Yleistetyt lineaariset latenttimuuttujamallit – sovelluksena lajiyhteisöjen mallinnus

Tilastotieteen pro gradu -tutkielma

26. tammikuuta 2022
Markus Kulmala
Matematiikan ja tilastotieteen laitos
Jyväskylän yliopisto

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos
Kulmala, Markus: *Yleistetyt lineaariset latenttimuuttujamallit – sovelluksena lajiyhteisöjen mallinnus*
Tilastotieteen pro gradu -tutkielma, 42 sivua, 1 liite (4 sivua)
26. tammikuuta 2022

Tiivistelmä

Lajiyhteisöjen mallintamiseen liittyvillä menetelmillä pystytään saamaan tietoa ekologisista vuorovaikutussuhteista ja ennustamaan ympäristökovariaattien muutosten vaikutusta lajiyhteisöihin. Tällaiset kysymykset ovat nykyisin erittäin keskeisiä, kun tutkitaan esimerkiksi ilmastonmuutoksen vaikutusta lajien esiintyvyyteen ja lajiyhteisöjen koostumukseen.

Lajiyhteisöjä voidaan mallintaa sekä frekventistisen tilastotieteen että Bayes-menetelmien avulla. Tässä työssä tutkitaan kahden vaihtoehtoisen mallinnustavan eroja ja samankaltaisuuksia sekä teoreettisesti että empiirisesti. Vertailun kohteena ovat frekventistinen yleistetty lineaarinen latenttimuuttujamalli (*generalized linear latent variable models, GLLVM*) ja bayesilaisittain sovitettu hierarkkinen yleistetty lineaarinen sekamalli (*hierarchical modelling of species communities, HMSC*).

Teoreettisen tarkastelun painopiste on mallien sovitustavassa sekä mallien tavassa ottaa huomioon lajien välinen korrelaatiorakenne. Lisäksi osoitamme, että tulkinta mallien tavasta hyödyntää lajikovariaatteja on yhtäläinen. Analyysissä tutkimme mallien parametriestimaatteja sekä selitys- ja ennustevoimaa diskriminaation, tarkkuuden ja kalibraation avulla. HMSC-malli suoriutui tarkasteltujen suureiden osalta keskimäärin GLLVM-mallia paremmin, erojen ollessa suurimpia eniten havaituilla lajeilla. GLLVM-malli kompensoi tuloksia huomattavasti pienemmällä sovitusajalla.

Avainsanat: hierarkkinen yleistetty lineaarinen sekamalli, lajiyhteisön mallinnus, mallidiagnostiikka, MCMC-algoritmi, variaatioapproksimaatio, yleistetty lineaarinen latenttimuuttujamalli

Sisällys

1 Johdanto

Lajiyhteisöjen mallintamiseen -ja ympäristötekijöiden vaikutuksen lajiyhteisöihin ymmärtämiseen tähtäävät tilastolliset menetelmät ovat laajan kiinnostuksen kohteena ekologisessa tutkimuksessa. Niiden avulla pystytään saamaan tietoa ekologisista vuorovaikutussuhteista ja ennustamaan ympäristökovariaattien muutosten vaikutusta lajiyhteisöihin sekä niiden rakenteeseen. Tänä päivänä kysymys on erittäin keskeinen esimerkiksi, kun tutkitaan ilmastonmuutoksen vaikutusta lajien esiintyvyyteen ja lajiyhteisöjen koostumukseen.

Viime vuosina malliperusteiset menetelmät ovat nostaneet suosiota lajiyhteisöjen tilastollisessa tutkimuksessa niiden tulkittavuuden, joustavuuden sekä tehokkuuden ansiosta. Malliperusteiset menetelmät ovat monella tapaa objektiivisempi vaihtoehto lajiyhteisöjen tutkimiseen verrattuna lajien ordinaatioon perustuviin menetelmiin, sillä käytössä ovat perinteiset työkalut esimerkiksi mallinvalinnan tarkasteluun ja tulosten tulkitsemiseen (Warton et al., 2015). Malliperusteisten ratkaisujen vahvuutena on myös niiden kyky sisällyttää malliin laajasti ilmiöön liittyvää tietoa ymmärrettävällä tavalla. Perinteisten ympäristökovariaattien lisäksi pystytään tutkimaan esimerkiksi lajiyhteisön lajikohtaisten muuttujien, kuten lajin painon tai ruokaketjun sijoituksen vaikutusta sekä tutkittavaan lajiyhteisöön että lajin suhtautumiseen ympäristökovariaatteihin (Ovaskainen ja Abrego, 2020; Niku, 2020). Esimerkiksi lajiyhteisöissä usein esiintyvä spatiaalinen autokorrelaatio tai aineiston hierarkkinen luonne voidaan myös ottaa mallinnuksessa huomioon.

Kuten malliperusteisissa menetelmissä yleisesti, voidaan lajijakaumia mallintaa sekä Bayes-tilastotieteen että frekventistisen tilastotieteen menetelmien avulla. Tässä työssä tutkitaan frekventististen yleistettyjen moniulotteisten lineaaristen latenttimuuttujamallien, alkuperäiseltä nimeltä GLLVM (*generalized linear latent variable models*) (Niku, 2020) ja Bayes-menetelmillä sovitettujen moniulotteisten hierarkkisten yleistettyjen lineaaristen sekamallien, alkuperäiseltä nimeltä HMSC (*hierarchical modelling of species communities*) (Ovaskainen ja Abrego, 2020) eroavaisuuksia, kun mallinnetaan lajien läsnäolodataa. Erona frekventistiseen mallinnukseen Bayes-estimoinnissa on

inferenssin perustuminen parametrien posteriorijakaumiin. Siinä missä frekventistisellä mallinnuksella saadaan parametreille piste-estimaatit ja vastaavat estimaattien luottamusvälit, Bayes-mallinnuksessa parametreille saadaan posteriorijakaumaestimaatit sekä niihin perustuvat jakaumakeskiarvot sekä todennäköisyysvälit.

Tässä työssä kiinnostuksen kohteena on mallidiagnostiikka mallien selitys- ja ennustevoimalle lajikohtaisella tasolla, sekä mallien avulla tehtävä inferenssi ympäristökovariaattien vaikutuksesta lajiyhteisöön. Käytettävien mallien vertailun mielekkyyden vuoksi sovitamme käytettävissä olevaan aineistoon tutkimusasetelman näkökulmasta tarkoituksenmukaiset sekä mahdollisimman samankaltaiset mallit. Esittelemme luvussa 2 analyysissä käytettävät mallit sekä niihin liittyvän teorian ja merkinnät. Luvussa 3 tarkastelemme mallien sovitusta ja siihen liittyvää teoriaa sekä esittelemme mallien sovittamiseen tarvittavat hyperparametrit. Mallien selitys- ja ennustevoimaa tarkastellaan diskriminaation, tarkkuuden sekä kalibraation avulla, joiden teoria esitellään luvussa 4.

Tässä työssä käytämme esimerkkiaineistona tutkimuksen Elo et al. (2016) putkilokasviaineistoa. Aineisto koostuu 120 eri suosta (myöhemmin palsta), joista jokainen sisältää 10 havaintopaikkaa. Näistä jokaisesta on tutkittu 131 putkilokasvilajin esiintyvyydet. Aineisto sekä mallinnuksen tulokset esitellään tarkemmin luvussa 5. Lopuksi luvussa 6 käymme läpi pohdintaa työstä ja työssä tehdyistä valinnoista, sekä esitämme mahdollisia jatkotutkimuskysymyksiä työn aiheeseen liittyen.

2 Ekologian lajiyhteisöjen mallinnus

Tässä luvussa tarkastelemme lajiyhteisöjen mallinnukseen käytettäviä GLLVM- ja HMSC-malleja sekä niiden teoriaa. Aloitamme esittelemällä mallinnuksessa käytettävän notaation ja termistön, joka on yhteinen molemmille malleille. Tämän jälkeen esittelemme GLLVM- ja HMSC-mallit ja niihin liittyvän teorian työn kannalta oleellisin osin.

Yleisesti ekologian runsausdata voidaan koota $n \times m$ havaintomatriisiin $\mathbf{Y}$, missä alkio y_{ij} on havaintopaikalta $i = 1, \ldots, n$ havaittu laji $j = 1, \ldots, m$.

Ekologian runsausdatan tutkimusasetelmassa havaintopaikat $i = 1, \ldots, n$ kuuluvat usein korkeampitasoisiin palstoihin $p(i)$, jossa yhdellä palstalla on useampi havaintopaikka. Asetelma on verrattavissa yleistajuiseen luokka–oppilas-hierarkiaan.

Lajiyhteisöjen mallinnuksessa on oleellista, että mallinnuksen avulla voidaan tehdä päätelmiä useammalle kuin yhdelle lajille. Tämän vuoksi mallinnamme samanaikaisesti $m \times 1$ havaintovektoria $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})'$, missä $i = 1, \ldots, n$. Tässä työssä käsittelemme aineistoja, joissa lajista j tiedetään, onko sitä havaittu vai ei. Vastemuuttuja y_{ij} saa arvon yksi, jos laji j on havaittu havaintopaikalla i, ja arvon nolla, jos lajia ei ole havaittu.

Malleissa merkitään $p \times 1$ vektorilla $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ ympäristökovariaatteja, missä alkion x_{ik} indeksi $k = 1, \ldots, p$ viittaa ympäristökovariaatin järjestyslukuun ja $i = 1, \ldots, n$ havainnon järjestyslukuun. Ympäristökovariaatit kerätään $n \times p$ matriisiin $\mathbf{X}$. Lajikovariaatteja merkitään $q \times 1$ vektorilla $\boldsymbol{t}_j = (t_{j1}, \ldots, t_{jq})'$, missä alkio t_{jl} viittaa lajin $j = 1, \ldots, m$ kovariaattiin $l = 1, \ldots, q$.

Tilastollisessa mallinnuksessa käytämme linkkifunktiota, jonka tarkoitus on kuvata lineaarinen prediktori mielekkäälle intervallille. Dikotomisen datan vuoksi emme voi käyttää mallinnuksessa identiteettilinkkifunktiota $g(x) = x$, sillä lineaarinen prediktori saa arvoja välillä $(-\infty, \infty)$ ja vastemuuttuja y_{ij} puolestaan arvoja $\{0, 1\}$. Sen sijaan toimivia linkkifunktioita ovat esimerkiksi logit- ja probit-linkkifunktio. Tässä työssä käytämme molemmissa malleissa probit-linkkifunktiota sen laskennallisten ominaisuuksien vuoksi (Ovaskainen ja Abrego, 2020; Hui et al., 2017). Probit-linkkifunktio on muotoa $g(x) = \Phi(x)$, missä $\Phi(x)$ viittaa standardinormaalijakauman kertymäfunktioon lineaariprediktorin arvolla x.

Yhteistä GLLVM- ja HMSC-malleille on myös niiden tapa mallintaa useita vastemuuttujia samanaikaisesti. Tämänkaltaisia malleja kutsutaan yleistäen sateenvarjotermillä JSDM (*joint species distribution model*). JSDM-mallit eroavat niin kutsutuista SSDM-malleista (*stacked species distribution model*), jotka mallintavat useaa vastemuuttujaa (lajia), mutta yhtä kerrallaan ja erikseen. Lopuksi näiden erillisten mallien tulokset kootaan yhteen. JSDM-mallien oleellinen hyöty SSDM-malleihin verrattuna on, että mallin-

nuksessa pystytään ikään kuin lainaamaan informaatiota lajien välillä, mikäli lajit suhtautuvat käytettyihin kovariaatteihin tarpeeksi samankaltaisesti. Tämä vuorostaan mahdollistaa tarkempien tulosten saamisen mallinnuksesta SSDM-malleihin verrattuna (Elith et al., 2006; Warton et al., 2015) sekä suuremman voiman havaita aineistossa piileviä lainalaisuuksia (Wang et al., 2012).

2.1 Yleistetyt lineaariset latenttimuuttujamallit

Tässä luvussa rakennamme GLLVM-mallin aloittaen yksinkertaisimmasta mahdollisesta lajiyhteisön mallinnustavasta eli yleistetystä lineaarisesta mallista. Mallin vaikeusastetta lisätään askel askeleelta ottamalla mukaan GLLVM-malliperheessä esiintyviä termejä, joiden tarpeellisuus ja hyöty perustellaan tutkimuskysymyksen ja teorian näkökulmasta. Seuraamme mallien esityksessä väitöskirjan Niku (2020) esitystapaa.

2.1.1 Yleistetty lineaarinen malli

Yleistetty lineaarinen malli on perinteisen lineaarisen regression yleistys tilanteisiin, joissa vastemuuttujat eivät ole jatkuva-arvoisia eikä vastemuuttujan mallinnus siten ole mielekästä ilman linkkifunktiota. Yleistetty lineaarinen malli moniulotteiselle runsausdatalle voidaan esittää muodossa

$$g(\mu_{ij}) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j,$$

missä $g(\cdot)$ on vastemuuttujan ja kovariaattien suhdetta kuvaava linkkifunktio, β_{0j} sisältää lajikohtaiset tasoparametrit ja $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})'$ on ympäristökovariaatteja vastaavat regressiokertoimet sisältävä $p \times 1$ vektori, missä vektorin alkio β_{jk} viittaa lajin $j = 1, \ldots, m$ ja ympäristökovariaatin $k = 1, \ldots, p$ väliseen regressiokertoimeen. Mallissa $\mu_{ij} = E[y_{ij}|\boldsymbol{x}_i] = g^{-1}(\beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j)$.

2.1.2 Yleistetty lineaarinen sekamalli

Usein vastemuuttujat ovat keskenään korreloituneita. Esiintyvä korrelaatiorakenne pitää ottaa mallintamisessa huomioon, jotta tulosten avulla tehtävä tulkinta ja sen pohjalta saatu ymmärrys tutkittavasta ilmiöstä ei ole virheellinen. Tällainen korrelaatiorakenne syntyy, kun samalta havaintopaikalta tehdään useita havaintoja; esimerkiksi samalta suolta rekisteröidään useaan otteeseen kasvien lukumäärät tai esiintyvyydet.

Korrelaatiorakenne voidaan ottaa mallinnuksessa huomioon esimerkiksi lisäämällä malliin havaintopaikkakohtainen satunnaistermi. Tällöin säästytään lisäämästä kategorista kiinteän vaikutuksen selittäjää malliin. Havaintopaikkojen välisistä eroista ei yleensä olla erityisen kiinnostuneita, pelkästään niiden aikaansaama korrelaatiorakenne halutaan ottaa huomioon. Lisäämällä satunnaistermi kiinteän vaikutuksen sijasta mallin vapausasteiden määrä on myös pienempi, ja mallin estimaatit tulevat tarkemmiksi varsinkin ryhmille, joissa on vain vähän havaintoja (Gelman ja Hill, 2006; Harrison et al., 2018).

Korrelaatiorakenteen huomioiva malli voidaan kirjoittaa muodossa

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j, \tag{1}$$

missä satunnaisvaikutus $\alpha_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$. Havaintopaikkakohtaisen satunnaisvaikutuksen lisääminen malliin saa aikaan tilanteen, jossa havaintopaikan vaikutus on samanlainen kaikkiin siltä paikalta havaittuihin lajeihin (Niku, 2020). Toisin sanoen kaikkien lajien korrelaatio on positiivinen havaintopaikalla i. Tämä ei luonnollisesti ole validi oletus havaintojen mallintamiseen, sillä yhden lajin esiintyminen voi tehdä toisen lajin olemassaolon lähes mahdottomaksi samalla havaintopaikalla.

