LAJIEN ESIINTYMISEN MALLINTAMINEN LOGISTISELLA REGRESSIOLLA

LuK-tutkielma 2010

Henna Kettunen

Helsingin yliopisto
Bio- ja ympäristötieteellinen tiedekunta
Biotieteiden laitos, pääaine kasvibiologia

Työn ohjaaja: dos. Tomas Roslin

Kandidaatintutkielman tiivistelmä

Tiedekunta/Osasto – Fakultet/Sektion – Faculty	Laitos – Institution – Department	
Bio- ja ympäristötieteellinen tiedekunta	Biotieteiden laitos	
Tekijä – Författare – Author		
Henna Kettunen		
Työn nimi – Arbetets titel – Title		
Lajien esiintymisen mallintaminen logistisella regressiolla		
Oppiaine – Läroämne – Subject		
Kasvibiologia		
Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
LuK-tutkielma	Lokakuu 2010	22

Tiivistelmä – Referat – Abstract

Elinympäristön vaikutusta eliölajien esiintymiseen voidaan tutkia habitaattimallinnuksen keinoin. Habitaattimallit ovat tilastollisia malleja, jotka suhteuttavat havainnot lajin esiintymisestä ympäristön ominaisuuksiin. Habitaattimalleja voidaan käyttää apuna esimerkiksi luonnonsuojelualueiden rajaamisessa: jos havaintotiedot eivät ole kattavia, lajien esiintymistä voidaan ennustaa ympäristön ominaisuuksien perusteella. Habitaattimallinnukseen on useita eri menetelmiä, joista tässä työssä käsittelen logistista regressiota. Logistiset regressiomallit ovat yleistettyjä lineaarisia malleja binäärisille vastemuuttujille. Niitä voidaan käyttää habitaattimallinnukseen tilanteissa, joissa mallin sovittamiseen käytettävä lajihavaintoaineisto sisältää sekä läsnäolo- että puuttumishavaintoja.

Habitaattimallinnuksen taustalla on oletus, että lajille soveltuvien elinympäristöjen esiintyminen määrää suoraviivaisesti lajin esiintymisen. Siten habitaattimallien ennusteet ovat luotettavia ainoastaan silloin, kun lajin esiintyminen on tasapainossa ympäristön kanssa. Toisaalta monet muutkin tekijät, kuten väärin valitut selittävät muuttujat tai vaikeasti mallinnettavat bioottiset vuorovaikutukset, voivat heikentää habitaattimallin ennusteita. Siksi habitaattimallin selityskykyä eli sitä, kuinka hyvin ennusteet vastaavat lajin todellista esiintymistä, on syytä arvioida. Mallin selityskykyä voidaan arvioida esimerkiksi ROC-käyrän alaisella pinta-alalla tai ristiinvalidoimalla.

Kun habitaattimallien perusteella tehdään johtopäätöksiä lajien elinympäristövaatimuksista tai esiintymisestä, täytyy aina muistaa, ettei malli vastaa täydellisesti todellisuutta. Kuvaan kahden esimerkkitutkimuksen avulla muutamia tähän liittyviä näkökohtia, jotka habitaattimallien tulkinnassa on hyvä huomioida. Pienen lounaissuomalaisen Wattkast-saaren tammille luodun habitaattimallin heikko selityskyky on hyvä esimerkki siitä, kuinka tärkeää on varmistua aina habitaattimallin taustaoletusten toteutumisesta. Neljälle australialaiselle kasvilajille luotu habitaattimalli taas osoittaa, että menetelmä, jolla habitaattimallilla arvioidut esiintymistodennäköisyydet muunnetaan ennusteiksi lajin esiintymisestä, voi vaikuttaa suuresti mallin perusteella tehtäviin johtopäätöksiin.

Kaiken kaikkiaan habitaattimallinnus on oikein sovellettuna erittäin tärkeä työkalu eliöiden ja luonnon monimuotoisuuden esiintymisen arvioinnissa. Erityisesti ilmastonmuutoksen vaikutuksia ennakoitaessa on välttämätöntä turvautua lajien levinneisyyden mallintamiseen, koska tulevaisuudesta ei luonnollisesti ole saatavilla havaintoaineistoja. Havainnoivan ekologisen perustutkimuksen tärkeyttä ei pidä silti unohtaa: esiintymismallit tuottavat järkeviä ennusteita vain, mikäli ne perustuvat järkeville oletuksille eliöiden ja niiden ympäristön välisistä suhteista. Koska ympäristö muuttuu jatkuvasti, myös tällä saralla riittää koko ajan tutkittavaa.

Avainsanat – Nyckelord – Keywords	
habitaattimallinnus, logistinen regressio, ristiinvalidointi, ROC-käyrä, yleistetty lineaarinen malli	
Säilytyspaikka – Förvaringställe – Where deposited	
Kasvibiologian käsikirjasto	
Muita tietoja – Övriga uppgifter – Additional information	

SISÄLLYS

1. ALUKSI

Eliöiden levinneisyyden ja runsauden selittäminen on yksi ekologian perustehtävistä (Krebs 1985: 4). Tässä työssä tarkastelen lajien esiintymisen mallintamista logistisen regression avulla. Esimerkkeinä käytän 1) omaan pro gradu -työhöni sisältyvää logistista habitaatimallia, jolla kuvasin tammen alueellista levinneisyyttä pienessä lounaissuomalaisessa saaressa, ja 2) australialaista tutkimusta neljän kasvilajin potentiaalisten esiintymisalueiden mallintamisesta suojelutarkoituksessa.

2. HABITAATTIMALLINNUS

Tietyn lajin esiintymiseen keskeisesti vaikuttavien tekijöiden tunnistaminen on usein vaikeaa, koska sekä lajin esiintymiskuvio että siihen vaikuttavat ympäristötekijät ja muut prosessit voivat vaihdella ajassa (vrt. Hanski 1999). Apuna esiintymisen tutkimisessa ja ennustamisessa voidaan käyttää habitaatti- eli elinympäristömallinnusta (Elith & Leathwick 2009b). Habitaattimallit ovat tilastollisia malleja, jotka suhteuttavat havainnot lajin esiintymisestä tai puuttumisesta maiseman eri osissa näillä paikoilla vallitseviin ympäristöoloihin (Schultz ym. 2003, Elith & Leathwick 2009a, b). Habitaattimallien avulla voidaan siis arvioida tietyn ympäristötekijäyhdistelmän luonnehtiman elinympäristön soveltuvuutta lajille (Beutel ym. 1999). GIS-tietokantoihin yhdistetyillä habitaattimalleilla voidaan lisäksi luoda ennustekarttoja eliölajien potentiaalisesta levinneisyydestä eri maantieteellisissä mittakaavoissa.

Habitaattimallit ovat hyödyllisiä etenkin tilanteissa, joissa joudutaan rajaamaan suojelualueita puutteellisten tai epäluotettavien esiintymistietojen perusteella (Elith & Leathwick 2009a). Suojeluun käytettävissä olevat resurssit ovat yleensä rajalliset, joten on tärkeää varmistua siitä, että suojeltaviksi valitaan juuri ne alueet, jotka oikeasti mahdollistavat harvinaisten lajien säilymisen. Habitaattimallien avulla voidaan arvioida potentiaalisten alueiden soveltuvuutta tarkasteltaville lajeille, jos alueiden oleelliset ympäristöolot tunnetaan esimerkiksi kaukokartoituksen perusteella (vrt. Roslin ym. 2009). Toisaalta habitaattimalleja voidaan myös käyttää lajien elinympäristövaatimusten tunnistamiseen ilman ennustustarkoitusta (Schultz ym. 2003, Elith & Leathwick 2009b).

Habitaattimallinnukseen on useita, osittain eri käyttötarkoituksiin soveltuvia menetelmiä (Elith & Leathwick 2009a, b). Lajien levinneisyyttä suhteessa ympäristötekijöihin voidaan mallintaa mm. regressiopohjaisilla malleilla, monimuuttujamenetelmillä, ilmastollisilla profiileilla tai eri-

laisilla koneoppimismenetelmillä (ks. Elith & Burgman 2003). Jos käytetään ulkopuolisesta lähteestä saatua aineistoa, mallinnusmenetelmän valintaan vaikuttaa paitsi mallin käyttötarkoitus, myös mallin luomisessa käytettävän havaintoaineiston tyyppi. Eri menetelmien edellytyksiä ja tyypillisiä sovelluskohteita käsittelevät esimerkiksi Elith & Burgman (2003) sekä Elith & Leathwick (2009a).

