: Indeksien poikkeamat oikeista arvoista eri korjausmenetelmillä, kun sekaannusmatriisin koko on 300 ja otoksen 1000. Punainen poikkiviiva vastaa harhatonta indeksin arvoa. Luokittelu tehtiin käyttäen satunnaista metsää. Sekaannusmatriisin koko vaikuttaa enemmän korjauksen onnistumiseen kuin luokiteltavan otoksen koko.

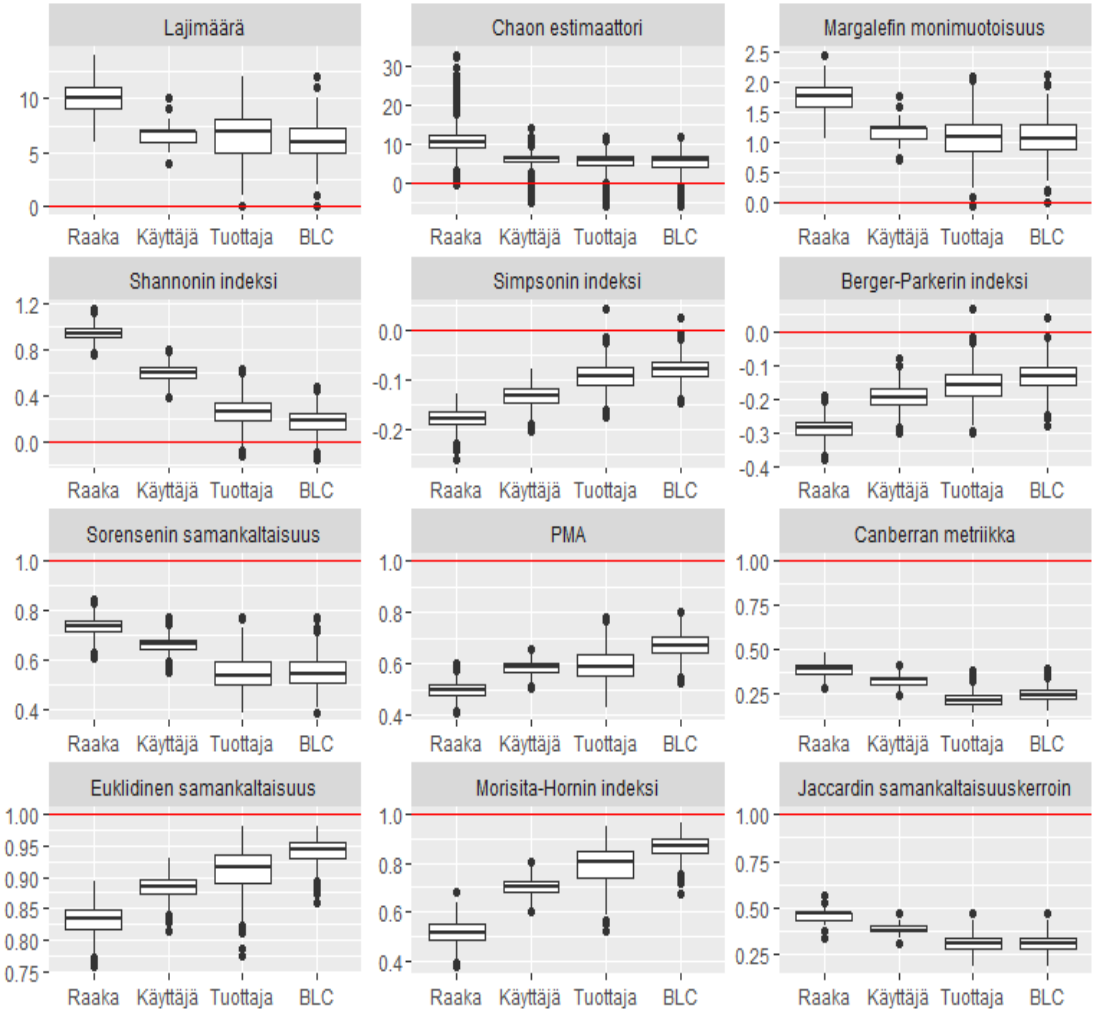
Liite C: Luokittelijan ja otoskoon vaikutus, kun sekaannusmatriisi on estimoitu erilaisesta populaatiosta



Liite C1: Sekaannusmatriisi on estimoitu erilaisesta jokityypistä, kuin mihin sitä käytetään (**p**-vektori on populaatioissa erilainen). Satunnaisella metsällä ja 300 yksilön otoskoolla lasketut indeksien arvot on esitetty poikkeamana oikeista indeksien arvoista. Punainen poikkiviiva vastaa indeksin oikeaa arvoa, johon raakaestimaatteja ja korjausmenetelmiä verrataan.



Liite C2: Sekaannusmatriisi on estimoitu erilaisesta jokityypistä, kuin mihin sitä käytetään (**p**-vektori on populaatioissa erilainen). Naiivi Bayes luokittelijalla ja 1000 yksilön otoskoolla lasketut indeksien arvot on esitetty poikkeamana oikeista indeksien arvoista. Punainen poikkiviiva vastaa indeksin oikeaa arvoa, johon raakaestimaatteja ja korjausmenetelmiä verrataan.



Liite C3: Indeksien poikkeamat oikeista arvoista, kun on käytetty raakaestimaatteja ja korjausmenetelmiä. Punainen poikkiviiva viittaa harhattomaan estimointiin. Jakaumat on laskettu naiivi Bayes -menetelmällä, joka on tässä tapauksessa huono luokittelija, ja otoskoon ollessa 300. Sekaannusmatriisi on estimoitu erilaisesta jokityypistä kuin mistä näyte on saatu.