Ratkaisuna lisätään malliin (1) havaintopaikan i lajiin j liittyvä satunnaisvaikutus α_{ij}. Malli kirjoitetaan nyt muodossa

$$g(\mu_{ij}) = \alpha_{ij} + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j, \tag{2}$$

missä $\alpha_{ij} \sim N(0, \sigma_{ij}^2)$, $i = 1, \ldots, n$ ja $j = 1, \ldots, m$. Havaintopaikasta ja

lajista riippuvan satunnaisvaikutuksen α_{ij} lisääminen malliin generoi lajien välisen korrelaatiomatriisin, joka ei ole lainkaan rajoitettu. Nyt lajien välistä korrelaatiota kuvastaa yleinen $m \times m$ matriisi $\mathbf{\Omega}$, joka sisältää lajien väliset korrelaatiot. Matriisi $\mathbf{\Omega}$ sisältää kuitenkin $m(m+1)/2$ estimoitavaa parametria, minkä vuoksi mallin sovittaminen melko vähäiselläkin lajimäärällä tulee laskennallisesti vaativaksi (Niku, 2020).

2.1.3 Yleistetty lineaarinen latenttimuuttujamalli

Lajien välistä korrelaatiorakennetta voidaan myös mallintaa yksinkertaisemmin lisäämällä malliin d-dimensioinen latenttimuuttujatermi $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{id})'$, missä $d \ll m$. Aikaisempi korrelaatiorakenteen huomioiva malli (2) voidaan nyt kirjoittaa muodossa

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{\eta}_i'\boldsymbol{\gamma}_j, \tag{3}$$

missä lajikohtainen korrelaatiorakenne huomioidaan lisäämällä latenttimuuttujille $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{id})'$ kertoimet $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jd})'$, jotka ovat (havaintopaikkakohtaisten) latenttimuuttujien lajikohtaiset lataukset. Lataukset voidaan esittää matriisimuodossa $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m)'$, jonka dimensio on $m \times d$. Nyt lajien välistä residuaalikorrelaatiota kuvaava matriisi voidaan kirjoittaa muodossa $\mathbf{\Omega} = \mathbf{\Gamma}\mathbf{\Gamma}'$, jonka dimensio on $m \times m$ ja estimoitavien parametrien määrä on dm (Niku, 2020). Mallin identifioituvuuden vuoksi matriisin $\mathbf{\Gamma}$ yläkolmion arvot pitää asettaa nollaksi sekä diagonaalit itseisarvoiksi (Huber et al., 2004).

Mallissa (3) parametri α_i voi olla joko kiinteä- tai $N(0, \sigma^2)$-jakautunut satunnaisvaikutus. Mallin sovittamista lajien levinneisyysaineistoon suositellaan siten, että parametria α_i käsitellään satunnaisena, sillä kiinteänä vaikutuksena se voi antaa harhaisia tuloksia mallinnuksessa (Warton et al., 2015; Hui et al., 2014). Vaihtoehtoisesti parametrin α voi asettaa myös korkeammalle hierarkkiselle tasolle kuin havaintopaikalle i. Tässä työssä asetamme palstakohtaisen satunnaistermin $\alpha_{p(i)}$, missä $p(i)$ osoittaa, mihin palstaan ha-

vaintopaikka i kuuluu. Nyt malli (3) voidaan esittää muodossa

$$g(\mu_{ij}) = \alpha_{p(i)} + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{\eta}_i'\boldsymbol{\gamma}_j, \tag{4}$$

Malleissa (3) – (4) latenttimuuttujat $\boldsymbol{\eta}_i$ noudattavat d-ulotteista standardinormaalijakaumaa ja ovat riippumattomia havaintopaikkojen i välillä.

2.1.4 Yleistetty lineaarinen latenttimuuttujamalli lajikovariaateilla

Mikäli aineistossa on mukana myös lajikohtaista informaatiota, kuten esimerkiksi lajin keskimääräinen paino tai tieto lajin ruokavaliosta, voidaan sitä hyödyntää lisäämällä malliin lajikovariaatteja selittäjämuuttujiksi. Lajikovariaatit voivat olla joko jatkuvia tai kategorisia, samaan tapaan kuin ympäristökovariaatit. Ilman lajikovariaatteja ympäristökovariaattien vaikutus lajin havaitsemistodennäköisyyteen oletetaan lähtökohtaisesti vakioksi kaikille lajeille, mikä ei usein ole realistinen oletus.

Esimerkkinä voidaan ajatella tilannetta, jossa maaperän pH-arvo on kiinnostuksen alainen ympäristökovariaatti. Sienilajien ja bakteerien tutkimuksesta tiedetään, että alhaiset pH-arvot pienentävät bakteerikasvustoa huomattavasti, kun taas vaikutus on päinvastainen sienikasvustolle (Rousk et al., 2009). Tällöin mallinnuksen kannalta pH-arvon aleneminen aiheuttaa bakteerien esiintymistodennäköisyyden pienenemisen, kun taas sienilajeille esiintymistodennäköisyyden merkittävän kasvun. Lajikovariaattien avulla voimme siis selittää lajien välistä vaihtelua reaktiossa ympäristökovariaatteihin (Niku, 2020).

Lajikovariaatit lisätään malliin ottamalla päävaikutustermit ja interaktiotermi ympäristö- ja lajikovariaateista (Brown et al., 2014). Tällöin malli ilman satunnaisefektiä tai latenttimuuttujatermiä voidaan kirjoittaa muodossa

$$g(\mu_{ij}) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}_e + \mathbf{t}_j'\boldsymbol{\beta}_t + \mathrm{vec}(\mathbf{B}_{te})'(\boldsymbol{x}_i \otimes \boldsymbol{t}_j), \tag{5}$$

missä $\boldsymbol{x}_i \otimes \boldsymbol{t}_j$ on ympäristö- ja lajikovariaattien interaktiotermit sisältävä vektori. Operaattori $\otimes$ viittaa matriisien $\mathbf{A}$ ja $\mathbf{B}$ väliseen Kroneckerin tuloon

$\mathbf{A} \otimes \mathbf{B}$. Matriisi $\mathbf{B}_{te}$ on ympäristö- ja lajikovariaattien interaktiotermien $p \times q$ parametrimatriisi. Operaattori vec viittaa matriisin vektorisaatioon.

Kun malliin (5) lisätään havaintopaikkakohtainen satunnaisvaikutustermi α_i ja latenttimuuttujatermi $\boldsymbol{\eta}_i'\boldsymbol{\gamma}_j$ voidaan malli kirjoittaa muodossa

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_e + \text{vec}(\mathbf{B}_{te})'(\boldsymbol{x}_i \otimes \boldsymbol{t}_j) + \boldsymbol{\eta}_i'\boldsymbol{\gamma}_j.$$

Huomataan, että edellä tasoparametri β_{0j} pitää sisällään lajikovariaattien päävaikutustermin (Niku et al., 2019b).

2.2 Hierarkkiset Bayes-latenttimuuttujamallit

Käsittelemme tässä luvussa Bayes-menetelmillä estimoituja hierarkkisia yleistettyjä latenttimuuttujamalleja. Kutsumme tässä työssä näitä malleja nimellä HMSC kirjan Ovaskainen ja Abrego (2020) terminologian mukaisesti. Tässä luvussa esittelemme HMSC-malleihin liittyvän teorian oleellisilta osilta ja tuomme esiin yhtäläisyyksiä sekä eroavaisuuksia GLLVM-malliin nähden.

Tässä työssä teorian osalta keskitytään mallien vertailukelpoisuuden vuoksi HMSC:n osalta malleihin, joissa lajiyhteisöä mallinnetaan ympäristö- ja lajikovariaateilla, vaikkakin HMSC-malliperhe mahdollistaa monenlaisen informaation sisältämisen mallinnukseen. Ympäristö- ja lajikovariaattien lisäksi täydessä HMSC-mallissa voi olla mukana myös tietoa esimerkiksi havaintopaikan spatiaalisista koordinaateista tai lajien fylogenetiikasta (Ovaskainen ja Abrego, 2020).

HMSC hyödyntää GLLVM:n tapaan latenttimuuttujia mallinnettaessa vastemuuttujien välistä korrelaatiorakennetta, mikä niin ikään pienentää huomattavasti estimoitavien parametrien määrää ja siten myös mallin sovittamiseen kuluvaa aikaa. Ilman latenttimuuttujatermiä estimoitavien parametrien määrä skaalautuu suhteessa lajien määrään m^2, kun taas latenttimuuttujatermien kanssa estimoitavien parametrien määrä skaalautuu suhteessa dm (Ovaskainen ja Abrego, 2020).

Yksinkertaisimmillaan HMSC-malli voidaan kirjoittaa muodossa

$$g(\mu_{ij}) = L^F_{ij} + L^R_{ij}, \tag{6}$$

missä L^F_{ij} viittaa lineaarisen prediktorin kiinteään osaan ja L^R_{ij} lineaarisen prediktorin satunnaiseen osaan, $i = 1, \ldots, n$ viittaa havaintopaikkoihin ja $j = 1, \ldots, m$ lajeihin. Kiinteä osa voidaan kirjoittaa muodossa

$$L^F_{ij} = \boldsymbol{x}'_i \boldsymbol{\beta}_j, \tag{7}$$

missä $\boldsymbol{\beta}_j \sim N(\boldsymbol{\mu}, \mathbf{V})$ ja $\boldsymbol{\beta}_j$ on kaikki lajin j regressiokertoimet vakio mukaanlukien sisältävä $(p+1) \times 1$ vektori ja $\mathbf{V}$ on näiden $(p+1) \times (p+1)$ kovarianssimatriisi.

Mallin (6) satunnaisosa L^R_{ij} voidaan kirjoittaa muodossa

$$L^R_{ij} = \boldsymbol{\eta}'_i \boldsymbol{\gamma}_j = \sum_{h=1}^{d} \eta_{ih} \gamma_{hj},$$

missä $h = 1, \ldots, d$ viittaa latenttimuuttujien määrään ja lajikohtainen korrelaatiorakenne huomioidaan lisäämällä latenttimuuttujille $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{id})'$ kertoimet $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jd})'$, jotka ovat (havaintopaikkakohtaisten) latenttimuuttujien lajikohtaiset lataukset. Voimme nyt kirjoittaa mallin (6) auki muodossa

$$g(\mu_{ij}) = \boldsymbol{x}'_i \boldsymbol{\beta}_j + \boldsymbol{\eta}'_i \boldsymbol{\gamma}_j. \tag{8}$$

Erona GLLVM-malleihin on, että HMSC-malleissa kaikkien satunnaisefektien implementointi tapahtuu latenttimuuttujien avulla. HMSC-malleissa satunnaistermiä (vrt. GLLVM mallin (3) α_i-termi) ei ole (Ovaskainen ja Abrego, 2020). Sen sijaan HMSC-malleissa on mahdollista asettaa latenttimuuttujatermi $\boldsymbol{\eta}$ korkeammalle hierarkkiselle tasolle kuin havaintopaikoille i. Palstakohtainen latenttimuuttuja $\boldsymbol{\eta}_{p(i)} = (\eta_{p(i)1}, \ldots, \eta_{p(i)d})'$ voidaan esittää muodossa $L^R_{ij} = \sum_{h=1}^{d} \eta_{p(i)h} \gamma_{hj}$, missä $p(i)$ viittaa palstaan, johon havaintopaikka i kuuluu. Malli (8) palstakohtaisella latenttimuuttujatermillä voidaan

nyt esittää muodossa

$$g(\mu_{ij}) = \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{\eta}_{p(i)}'\boldsymbol{\gamma}_j. \tag{9}$$

2.2.1 Hierarkkinen latenttimuuttujamalli lajikovariaateilla

Lajikovariaatit lisätään mallin kiinteään osaan hieman eri tavalla kuin GLLVM-malleissa. Saadaksemme selkeän kuvan prosessista todetaan ensin, että mallin (7) oletuksella ympäristökovariaattien vaikutus on lähtökohtaisesti vakio kaikille lajeille j, mikä ei luonnollisesti ole realistinen oletus. Esimerkkinä voidaan ajatella tilannetta, jossa havaintopaikkaa koskeva ympäristökovariaatti kuvastaa havaintopaikan tietyn resurssin runsautta. Tällöin tieto siitä, käyttääkö laji ravinnokseen juuri tämänkaltaista resurssia, vaikuttaa vahvasti siihen, millä tavalla havaintopaikan resurssirikkaus vaikuttaa tämän lajin havaitsemistodennäköisyyteen. Jos laji ei käytä kyseistä resurssia, voidaan ajatella, että resurssin runsaus ei vaikuta lajin havaitsemistodennäköisyyteen, kun taas muussa tapauksessa resurssin rikkaus kasvattaa lajin havaitsemistodennäköisyyttä.

Lajikovariaattien vaikutus lisätään HMSC-mallin kiinteään osaan sallimalla $\boldsymbol{\beta}_j$ parametrien odotusarvon vaihtelevuus lajikohtaisesti. Tämän seurauksena ympäristökovariaattien kertoimet noudattavat jakaumaa

$$\boldsymbol{\beta}_j \sim N(\boldsymbol{\mu}_j, \mathbf{V}),$$

missä vektorin $\boldsymbol{\mu}_j$ alkio μ_{kj} määräytyy lajikovariaatin t_{jl} arvon ja ympäristökovariaatin k välistä yhteyttä kuvaavan λ_{kl} tulojen summana

$$\mu_{kj} = \sum_{l=1}^{q} t_{jl}\lambda_{kl},$$

ja matriisi $\mathbf{V}$ on regressiokertoimien kovarianssimatriisi.

Mallin (8) ja $\boldsymbol{\beta}_j$ lajikohtaisen vaihtelun sallimisen perusteella ei ole vielä täysin selvää, onko lajikovariaattien vaikutusmekanismi HMSC-mallissa vastaavanlainen interaktiotermi kuin GLLVM-mallissa. Osoitetaan seuraavaksi, että lajikovariaattien tulkinta on yhtäläinen HMSC- ja GLLVM-malleissa.

Tiedetään, että GLLVM-mallissa lajikovariaattien lisääminen malliin saa aikaan laji- ja ympäristökovariaattien välisen interaktiotermin. Mikäli HMSC-mallin lineaarisen prediktorin L_{ij}^F odotusarvovektorissa regressiokertoimien β_{kj} yli on laji- ja ympäristökovariatin välinen tulo, niin se voidaan tulkita interaktioterminä. Odotusarvo saa muodon

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}_j}[L_{ij}^F] &= \mathbb{E}_{\boldsymbol{\beta}_j}[\sum_{k=1}^{p} x_{ik}\beta_{kj}] = \sum_{k=1}^{p} x_{ik}\mu_{kj} \\ &= \sum_{k=1}^{p}(x_{ik}\sum_{l=1}^{q} t_{jl}\lambda_{kl}) = \sum_{k=1}^{p}\sum_{l=1}^{q} x_{ik}t_{jl}\lambda_{kl},\end{aligned}$$

mikä vastaa laji- ja ympäristökovariaatin välistä tuloa regressiokertoimella λ_{kl}. Lajikovariaattien tulkinta on siten yhtäläinen GLLVM- ja HMSC-malleissa.