Habitaattimallin lähtöaineisto voi joko sisältää pistehavaintoja sekä lajin läsnäolosta että puuttumisesta eri puolilla maisemaa tai, kuten perinteiset, herbaariokokoelmiin perustuvat kasvitieteelliset havaintoaineistot, pelkästään tietoja lajin havaintopaikoista (Gibson ym. 2007). Pisteaineistot ovat yleisesti melko luotettavia, mutta niihin liittyy muutamia tyypillisiä ongelmia, kuten havaintojen maantieteellinen niukkuus, väärät puuttuvat havainnot (laji on läsnä mutta jäänyt havaitsematta) sekä havaintojen painottuminen helppokulkuisille alueille (Elith & Leathwick 2009a). Habitaattimallin lähtöaineistona on mahdollista käyttää myös tietoja lajin runsausvaihteluista erilaisissa elinympäristöissä (Elith & Burgman 2003). Runsausaineistojen kerääminen on kuitenkin huomattavasti työläämpää, mistä johtuen niitä on saatavilla pisteaineistoja niukemmin.

Habitaattimalleissa käytetään selittäjinä ympäristötekijöitä ja joskus myös maantieteellisiä muuttujia (Elith & Leathwick 2009b). Se, että selittäjät ovat ekologisesti oleellisia – että ne oikeasti vaikuttavat lajin esiintymiseen – on ratkaisevaa habitaattimallin yleistettävyydelle (Elith & Leathwick 2009a). Yleensä proksimaaliset eli suoraan lajin elinympäristövaatimuksiin liittyvät selittäjät tuottavat parempia ennusteita kuin maantieteelliset muuttujat (esim. korkeus merenpinnasta), jotka ovat vain epäsuorasti yhteydessä lajin esiintymiseen (Guisan & Zimmermann 2000). Oleellisten selittäjien valinta potentiaalisten selittäjien joukosta on melkeinpä oma tieteenalansa, johon on tarjolla lukemattomia tilastollisia apuvälineitä (ks. Elith & Leathwick 2009b). En käsittele tässä työssä juurikaan selittäjien valintaa.

Habitaattimallien taustaoletuksena on, että mallinnettavien lajien esiintyminen on ajallisesti vakaata ja että elinympäristön esiintyminen määrää suoraviivaisesti sen, missä lajit esiintyvät (Hanski 2007: 38). Jos lajin nykyinen levinneisyys on seurausta historiallisista ympäristönmuutoksista tai monimutkaisista häiriökuvioista, laji voi puuttua monilta ympäristöoloiltaan suotuisilta paikoilta, jolloin potentiaalisten esiintymisalueiden mallintaminen on vaikeaa (vrt. Elith & Leathwick 2009a). Toinen vaikeasti mallinnettava lajiryhmä ovat invasiiviset tulokaslajit, joiden levinneisyys ei ole ajallisesti vakaata vaan laajenee jatkuvasti (Elith & Burgman 2003). Taustaoletukset eivät päde myöskään niillä lajeilla, joiden esiintymiseen liittyy metapopulaatiodynamiikkaa. Metapopulaation sisällä on aina ympäristöoloiltaan sopivia

mutta asuttamattomia elinympäristölaikkuja, ja lisäksi lajin esiintyminen yksittäisellä elinympäristölaikulla vaihtelee ajan funktiona (Hanski 2007: 38).

Mikäli habitaattimallin taustaoletuksista poiketaan suuresti, mallin perusteella voidaan tehdä virheellisiä johtopäätöksiä (vrt. Elith & Burgman 2003, Elith & Leathwick 2009a, b). Ennen kuin habitaattimallin ennusteita käytetään minkäänlaisen päätöksenteon pohjana, on siten suositeltavaa varmistua siitä, että taustaoletukset toteutuvat. Habitaattimallien ennusteet ovat luotettavimmillaan silloin, kun lajien levinneisyyttä tarkastellaan suuressa mittakaavassa, tarkasteltavat lajit ovat suhteellisen yleisiä ja lajien elinympäristövaatimukset voidaan kiteyttää muutamaan keskeiseen ympäristötekijään (Hanski 2007: 38).

3. ELINYMPÄRISTÖN MALLINTAMINEN LOGISTISELLA REGRESSIOLLA

3.1. Logistinen regressiomalli

Regressiomallit ovat korrelaatiopohjaisia tilastollisia malleja, jotka kuvaavat yhden tai useamman selittävän muuttujan vaikutusta selitettävään vastemuuttujaan (ks. esim. Ranta 2005: 365). Logistinen regressio on regressiomallien erikoistapaus binääriselle eli kaksiarvoiselle vastemuuttujalle (Collett 2003: 59). Logistinen regressiomalli estimoi selittävien muuttujien (selittäjien) arvojen perusteella todennäköisyyden tarkasteltavalle tapahtumalle. Vastetodennäköisyyden logit-muunnettu[1] estimaatti saadaan selittäjien arvojen lineaarikombinaationa (vrt. kuva 3.1a):

$$logit\,(p_i) = a + b_1X_1 + b_2X_2 + \ldots \text{ , jossa} \qquad (1)$$

$p_i =$ vastetodennäköisyys
$a =$ regressiovakio
$b_j =$ *j.* selittäjän regressiokerroin
$X_j =$ *j.* selittäjän arvo

Kaavasta (1) voidaan johtaa edelleen vastetodennäköisyyden sigmoidinen yhteys (2) selittäjien arvoihin (vrt. kuva 3.1b). Sigmoidista funktiota kutsutaan myös logistiseksi funktioksi.

[1] Logit-funktio on logistisen (sigmoidisen) funktion käänteisfunktio: $\mathrm{logit}\,(p) = \log\left(\frac{p}{1-p}\right)$, missä p voi saada arvoja nollan ja yhden väliltä.

$$p_i = \frac{e^{a + b_1X_1 + b_2X_2 + \ldots}}{1 + e^{a + b_1X_1 + b_2X_2 + \ldots}} \ . \qquad (2)$$

Koska vasteen ja selittäjien yhteys ei ole lineaarinen, yksittäisen selittäjän vaikutuksen suuruus vasteeseen riippuu selittäjän arvosta. Siten regressiokertoimelle ei logistisessa regressiossa ole yhtä suoraviivaista tulkintaa kuin lineaarisessa regressiossa, jossa selittäjän kerroin ilmaisee selittäjän lähtöarvosta riippumatta vasteen muutoksen silloin, kun selittäjän arvo muuttuu yhden yksikön. Logistisen regressiomallin kertoimille ja niiden keskivirheille ei myöskään ole mahdollista muodostaa analyyttisia lausekkeita. Parametrit estimoidaan yleensä numeerisesti suurimman uskottavuuden menetelmällä (Collett 2003: 60).

Tulkinnan helpottamiseksi logistisen regressiomallin kertoimet voidaan ilmaista vetosuhteen (*odds ratio*) avulla (Rita & Komonen 2008). Vetosuhde on mitta kahden osuuden väliselle etäisyydelle ja siten luonteva suure kuvaamaan sitä, miten selittäjän arvon muuttuminen muuttaa logistisen mallin vastetodennäköisyyttä. Vetosuhde ilmoittaa vastetodennäköisyyttä p_i vastaavassa vedossa (*odds*) tapahtuvan muutoksen, kun selittävän muuttujan arvo muuttuu yhden yksikön. Veto on vasteen tapahtumisen ja tapahtumatta jäämisen todennäköisyyksien suhde (3) ja siten aina positiivinen luku.

$$v(p_i) = \frac{p_i}{1 - p_i} \ . \qquad (3)$$

Vetosuhde (VS) taas on nimensä mukaisesti kahden vedon suhde (4). On tärkeää huomata, että vetosuhde ei ole todennäköisyyksien (tai osuuksien) suora suhde – tätä kutsutaan riskisuhteeksi (*risk ratio*).

$$VS\ (p_1,\ p_2) = \frac{v(p_1)}{v(p_2)} = \frac{p_1\ /\ (1 - p_1)}{p_2\ /\ (1 - p_2)} \ . \qquad (4)$$

Logistisen regressiomallin selittäjän regressiokerrointa vastaava vetosuhde saadaan laskemalla kaikkiin peräkkäisiin selittäjän arvoihin liittyvät vetosuhteet ja ottamalla näistä keskiarvo. Siis:

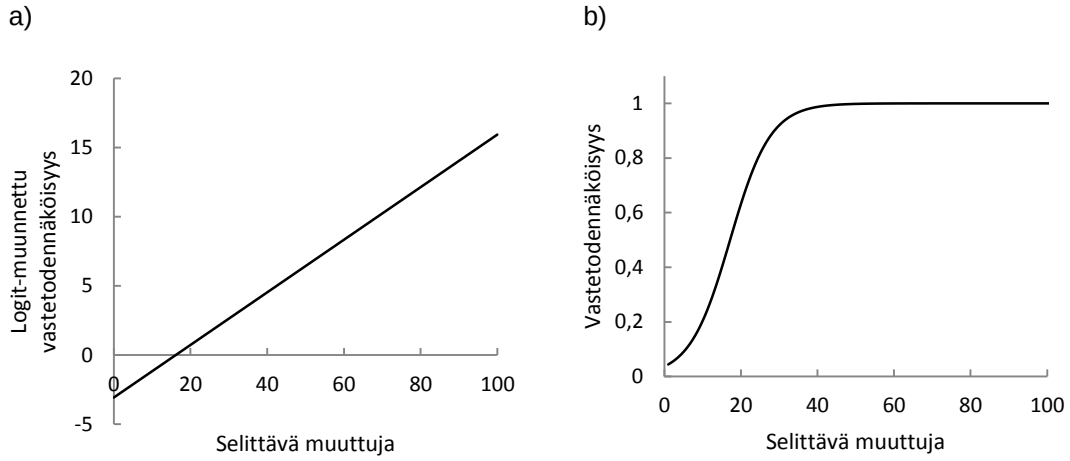
1) Kullakin selittäjän arvoparilla x ja $x + 1$ (esim. 1 ja 2, 2 ja 3, 3 ja 4 jne.) lasketaan vastetodennäköisyydet $p(x)$ ja $p(x + 1)$ sekä näitä vastaavat vedot $v[p(x)]$ ja $v[p(x + 1)]$.
2) Kullekin vetoparille $v[p(x)]$ ja $v[p(x + 1)]$ lasketaan vetosuhde $VS\ [p(x + 1),\ p(x)] = v[p(x + 1)]\ /\ v[p(x)]$.

3) Kaikista näin saaduista vetosuhteista lasketaan keskiarvo, joka on tarkasteltavan selittäjän regressiokerroin vetosuhdemuodossa ilmaistuna.

Voidaan osoittaa (ks. Rita & Komonen 2008: 69), että keskiarvona lasketun vetosuhteen VS ja tarkasteltavan selittäjän regressiokertoimen b välillä on yhteys

$$VS = e^b \ . \qquad (5)$$

Tästä nähdään, että vetosuhde voi saada minkä tahansa positiivisen arvon. Jos b on positiivinen, $VS > 1$ ja selittäjän arvon kasvaminen kasvattaa vastetodennäköisyyttä. Jos taas b on negatiivinen, pätee $0 < VS < 1$ ja selittäjän arvon kasvaminen pienentää vastetodennäköisyyttä. Rita (2004) sekä Rita & Komonen (2008) suosittelevat, että ykköstä pienemmät vetosuhteen arvot ilmoitetaan käänteislukuina; esimerkiksi 0,5 tulisi siten ilmoittaa muodossa 2^{-1}. Tämä käytäntö helpottaa erisuuruisten (ykköstä pienempien ja suurempien) vetosuhteiden suuruusluokan vertailua säilyttäen kuitenkin samalla tiedon vaikutuksen suunnasta.



Kuva 3.1. Yhden selittäjän logistinen regressiomalli a) lineaarisessa ja b) sigmoidisessa muodossaan (kuvitteellinen aineisto). Tässä regressiovakio a = -3,07 ja selittäjän kerroin b = 0,19. Regressiokerrointa vastaava vetosuhde VS = 1,21.

3.2. Logistinen regressio habitaattimallinnuksessa

Kun lajien esiintymistä suhteessa ympäristötekijöihin mallinnetaan logistisella regressiolla, vasteena ovat havainnot tarkasteltavan lajin esiintymisestä ja puuttumisesta eri puolilla maisemaa tai laajempaa maantieteellistä aluetta (vrt. Elith & Leathwick 2009a). Logistista

habitaattimallia ei siis voida luoda pelkkiä läsnäolohavaintoja sisältävästä aineistosta. Suurin osa havaintoaineistoista (mm. kasvitieteelliset museokokoelmat) kuitenkin käsittää pelkkiä läsnäolohavaintoja, jolloin puuttumishavainnot on mahdollista korvata ottamalla satunnaisotanta ympäristötekijöistä tarkastelualueella ja käsittelemällä otannan kattamia paikkoja mallissa havaintoina lajin puuttumisesta (engl. *pseudo-absence*) (Elith & Burgman 2003, Elith & Leathwick 2007, Gibson ym. 2007, Elith & Leathwick 2009a).

Elith & Leathwick (2009a) painottavat, että valittaessa habitaattimallinnuksessa käytettävää menetelmää on syytä kiinnittää huomiota siihen, kuinka helposti ja nopeasti eri menetelmät ovat opittavissa ja voidaanko niitä soveltaa myös muissa yhteyksissä. Logistinen regressio on yleisesti tunnettu tilastollinen menetelmä, joka sisältyy useimpiin tilasto-ohjelmistoihin. Lisäksi logistiset regressiomallit kuuluvat yleistettyihin lineaarisiin malleihin (*generalized linear models*; ks. McCullagh & Nelder 1989), jotka muodostavat ekologisessa tutkimuksessa erittäin monikäyttöisen työkalupakin, koska vasteen muoto (ns. linkkifunktio) ja virhevaihtelun jakauma voidaan valita niissä aineistolle sopivaksi. Kaiken kaikkiaan logistinen regressio on suhteellisen helposti opittava ja joustava habitaattimallinnuksen menetelmä, joka lisäksi toimii porttina koko yleistettyjen lineaaristen mallien patteristoon.

4. LOGISTISEN HABITAATTIMALLIN SELITYSKYVYN ARVIOIMINEN

Malli on aina yksinkertaistus todellisuudesta, eikä tilastollinen malli vastaa koskaan täydellisesti kaikkia perusjoukon mahdollistamia havaintoja. Siksi mallin selityskykyä – sitä kuinka hyvin malli vastaa tehtyjä havaintoja – on syytä arvioida. Ennustustarkoitukseen luodun habitaattimallin tapauksessa kiinnostavaa on erityisesti mallin yleistettävyys eli se, kuinka hyvin malli ennustaa lajin esiintymisen uusilla alueilla. Mallin selityskyky sen sovittamiseen käytetyssä aineistossa on vähemmän oleellista ja voi lähinnä paljastaa ongelmia joko mallin rakenteessa tai itse aineistossa (Elith & Leathwick 2009a).

4.1. ROC-käyrä

Binäärisenä luokittelijana käytettävän mallin erottelukyky kertoo, kuinka hyvin malli kykenee erottelemaan positiiviset ja negatiiviset havainnot (habitaattimallien tapauksessa havainnot lajin läsnäolosta ja puuttumisesta eri puolilla maisemaa) toisistaan. Jos mallin antamat ennusteet vastaavat havaintoja, mallin erottelu- ja siten myös selityskyky on hyvä. Mallin erot-

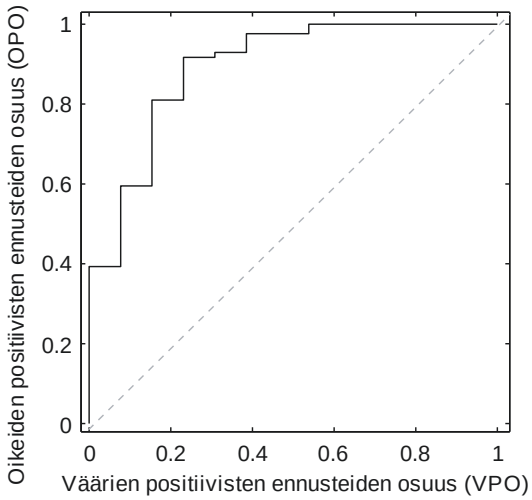
telukykyä voidaan arvioida ROC-käyrän (*Receiver Operator Characteristic*; ks. esim. Fawcett 2006) avulla.