3 Mallien sovitus

Frekventistiset- ja Bayes-menetelmät eroavat mallin sovittamistavassa, minkä vuoksi mallien parametriestimaattien ja niiden hajonnan tulkinta eroaa. Tässä luvussa käymme läpi näiden menetelmien sovitustavat sekä oleellisimmat eroavaisuudet tämän työn kannalta.

3.1 Uskottavuuspäättely

Uskottavuuspäättely pohjautuu uskottavuusfunktioon ja sen maksimointiin parametrien suhteen. Tällä tavoin saamme piste-estimaatit parametreille, jotka maksimoivat uskottavuusfunktion arvon parametrien suhteen. Konseptuaalisesti uskottavuusfunktio voidaan kirjoittaa havaintojen riippumattomuuden vallitessa tulona aineiston havaintojen todennäköisyyksien yli, kun havaintoja x_i, $i = 1, \ldots, n$, pidetään vakioina uskottavuuden suhteen ja parametreja θ käsitellään muuttujana. Yksinkertaisimmillaan uskottavuusfunktio

voidaan kirjoittaa yleisessä muodossa

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

missä $f(x_i|\theta)$ on tiheysfunktion arvo, kun parametrien θ arvot on kiinnitetty. Käytännössä uskottavuusfunktion sijaan maksimoidaan log-uskottavuus $l(\theta) = \log(L(\theta))$ sen helpomman käsiteltävyyden vuoksi. Kiinnostuksen kohteena olevan uskottavuusfunktion, ja siten myös log-uskottavuusfunktion, maksimi $\hat{\theta}$ voidaan antaa muodossa

$$\hat{\theta} = \arg\max l(\theta),$$

joka löytyy uskottavuusfunktion differentioituvuuden ollessa voimassa logaritmin derivaatan $l(\theta)^{'}$ nollakohdasta, kun toinen derivaatta $l(\theta)^{''}$ on negatiivista.

3.1.1 Uskottavuuspäättely GLLVM-malliperheelle

GLLVM-malleissa uskottavuusfunktio koostuu kiinteästä osasta ja satunnaisosasta. Tässä luvussa tarkastelemme GLLVM-malleja, jotka sisältävät havaintopaikkakohtaisen satunnaistermin, latenttimuuttujia sekä ympäristö- ja lajikovariaatteja. Seuraamme GLLVM-mallin uskottavuuspäättelyn esityksessä väitöskirjan Niku (2020) esitystapaa. Uskottavuusfunktion kirjoittamista varten kootaan malliparametrit vektoreihin seuraavasti. Havaintopaikkakohtaiset satunnaistermit kootaan vektoriin $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$ ja lajikohtaiset tasoparametrit vektoriin $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0m})'$. Kovariaattikohtaiset parametrit kootaan $1 \times pm$ vektoriin $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_m')'$. Lopuksi parametrit kerätään vektoriin $\boldsymbol{\Psi} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}, \mathrm{vec}(\boldsymbol{\Gamma}))'$, missä $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m)'$ sisältää latenttimuuttujien lataukset. Latenttimuuttujatermit ovat $1 \times nd$ vektorissa $\mathbf{H} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_n')'$.

Uskottavuusfunktio GLLVM-malliperheelle voidaan nyt kirjoittaa yleises-

sä muodossa (Niku, 2020)

$$L(\boldsymbol{\Psi}, \boldsymbol{\alpha}, \mathbf{H}) = \prod_{i=1}^{n} \left(\prod_{j=1}^{m} f(y_{ij}|\alpha_i, \boldsymbol{\eta}_i, \boldsymbol{\Psi}) \right) f(\alpha_i) f(\boldsymbol{\eta}_i), \tag{10}$$

missä funktio $f(\cdot)$ kuvaa järjestyksessä vaste,- satunnais- ja latenttimuuttujien tiheysjakaumia. Uskottavuusfunktion (10) logaritmi voidaan esittää yleisessä muodossa

$$l(\boldsymbol{\Psi}, \boldsymbol{\alpha}, \mathbf{H}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \log f(y_{ij}|\alpha_i, \boldsymbol{\eta}_i, \boldsymbol{\Psi}) + \log f(\boldsymbol{\eta}_i) + \log f(\alpha_i) \right).$$

Tavoitteena on maksimoida uskottavuusfunktio parametrien $\boldsymbol{\Psi}$ suhteen. Yhtälön maksimoiva $\hat{\boldsymbol{\Psi}}$ voidaan esittää muodossa

$$\hat{\boldsymbol{\Psi}} = \arg\max \log(L(\boldsymbol{\Psi})).$$

Huomataan, että satunnaistermi α_i sekä latenttimuuttujatermi $\boldsymbol{\eta}_i$ eivät ole havaittua tietoa, joten ne pitää integroida pois uskottavuudesta $L(\boldsymbol{\Psi}; \boldsymbol{\alpha}, \mathbf{H})$. Tällöin käsittelemme mallin marginaaliuskottavuutta. Uskottavuusfunktion integraaleille ei kuitenkaan ole suljetun muodon ratkaisua, kun vastemuuttujan normaalijakaumaoletus ei päde ja linkkifunktiona ei voida käyttää identtistä linkkiä (Hui et al., 2017). Lähestymistapoja ongelman ratkaisuun on monia. Tässä työssä käsittelemme variaatioapproksimaatioita uskottavuuden logaritmille.

3.1.2 Variaatioapproksimaatio

Variaatioapproksimaatiossa perusideana on löytää alkuperäiselle uskottavuusfunktiolle suljetun muodon approksimaatio. Tämä tapahtuu etsimällä alkuperäiselle log-uskottavuudelle alaraja, joka voidaan esittää suljetussa muodossa sekä maksimoida. Maksimoimalla log-uskottavuuden alarajan, etäisyys aitoon integraaliin minimoituu (Niku, 2020).

Variaatioalaraja mallin (3) marginaaliselle log-uskottavuudelle voidaan

esittää yleisessä muodossa (Niku et al., 2019a)

$$\underline{l}(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \int \log \left(\frac{f(\boldsymbol{y}_i|\boldsymbol{\eta}_i^*, \boldsymbol{\Psi})}{q(\boldsymbol{\eta}_i^*|\boldsymbol{\xi})} f(\boldsymbol{\eta}_i^*) \right) q(\boldsymbol{\eta}_i^*|\boldsymbol{\xi}) d\boldsymbol{\eta}_i^*, \tag{11}$$

missä $\boldsymbol{\eta}_i^* = (\boldsymbol{\eta}_i', \alpha_i)'$ ja $q(\boldsymbol{\eta}_i^*|\boldsymbol{\xi})$ on jokin variaatiojakauma parametreilla $\boldsymbol{\xi}$.

Alarajan ja variaatioapproksimaation GLLVM-malliperheelle esitti ensimmäisenä Hui et al. (2017), jonka esitystä seuraamme Niku et al. (2019a) ohella probit-mallin VA-uskottavuuden esityksessämme. Alarajan muodostamiseksi probit-mallille esitämme uskottavuuden (10) apumuuttujan $z_{ij} \sim N(\upsilon_{ij}, 1)$ avulla, missä υ_{ij} on valitun mallin lineaarinen prediktori. Lisäksi apumuuttujalle z_{ij} pätee $y_{ij} = 1$, kun $z_{ij} \geq 0$ ja $y_{ij} = 0$ muulloin. Nyt malli dikotomiselle vastemuuttujalle probit-linkkifunktiolla voidaan kirjoittaa muodossa

$$f(y_{ij}|z_{ij}, \alpha_i, \boldsymbol{\eta}_i, \boldsymbol{\Psi}) = I(z_{ij} \geq 0)^{y_{ij}} I(z_{ij} < 0)^{1-y_{ij}},$$

missä $I(\cdot)$ on indikaattorifunktio.

Apumuuttujan z_{ij} lisääminen malliin mahdollistaa variaatioapproksimaation laskemisen suljetussa muodossa. Nyt marginaalinen log-uskottavuus probit-mallille on muotoa

$$\underline{l}(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left(\int \int \int \prod_{j=1}^{m} f(y_{ij}|z_{ij}, \alpha_i, \boldsymbol{\eta}_i, \boldsymbol{\Psi}) f(z_{ij}) f(\alpha_i) f(\boldsymbol{\eta}_i) d\boldsymbol{z}_i d\alpha_i d\boldsymbol{\eta}_i \right).$$

Variaatioapproksimaatiota varten valitsemme havaintopaikkakohtaiselle satunnaistermille α_i ja latenttimuuttujille $\boldsymbol{\eta}_i$ variaatioapproksimaatiojakaumaksi $q(\boldsymbol{\eta}_i^*|\boldsymbol{\xi})$ normaalijakauman $N(\mathbf{a}_i, \mathbf{A}_i)$, missä $\mathbf{a}_i$ on $d+1$ -pituinen odotusarvovektori ja $\mathbf{A}_i = \text{bdiag}(\text{A}_{\alpha_i}, \mathbf{A}_{\boldsymbol{\eta}_i})$, missä A_{α_i} kuvaa satunnaismuuttujan varianssia, $\mathbf{A}_{\eta_i}$ on latenttimuuttujien rajoittamaton $d \times d$ kovarianssimatriisi ja bdiag on lohkodiagonaalioperaattori (Niku et al., 2019a). Variaatioparametrit kootaan vektoriin $\boldsymbol{\xi} = (\mathbf{a}_i', \text{vec}(\mathbf{A}_i)')'$.

Apumuuttujalle z_{ij} variaatioapproksimaatiojakaumaksi $q(z_{ij})$ valitsemme satunnaistermistä ja latenttimuuttujista riippumattoman katkaistun nor-

maalijakauman, jossa lokaatioparametrina on uskottavuusfunktion variaatioapproksimaation lineaarinen prediktori, täydessä lajikovariaatit ja latenttimuuttujat sisältävässä mallissa $\tilde{v}_{ij} = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_e + \text{vec}(\mathbf{B}_{te})'(\boldsymbol{x}_i \otimes \boldsymbol{t}_j) + \mathbf{a}_i'\boldsymbol{\gamma}_j^*$, missä $\boldsymbol{\gamma}_j^* = (\boldsymbol{\gamma}_j', 1)'$ ja jakauman skaalaparametrina 1. Katkaistun normaalijakauman rajoina ovat $(-\infty, 0)$ kun $y_{ij} = 0$ ja $(0, \infty)$ kun $y_{ij} = 1$. Tällöin yhtälö (11) ratkeaa suljetussa muodossa dikotomisen vasteen ja probit-linkkifunktion tapauksessa muotoon

$$\begin{aligned}\underline{l}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = &\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}\log(\Phi(\tilde{v}_{ij})) + (1 - y_{ij})\log(1 - \Phi(\tilde{v}_{ij}))) \\ &+ \frac{1}{2}\sum_{i=1}^{n}(\log\det(\mathbf{A}_i) - \text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{A}_i) - \mathbf{a}_i'\boldsymbol{\Sigma}^{-1}\mathbf{a}_i - \log\det(\boldsymbol{\Sigma})) \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{\gamma}_j^{*'}\mathbf{A}_i\boldsymbol{\gamma}_j^*,\end{aligned}$$

missä $\boldsymbol{\Sigma}$ on lohkodiagonaalimatriisi $d \times d$ identiteettimatriisista ja satunnaistermin α varianssista σ^2. Samankaltainen formulointi on esitetty eksponentiaaliselle jakaumaperheelle artikkelissa Niku et al. (2019a).

Variaatioapproksimoinnissa maksimoimme $\underline{l}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ malliparametrien $\boldsymbol{\Psi}$ ja variaatioparametrien $\boldsymbol{\xi}$ suhteen. Variaatioparametrien estimaatit $\hat{\mathbf{a}}_i$ antavat latenttimuuttujien $\boldsymbol{\eta}_i$ ja satunnaisvaikutusten α_i ennusteet ja $\hat{\mathbf{A}}_i$ näille ennusteille varianssi-kovarianssirakenteen. Malliin pohjautuva inferenssi toteutuu samalla tavalla kuin perinteisellä suurimman uskottavuuden menetelmällä.

3.2 Bayes-estimointi

Bayes-estimointi perustuu parametrien posteriorijakaumiin. Siinä missä frekventistisessä mallinnuksessa estimoidaan parametreille piste-estimaatit ja luottamusvälit, niin bayesiläisessä mallinnuksessa approksimoidaan parametrien todennäköisyysjakaumat ja näiden avulla parametrien keskiarvot sekä todennäköisyysvälit.

Konseptuaalisesti bayesiläinen mallintaminen perustuu Bayesin kaavaan

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

missä A ja B ovat tapahtumia, $P(A|B)$ on todennäköisyys, että A tapahtuu ehdolla B, $P(B|A)$ on todennäköisyys, että B tapahtuu ehdolla A ja $P(A)$ on tapahtuman A todennäköisyys. Luonnollisesti oletetaan myös, että $P(B) \neq 0$.

Edellä oleva kaava voidaan kirjoittaa havaittuun aineistoon x ja parametreihin θ perustuvassa inferenssissä yleisessä muodossa

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)},$$

missä termillä $f(\theta|x)$ viitataan parametrien θ posteriorijakaumaan, eli parametrien yhteistiheysjakaumaan ehdolla havaittu aineisto x ja termillä $f(x|\theta)$ aineiston x yhteistiheysjakaumaan ehdolla parametrien arvo θ, eli niin sanottuun otostodennäköisyyteen tai uskottavuuteen. Priorijakauma $f(\theta)$ kuvastaa parametrien todennäköisyysjakaumaa ilman tietoa aineistosta. Termi $f(x)$ kuvastaa havaintojen x marginaalijakaumaa, eli integraalia parametrien θ yli

$$f(x) = \int_\theta f(x|\theta)f(\theta)\, d\theta.$$

Reunajakauman integraalin laskeminen suljetussa muodossa on mahdollista vain konjugaattiprioreilla. Tämän vuoksi Bayes-estimoinnissa käytetään usein numeerisia menetelmiä parametrien θ empiirisen posteriorijakauman laskemiseksi. Eräs näistä menetelmistä on *Markov Chain Monte Carlo*, eli MCMC-menetelmä, jota myös tässä työssä käytetään posteriorijakaumien approksimointiin.

Merkittävänä erona frekventistisen mallintamisen työnkulkuun Bayes-estimoinnissa on asetettava ennen mallintamista priorijakaumat kaikille estimoitaville parametreille ja tarpeen vaatiessa myös hyperparametreille. Usein Bayes-estimoinnissa suositaan epäinformatiivisia priorijakaumia, jotka mahdollistavat mahdollisimman muovautuvat posteriorijakaumat parametreille.

Myös tässä työssä suosimme epäinformatiivisia priorijakaumia.