Usein ennusteiden luokittelun ongelmana on vaihtokauppa: mitä useampi positiivinen tapahtuma luokitellaan oikein eli positiiviseksi, sitä useampi negatiivinen tapahtuma luokitellaan myös positiiviseksi eli väärin. Optimaalista luokittelua voidaan etsiä tarkastelemalla mallin erottelukykyä eri kynnysarvoilla. Kynnysarvo on ennalta määrätty arvo, jonka perusteella mallin estimoima todennäköisyys positiiviselle tapahtumalle tulkitaan positiiviseksi tai negatiiviseksi ennusteeksi. Käytännössä aineisto mahdollistaa niin monta erilaista kynnysarvoa kuin malli estimoi erisuuria positiivisen tapahtuman todennäköisyyksiä kaikille aineiston havaintoyksiköille.

ROC-käyrän muodostamiseksi määritetään kullakin aineiston mahdollistamalla kynnysarvolla oikeiden positiivisten ennusteiden osuus positiivisista havainnoista (OPO) ja väärien positiivisten ennusteiden osuus negatiivisista havainnoista (VPO). Saadut osuudet piirretään hajontakuvana (OPO pystyakselilla, VPO vaaka-akselilla), jossa kutakin kynnysarvoa vastaa yksi piste. Kun pisteet yhdistetään janoilla, muodostuu ROC-käyrä (kuva 4.1). Mitä lähempänä käyrän piste on kuvaajan vasenta yläkulmaa (iso OPO, pieni VPO) sitä parempi mallin erottelukyky tällä kynnysarvolla on. Käytännössä optimaalisen kynnysarvon valintaan vaikuttaa kuitenkin myös luokittelun tavoite: lääketieteellisellä testillä on tärkeää havaita kaikki positiiviset tapahtumat (taudin kantajat), mutta luonnonsuojelualueiden valinnassa tärkeämpi tavoite on rajata negatiiviset tapahtumat (soveltumattomat alueet) pois potentiaalisten alueiden joukosta.

Mallin yleistä erottelukykyä voidaan kuvata ROC-käyrän alaisella pinta-alalla (KAP). Mitä suurempi pinta-ala, sitä erottelukykyisempi malli kokonaisuudessaan on. Täsmällisemmin sanottuna KAP kertoo todennäköisyyden sille, että malli liittää satunnaisesti valittuun positiiviseen havaintoon positiivisemman ennusteen kuin satunnaisesti valittuun negatiiviseen havaintoon (Fawcett 2006). Jos KAP = 0,5, eli käyrä on lävistäjäsuora, mallin erottelukyky on yhtä huono kuin satunnaisen arvauksen. Jos taas KAP = 1, malli erottelee positiiviset ja negatiiviset havainnot toisistaan täydellisesti.

Malli on optimoitu sen sovittamisessa käytetylle aineistolle. Siksi ROC-käyrä on suositeltavaa muodostaa testaamalla mallia erilliseen havaintojoukkoon, ns. testijoukkoon, jota ei ole käytetty lainkaan mallin sovittamisessa (Beutel ym. 1999, Elith & Leathwick 2009a).



Kuva 4.1. Esimerkki binäärisenä luokittelijana toimivan mallin ROC-käyrästä (kuvitteellinen aineisto). Jokainen käyrän kulmapiste vastaa yhtä luokittelun kynnysarvoa. Mallin erottelukykyä voidaan kuvata ROC-käyrän alaisella pinta-alalla (KAP), jonka suuri arvo tarkoittaa hyvää erottelukykyä. Tässä KAP = 0,89; mallin erottelukyky on siis hyvä.

4.2. Ristiinvalidointi

Jos aineisto on hyvin pieni tai positiivisten ja negatiivisten havaintojen jakauma aineistossa on huomattavan vino, aineistoa ei välttämättä ole mahdollista jakaa erillisiksi sovitus- ja testijoukoksi. Tällöin mallin selityskyvyn arviointiin voidaan käyttää esimerkiksi ristiinvalidointia (Elith & Leathwick 2009a). Ristiinvalidointi on uudelleenotantamenetelmä, jossa aineisto jaetaan muutamaan (yleensä viiteen tai kymmeneen) pienempään, toisensa poissulkevaan joukkoon. Kukin näistä joukoista poistetaan vuorollaan, malli estimoidaan uudelleen jäljelle jääneiden havaintojen perusteella ja lopuksi mallin selityskykyä arvioidaan vertailemalla sen ennusteita poisjätettyihin havaintoihin. Mallin keskimääräinen selityskyky saadaan kaikkien toistojen keskiarvona.

5. ESIMERKKI 1: MIKÄ RAJOITTAA TAMMEN LEVINNEISYYTTÄ?

Pro gradu –työssäni (H. Kettunen, julkaisematon) tutkin, mikä rajoittaa tammen alueellista levinneisyyttä pienessä Wattkastin saaressa (60°11´N, 21°37´E) Länsi-Turunmaalla. Työni pohjana oli Wattkastissa tehty tammen istutuskoe, jonka avulla tarkastelin tammen menestymiseen – siis selviytymiseen ja kuntoon – vaikuttavia tekijöitä. Voidakseni selvittää, vaikuttavatko tammen luontaiseen esiintymiseen samat tekijät kuin istutettujen puiden menestymiseen, käytin työssäni lisäksi toista aineistoa – kartoitusta tammen luontaisesta esiinty-

misestä Wattkastissa. Loin sekä tämän havaintoaineiston että istutuskokeen pohjalta habitaattimallin tammen nykyiselle levinneisyydelle saaressa.

Alkuperäiset hypoteesini olivat,

1) että tammen levinneisyyttä rajoittaa sille soveltuvien kasvupaikkojen tilajakauma, jolloin habitaattimallin selityskyky on oletettavasti hyvä,
   TAI
2) että tammen levinneisyyttä rajoittaa sen heikko leviämiskyky suhteessa paikalliseen häviämisnopeuteen, jolloin habitaattimallin selityskyky on oletettavasti alhaisempi.

Tässä osassa tarkastelen habitaattimallinnuksen roolia kyseisessä tammitutkimuksessa.

5.1. Aineisto

Wattkastissa kasvaa yli 1800 vähintään 50 cm pitkää tammea, joiden sijainnit on kartoitettu muutaman metrin tarkkuudella vuosina 2003–2004 (ks. Gripenberg & Roslin 2005). Osana Wattkastissa toteutettavia hyönteistutkimuksia eri puolille saarta istutettiin vuoden 2004 toukokuussa 104 pientä tammea, jotka muodostivat tutkimukseni kokeellisen aineiston.

Kartoitin syksyllä 2009 istutettujen tammien selviytymisen ja kunnon sekä muutamia puun kasvupaikan ympäristöoloja kuvaavia tunnuksia (mm. topografia, avoimuus, latvuksen paikallinen aukkoisuus, maaperän kivisyys, humuskerroksen pH, aluskasvillisuuden keskikorkeus ja nisäkäslaidunnuksen vaikutus puuhun), joiden yhteyttä puun menestymiseen ja tammen luontaiseen esiintymiseen myöhemmin tutkin. Onnistuin paikantamaan istutetuista tammista kaikki paitsi kuusi. Habitaattimallia varten jaoin istutettujen puiden kasvupaikat niihin, joissa kasvaa luonnonvarainen tammi, ja niihin, jotka eivät ole tammen nykyisiä esiintymispaikkoja. Käytin ehtona kasvupaikan luontaisuudelle 25 m:n maksimietäisyyttä lähimpään luonnonvaraiseen tammeen, jolloin 19 kasvupaikkaa tulkittiin luontaisiksi ja 79 paikkaa ei-luontaisiksi.

5.2. Habitaattimalli tammen esiintymiselle Wattkastissa

Mallinsin tammen nykylevinneisyyttä suhteessa ympäristötekijöihin yleistetyllä lineaarisella mallilla. Koska vastemuuttuja (kasvupaikan luontaisuus) oli binäärinen – paikka joko oli tai ei ollut tammen luontainen kasvupaikka – oletin vasteen virhevaihtelun noudattavan binomi-

jakaumaa ja käytin logistista linkkifunktiota. Näin muodostettu yleistetty lineaarinen malli on logistinen regressiomalli (vrt. Collett 2003: 58).

Koska tarkoitukseni oli selvittää, voidaanko tammen luontainen levinneisyys selittää samoilla tekijöillä kuin istutettujen puiden menestyminen, käytin mallissa selittäjinä kasvupaikan avoimuutta, maaperän pH:ta, aluskasvillisuuden keskikorkeutta sekä istutetusta tammesta määritettyä laidunnusastetta, joissa oli graafisessa tarkastelussa havaittu viitteitä yhteydestä istutetun puun selviytymiseen. Testasin selittäjien merkitsevyyden tyypin 3 analyysilla, joka huomioi selittäjien järjestyksen mallissa (ks. SAS Institute Inc. 2004: 1615). Lopulliseen habitaattimalliin valitsin selittäjiksi kasvupaikan avoimuuden ja aiempien vuosien laidunnusasteen, jotka olivat ensimmäisessä mallissa ainoat tilastollisesti merkitsevät selittäjät. Lopullisessa mallissa selittäjien vaikutus vasteeseen estimoitiin uudelleen. Sovitin habitaattimallin SAS/STAT-ohjelmiston (versio 9.1.3; SAS Institute Inc., Cary, Pohjois-Carolina, Yhdysvallat) GENMOD-proseduurilla.

Habitaattimallin estimoimat yhteydet olivat odotusten vastaisia: kasvupaikan sulkeutuneisuus ja korkea istutettuihin puihin kohdistunut laidunnusaste lisäsivät mallin mukaan tammen luontaisen esiintymisen todennäköisyyttä (taulukko 5.1). Istutetuilla puilla kasvupaikan sulkeutuneisuus (mitattu latvuksen aukkoisuutena) kuitenkin pienensi puun selviytymistodennäköisyyttä, mikä on biologisesti mielekkäämpi vaikutussuunta (vrt. Drakenberg ym. 1991: 27). Myös laidunnuksella on yleensä negatiivinen vaikutus tammen kasvuun (Rainio 1986).

Taulukko 5.1. Lopullinen habitaattimalli Wattkastin tammille (n = 94). Mallin devianssi = 82,7; df = 91.

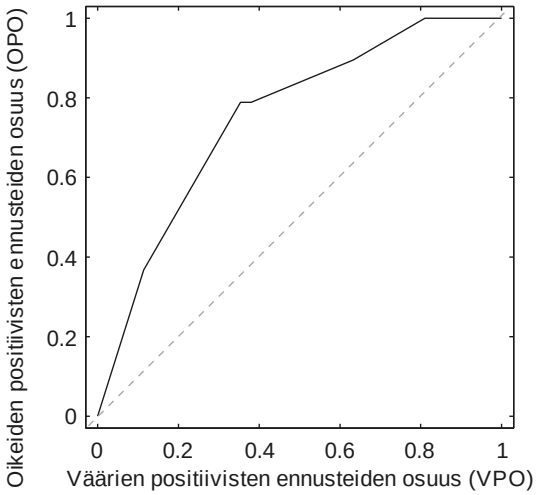
Selittäjä	Parametrit			Merkitsevyys		
	Kerroin	Keskivirhe	Vetosuhde	χ^2	df	p
(Vakio)	-2,86	1,51	-	-	-	-
Avoimuus	-0,96	0,38	1,46	7,35	1	0,007
Laidunnusaste (aiemmat vuodet)	1,38	0,77	2,17	4,29	1	0,038

5.3. Habitaattimallin selityskyvyn arvioiminen

Arvioin habitaattimallin selityskykyä määrittämällä mallille ROC-käyrän alaisen pinta-alan (KAP) sekä yleistetyn selitysasteen (ks. Nagelkerke 1991). Yleistetty selitysaste (R^2) kuvaa

sitä osuutta vastemuuttujan vaihtelusta, jonka malli selittää. Habitaattimallin selityskyky oli heikko: käyrän alainen pinta-ala KAP = 0,75 (kuva 5.1) ja selitysaste R^2 = 0,19.

On tärkeää huomata, että arvioin habitaattimallin selityskykyä samasta aineistosta kuin olin käyttänyt mallin sovittamisessa, joten estimaatit mallin selityskyvystä ovat todennäköisesti liioiteltuja (vrt. Elith & Leathwick 2009a). Käytin samaa aineistoa sekä mallin sovittamiseen että arviointiin, koska aineistoni oli sen verran pieni ja vastemuuttujan (kasvupaikan luontaisuus) arvojakauma niin epätasainen, etten pystynyt jakamaan aineistoa erillisiksi sovitus- ja testijoukoiksi. Koska käytin habitaattimallia ainoastaan tammen nykyisen levinneisyyden tutkimiseen rajatulla alueella – en siis lainkaan ekstrapoloimiseen tilassa tai ajassa – ja mallin tarkkuus oli joka tapauksessa heikko, on menettelytapa jokseenkin perusteltu. Toinen mahdollisuus olisi ollut mallin selityskyvyn arvioiminen ristiinvalidoimalla.



Kuva 5.1. Mallin erottelukykyä kuvaava ROC-käyrä (*Receiver Operator Characteristic*) habitaattimallille. Käyrän alainen pinta-ala KAP = 0,75 (SE = 0,06), eli habitaattimallin erottelukyky on alhainen. Käyrä on tuotettu SPSS Statistics -ohjelmalla (versio 17.0; SPSS Inc., Chicago, Illinois, Yhdysvallat).

5.4. Tammen esiintymiskuvio ei ole tasapainoinen

Habitaattimalli selitti heikosti tammen nykylevinneisyyden Wattkastissa ja lisäksi estimoidut yhteydet tammen esiintymisen ja ympäristöolojen välillä olivat biologisesti kyseenalaisia. Yhdessä istutettujen puiden hyvän keskimääräisen selviytymisen kanssa tämä osoittaa, että tammen alueellista levinneisyyttä Wattkastissa rajoittaa tammen leviämiskyky, joka on heikko suhteessa lajin paikalliseen häviämisnopeuteen.

Habitaattimallin huono selityskyky voi johtua joko mallin taustaoletuksen toteutumattomuudesta – siitä että lajin esiintymiskuvio ei vastaa lajille soveltuvien elinympäristöjen esiintymistä – tai toisaalta väärin valituista selittävistä ympäristötekijöistä (vrt. Elith & Leathwick 2009a). Käyttämällä habitaattimallin rinnalla kokeellista aineistoa (istutuskoe) pystyin osoittamaan, että Wattkastin tammilla huonon selityskyvyn syy on nimenomaan esiintymiskuvion ja elinympäristöjen tilajakauman välinen epätasapaino. Tähän viittaa erityisesti se, että tammi menestyi istutettuna myös sellaisissa paikoissa, joissa se ei nykyisin esiinny. Maisemassa on siis selvästi jonkin verran sellaisia kasvupaikkoja, jotka kyllä soveltuisivat tammelle mutta ovat tällä hetkellä kuitenkin asuttamattomia.

Syy siihen, että ympäristötekijät vaikuttivat eri tavoin istutettuihin ja luonnonvaraisiin puihin, on todennäköisesti istutuskokeen koeasetelmassa. Koepuut olivat istutushetkellä ohittaneet jo taimivaiheen, joten istutuskokeen perusteella on mahdotonta arvioida kasvupaikan ympäristöolojen vaikutusta esimerkiksi tammen itämiseen tai taimivaiheen kasvuun. Yleisesti habitaattimalli huomioi istutuskoetta paremmin ympäristötekijöiden vaikutuksen puun koko elämänkaareen. Tässä tapauksessa istutuskoe antaa kuitenkin luotettavamman kuvan tammen menestymisestä erilaisilla kasvupaikoilla, koska habitaattimallin taustaoletuksesta on poikettu huomattavasti. Puiden kohdalla on lisäksi muistettava, että vanhan puun kasvupaikalla tällä hetkellä vallitsevat ympäristöolot eivät välttämättä ole samanlaiset kuin puun ollessa nuori, mikä myös saattoi heikentää tammelle luodun habitaattimallin selityskykyä.

Kaiken kaikkiaan tutkimukseni osoitti, miten tärkeää lajien elinympäristövaatimuksia mallinnettaessa on selvittää, vastaako lajin esiintyminen sille soveltuvien elinympäristöjen tilajakaumaa maisemassa. Jos lajin esiintyminen kuvastaa dynaamisia häviämis- ja asuttamisprosesseja, jolloin maisemassa on lajille soveltuvia mutta asuttamattomia elinympäristöjä, ympäristöoloihin pohjautuva malli liioittelee lajin levinneisyyttä. Tutkimukseni perusteella havainnoivan ja kokeellisen tutkimuksen yhdistäminen on tällaisessa tapauksessa tehokas keino lajin esiintymisen taustalla olevien prosessien paljastamiseen.