3.2.1 Bayes-estimointi HMSC-malliperheelle

Verrattuna frekventistiseen uskottavuuspäättelyyn myös mallien sovittamistavat eroavat. Kun GLLVM-malleissa preferoimme erilaisten approksimaatioiden laskemista lopulliselle uskottavuusfunktiolle, niin HMSC-mallin sovittaminen tapahtuu parametrien empiiristen posteriorijakaumien estimoinnilla simuloinnin avulla. Simuloimme MCMC simulointiketjuja n_{ketju} kappaletta, joiden avulla saamme empiiriset posteriorijakaumat parametreille. Simuloinnilla pystymme saamaan periaatteessa niin tarkat tulokset kuin haluamme, jos simuloinnin annetaan jatkua tarpeeksi pitkään. Toisaalta laskennallisesti raskas simulointi johtaa siihen, että yhtä tarkkojen tuloksien saaminen kuin frekventistisillä menetelmillä kestää usein kauemmin.

Asetamme simulaatioiden määräksi mielivaltaisen luvun n_{sim} jolloin saamme n_{sim} kappaletta simuloituja arvoja jokaiselle kiinnostuksen kohteena olevalle parametrille. Tätä parametrien simuloitujen arvojen joukkoa kutsutaan parametrin empiiriseksi posteriorijakaumaksi. Posteriorijakauman avulla voimme laskea haluamamme jakaumatunnusluvut parametreille ja tehdä näiden avulla tilastollista päättelyä mallista.

Kun tarkastelemme HMSC-mallinnuksen tuloksia, on ensin tarkasteltava MCMC-ketjujen konvergenssia. Saadaksemme simuloinnista luotettavia tuloksia parametrin posteriorijakauman kannalta, tulee simuloitujen MCMC-ketjujen olla konvergoituneita. Mikäli toisistaan riippumattomasti simuloidut ketjut antavat samankaltaisia tuloksia, voidaan olettaa, että on näyttöä ketjujen konvergoitumisesta. Mikäli ketjut antavat keskenään selkeästi erilaisia tuloksia, niin luultavasti ketjut eivät ole konvergoituneet ja erilaiset tulokset johtuvat divergenssin muodostamasta vaihtelusta. Tässä työssä tarkastelemme MCMC-ketjujen konvergenssin määrittämiseksi Gelman-Rubin $\hat{R}$-tunnusluvun (*potential scale reduction*) diagnostiikkaa, jonka ideana on verrata toisistaan riippumattomasti simuloitujen MCMC-ketjujen samankaltaisuutta keskenään (Gelman ja Rubin, 1992). Tyypillisesti Gelman-Rubin $\hat{R}$-tunnusluvun arvoja $\hat{R} < 1.1$ pidetään indikaattorina MCMC-ketjujen kon-

vergenssista (Gelman, 1995), ja käytämme sitä myös tässä työssä indikaattorina ketjujen konvergenssista.

MCMC-ketjussa esiintyy käytännön kannalta aina hieman autokorrelaatiota estimoidussa posteriorijakaumassa, joten hyödynnämme simuloinnissa harvennusta. Harvennuksen idea on ottaa MCMC-ketjusta lopulliseen posteriorijakaumaan joka n_h:nnes arvo. Tämä vähentää empiirisen posteriorijakauman peräkkäisten arvojen välistä korrelaatiota (Ovaskainen ja Abrego, 2020), mikä usein tarkoittaa myös luotettavampaa tulosta posteriorijakaumalle. Lisäksi usein asetetaan suunnilleen ensimmäinen kolmannes simuloituja arvoja sisäänajojaksoksi n_s. Näitä arvoja ei sisällytetä lopulliseen posteriorijakaumaan, koska MCMC-ketju ei ole ensimmäisillä askelilla vielä konvergoitunut, joten sen arvot ovat harhaisia. Tällöin näiden arvojen sisällyttäminen posteriorijakaumaan saa empiirisestä posteriorijakaumasta myös harhaisen. Posteriorijakauman otoskoko määräytyy siten edellä mainittujen parametrien mukaisesti kaavalla

$$n_{posteriori} = \frac{n_{ketju}(n_{sim} - n_s)}{n_h}.$$

MCMC-ketjun arvojen simuloinnille on monia vaihtoehtoisia algoritmeja. HMSC-malliperheen simuloinnissa käytetään Gibbsin menetelmää, jossa estimoitavat parametrit jaetaan ryhmiin siten, että jokaisessa ryhmässä parametrin arvo simuloidaan yksi toisensa jälkeen sen ehdollisesta jakaumasta ottaen huomioon muiden parametrien arvot (Ovaskainen ja Abrego, 2020). Tässä työssä ei tarkastella simuloinnin toteutusta tarkemmin.

3.2.2 Priorijakaumat

Seuraamme HMSC-mallin priorijakaumien esittämisessä kirjan Ovaskainen ja Abrego (2020) esitystä. Merkitään kaikkien priorijakaumien yhteistä todennäköisyysjakaumaa $p(\boldsymbol{\theta})$. On tärkeä huomata, että parametri $\boldsymbol{\theta}$ koostuu kaikista niistä parametreista, joille asetetaan mallissa priorijakauma. Karkeasti yhteinen priori Bernoulli-jakautuneessa probit-linkkifunktiota käyttä-

vässä mallissa voidaan hajottaa osiin seuraavasti

$$p(\boldsymbol{\theta}) = p(\mathbf{B}, \boldsymbol{\Lambda}, \mathbf{V})p(\mathbf{H}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\delta}). \tag{12}$$

Kaavassa (12) priori $p(\mathbf{B}, \boldsymbol{\Lambda}, \mathbf{V})$ viittaa mallin kiinteän osan parametreihin, ja se voidaan jakaa edelleen osiin $p(\mathbf{B}, \boldsymbol{\Lambda}, \mathbf{V}) = p(\mathbf{B}|\boldsymbol{\Lambda}, \mathbf{V})p(\boldsymbol{\Lambda})p(\mathbf{V})$. Tässä $p \times m$ matriisi $\mathbf{B}$ viittaa ympäristökovariaattien regressiokertoimiin ja $\boldsymbol{\Lambda}$ on lajikovariaattien suhdetta ympäristökovariaattien vaikutukseen kuvaava $p \times q$ matriisi. Viimeisin $(p+1) \times (p+1)$ matriisi $\mathbf{V}$ kuvastaa kiinteän osan parametrien kovarianssirakennetta. Nämä asettavat implisiittisesti priorijakauman ympäristökovariaattien regressiokertoimille $\mathbf{B}$ (Ovaskainen ja Abrego, 2020).

Oletamme matriisille $\boldsymbol{\Lambda}$ priorijakauman

$$\mathrm{vec}(\boldsymbol{\Lambda}) \sim N(\boldsymbol{\mu}_\lambda, \mathbf{U}_\lambda),$$

missä mallin sovituksessa valitaan odotusarvovektorille $\boldsymbol{\mu}_\lambda$ ja kovarianssimatriisille $\mathbf{U}_\lambda$ arvot. Luontainen epäinformatiivinen valinta on $\boldsymbol{\mu}_\lambda = \mathbf{0}$ ja $\mathbf{U}_\lambda = \mathbf{I}$, missä $\mathbf{0}$ on pq-pituinen nollavektori ja $\mathbf{I}$ on $pq \times pq$ identiteettimatriisi. Kiinteävaikutuksen regressiokertoimien kovarianssirakennetta kuvaavalle matriisille $\mathbf{V}$ oletetaan priorijakauma

$$\mathbf{V} \sim W^{-1}(\mathbf{V}_0, f_0),$$

missä $W^{-1}(\cdot)$ on käänteinen Wishart-jakauma. Matriisi $\mathbf{V}_0$ valitaan $p \times p$ identiteettimatriisiksi $\mathbf{I}$ ja vapausasteiden määräksi valitaan $f_0 = p+1$, missä p on ympäristökovariaattien lukumäärä.

Kaavassa (12) termi $p(\mathbf{H}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\delta})$ viittaa mallin satunnaisosan priorijakaumiin, missä $\mathbf{H}$ on latenttimuuttujamatriisi, jonka elementti η_{ih} kuvaa havaintopaikan i latenttimuuttujan h arvoa, ja matriisin $\boldsymbol{\Gamma}$ alkio γ_{jh} viittaa lajin j ja latenttimuuttujan h väliseen lataukseen. Matriisin $\boldsymbol{\Gamma}$ priorin määrittämiseksi tarvitsemme apumuuttujat $\boldsymbol{\Phi}$ ja $\boldsymbol{\delta}$. Matriisi $\boldsymbol{\Phi}$, jonka elementti on ϕ_{hj}, kuvaa lajien latausten lokaalia kutistumista. Parametri $\boldsymbol{\delta}$, jonka elementti on δ_h, viittaa lajien latausten globaaliin kutistumiseen (Bhattacharya

ja Dunson, 2011). Havaintopaikan i ja latenttimuuttujan h välistä latausta kuvaavan matriisin $\mathbf{H}$ alkioille η_{ih} asetamme prioriksi $\eta_{ih} \sim N(0,1)$. Lajien ja latenttimuuttujien suhdetta kuvaavalle satunnaisvaikutustermille $\mathbf{\Gamma}$ asetetaan priorijakauma kaavan

$$p(\mathbf{\Gamma}, \mathbf{\Phi}, \boldsymbol{\delta}) = p(\mathbf{\Gamma}|\mathbf{\Phi}, \boldsymbol{\delta})p(\mathbf{\Phi})p(\boldsymbol{\delta})$$

avulla. Siinä missä muut HMSC-mallin termit eivät ole järin sensitiivisiä priorin valinnalle, niin $\mathbf{\Gamma}$ priorin valinnalla voi olla suuri vaikutus koko analyysin tuloksiin (Ovaskainen ja Abrego, 2020). Matriisin $\mathbf{\Gamma}$ priorin asettaminen riittää määrittämään lajien välistä assosiaatiota kuvaavan matriisin $\mathbf{\Omega}$ priorin, sillä $\mathbf{\Omega} = \mathbf{\Gamma}'\mathbf{\Gamma}$ (Ovaskainen ja Abrego, 2020). Priorijakaumat $\mathbf{\Gamma}$, $\mathbf{\Phi}$ ja $\boldsymbol{\delta}$ parametreille voidaan esittää muodossa

$$\gamma_{hj}|\phi_{hj}, \delta \sim N(0, \phi_{hj}^{-1}\boldsymbol{\tau}_h^{-1}), \boldsymbol{\tau}_h = \prod_{l=1}^{h} \delta_l \tag{13}$$

$$\phi_{hj}|\upsilon \sim Gamma(\frac{\upsilon}{2}, \frac{\upsilon}{2}) \tag{14}$$

$$\delta_1|a, b \sim Gamma(a_1, b_1), \delta_h|a, b \sim Gamma(a_2, b_2), h \geq 2. \tag{15}$$

Kaavoissa (13) - (15) ϕ_{hj} kuvaa lajien ja latenttimuuttujien välistä lokaalia kutistumista, ja δ_h kuvaa lajien ja latenttimuuttujien välistä globaalia kutistumista (Bhattacharya ja Dunson, 2011). Priorin määrittämisessä täytyy asettaa alkuarvot parametreille υ, a ja b. Mallin sovitukseen käytettävän R-paketin `Hmsc` nimikkofunktio käyttää oletusarvoisesti arvoja $\upsilon = 3$, $a = (50, 50)$, $b = (1, 1)$ (Tikhonov et al., 2021), joita käytämme myös tässä työssä. Varsinkin parametrin a arvojen kanssa tulee käyttää harkintaa, sillä ne säätävät residuaalikovarianssia kuvaavan matriisin $\mathbf{\Omega}$ kutistuneisuuden määrää (Ovaskainen ja Abrego, 2020).

Kun priorijakaumat ovat jokaiselle parametrille erikseen määritelty, saamme priorien yhteisjakaumaksi mallin (12) mukaisen priorin. Mallioletuksena on priorijakaumiin liittyvien muuttujien keskinäinen riippumattomuus.

4 Mallien vertailu

Tässä luvussa tarkastellaan mallien vertailuun käytettäviä menetelmiä ja tunnuslukuja. Aloitamme määrittelemällä, mitä tarkoitamme mallien selitys- ja ennustevoimalla. Tämän jälkeen esittelemme menetelmät, joilla mallien selitys- ja ennustevoiman diskriminaatiota, tarkkuutta ja kalibraatiota, sekä mallien parametriestimaatteja tutkitaan ja vertaillaan.

Tämän työn kiinnostuksen kohteena on tutkia GLLVM- ja HMSC-mallien välisiä eroja sekä selitys- että ennustevoimassa. Erityisenä kiinnostuksen kohteena on lajin prevalenssin vaikutus tutkittaviin suureisiin sekä varsinkin harvinaisten lajien erot mallien välillä.

Mallien selitysvoimalla tarkoitamme tässä työssä tilannetta, jossa malli sovitetaan koko aineistoon $\mathbf{X}$ ja $\mathbf{Y}$. Näin saamme jokaisen lajin j parametrille k parametriestimaatit $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jp})'$, sekä ennusteet mallin satunnaistermeille, joiden avulla saamme laskettua ennusteet lajin j havaitsemiselle havaintopaikalla i.

Mallien ennustevoiman tutkimista varten jaamme aineiston satunnaisesti kahteen yhtäsuureen osaan, opetusdataan $\mathbf{X}^{(o)}$, $\mathbf{Y}^{(o)}$ ja testidataan $\mathbf{X}^{(t)}$, $\mathbf{Y}^{(t)}$. Tässä työssä toteutamme ennustevoiman tutkimisen käyttäen kaksinkertaista ristiinvalidointia mallien sovittamiseen kuluvan pitkähkön ajan vuoksi. Useampikertainen ristiinvalidointi on teknisesti mahdollista ja jatkotutkimusten kannalta sekä suotavaa että mielenkiintoista. Malli sovitetaan opetusdatalla, josta saadaan opetusdatan rivejä vastaavat parametriestimaatit $\hat{\boldsymbol{\beta}}_j^{(o)}$ sekä vastaavat satunnaistermien ennusteet. Parametriestimaatteja $\hat{\boldsymbol{\beta}}_j^{(o)}$ ja satunnaistermien ennusteita käytetään testidatan $\mathbf{X}^{(t)}$ ennustamiseen, josta saadaan todennäköisyysennusteet lajin j havaitsemiselle havaintopaikoilla $i^{(t)}$. Toistamalla mallinnus niin, että testidatalla ennustetaan opetusdataa, saadaan täydellinen ennustematriisi lajin j havaitsemiselle havaintopaikoilla i.

Vertailuja toteutetaan lajikohtaisella tasolla sekä yksittäisiä lajeja tarkastellen, että jakaen lajit prevalenssin mukaan kymmeneen likimain yhtä suureen ryhmään. Näille lajin prevalenssia kuvaaville ryhmille lasketaan tarkasteltavan suureen keskiarvo ja ryhmien keskiarvoja jälleen vertaillaan toisiin-

sa. Kaavat mallien tarkkuuden, diskriminaation ja kalibraation laskemiseen ovat samat sekä mallien selitys- että ennustevoimalle. Parametriestimaattien vertailu toteutetaan pelkästään mallien selitysvoimalle.

4.1 Diskriminaatio

Mallin diskriminaatiolla tarkoitetaan mallin kykyä erotella ennustetodennäköisyyttä eri suuruisten havaitsemistodennäköisyyksien välillä: esimerkiksi korkean diskriminaation malli ennustaa keskimäärin suurempia havaitsemistodennäköisyyksiä havaintopaikalle, jossa esiintyy enemmän lajeja, tai lajille, joka esiintyy usealla havaintopaikalla. Sen sijaan alhaisen diskriminaatiovoiman malli antaa likimain yhtäsuuria ennustetodennäköisyyksiä havaintopaikalle, jossa esiintyy keskimääräistä enemmän lajeja, tai lajille, joka esiintyy keskimääräistä useammalla havaintopaikalla.