6. ESIMERKKI 2: NELJÄN AUSTRALIALAISEN KASVILAJIN ESIINTYMISEN ENNUSTAMINEN SUOJELUTARKOITUKSESSA

Suojelualueiden rajaamiseen tarvitaan tarkkoja tietoja lajien levinneisyydestä, mutta usein tietojen alueellisessa kattavuudessa on puutteita. Puutteita voidaan paikata ennustamalla lajien potentiaalista esiintymistä habitaattimallinnuksen keinoin. Kun logistista habitaattimallia

käytetään apuna suojelualueiden rajauksessa, mallin antamat esiintymistodennäköisyydet täytyy ensin luokitella binäärisiksi ennusteiksi lajin esiintymisestä ja puuttumisesta eri tarkastelupaikoilla. Tähän on useita menetelmiä. Wilson ym. (2005) tutkivat suojelualue-verkoston suunnittelun lopputuloksen riippuvuutta esiintymistodennäköisyyksien luokitteluun käytetystä menetelmästä, kun alueiden rajaus tehtiin neljälle kasvilajille luodun habitaattimallin avulla. Tässä osassa havainnollistan Wilsonin ym. tutkimuksen avulla erilaisia tapoja muuntaa logistisen habitaattimallin estimoimat todennäköisyydet positiivisiksi ja negatiivisiksi ennusteiksi lajin esiintymisestä.

6.1. Aineisto

Tutkimus toteutettiin Australiassa Victorian osavaltiossa, alueella joka käsittää yli 5 000 luontaisen kasvillisuuden ripettä (yhteispeittävyys yli 300 000 ha). Tarkasteltavat kasvilajit olivat *Eucalyptus tricarpa*, *Pultenaea largiflorens*, *Hibbertia exutacies* ja *Acacia ausfeldii*, joista kolme ensimmäistä on yleisiä ja viimeinen on uhanalainen laji. Lajien esiintymisestä ei ollut saatavissa havaintotietoja koko alueen laajuudelta, joten niiden esiintymiselle luotiin ympäristötekijöihin pohjautuva logistinen habitaattimalli, jonka avulla havaintotiedot ekstrapoloitiin myös katvealueille. Havaintojen tarkkuuden ja kasvillisuuden vaihtelun perusteella esiintymisennusteet kohdistettiin yhden hehtaarin kokoisille karttaruuduille. Habitaattimallia ja siinä käytettyjä selittäjiä ei ole kuvattu artikkelissa tarkemmin.

6.2. Todennäköisyyksien luokittelu binäärisiksi ennusteiksi

Habitaattimallin perusteella saatujen esiintymistodennäköisyyksien luokitteluun vaikuttaa se, miten ennusteita on tarkoitus hyödyntää. Tutkimuksessa suojelualueverkoston luomista lähestyttiin ns. *minimum-set*-ongelmana: tavoitteena oli minimoida suojeltava pinta-ala siten, että suojelulliset tavoitteet (määritelty kullekin lajille erikseen) kuitenkin täyttyvät. Tämä tavoite on otettava huomioon mallilla estimoitujen todennäköisyyksien binäärisessä luokittelussa. Todennäköisyyksien muuntamista esiintymisennusteiksi tutkittiin sekä käyttämällä kiinteitä kynnysarvoja että suoria esiintymistodennäköisyyksiä.

6.2.1. Kynnysarvomenetelmät

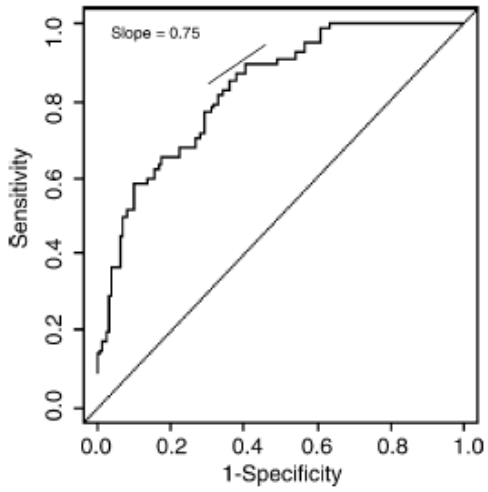
Tutkimuksessa vertailtiin kolmea erilaista kynnysarvon käyttämiseen perustuvaa lähestymistapaa. Kynnysarvo on se rajatodennäköisyys, jonka perusteella lajin esiintymiselle estimoidut todennäköisyydet luokitellaan esiintymis- tai puuttumisennusteiksi. Tutkimuksessa käytetyt lähestymistavat olivat:

1) **Etukäteen valittu kiinteä kynnysarvo 0,5.** Kaikki habitaattimallista saadut tätä suuremmat todennäköisyydet tulkittiin ennusteiksi lajin läsnäolosta ja pienemmät ennusteiksi puuttumisesta. Suurin tähän lähestymistapaan liittyvä ongelma on, että läsnäolo- ja puuttumishavaintojen jakauma mallin sovitusaineistossa vaikuttaa estimoitujen todennäköisyyksien suuruuteen. Jos puuttumishavaintoja on huomattavasti enemmän kuin läsnäolohavaintoja, estimoidut todennäköisyydet painottuvat kohti nollaa ja vaarana on, että kaikki todennäköisyydet tulkitaan ennusteiksi lajin puuttumisesta.

2) **Jälkikäteen määrätty kynnysarvo, joka on valittu ennusteiden halutun sensitiivisyyden perusteella.** Ennusteiden sensitiivisyys tarkoittaa oikein luokiteltujen esiintymisennusteiden osuutta havainnoista (= ROC-käyrän OPO). Suuren sensitiivisyyden kääntöpuolena on, että ennusteiden spesifisyys (oikein luokiteltujen puuttumisennusteiden osuus havainnoista) tyypillisesti laskee. Suojelualueiden rajaamisessa tavoitteena on ennen kaikkea soveltumattomien alueiden poissulkeminen, joten kynnysarvo täytyy valita siten, että ennusteiden sensitiivisyys on matala ja spesifisyys korkea. Tutkimuksessa sensitiivisyydeksi määrättiin 10 % ja spesifisyydeksi 90 %, jolloin lajikohtaiset kynnysarvot olivat 0,60 (*Acacia ausfeldii*), 0,62 (*Eucalyptus tricarpa*), 0,77 (*Hibbertia exutacies*) ja 0,72 (*Pultenaea largiflorens*).

3) **Jälkikäteen määrätty kynnysarvo, jonka määrittämisessä on huomioitu virheellisiin ennusteisiin liittyvät kustannukset sekä lajien prevalenssi eli vallitsevuus (läsnäolohavaintojen osuus havaintoaineistossa).** Vääriin positiivisiin ennusteisiin liittyy ylimääräisiä kustannuksia, koska myös lajille soveltumattomia alueita suojellaan. Vastaavasti vääriin negatiivisiin ennusteisiin liittyy ylimääräisiä kustannuksia, koska osaa niistä alueista, joilla laji esiintyy, ei suojella. Suojelualueita rajattaessa toivottava tilanne on, että vääriin positiivisiin ennusteisiin liittyvät kustannukset ovat suuremmat kuin vääriin negatiivisiin ennusteisiin liittyvät kustannukset. Tällä varmistetaan, että suojelun piiriin otetaan vain lajin erittäin todennäköiset esiintymisalueet. Halutun kustannussuhteen mahdollistavan kynnysarvon määrittämisessä tulee lisäksi

huomioida lajin vallitsevuus havaintoaineistossa. Jos vallitsevuus on pieni, kynnysarvon on oltava alhainen, koska esiintymiselle estimoidut todennäköisyydet ovat keskimäärin alhaisia (vrt. kynnysarvomenetelmä 1). Sopiva kynnysarvo saadaan selville ROC-käyrän avulla (kuva 6.1). Tällä menetelmällä määritetyt kynnysarvot tarkastelulajeille olivat 0,62 (*Acacia ausfeldii*), 0,64 (*Eucalyptus tricarpa*), 0,69 (*Hibbertia exutacies*) ja 0,71 (*Pultenaea largiflorens*). *Acacia*-lajilla on pienin kynnysarvo, koska sillä oli lajeista alhaisin vallitsevuus.



Kuva 6.1. Kynnysarvon määrittäminen ROC-käyrän avulla (kynnysarvomenetelmä 3). Käyrän vieressä oleva suora (vain osa piirretty) kuvaa väärien positiivisten (VPK) ja väärien negatiivisten ennusteiden kustannuksien (VNK) suhdetta, kun lajin vallitsevuus (v) on huomioitu. Suoran kulmakerroin saadaan kaavalla $VPK\ /\ VNK \times [(1-v)\ /\ v]$. Kun suoraa liikutetaan kuvan vasemmasta ylänurkasta alaspäin, se kohta, jossa suora ensiksi koskettaa ROC-käyrää, määrää optimaalisen kynnysarvon (arvo ei luettavissa suoraan kuvasta). Sensitiivisyys (pystyakselilla) tarkoittaa oikeiden positiivisten ennusteiden osuutta positiivisista havainnoista. 1 - spesifisyys (vaaka-akselilla) on väärien positiivisten ennusteiden osuus. Huom., että suhteen $VPK\ /\ VNK$ arvo on valittu tässä mielivaltaisesti. Kuva: Wilson ym. (2005).

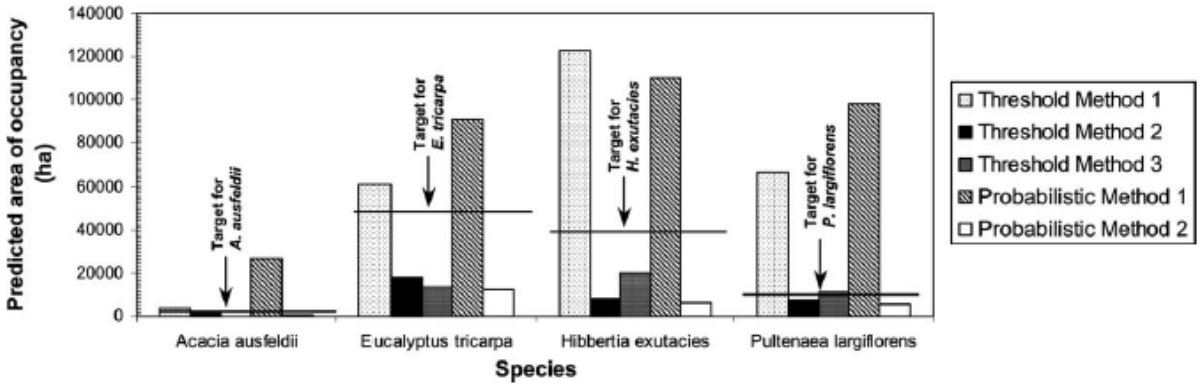
6.2.2. Esiintymistodennäköisyyttä suoraan hyödyntävät menetelmät

Suojelualueiden suunnittelussa voidaan myös hyödyntää habitaattimallilla estimoituja esiintymistodennäköisyyksiä sellaisinaan. Artikkelissa vertailtiin kahta suoriin todennäköisyyksiin perustuvaa menetelmää (**todennäköisyysmenetelmät 1 ja 2**). Molemmissa menetelmissä laskettiin odotusarvot lajien esiintymispinta-alalle (hehtaareina) kullakin potentiaalisella suojelualueella. Odotusarvot saatiin laskemalla yhteen esiintymistodennäköisyydet yksittäisillä hehtaariruuduilla. Tavoitteena oli löytää sellainen suojelualueverkosto, joka minimoi suojeltavan kokonaispinta-alan niin, että ennalta määritellyt lajikohtaiset suojelutavoitteet kui-

tenkin täyttyvät. Menetelmässä 1 huomioitiin kaikki todennäköisyydet, menetelmässä 2 ainoastaan kynnysarvomenetelmällä 2 määritettyä kynnysarvoa suuremmat todennäköisyydet.

6.3. Luokittelutapa vaikuttaa suojeluratkaisuihin

Eri menetelmillä arvioidut lajikohtaiset esiintymispinta-alaennusteet erosivat huomattavasti toisistaan, ja muutamalla menetelmällä ei pystytty saavuttamaan lajikohtaisia suojelutavoitteita (kuva 6.2). Kynnysarvomenetelmällä 1 arvioidut esiintymispinta-alat olivat muita kynnysarvomenetelmiä korkeampia, koska sillä määritetyt lajikohtaiset kynnysarvot olivat matalia ja siten suurempi osuus hehtaariruuduista tulkittiin lajeille sopiviksi kuin muilla menetelmillä. Vastaavasti todennäköisyysmenetelmällä 1 saatiin suurempi esiintymispinta-ala-arvio kuin todennäköisyysmenetelmällä 2, koska menetelmässä 2 vain osa todennäköisyyksistä tulkittiin suotuisiksi lajien esiintymiselle. Eri menetelmiin pohjautuvissa suojelualueverkostoratkaisuissa oli huomattavia eroja.



Kuva 6.2. Eri menetelmillä ennustetut kokonaisesiintymispinta-alat tarkastelulajeille. Lajikohtaiset suojelutavoitteet on esitetty poikkiviivoilla. Osalla menetelmistä ei saavutettu suojelutavoitteita, koska ennuste esiintymisen kokonaispinta-alalle oli alhaisempi kuin suojelutavoite. *Treshold method* on käännetty tekstissä kynnysarvomenetelmäksi ja *probabilistic method* todennäköisyysmenetelmäksi. Kuva: Wilson ym. (2005).

Suojelualueverkostoja suunniteltaessa pyritään suojelemaan ainoastaan tarkastelulajeille erittäin todennäköisesti soveltuvat alueet, koska suojeluun käytettävissä olevat resurssit (määrärahat, maapinta-ala) ovat poikkeuksetta niukat. Habitaattimallin esiintymisennusteita tulkittaessa tämä tarkoittaa sitä, että lajin esiintymiselle pyritään valitsemaan mieluummin korkea kuin matala kynnysarvo – vain erittäin korkeiden todennäköisyyksien oletetaan siis olevan merkki lajin läsnäolosta. Siten kynnysarvomenetelmä 1 ja todennäköisyysmenetelmä 1, joissa luokitteluperusteet olivat väljät, soveltuvat heikosti suojelusovelluksiin.

Toisaalta myös havaintoaineiston koostumus on huomioitava estimoitujen esiintymistodennäköisyyksien tulkinnassa. Jos läsnäolohavaintoja on puuttumishavaintoihin nähden niukasti, korkea kynnysarvo voi johtaa esiintymisen aliarviointiin ja monia lajien oikeita esiintymisalueita jää todennäköisesti suojelematta. Kuvasta 6.2 ei kuitenkaan voida päätellä suoraan, aiheuttiko tämä eräiden menetelmien tapauksessa esiintymisarvion jäämisen alle suojelutavoitteen. Kaiken kaikkiaan Wilsonin ym. tutkimus osoittaa, että suojelualueita habitaattimallinnuksen avulla rajattaessa tulee kiinnittää huomiota menetelmään, jolla mallin estimoimat todennäköisyydet muunnetaan esiintymisennusteiksi.

7. LOPUKSI

Habitaatti- eli elinympäristömallit ovat tilastollisia malleja, jotka kuvaavat lajien esiintymistä suhteessa ympäristötekijöihin. Habitaattimallien avulla voidaan tutkia lajien esiintymiseen vaikuttavia tekijöitä ja luoda ennusteita lajien levinneisyydelle. Habitaattimallin luomiseksi tarvitaan havaintoja lajin esiintymisestä erilaisissa ympäristöissä. Jos käytettävissä on sekä läsnäolo- että puuttumishavaintoja, lajin esiintymistä voidaan mallintaa logistisella regressiolla.

Habitaattimallien ennusteet eivät vastaa täydellisesti todellisuutta. Siksi habitaattimallin selityskykyä eli mallin ennusteiden ja lajin todellisen esiintymisen vastaavuutta on syytä arvioida. Erityisen kiinnostavaa on mallin yleistettävyys eli se, kuinka hyvin habitaattimalli ennustaa lajin esiintymisen uusilla alueilla. Logistisen habitaattimallin selityskykyä voidaan arvioida erillisestä testiaineistosta ROC-käyrän alaisella pinta-alalla tai suoraan sovitusaineistosta ristiinvalidoimalla.