Käytämme mallien diskriminaation vertailuun Tjur R^2 -indeksiä, joka määritellään

$$R^2_{Tjur,j} = \frac{1}{n_1}\sum_{i=1}^{n_1} p(y_{ij} = 1) - \frac{1}{n_0}\sum_{i=1}^{n_0} p(y_{ij} = 0),$$

missä n_1 viittaa siihen, kuinka monta kertaa laji j on tullut havaituksi aineistossa ja n_0 kuinka monta kertaa laji on jäänyt havaitsematta. Todennäköisyydellä $p(y_{ij} = 1)$ ja $p(y_{ij} = 0)$ viitataan estimoituihin todennäköisyyksiin havaituille ja ei-havaituille lajille j havaintopaikassa i (Tjur, 2009).

Huomataan, että R_{Tjur} arvo rajoittuu välille $-1 \leq R^2_{Tjur} \leq 1$. Arvot lähellä yhtä indikoivat suuresta diskriminaatiovoimasta ja nollaa lähellä olevat arvot mallin heikosta kyvystä tuottaa havaituille lajeille suurempia todennäköisyysennusteita kuin ei-havaituille lajeille (Tjur, 2009). Sen sijaan negatiiviset R^2_{Tjur} arvot kertovat, että malli antaa ei-havaituille lajeille suurempia todennäköisyysennusteita kuin havaituille lajeille.

4.2 Tarkkuus

Mallin tarkkuuden voidaan ajatella kertovan siitä, kuinka lähelle oikeaa suhteellista frekvenssiä malli ennustaa lajin j havaitsemisen keskimäärin. Perinteisesti mallin tarkkuutta mitataan esimerkiksi keskineliövirheellä (RMSE), mutta dikotomisen vasteen tapauksessa keskineliövirhe ei tuota informatiivisia tuloksia.

Tämän vuoksi otamme tässä työssä mallin tarkkuuden mittaamiseksi lähestymistavan, jossa tarkastelemme mallien devianssia lajikohtaisesti. Mallin lajikohtainen devianssi voidaan määritellää yleisellä kaavalla

$$D_j = -2l(\hat{\theta}, y_j) + 2l(\hat{\theta}_s, y_j),$$

missä $l(\hat{\theta}, y_j)$ viittaa mallin log-uskottavuuteen parametriestimaatilla $\hat{\theta}$ ja $l(\hat{\theta}_s, y_j)$ viittaa saturoidun mallin log-uskottavuuteen. Termi y_j viittaa lajin j havaintovektoriin. Dikotomisen vasteen tapauksessa kaava sievenee muotoon

$$D_j = -2\sum_{i=1}^{n}(y_{ij}\log p_{ij} + (1 - y_{ij})\log(1 - p_{ij})),$$

sillä saturoidun mallin log-uskottavuus $l(\hat{\theta}_s, y_j) = \log 1 = 0$. Todennäköisyydet p_{ij} viittaavat mallin antamaan ennustetodennäköisyyteen lajille y_{ij}. Devianssin arvot ovat aina positiivisia ja mitä lähempänä devianssin arvo on nollaa, sitä parempi mallin tarkkuus on.

4.3 Kalibraatio

Kalibraatiolla mitataan tilastollisen mallin konsistenttiutta. Mallin konsistenttius voidaan ajatella absoluuttisena erotuksena ennustettujen havaitsemistodennäköisyyksien ja aineistosta laskettujen aitojen havaitsemistodennäköisyyksien välillä. Tällöin nollaa lähellä olevat arvot kertovat tämän myötä hyvin kalibroituneesta mallista (Gneiting ja Raftery, 2007). Hyvin kalibroituneessa mallissa esimerkiksi ennustetodennäköisyyden 0.10 saanut laji esiintyy keskimäärin noin 10 % havaintopaikoista.

Seuraamme kalibraation tarkastelussa tutkimuksen Norberg et al. (2019) menettelytapaa, jossa mallin tuottamat ennustetodennäköisyydet jaetaan ensin kymmeneen yhtäsuureen kvantiiliin, minkä jälkeen jokaisesta kvantiilista lasketaan havaitun prevalenssin ja ennustetun prevalenssin keskiarvo. Lopuksi piirrämme lasketuista pisteistä kuvaajan, jossa hyvin kalibroitunut malli noudattaa mahdollisimman tarkasti suoraa $y = x$. Huomion arvoista on, että toisin kuin mallin diskriminaation ja tarkkuuden tutkimisessa, tarkastelemme kalibraatiota kaikille lajeille samanaikaisesti ja jako kvantiiliryhmiin tapahtuu ennustetodennäköisyyksien avulla aineistosta laskettujen prevalenssien sijasta.

4.4 Parametriestimaattien vertailu

Parametriestimaattien vertailussa mielenkiinnon kohteena on suon käsittelyn (luonnontilainen vs. ojitettu) vaikutus lajin havaitsemistodennäköisyyteen. Suon käsittelyyn liittyvän parametrin β vertailu toteutetaan vertailemalla mallien tuottamia regressiokertoimia ja niiden luottamus- ja todennäköisyysvälejä silmämääräisesti. Tämän jälkeen tarkastelemme mallien tuottamien regressiokertoimien yhteyttä lajin prevalenssiin. Jaottelemme molempien mallien tuottamat regressiokertoimet kolmeen ryhmään sen mukaan, oliko parametriestimaatin 95 %:n luottamusväli GLLVM-mallin parametriestimaateille ja 95 %:n todennäköisyysväli HMSC-mallin parametriestimaateille, positiivista, negatiivista vai sisältyikö nolla vastaavaan 95 %:n väliin. Kolmiluokkaisen jaottelun avulla tutkimme, onko taulukossa 4 esitetyllä lajien prevalenssien tasajaolla yhteyttä suon käsittely -muuttujan parametriestimaattiin.

5 Putkilokasviaineiston analyysi

Tässä luvussa toteutetaan aiemmissa luvuissa kuvattu analyysi putkilokasviaineistoon, joka on esitelty alunperin tutkimuksessa Elo et al. (2016). Esittelemme analysoitavan aineiston, minkä jälkeen esitämme sovitetut mallit ja sovitukseen liittyvät valinnat. Lopuksi esittelemme tulokset tulkintoineen.

5.1 Aineiston kuvailu

Putkilokasviaineisto on kerätty vuosina 2007–2010 ja koostuu 120 eri suosta (myöhemmin palsta). Jokaisesta suosta valittiin kymmenen yhden neliömetrin suuruista havaintoruutua, joista kustakin tutkittiin 131 putkilokasvilajin esiintyvyydet. Aineisto sisältää siis yhteensä $n = 1200$ havaintopaikkaa, josta jokaisesta on tehty havainnot $m = 131$ putkilokasvilajista. Putkilokasviaineisto on alun perin kerätty peittävyysaineistona, joka saa arvoja $[0, 1]$ väliltä sen mukaan, kuinka paljon kyseessä oleva kasvi peittää näytteen havaintoruudun pinta-alasta. Kasvin peittävyys havaintoruudun pinta-alasta määriteltiin silmämääräisesti prosentin tarkkuudella. Havaintopaikat sijaitsevat eteläisellä, keskisellä ja pohjoisella boreaalisella kasvimaantieteellisellä alueella. Tätä työtä varten aineisto on muokattu kaksitasoiseksi seuraavasti: jos lajin j peittävyys havaintopaikalla i on ollut suurempaa kuin nolla, niin uusi lajimuuttuja y_{ij} saa arvon 1, muutoin arvon 0.

Alkuperäisessä aineistossa useita lajeja havaitaan yhteensä alle viisi kertaa. Vaikka tämän työn yhtenä kiinnostuksen kohteena on inferenssi ja mallitarkastelu varsinkin harvoin havaituille lajille, on mallien sovittamiseksi pakko karsia kaikista harvinaisimpia lajeja. Tässä työssä sisällytämme analyysissä käytettyyn dataan vain vähintään 5 kertaa havaitut putkilokasvilajit. Tällöin lajien määrä on $m = 91$. Taulukossa 1 esittelemme lajien määrät prevalenssin mukaan. Huomataan että noin 36 % kerätyistä lajeista esiintyy korkeintaan 5 %:lla havaintopaikoista. Lisäksi suurin määrä lajeista havaitaan aineistossa $5 - 10$ %:lla havaintopaikoista ja vain 11 lajia, eli noin 12 % lajeista esiintyy aineistossa vähintään 20 %:lla havaintopaikoista.

Taulukko 1: Putkilokasvilajien määrät (N) ja osuudet aineistossa prevalenssin mukaan. Prevalenssiluokat ovat jaoteltu usein eksploratiivisen tarkastelun tapauksessa mielenkiinnon kohteena oleviin intervalleihin.

Prevalenssi	N	Kum. N	Osuus	Kum. osuus
(0, 0.01]	25	25	0.275	0.275
(0.01, 0.05]	8	33	0.088	0.363
(0.05, 0.10]	35	68	0.385	0.747
(0.10, 0.20]	12	80	0.132	0.879
(0.20, 0.30]	6	86	0.066	0.945
(0.30, 0.40]	1	87	0.011	0.956
(0.40, 0.50]	1	88	0.011	0.967
(0.50, 0.60]	3	91	0.033	1.000

Lisäksi aineistoon sisältyy jokaiselta havaintopaikalta kerättyjä ympäristökovariaatteja. Analyyseissä käytämme samoja ympäristökovariaatteja kuin tutkimuksen Elo et al. (2016) analyyseissä, eli suotyyppiä, suon ravinteisuutta sekä suon käsittelyä. Kaikki edellä mainitut kovariaatit ovat nominaalisia faktoreita, joiden frekvenssit ovat esitettynä taulukossa 2 siten, että muuttujan järjestyksessä ensimmäistä tasoa käsitellään analyyseissä referenssiluokkana. Taulukosta 2 huomataan, että havaintoja on jokaisesta luokasta riittävä määrä, joten muuttujien muokkaamiselle, kuten luokkien yhdistämiselle, ei ole tarvetta. Lisäksi taulukoista 1 ja 2 huomataan, että aineistossa ei ole puuttuvuutta ympäristökovariaattien tai lajien suhteen.

Taulukko 2: Analyyseissä käytettyjen ympäristökovariaattien frekvenssit putkilokasviaineistossa.

Muuttuja	Taso	N
Ravinteisuus	karu	590
	rehevä	610
Suotyyppi	korpi	400
	räme	400
	avosuo	400
Käsittely	luonnontilainen	600
	ojitettu	600

5.2 Mallien sovitus aineistoon

Rakennamme GLLVM- ja HMSC-mallit mahdollisimman samankaltaisiksi sen mukaan minkälaisten termien asettaminen on mahdollista ja tarkoituksenmukaista tutkimusasetelman kannalta. Tämän pohjalta sovitamme aineistoon luvussa 2 esitellyt mallit. GLLVM-malli sovitetaan kaavan (4) ja HMSC-malli kaavan (9) mukaisesti, sekä luvussa 3.2.2 esitellyillä priorijakaumilla ja hyperparametrien alkuarvoilla. Molempiin malleihin asetamme latenttimuuttujien määräksi kaksi ja ympäristökovariaateiksi edellisessä luvussa mainitut muuttujat. Erona sovitettujen mallien välillä on, että HMSC-mallissa latenttimuuttujat ovat palstakohtaisia, kun taas GLLVM-mallissa latenttimuuttujat ovat havaintopaikkakohtaisia. Lisäksi sovitetussa GLLVM-mallissa on latenttimuuttujien lisäksi erikseen palstakohtainen satunnaistermi, jota HMSC-mallissa ei ole lainkaan. Syy palstakohtaisten latenttimuuttujien sovittamiselle HMSC-malliin on tutkittavan aineiston hierarkkisuudessa, sillä yhdeltä palstalta tehdään 10 havaintoa putkilokasvilajien esiintyvyydestä. Samasta syystä GLLVM-malliin lisätään palstakohtainen satunnaistermi, jota HMSC-malliin ei voida erikseen lisätä. Tämän työn tekohetkellä R-kirjasto `gllvm` (Niku et al., 2021a) ei sisältänyt mahdollisuutta sovittaa latenttimuuttujia palstakohtaisina.

Uskottavuuspäättely tehdään GLLVM-mallille variaatioapproksimaatiolla ja HMSC-mallille Bayes-estimointi MCMC-simuloinnin avulla. Simuloinnissa sovitamme kaksi ketjua, joille molemmille asetamme simulaatioiden määräksi $n_{sim} = 133000$, harvennusväliksi $n_h = 100$ ja sisäänajoksi $n_s = 33000$. Näin lopulliseen empiiriseen posteriorijakaumaan tulee $2 \times 1000 = 2000$ simulaatiopistettä. Tämän jälkeen tarkastelemme HMSC-mallin ympäristökovariaattikohtaisten $\boldsymbol{\beta}$-parametrien ja ympäristökovariaattien kovarianssia kuvaavien parametrien $\mathbf{V}$ konvergenssia Gelman-Rubin diagnostiikan avulla.

Sovitamme mallit käyttäen R-ohjelmiston (R Core Team, 2021) kirjastoja `gllvm` ja `Hmsc`. Mallikoodit ovat liitteessä A. Mallit sovitettiin käyttäen laskentateholtaan tyypillistä PC-konetta ja sovittamiseen kuluneet ajat ovat taulukossa 3. Johtuen HMSC-mallin käyttämästä MCMC-simuloinnista, ovat HMSC-mallin sovitukseen kuluneet ajat merkittävästi GLLVM-mallia suuremmat. Ero on odotettavissa ja esimerkiksi tutkimuksissa Niku et al. (2019b); Ovaskainen ja Abrego (2020) ja Niku (2020) on havaittu vastaavia tuloksia.

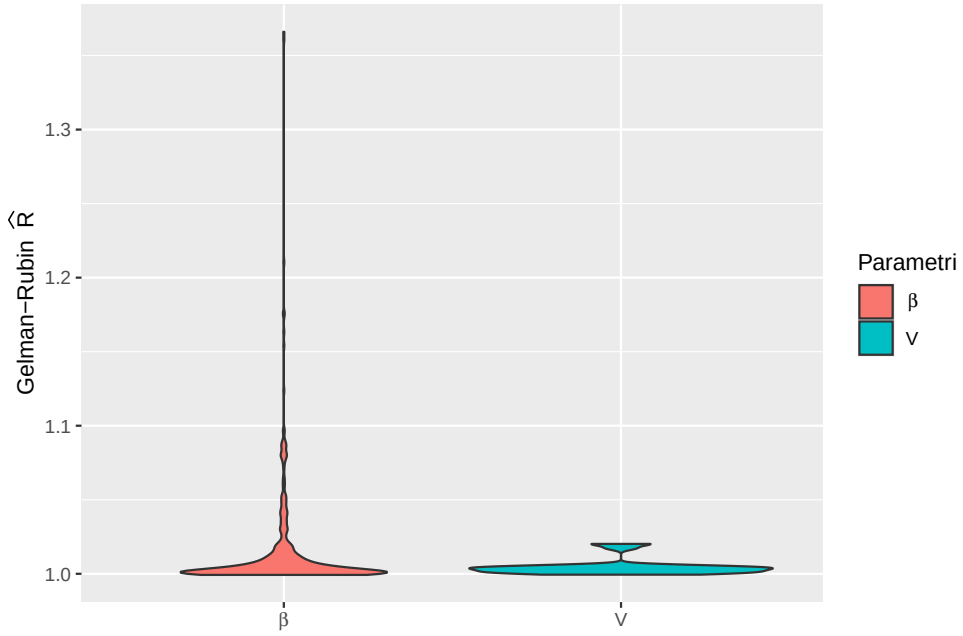
Taulukko 3: Mallien sovittamiseen kulunut aika PC-koneella (Intel Core i5-9400F). Selitysvoimalla viitataan tilanteeseen, jossa malli sovitetaan kerran koko käytettävissä olevaan dataan, ennustevoimalla tilanteeseen, jossa malli sovitetaan 2-kertaisella ristiinvalidoinnilla.

	GLLVM	HMSC
Selitysvoima	14min	14h 19min
Ennustevoima	7min	21h 31min

Mallien sovittamisen jälkeen laskemme luvun 4 esitystä seuraten sekä selitysvoimaa että ennustevoimaa kuvaavat $n \times m$ prediktorimatriisit $\mathbf{P}$ lajien havaitsemiselle parametriestimaattien avulla. Prediktorimatriiseja käytetään jatkossa tutkittaessa mallien kalibraatiota, diskriminaatiota ja tarkkuutta.

5.2.1 HMSC-mallin konvergenssi

Tarkastelemme kuvassa 1 mallin konvergenssia sekä ympäristökovariaattikohtaisille parametreille $\boldsymbol{\beta}$ että näiden välistä kovarianssirakennetta kuvaaville $\mathbf{V}$ parametreille. Huomataan, että $\boldsymbol{\beta}$-parametrien tapauksessa lähes kaikkien parametrien Gelman-Rubin $\hat{R}$ -tunnusluku on alle asetetun 1.1-rajan.



Kuva 1: Violin plot -kuvaaja sovitetun HMSC-mallin parametrien $\boldsymbol{\beta}$ ja $\mathbf{V}$ Gelman-Rubin $\hat{R}$ -tunnusluvuille.

Huomataan, että suurin osa tunnusluvuista on jopa alle 1.05, joka antaa lisää näyttöä siitä, että mallin $\boldsymbol{\beta}$-parametrien konvergenssi on pääasiallisesti hyvällä tasolla muutamista poikkeamista huolimatta. Lisäksi kaikki $\mathbf{V}$ parametrin Gelman-Rubin $\hat{R}$ -tunnusluvut ovat selkeästi alle 1.1-rajan. Voimme siis pitää estimoituja parametrien posteriorijakaumia pääosin luotettavina.

5.3 Mallien vertailu eri suureiden avulla

Vertailun tulkittavuuden helpottamiseksi tarkastelemme mallien selitys- ja ennustevoimaa kuvaavien tunnuslukujen välisiä eroja lajien prevalenssien

suhteen ryhmittäin. Jaamme ensin lajit prevalenssin mukaan kymmeneen likimain yhtäsuureen kvantiiliin, jotka on esitetty taulukossa 4. Tarkastelemme mallien välisiä eroja eri suureissa laskemalla jokaiselle kvantiilille tarkasteltavan suureen keskiarvon ja piirtämällä kuvan kvantiilikeskiarvoista ja suurekeskiarvoista.

Taulukko 4: Putkilokasvilajien prevalenssit jaettuna kymmeneen likimain yhtäsuureen kvantiiliin sekä luokkia vastaavat lajien määrät ja osuudet. Keskiarvo viittaa prevalenssiluokan ala- ja ylärajasta laskettuun keskiarvoon.

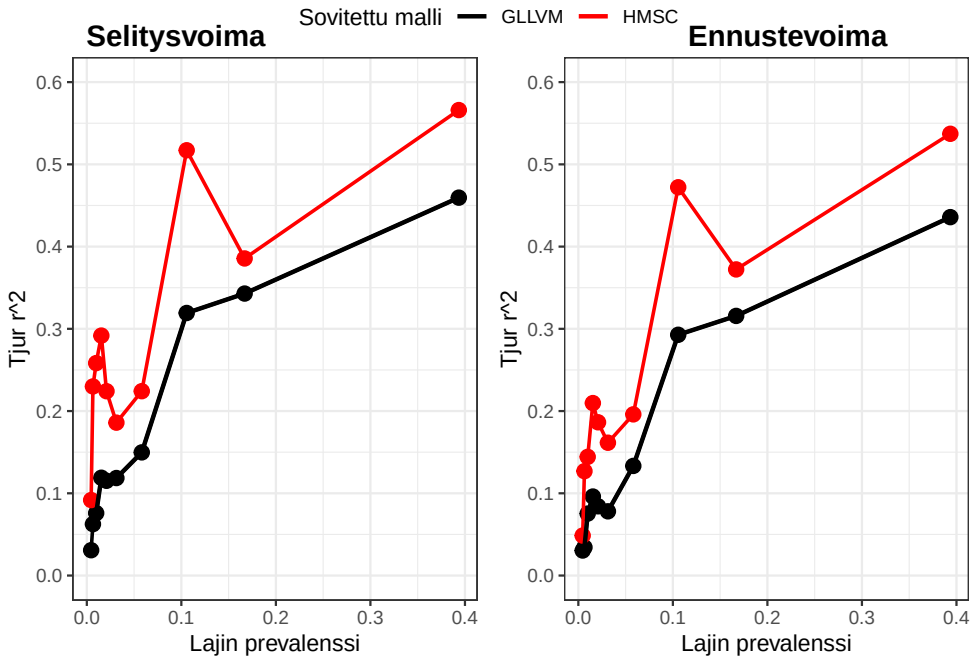
Prevalenssi	Keskiarvo	N	Kum. N	Osuus	Kum. osuus
[0.004, 0.006)	0.005	10	10	0.110	0.110
[0.006, 0.008)	0.006	9	19	0.099	0.209
[0.008, 0.013)	0.010	9	28	0.099	0.308
[0.013, 0.018)	0.015	9	37	0.099	0.407
[0.018, 0.025)	0.021	9	46	0.099	0.505
[0.025, 0.041)	0.031	9	55	0.099	0.604
[0.041, 0.082)	0.058	9	64	0.099	0.703
[0.082, 0.139)	0.106	9	73	0.099	0.802
[0.139, 0.222)	0.167	9	82	0.099	0.901
[0.222, 0.559]	0.394	9	91	0.099	1.000

Näemme että harvinaisten lajien prevalenssien kvantiilikeskiarvot ovat hyvin lähellä toisiaan. Tämän vuoksi osa keskiarvopisteistä jää kuvissa 2–4 osittain piiloon.

5.3.1 Diskriminaatio

Kuvassa 2 esitetään prevalenssin mukaan järjestetyt lajikohtaiset Tjur R^2 arvot sovitetuille malleille selitys- ja ennustevoiman tapauksessa. Kuvasta 2 huomataan, että molempien mallien diskriminaatiovoima kasvaa lajin prevalenssin kasvaessa. Tämä on odotettavissa, sillä samankaltaisia tuloksia on

raportoitu esimerkiksi julkaisussa Ovaskainen ja Abrego (2020). GLLVM-mallin diskriminaatiovoima on systemaattisesti hieman heikompi HMSC-malliin nähden sekä selitys- että ennustevoimalla, mutta sen kasvu on myös tasaisempaa. Erot diskriminaatiovoimassa tulevat esiin sekä harvoilla että yleisillä lajeilla.



Kuva 2: Tjur R^2 arvot mallien selitys- ja ennustevoimalle. Pisteet vastaavat lajikohtaisen prevalenssin ja vastaavan ryhmän Tjur R^2:n keskiarvopisteitä.

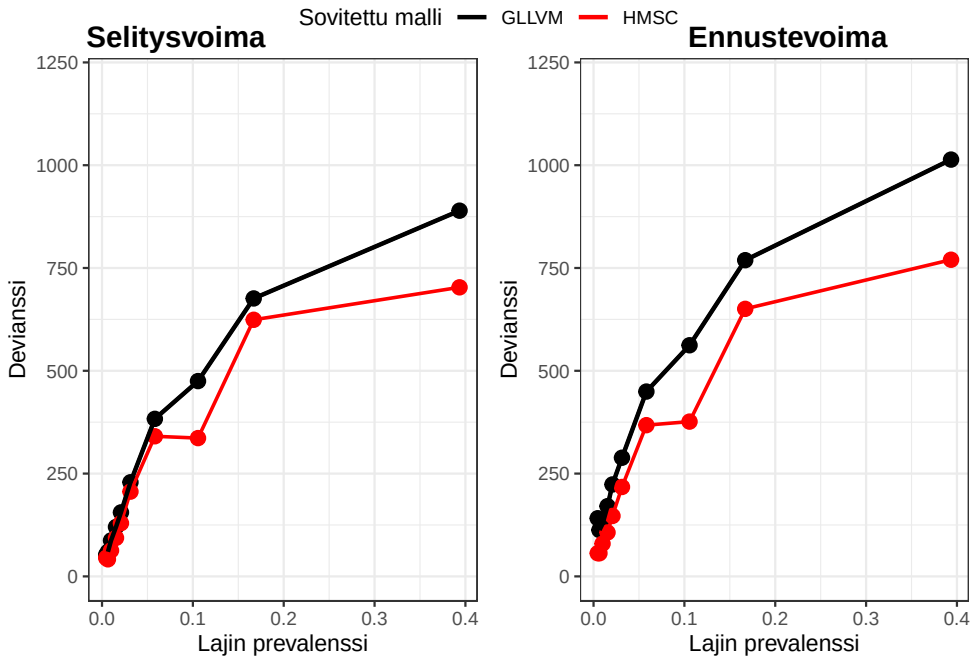
HMSC-mallin Tjur R^2 arvon käyttäytyminen on paljon epätasaisempaa ja vaihtelee paljon suuremmalla välillä lajin havaitun prevalenssin suhteen kuin GLLVM-mallin tuottama arvo. HMSC-mallin Tjur R^2 arvo saavuttaa lokaalin huippukohdan noin 1.5 % ja 10 % havaintopaikoista havaituilla lajeilla. Molempien huippukohtien jälkeen diskriminaatiovoima laskee hetkellisesti tarkasteltaessa seuraavia havaintopisteitä. GLLVM-mallissa tällaista käyttäytymistä ei ole, vaan malli tuottaa systemaattisesti keskimäärin hieman suurempia Tjur R^2 lukuja siirryttäessä prevalenssiltaan suurempiin lajeihin.

Tjur R^2 arvon käyttäytyminen on samanlaista molempien mallien kohdalla verrattaessa saman mallin selitysvoimaa ennustevoimaan. Molemmil-

la malleilla ennustevoiman diskriminaatio on heikompaa selitysvoimaan verrattuna, mikä on odotettavissa. Kuitenkin HMSC-mallin diskriminaatiovoima laskee huomattavasti enemmän suhteessa selitysvoiman diskriminaatioon kuin GLLVM-mallilla. Lisäksi HMSC-mallin diskriminaatio on sekä selitysettä ennustevoiman tapauksessa huomattavasti GLLVM-mallia suurempi varsinkin harvinaisilla lajeilla. Yleisesti ottaen kuitenkin molempien mallien Tjur R^2 arvot ovat verrattain hyviä, kun lajin prevalenssi on yli 0.10.

5.3.2 Tarkkuus

Mallien selitys- ja ennustevoiman tarkkuutta tutkitaan luvun 4.2 esityksen mukaan devianssin avulla. Kuvassa 3 esitetään prevalenssin mukaan järjestetyt lajikohtaiset devianssin arvot sovitetuille malleille selitys- ja ennustevoiman tapauksessa. Kuvasta 3 nähdään, että mallien devianssissa on havaittavissa melko identtistä käyttäytymistä lajin prevalenssin suhteen sekä selitysettä ennustevoiman tapauksessa. HMSC-mallin devianssi on melko järjestelmällisesti pienempää verrattuna GLLVM-mallin devianssiin, mikä kertoo HMSC-mallin paremmasta tarkkuudesta. Tämä on havaittavissa sekä mallien selitys- että ennustevoimalla. Mallien deviansseissa on havaittavissa myös nouseva trendi lajin prevalenssiin suhteen, mikä kertoo siitä että mitä useammin laji on havaittu aineistossa, sitä epätarkempia estimaatteja mallit tuottavat.



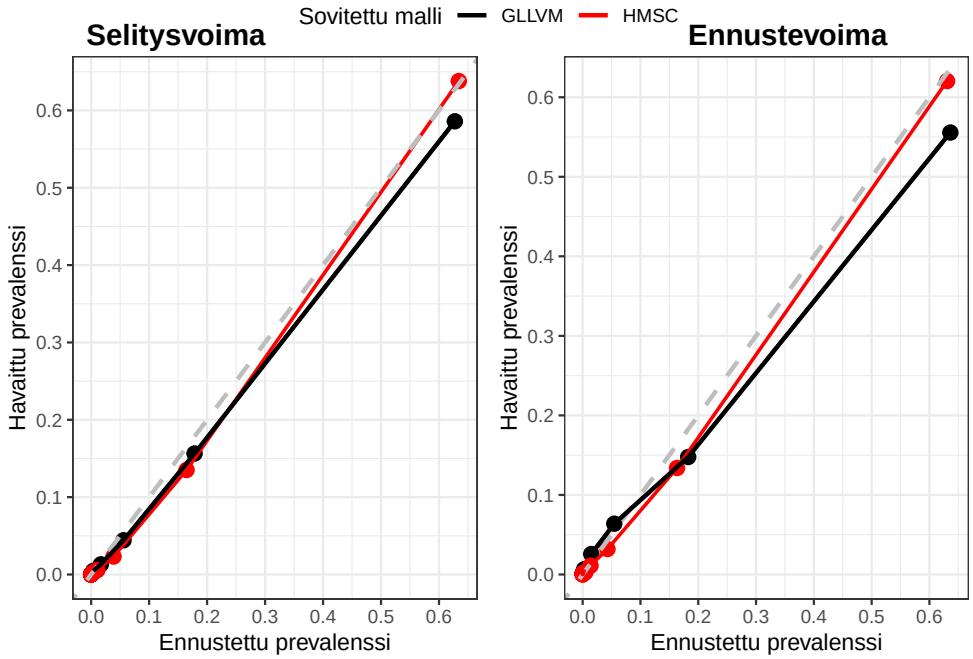
Kuva 3: Mallien devianssit selitys- ja ennustevoimalle lajien prevalenssin suhteen. Pisteet vastaavat lajikohtaisen prevalenssin ja vastaavan ryhmän devianssin keskiarvopisteitä.

Verrattaessa mallien selitysvoiman- ja ennustevoiman devianssia keskenään on havaittavissa ainoastaan tasoeroja. Devianssin käyttäytyminen lajin prevalenssin suhteen on hyvin samankaltaista selitys- ja ennustevoiman välillä. Poikkeuksena on HMSC-mallin selitysvoiman devianssi siirryttäessä 6 %:lla havaintopaikoista havaituista lajeista 10 %:lla havaintopaikoista havaittuihin lajeihin, jolloin devianssi pienenee. Ero näiden kahden pisteen välillä on kuitenkin niin pieni, että se voi myös olla satunnaisuuden aiheuttamaa. Molemmilla malleilla selitysvoima tuottaa tarkempia estimaatteja ennustevoimaan nähden.