Habitaattimallien taustalla on oletus, että lajien esiintyminen kuvastaa niille sopivien elinympäristöjen esiintymistä. Kasvupaikan ympäristöoloihin perustuva levinneisyysmalli luo tarkan esiintymisennusteen ainoastaan tilanteessa, jossa elinympäristöjen tilajakauma määrää suoraviivaisesti lajin esiintymiskuvion eli esiintyminen on tasapainossa ympäristön kanssa. Ensimmäisen esimerkkitutkimus (Wattkastin tammet) kuvaa hyvin tasapaino-oletuksesta poikkeamisen seurauksia: lajin esiintymiseen vaikuttavista ympäristötekijöistä ei voida tällaisessa tapauksessa tehdä habitaattimallin perusteella luotettavia johtopäätöksiä.

Habitaattimallit ovat hyödyllinen apuväline mm. suojelualueverkostojen suunnittelussa. Lajihavaintoaineistoissa on yleensä alueellisia aukkoja, mutta habitaattimallinnuksella on mahdollista arvioida lajien esiintymistä myös katvealueilla. Kuten toinen esimerkkitutkimus

(habitaattimalli neljälle australialaiselle kasvilajille) osoitti, esiintymisennusteisiin liittyy aina epävarmuutta, joka on huomioitava habitaattimallin ennusteita tulkittaessa. Mallilla estimoitujen todennäköisyyksien esiintymisennusteiksi muuntamiseen on useita eri tapoja, ja muuntamistapaa valitessa tulee kiinnittää huomiota siihen, millaisia vääriin esiintymisennusteisiin liittyviä riskejä ollaan valmiit ottamaan. Suojelualueiden valinnassa halutaan yleensä ennen kaikkea minimoida riski, joka liittyy lajille soveltumattomien alueiden suojelemiseen.

Lajien esiintymistä mallinnettaessa täytyy myös muistaa, että esiintymismalli on korkeintaan yhtä mielekäs kuin sen tausta-aineisto. Hyvän esimerkin tästä tarjoavat Lozier ym. (2009), jotka selvittivät pohjoisamerikkalaiseen kansantarustoon kuuluvan apinamiehen, isojalan, levinneisyyttä. Kun raportoidut isojalkahavainnot suhteutettiin pieneen joukkoon bioklimaattisia ympäristötekijöitä, isojalalle pystyttiin luomaan tarkka ja täysin uskottava levinneisyyskartta – joka vastasi melko hyvin mustakarhun levinneisyysaluetta. Lozier ym. korostavatkin, että esiintymismallit pitäisi perustaa vain ja ainoastaan luotettaviin havaintoaineistoihin. Havaintojen taksonomisen oikeellisuuden tärkeys korostuu tulevaisuudessa entisestään, kun luonnontieteellisten museokokoelmien tietoja siirretään helposti saataville Internetiin. Museoaineistojen luotettavuutta ei välttämättä osata kyseenalaistaa, vaikka hyvin läheisten taksonien tai kryptisten lajien tapauksessa olisi usein syytä.

Kaiken kaikkiaan habitaattimallinnus on oikein sovellettuna erittäin tärkeä työkalu eliöiden ja luonnon monimuotoisuuden esiintymisen arvioinnissa. Erityisesti ilmastonmuutoksen vaikutuksia ennakoitaessa on välttämätöntä turvautua lajien levinneisyyden mallintamiseen, koska tulevaisuudesta ei luonnollisesti ole saatavilla havaintoaineistoja. Havainnoivan ekologisen perustutkimuksen tärkeyttä ei pidä silti unohtaa: esiintymismallit luovat järkeviä ennusteita vain, mikäli ne perustuvat järkeville oletuksille eliöiden ja niiden ympäristön välisistä suhteista. Koska ympäristö muuttuu jatkuvasti, myös tällä saralla riittää koko ajan tutkittavaa.

KIITOKSET

Kiitän ohjaajaani Tomas Roslinia kärsivällisyydestä ja hyvistä neuvoista. Kaikkia neuvoja en ikävä kyllä noudattanut, joten vastuu työn puutteista on yksin minun.

LÄHTEET

Beutel, T. S., Beeton, R. J. S. & Baxter, G. S. 1999: Building better wildlife-habitat models. — *Ecography* 22: 219-219.

Collett, D. 2003: *Modelling binary data*. — Chapman & Hall/CRC, Boca Raton, FL. 387 s.

Drakenberg, B., Ehnström, B. A., Liljelund, L.-E. & Österberg, K. 1991: *Lövskogens naturvärden. Rapport* 2946. — Naturvårdsverket, Solna. 117 s.

Elith, J. & Burgman, M. A. 2003: Habitat models for population viability analysis. — Teoksessa: Brigham, C. A. & Schwartz, M. W. (toim.), *Population viability in plants: conservation, management, and modeling of rare plants*: 203-235. Springer-Verlag, Berliini. 362 s.

Elith, J. & Leathwick, J. 2007: Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. — *Diversity & Distributions* 13: 265–275.

Elith, J. & Leathwick, J. 2009a: Conservation prioritisation using species distribution models. — Teoksessa: Moilanen, A., Wilson, K. A. & Possingham, H. P. (toim.), *Spatial conservation prioritization: quantitative methods and computational tools*: 70–93. Oxford University Press, Oxford. 328 s.

Elith, J. & Leathwick, J. 2009b: Species distribution models: ecological explanation and prediction across space and time. — *Annual Review of Ecology, Evolution, and Systematics* 40: 677-697.

Fawcett, T. 2006: An introduction to ROC analysis. — *Pattern Recognition Letters* 27: 861–874.

Gibson, L., Barrett, B. & Burbidge, A. 2007: Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot. — *Diversity and Distributions* 13: 704-713.

Gripenberg, S. & Roslin, T. 2005: Host plants as islands: Resource quality and spatial setting as determinants of insect distribution. — *Annales Zoologici Fennici* 42: 335-345.

Guisan, A. & Zimmermann, N. E. 2000: Predictive habitat distribution models in ecology. — *Ecological Modelling* 135: 147-186.

Hanski, I. 2007: *Kutistuva maailma: elinympäristöjen häviämisen populaatioekologiset seuraukset*. — Gaudeamus, Helsinki. 295 s.

Hanski, I. 1999: *Metapopulation ecology*. — Oxford University Press, Oxford. 313 s.

Krebs, C. J. 1985: *Ecology: the experimental analysis of distribution and abundance*. — Harper & Row, New York. 800 s.

Lozier, J. D., Aniello, P. & Hickerson, M. J. 2009: Predicting the distribution of Sasquatch in western North America: Anything goes with ecological niche modelling. — *Journal of Biogeography* 36: 1623-1627.

McCullagh, P. & Nelder, J. A. 1989: *Generalized linear models*. — Chapman and Hall, Lontoo. 511 s.

Nagelkerke, N. J. D. 1991: A note on a general definition of the coefficient of determination. — *Biometrika* 78: 691-692.

Rainio, R. 1986: Jyrsijöiden ja hirvieläimien tammelle aiheuttamat tuhot. — *Sorbifolia* 17: 210-214.

Ranta, E., Rita, H. & Kouki, J. 2005: *Biometria: tilastotiedettä ekologeille*. — Yliopistopaino, Helsinki. 569 s.

Rita, H. 2004: Vetosuhde (odds ratio) ei ole todennäköisyyksien suhde. — *Metsätieteen aikakauskirja* 2004: 207-212.

Rita, H. & Komonen, A. 2008: Odds ratio: An ecologically sound tool to compare proportions. — *Annales Zoologici Fennici* 45: 66-72.

Roslin, T., Avomaa, T., Leonard, M., Luoto, M. & Ovaskainen, O. 2009: Some like it hot: microclimatic variation affects the abundance and movements of a critically endangered dung beetle. — *Insect Conservation and Diversity* 2: 232-241.

SAS Institute Inc. 2004: *SAS/STAT® 9.1 User's Guide.* — SAS Institute Inc., Cary, NC. 5136 s.

Schultz, A., Klenke, R., Lutze, G., Voss, M., Wieland, R. & Wilking, B. 2003: Habitat models to link situation evaluation and planning support in agricultural landscapes. — Teoksessa: Bissonette, J. A. and Storch, I. (toim.), *Landscape ecology and resource management: linking theory with practice*: 261-282. Island Press, Washington, D.C. 463 s.

Wilson, K. A., Westphal, M. I., Possingham, H. P. & Elith, J. 2005: Sensitivity of conservation planning to different approaches to using predicted species distribution data. — *Biological Conservation* 122: 99–112.