Tarkasteltaessa harvinaisimpia lajeja huomataan, että selitysvoiman tapauksessa molemmat mallit tuottavat keskimäärin yhtä tarkkoja estimaatteja. Ennustevoiman tapauksessa HMSC-malli tuottaa hieman tarkempia estimaatteja GLLVM-malliin nähden.

5.3.3 Kalibraatio

Kuvassa 4 tarkastelemme mallien kalibraatiota selitys- ja ennustevoimalle. Erona mallien diskriminaation ja devianssin tutkimiseen tarkastelemme kalibraatiota kaikilla lajeille samanaikaisesti siten, että jaamme aineiston kymmeneen ryhmään ennustettujen prevalenssien suhteen kaikkien ennustepisteiden yli. Tämän jälkeen tarkastelemme jokaisen ryhmän keskiarvopisteessä ryhmän aineistosta lasketun havaitsemistodennäköisyyden keskiarvoa.



Kuva 4: Kalibraatiot mallien selitys- ja ennustevoimalle. Pisteet vastaavat ennustetun prevalenssin mukaan ryhmiteltyjä ryhmien keskiarvopisteitä. Täydellisesti kalibroitunut malli seuraa harmaata katkoviivaa.

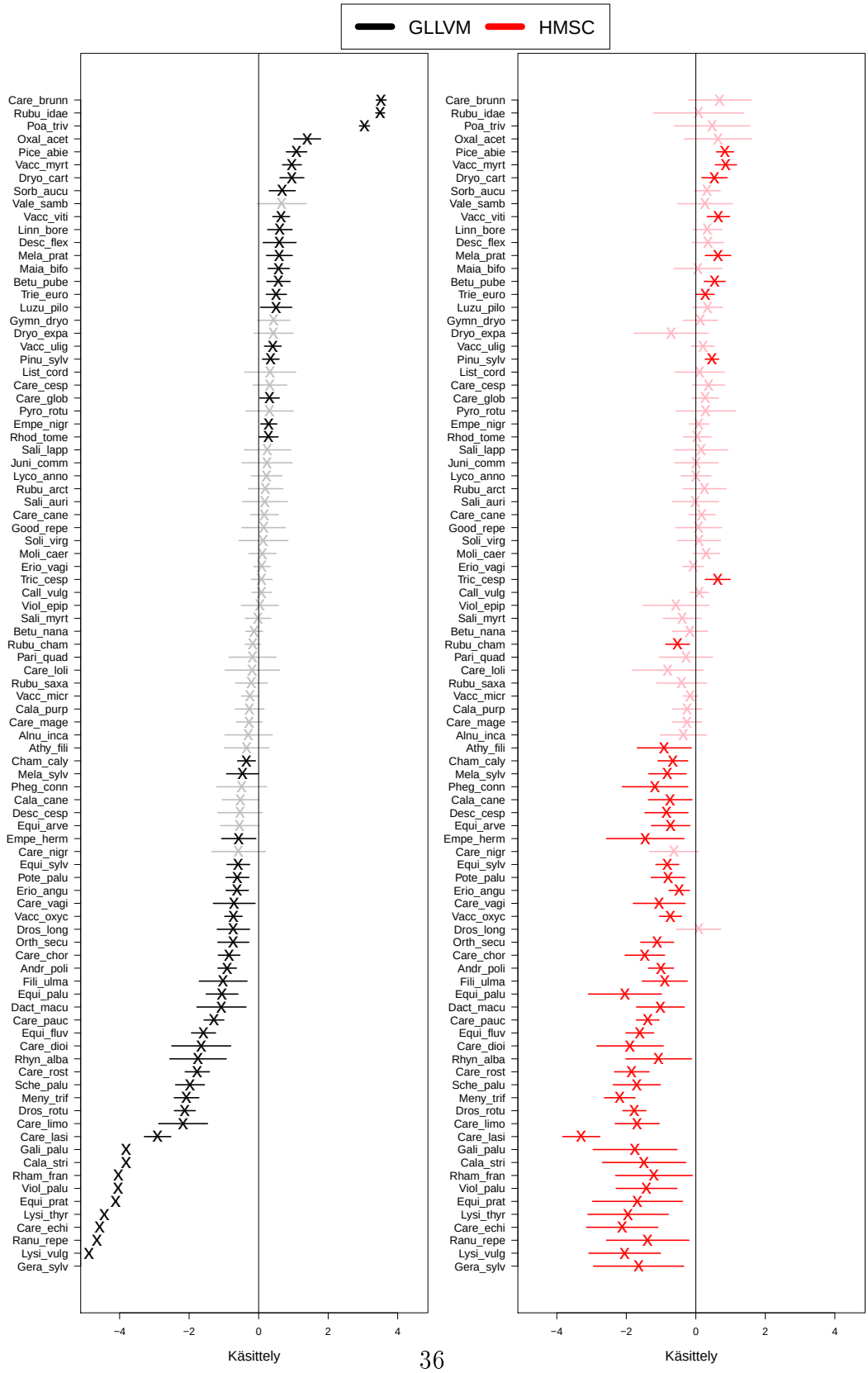
Kuvasta 4 nähdään, että sekä selitys- että ennustevoimassa lajeilla, joita on havaittu aineistossa alle 10 %:lla havaintopaikoista, GLLVM- ja HMSC-mallit ovat kalibroituneet likimain yhtä hyvin. Erona ennustevoiman tapauksessa edellä mainitulla välillä on, että GLLVM-malli tuottaa hieman liian pieniä ennustetodennäköisyyksiä, kun taas HMSC-malli tuottaa hieman liian suuria ennustetodennäköisyyksiä. Kun lajin havaittu prevalenssi nousee

yli 20 %:n, on HMSC-malli sekä selitys- että ennustevoiman tapauksessa paremmin kalibroitunut kuin GLLVM-malli. Lajeilla, joita on havaittu keskimäärin yli 35 % havaintopaikoista, HMSC:n kalibraatio on jo lähes nollassa, sillä se on hyvin lähellä harmaata diagonaalia. Samanaikaisesti GLLVM antaa systemaattisesti hieman liian suuria todennäköisyysennusteita.

5.3.4 Mallien parametriestimaattien vertailu

Kuvassa 5 tarkastelemme mallien suon käsittely -muuttujan β-parametrin estimaatteja sekä luottamus- ja todennäköisyysvälejä. Kuvasta 5 huomataan, että parametriestimaatit ovat mallien välillä pääosin samankaltaisia ja -suuruisia. Eroavaisuutena mallien välillä ilmenee kuitenkin GLLVM-mallin regressiokertoimien 95 %:n luottamusvälien ja HMSC-mallin tuottamien regressiokertoimien 95 %:n todennäköisyysvälien suuruuserot. HMSC-mallin parametriestimaattien 95 % todennäköisyysväli on huomattavasti GLLVM-mallia suurempaa, minkä seurauksena esimerkiksi lajien *Rhod_ tome* ja *Empe_ nigr* parametreissa GLLVM-malli tuottaa tilastollisesti merkitsevästi nollasta eroavan tuloksen, kun taas HMSC-mallin parametrin 95 %:n todennäköisyysväli sisältää nollan.

Nähdään myös, että varsinkin molempien ääripäiden estimaattien suuruusluokka eroaa. GLLVM tuottaa negatiivisessa ääripäässä selkeästi pienempiä ja positiivisessa ääripäässä selkeästi suurempia estimaateja HMSC-malliin verrattuna. Lisäksi negatiivisen ääripään estimaateissa GLLVM-mallin tuottaman parametriestimaatin keskivirhe estimoituu nollaan, kun taas HMSC-malli tuottaa verrattain suuren 95 %:n todennäköisyysvälin parametriestimaatille. Tämä selittyy ainakin GLLVM:n osalta sillä, että kyseessä olevia lajeja on havaittu pelkästään luonnontilaisilla soilla.



Kuva 5: (Jatkuu seuraavalla sivulla.)

Kuva 5: Suon käsittely -muuttujaan liittyvien lajikohtaisten regressiokertoimien estimaattien vertailu. Merkitsemme rastilla GLLVM-mallin parametrin piste-estimaattia ja HMSC-mallin posteriorikeskiarvoa. Viiva kertoo parametriestimaatin 95 %:n luottamusvälin GLLVM-mallille ja 95 %:n todennäköisyysvälin HMSC-mallille. Viivan haalea väri kertoo, että nolla sisältyy ko. väliin. Lajit ovat järjestetty GLLVM-mallin tuottaman piste-estimaatin mukaiseen suuruusjärjestykseen.

Pienemmästä vaihteluvälistä johtuen GLLVM-malli tuottaa selkeästi useamman lajin parametriestimaatille 95 %:n luottamusvälillä positiivisen parametriestimaatin HMSC-mallin 95 %:n todennäköisyysväliin verrattuna. Tämä saa aikaan osittain eriävän tulkinnan suon käsittelyn vaikutuksesta lajin havaitsemiseen GLLVM- ja HMSC-mallien välillä, kun GLLVM-mallin mukaan ojitettu suo nostaisi lajin havaitsemistodennäköisyyttä, mutta HMSC-mallin mukaan vastaavaa näyttöä ei ole. Yksi eroa selittävä tekijä on variaatioapproksimaatiolla sovitetun GLLVM-mallin liian kapeat luottamusvälit parametriestimaateille (Niku et al., 2021b).

Molempien mallien tuloksien avulla voidaan kutenkin sanoa, että enemmistölle lajeista ojitettu suo vähentää lajin todennäköisyyttä tulla havaituksi. Lopuille lajeista ojitettu suo joko lisää havaitsemistodennäköisyyttä tai ei vaikuta siihen tilastollisesti merkitsevästi. Verrattuna toisiinsa lajit vastaavat suon käsittelyyn selkeästi hyvin vaihtelevasti. Voidaan kuitenkin todeta, että molempien mallien avulla pystyttiin tekemään pääosittain samansuuntaista ja luotettavaa päättelyä suon käsittelyn vaikutuksesta lajin havaitsemistodennäköisyyteen.

Tarkastellaan seuraavaksi lajin prevalenssin yhteyttä suon käsittely -muuttujan parametriestimaattiin. Taulukoissa 5 ja 6 on esitettynä GLLVM- ja HMSC-mallin käsittely-parametrien estimaattien jakaumat lajin prevalenssin suhteen. Taulukoista 5–6 nähdään, että lajin prevalenssilla näyttäisi olevan molemmissa malleissa yhteys suon käsittely -muuttujan parametriestimaattiin.

Taulukko 5: GLLVM-mallin suon käsittely -muuttujan parametriestimaattien 95 %:n luottamusvälin jakautuminen lajin prevalenssin suhteen. Mikäli nolla sisältyy em. väliin, merkitään se nollaksi.

Lajin prevalenssi	<0 N(%)	=0 N(%)	>0 N(%)
[0.004, 0.006)	2 (20)	6 (60)	2 (20)
[0.006, 0.008)	3 (33.3)	6 (66.7)	0 (0.0)
[0.008, 0.013)	4 (44.4)	4 (44.4)	1 (11.1)
[0.013, 0.018)	4 (44.4)	4 (44.4)	1 (11.1)
[0.018, 0.025)	5 (55.6)	3 (33.3)	1 (11.1)
[0.025, 0.041)	2 (22.2)	6 (66.7)	1 (11.1)
[0.041, 0.082)	3 (33.3)	2 (22.2)	4 (44.4)
[0.082, 0.139)	5 (55.6)	0 (0.0)	4 (44.4)
[0.139, 0.222)	4 (44.4)	1 (11.1)	4 (44.4)
[0.222, 0.559]	3 (33.3)	3 (33.3)	3 (33.3)

Taulukko 6: HMSC-mallin suon käsittely -muuttujan parametriestimaattien 95 %:n todennäköisyysvälin jakautuminen lajin prevalenssin suhteen. Mikäli nolla sisältyy em. väliin, merkitään se nollaksi.

Lajin prevalenssi	<0 N(%)	=0 N(%)	>0 N(%)
[0.004, 0.006)	2 (20)	8 (80)	0 (0.0)
[0.006, 0.008)	4 (44.4)	5 (55.6)	0 (0.0)
[0.008, 0.013)	6 (66.7)	3 (33.3)	0 (0.0)
[0.013, 0.018)	6 (66.7)	3 (33.3)	0 (0.0)
[0.018, 0.025)	4 (44.4)	5 (55.6)	0 (0.0)
[0.025, 0.041)	2 (22.2)	6 (66.7)	1 (11.1)
[0.041, 0.082)	3 (33.3)	3 (33.3)	3 (33.3)
[0.082, 0.139)	5 (55.6)	4 (44.4)	0 (0.0)
[0.139, 0.222)	4 (44.4)	2 (22.2)	3 (33.3)
[0.222, 0.559]	4 (44.4)	3 (33.3)	2 (22.2)

Suurimmalla osalla harvinaisimmista lajeista suon käsittelyllä ei ole vaikutusta lajin havaitsemiseen. Kun siirrytään lajeihin, joita on havaittu suhteellisesti 0.008–0.041 havaintopaikoista, on suon käsittelyn vaikutus enimmäkseen havaitsemistodennäköisyyttä heikentävää, tai vaikutusta ei ole. Osalle lajeista GLLVM-malli tuottaa HMSC-malliin poiketen havaitsemistodennäköisyyttä nostavia estimaatteja. Tarkastellessa eniten havaittuja lajeja (prevalenssi 0.041 – 0.559) havaitaan, että lajien parametriestimaateissa on muihin lajeihin verrattuna enemmän lajin havaitsemistodennäköisyyttä kasvattavia estimaatteja.

GLLVM- ja HMSC-mallien välillä on pienehköjä eroja edellä mainituissa tuloksissa, mutta tulokset ovat pääasiassa samansuuntaiset. Voidaan siis tulkita, että suon ojitus keskimäärin pienentää harvinaisten lajien havaitsemistodennäköisyyttä, mutta usein havaittujen lajien kohdalla suon ojituksella voi jopa olla havaitsemista kasvattava vaikutus.

6 Pohdinta

Tämän työn tarkoituksena oli vertailla GLLVM- ja HMSC-malleja teoreettisesti. Lisäksi vertailimme mallien suorituskykyä lajiyhteisön mallinnuksessa. Sovitimme putkilokasvien havainnoista koostuvaan aineistoon mahdollisimman samankaltaiset mallit sen mukaan minkälaisten termien asettaminen oli kullakin mallilla mahdollista ja tarkoituksenmukaista tutkimusasetelman kannalta. Malleissa käytimme ympäristökovariaatteina suon ravinteisuutta, suotyyppiä ja suon käsittelyä. Lajikovariaatteja ei käytetty analyysivaiheessa, koska niitä ei ollut valmiiksi saatavilla työtä tehtäessä, ja työn painopiste oli menetelmien tilastotieteellisessä vertailussa.

Mallien teoreettisen tarkastelun ja vertailun tuloksena osoitimme eroavaisuudet sekä samankaltaisuudet niin mallien sovitustavassa kuin tavassa ottaa huomioon lajien välinen korrelaatiorakenne. Lisäksi osoitimme, että GLLVM- ja HMSC-mallin tapa ottaa mallinnuksessa huomioon lajikovariaatit on molemmilla malleilla tulkittavissa ympäristökovariaattien ja lajikovariaattien väliseksi interaktioksi.

Data-analyysien tuloksena saimme esitettyä mallien välisiä eroavaisuuk-

sia käytännössä. Tarkastellessa mallien diskriminaatiota, tarkkuutta ja kalibraatiota oli havaittavissa, että HMSC-malli suoriutui melko systemaattisesti GLLVM-mallia paremmin sekä selitys- että ennustevoiman tapauksessa. Havaittavissa oli myös tendenssi, että lajin prevalenssin kasvaessa erot mallien suoriutumisessa kasvoivat toisiinsa verrattuna. Tarkkaa syytä mallien välisille eroille on mahdoton paikantaa tämän työn avulla. Mahdollisia vaikuttavia tekijöitä ovat ainakin mallien eritasoiset latenttimuuttujatermit, GLLVM-mallin satunnaismuuttujatermin vaikutus sekä mahdolliset mallien väliset ns. luontaiset erot. Ottaen huomioon, että sovitettu GLLVM-malli sisälsi palstakohtaisen satunnaismuuttujatermin, jota HMSC-mallissa ei ollut, voidaan tulosten valossa sanoa, että palstakohtainen satunnaistermi ei pystynyt korvaamaan palstakohtaisten latenttimuuttujatermien selitysvoimaa. Palstakohtaiset latenttimuuttujatermit vaikuttivat olevan perusteltu valinta aineiston mallintamiseen.

Molempien mallien kalibraatio oli keskimääräisesti hyvä, ja eroavaisuuksia havaittiin mallien käyttäytymisessä harvojen ja usein havaittujen lajien välillä. GLLVM-malli tuotti harvinaisilla lajeilla liian pieniä todennäköisyysennusteita, kun taas HMSC-malli tuotti hieman liian suuria ennustetodennäköisyyksiä. Lajin prevalenssin kasvaessa HMSC tuotti paremmin kalibroituneita todennäköisyysestimaatteja, kun taas GLLVM tuotti systemaattisesti liian suuria todennäköisyysestimaatteja. Kummallakaan mallilla ei ollut juurikaan eroavaisuuksia kalibraatiossa selitys- ja ennustevoiman välillä.

Mallien diskriminaatiossa havaittiin myös suurehkoja eroja sekä lajin prevalenssin suhteen, että mallien käyttäytymisessä yleisesti. HMSC-mallin tuottamat Tjur R^2 -arvot vaihtelivat huomattavasti suuremmalla välillä kuin GLLVM-mallin vastaavat suureet. Selitys- ja ennustevoiman välillä HMSC-mallin suureet laskivat huomattavasti enemmän kuin GLLVM-mallin. Syyt näihin eroihin mallien välillä eivät ole täysin selkeät. Viitteitä näyttäisi olevan ainakin siitä, että käytettävissä olevan aineiston määrän pienentyessä, Bayes-estimointi alkaa tuottaa verrattain huonompia tuloksia selitysvoimaan verraten. Silti HMSC-mallin ennustevoima on systemaattisesti jopa suurempaa kuin GLLVM-mallin selitysvoima.

Mallien tarkkuutta tutkittaessa HMSC-mallin devianssi oli järjestelmälli-

sesti pienempää verrattuna GLLVM-mallin devianssiin, mikä kertoo HMSC-mallin paremmasta tarkkuudesta. Tämä on havaittavissa sekä mallien selitys- että ennustevoimalla. Mallien deviansseissa on havaittavissa myös nouseva trendi lajin prevalenssin suhteen, mikä kertoo siitä että mitä useammin laji on havaittu aineistossa, sitä epätarkempia estimaatteja mallit tuottavat.

Kun vertasimme GLLVM- ja HMSC-mallien parametriestimaatteja suon käsittely -muuttujalle, huomasimme että GLLVM-mallin kapeammat luottamusvälit aiheuttivat osittain eriävän tulkinnan suon käsittelyn vaikutuksesta lajin havaitsemistodennäköisyyteen HMSC-malliin verrattuna. GLLVM-malli tuotti useammalle lajille 95 %:n luottamusvälillä positiivisen estimaatin HMSC-mallin 95 %:n todennäköisyysväliin verrattuna. Kuitenkin molemmilla malleilla suurin osa parametriestimaateista oli ko. vaihteluvälillä negatiivista, mikä antaa näyttöä siitä, että ojitettu suo pienentää useiden putkilokasvilajien havaitsemistodennäköisyyttä.

Tarkoituksenmukaista jatkotarkastelua on mahdollista tehdä molempien mallien osalta. GLLVM-mallin sovittaminen palstakohtaisilla latenttimuuttujatermeillä olisi suora testaus hypoteesistä, että erot mallien välillä johtuivat palstakohtaisten latenttimuuttujien paremmasta sopivuudesta tarkasteltuun aineistoon. Myös esimerkiksi valinta approksimaatiomenetelmän käytöstä vaikuttanee tuloksiin. Tässä työssä käytetyn VA-approksimaation avulla tulokset saadaan laskennallisesti tehokkaasti, mutta kääntöpuolena tulosten hyvyys tarkastelluilla mittareilla vaikuttaa kärsivän. Tulosten toistettavuuden kannalta voisi myös olla järkevää tutkia, kuinka paljon HMSC-mallissa käytettyjen simulaatioiden määrä vaikuttaa saatuihin tuloksiin. Tässä työssä valitulla simulaatioiden määrällä saavutettiin lähes kaikkien parametrien konvergenssi, mutta posteriorijakauman koko 2×1000 ei ole perinteisessä Bayes-analyysissä kovin iso. Suuremman posteriorijakauman vaikutus parametriestimaattien 95 %:n todennäköisyysväliin olisi eräs mahdollinen jatkotarkastelun aihe. Lisäksi luvussa 3.2.2 esitettyjen sensitiivisten hyperparametrien arvojen vaikutusta on mahdollista tutkia. Tässä työssä käytetyt oletusarvoiset arvot parametreille a, b ja υ vaikuttivat toimivan kuitenkin verrattain hyvin.

Työstä pois jääneiden lajikovariaattien vaikutuksen tutkiminen olisi ver-

tailun kannalta mielenkiintoista. Esimerkiksi samantyyppisen vertailun tekeminen lajikovariaattien ja ympäristökovariaattien välisten interaktiotermien $\boldsymbol{\beta}$-kertoimille kuin tässä työssä on tehty suon käsittely -muuttujan $\boldsymbol{\beta}$-kertoimille voisi antaa uutta tietoa siitä, miten mallien perusteella tehty inferenssi eroaa toisistaan. Lienee mahdollista, että GLLVM- ja HMSC-mallien välillä löytyy eroja myös lajikovariaattien interaktiotermien estimaattien käyttäytymisessä, koska eroja löytyi myös ympäristökovariaattien estimaattien käyttäytymisessä.

Viitteet

Bhattacharya Anirban ja Dunson David B. Sparse Bayesian infinite factor models. *Biometrika*, sivut 291–306, 2011.

Brown Alexandra M, Warton David I, Andrew Nigel R, Binns Matthew, Cassis Gerasimos ja Gibb Heloise. The fourth-corner solution–using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution*, **5**, 344–352, 2014.

Elith Jane, H. Graham Catherine, P. Anderson Robert, Dudík Miroslav, Ferrier Simon, Guisan Antoine, J. Hijmans Robert, Huettmann Falk, R. Leathwick John ja Lehmann Anthony. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151, 2006.

Elo Merja, Kareksela Santtu, Haapalehto Tuomas, Vuori Hilja, Aapala Kaisu ja Kotiaho Janne S. The mechanistic basis of changes in community assembly in relation to anthropogenic disturbance and productivity. *Ecosphere*, **7**, e01310, 2016.

Gelman Andrew. *Markov Chain Monte Carlo in Practice*. CRC Press, 1995.

Gelman Andrew ja Hill Jennifer. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

Gelman Andrew ja Rubin Donald B. Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472, 1992.

Gneiting Tilmann ja Raftery Adrian E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378, 2007.

Harrison Xavier A, Donaldson Lynda, Correa-Cano Maria Eugenia, Evans Julian, Fisher David N, Goodwin Cecily ED, Robinson Beth S, Hodgson David J ja Inger Richard. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, **6**, e4794, 2018.

Huber Philippe, Ronchetti Elvezio ja Victoria-Feser Maria-Pia. Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 893–908, 2004.

Hui Francis K.C., Taskinen Sara, Pledger Shirley, Foster Scott D. ja Warton David I. Model-based approaches to unconstrained ordination. *British Ecological Society*, **6**, 399–411, 2014.

Hui Francis KC, Warton David I, Ormerod John T, Haapaniemi Viivi ja Taskinen Sara. Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, **26**, 35–43, 2017.

Niku Jenni. On modeling multivariate abundance data with generalized linear latent variable models. *JYU dissertations*, 2020.

Niku Jenni, Brooks Wesley, Herliansyah Riki, Hui Francis K. C., Taskinen Sara, Warton David I. ja van der Veen Bert. *gllvm: Generalized Linear Latent Variable Models*, 2021a. R package version 1.3.1.

Niku Jenni, Brooks Wesley, Herliansyah Riki, Hui Francis KC, Taskinen Sara ja Warton David I. Efficient estimation of generalized linear latent variable models. *PloS one*, **14**, e0216129, 2019a.

Niku Jenni, Hui Francis KC, Taskinen Sara ja Warton David I. gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, **10**, 2173–2182, 2019b.

Niku Jenni, Hui Francis KC, Taskinen Sara ja Warton David I. Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics*, **32**, e2683, 2021b.

Norberg Anna, Abrego Nerea, Blanchet F Guillaume, Adler Frederick R, Anderson Barbara J, Anttila Jani, Araújo Miguel B, Dallas Tad, Dunson David ja Elith Jane. A comprehensive evaluation of predictive performance of

33 Species distribution models at species and community levels. *Ecological Monographs*, **89**, 2019.

Ovaskainen Otso ja Abrego Nerea. *Joint Species Distribution Modelling: With Applications in R.* Cambridge University Press, 2020.

R Core Team . *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021.

Rousk Johannes, Brookes Philip C ja Baath Erland. Contrasting soil pH effects on fungal and bacterial growth suggest functional redundancy in carbon mineralization. *Applied and Environmental Microbiology*, **75**, 1589–1596, 2009.

Tikhonov Gleb, Ovaskainen Otso, Oksanen Jari, de Jonge Melinda, Opedal Oystein ja Dallas Tad. *Hmsc: Hierarchical Model of Species Communities*, 2021. R package version 3.0-11.

Tjur Tue. Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. *The American Statistician*, **63**, 366–372, 2009.

Wang YI, Naumann Ulrike, Wright Stephen T ja Warton David I. mvabund: an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, **3**, 471–474, 2012.

Warton David I, Foster Scott D, De'ath Glenn, Stoklosa Jakub ja Dunstan Piers K. Model-based thinking for community ecology. *Plant Ecology*, **216**, 669–682, 2015.

Liitteet

Liite A: GLLVM- ja HMSC-mallikoodit

```
# Sovitetaan ensin selitysvoimaa kuvaava GLLVM-malli.
# Vastemuuttujat ovat objektissa putkilo01 ja
# ympäristökovariaatit objektissa putkilo.kovariaatit.
# Ympäristökovariaatteina käytetään muuttujia "Suotyyppi",
# "Ravinteisuus" ja "Suon käsittely".
# Latenttimuuttujien määräksi asetetaan kaksi ja malliin
# asetetaan palstakohtainen satunnaistermi (ID).

require(gllvm)
fitGLLVM.selitysvoima <- gllvm(y = putkilo01,
                           x = putkilo.kovariaatit,
                           formula = ~Suotyyppi +
                           Ravinteisuus + Kasittely,
                           family = binomial(link= "probit"),
                           sd.errors = TRUE,
                           row.eff = ~(1|ID),
                           num.lv = 2)

# Sovitetaan ennustevoimaa kuvaava GLLVM-malli.

require(caTools)

# Jaetaan data ensin kahteen yhtäsuureen osaan satunnaisesti.
# Valitaan toistettavuuden takaamiseksi siemenluku.
set.seed(120)
sample <- sample.split(putkilo01, SplitRatio = .50)

# Lajien jako
```

```
train.putkilo01 <- subset(putkilo01, sample == TRUE)
test.putkilo01  <- subset(putkilo01, sample == FALSE)

# Kovariaattien jako
train.kovariaatti <- subset(putkilo.kovariaatit,
                     sample)
test.kovariaatti  <- subset(putkilo.kovariaatit,
                     !sample)

# Mallien sovitus opetusdatalla
fitGLLVM.train <- gllvm(y = train.putkilo01,
                  x = train.kovariaatti,
                  formula = ~ Suotyyppi +
                  Ravinteisuus + Kasittely,
                  family = binomial(link="probit"),
                  sd.errors = TRUE,
                  row.eff = ~(1|ID),
                  num.lv = 2)

# Mallien sovitus testidatalla
fitGLLVM.test <- gllvm(y = test.putkilo01,
                 x = test.kovariaatti,
                 formula = ~ Suotyyppi +
                 Ravinteisuus + Kasittely,
                 family = binomial(link="probit"),
                 sd.errors = TRUE,
                 row.eff = ~(1|ID),
                 num.lv = 2)

# Ennustetaan opetusdataa testidatalla.
GLLVM.pred.train.test <- predict.gllvm(fitGLLVM.train,
                         type = "response",
                         newX = test.kovariaatti)
```

```
# Ennustetaan testidataa opetusdatalla.
GLLVM.pred.test.train <- predict.gllvm(fitGLLVM.test,
                         type = "response",
                         newX = train.kovariaatti)

# Sovitetaan seuraavaksi HMSC-mallit.
# Aloitetaan selitysvoimaa kuvaavasta mallista.

require(Hmsc)

# Asetetaan latenttimuuttujat palstakohtaisiksi.
studyDesign <- data.frame(plot =
               as.factor(putkilo.kovariaatit$ID),
               stringsAsFactors = TRUE)
rL <- HmscRandomLevel(units = studyDesign$plot)

# Asetetaan latenttimuuttujien määrä täsmälleen kahdeksi.
rL$nfMin <- 2; rL$nfMax <- 2

# Mallin määrittäminen
model <- Hmsc(Y=putkilo01, XData=putkilo.kovariaatit,
         XFormula=~Suotyyppi + Ravinteisuus + Kasittely,
         distr = "probit", studyDesign=studyDesign,
         ranLevels=list(sample=rL))

# Mallin estimointi
model.est <- sampleMcmc(model, thin = 100,
             samples = 1000, transient = 100*330,
             nChains = 2, verbose = 50)

# Viimeiseksi lasketaan ennustevoimaa kuvaava HMSC-malli.
# Jaetaan data samaan kahteen osaan kuin GLLVM-mallin
```

```
# tapauksessa.

require(caTools)

# Datan jako
set.seed(120)
sample <- sample.split(putkilo01, SplitRatio = .50)
sample.num <- as.numeric(sample) +1

# Sovitetaan malli käyttäen hyödyksi koko aineistolle
# äsken sovitettua mallia model.est.
preds.HMSC.ennuste <- computePredictedValues(model.est,
                        partition=sample.num)
```