JYVÄSKYLÄN YLIOPISTO
MATEMATIIKAN JA TILASTO-
TIETEEN LAITOS

Pro gradu -tutkielma

Logistisen regressiomallin soveltaminen ekologisen tilan ennustamiseen

Eero Lehtonen

19. toukokuuta 2024

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos
Eero Lehtonen: Logistisen regressiomallin soveltaminen ekologisen tilan ennustamiseen
Pro gradu -tutkielma, tilastotiede ja datatiede, 26 sivua
19. toukokuuta 2024

Tiivistelmä

Tutkielmassa on tarkoitus selvittää vesimuodostumien kunnostustarvetta logistisella regressiomallilla. Vaste eli kunnostustarve on muunnos alkuperäisen aineiston viisiportaisesta ekologinen tila -muuttujasta. Haluttiin selvittää, voiko sitä ennustaa aineiston rekisterimuuttujilla, koska nämä ovat helposti saatavilla.

Koska aineistossa oli paljon puuttuvaa tietoa, käytettiin sen imputoimiseen moni-imputointia. Kaksi oleellista asiaa imputoinnin toteutuksen kannalta olivat, mitä muuttujia käytetään toisten muuttujien imputointiin ja mitä imputointimenetelmiä mihinkin muuttujaan sovelletaan. R-ohjelmiston `mice`-funktio tarjoaa vaihtoehtoja näiden ratkaisemiseen. Tutkimusongelmaa eli sitä, mitä muuttujia kunnostustarpeen ennustamiseen kannattaa käyttää, selvitettiin siten, että aineistoon sovitettiin erilaisia malleja, joita vertailtiin useilla kriteereillä. Esimerkiksi mallin antamaa tulosta siitä, tarvitseeko vesimuodostumaa kunnostaa, verrattiin todelliseen kunnostustarpeeseen, mikä oli tässä aineistossa tiedossa. Oleellista oli löytää ne rekisterimuuttujat, jotka parhaiten ennustavat kunnostustarvetta.

Imputointien perusteella kunnostustarpeen ennustamiseen kannattaa käyttää seuraavia muuttujia: leveysaste, keskisyvyys, suurin syvyys, kunnan pinta-ala, kunnan väkiluku, piiri, korkeus merenpinnasta, maatalousmaan osuus, suuralue Helsinki, suuralue pohjoinen, suuralue etelä, valuma-alueen peltoala ja valuma-alueen suhteellinen peltopinta-ala. Lisäksi kannattaa käyttää leveysasteen ja maatalousmaan osuuden yhteisvaikutusta. Luokitteluvirhe imputoiduille aineistoille on vielä 11.6 % ja useimmiten vielä niin päin, että malli ei löydä niitä vesimuodostumia, joilla on kunnostustarvetta. Jopa 52.5 % vesimuodostumista, jotka olivat kunnostustarpeessa, jäi löytämättä. Toisaalta toisin päin luokitteluvirhe oli parempi. Kunnostustarvetta ennustettiin 3.2 %:ssa tapauksia silloin, kun sitä ei ollut.

Avainsanat: ekologinen status, logistinen malli, luokitteluvirhe, moni-imputointi, puuttuva tieto, sensitiivisyysanalyysi

Sisällys

1 Johdanto

Suomessa pintavesien tilaa luokitellaan vesienhoitolain (1299/2004) soveltamista helpottamaan [1]. Vesienhoitolain mukaan elinkeino-, liikenne- ja ympäristökeskukset (ELY-keskukset) seuraavat ihmisen toiminnan vaikutusta vesien tilaan. Pintavesien hoidossa perusyksikkö on vesimuodostuma. Pintavesien vesimuodostuma on vesienhoitolain mukainen pintavesien erillinen osa kuten järvi, tekoallas, puro, joki, kanava. Nämä voidaan myös jakaa osiin, jos se on vesien hoidon kannalta perusteltua. Vesien hoitoa Suomessa suunnitellaan kuuden vuoden sykleissä.

Luvussa 2 käsitellään tutkielman aineistoa, joka on kerätty vuosina 2012-2017 ja tutkimusongelmaa eli vesienkunnostustarpeen ennustamista. Aineistossa on yhteensä 27 muuttujaa, joista valittiin osa lopulliseen malliin. Tutkielmassa ennustetaan vesimuodostumien kunnostustarvetta logistisella regressiomallilla, jota kuvaillaan luvussa 3. Oleellinen osa työtä oli valita sopivat muuttujat tähän malliin.

Aineistossa on puuttuvaa tietoa varsinkin syvyysmuuttujassa ja valumaalueiden tiedoissa. Tutkielmassa käytettiin imputointimenetelmiä, joilla pyrittiin saamaan hyödynnettyä myös ne havainnot, joista osa muuttujista puuttui. Imputointia toteutetaan R-ohjelmiston `mice`-funktiolla, josta kerrotaan luvussa 4. Moni-imputoinnissa aineistoja muodostettiin 50 kappaletta, mikä on kohtalaisen suuri määrä. Näin pystyttiin toimimaan, koska aineisto oli sen verran pieni, että simulointiin ei mennyt kohtuuttomasti aikaa. Näihin kaikkiin viiteenkymmeneen aineistoon sovitettiin logistinen malli. Luvussa 5 kerrotaan yksityiskohtaisemmin, miten edellä mainittuja menetelmiä sovellettiin tätä työtä tehtäessä ja miten menetelmät sopivat tutkimusongelmaan ja työn aineistoon. Luvussa 6 kerrotaan tarkemmin tuloksista ja siitä, miten valittuihin imputointimenetelmiin päädyttiin. Lisäksi kerrotaan siitä, miten muuttujat lopulliseen mallin valittiin. Luvussa on myös pohdintaa siitä, mitä asioita tutkielmassa olisi voinut käsitellä, että saataisiin tarkempia tuloksia ja mitä puutteita tutkielmassa ehkä on. Lisäksi pohditaan, mitä muita menetelmiä saman ongelman ratkaisuun olisi mahdollista käyttää.

2 Aineiston ja tutkimusongelman kuvaus

Tutkimusongelma oli selittää järvien kunnostustarvetta usealla eri muuttujalla ja selvittää, mitkä muuttujat vaikuttavat eniten kunnostustarpeeseen. Aineisto sisältää seuranta-aineiston lisäksi rekisteriaineistoja. Aineistot on esitelty käsikirjoituksessa [2]. Tutkielman seuranta-aineisto on kerätty vuosina 2012-2017. Se on osa aineistosta, jossa on luokiteltu kaikki Suomen pintavedet, ja sen on kerännyt Suomen ympäristökeskus. Aineisto koostuu järvistä tai isompien järvien osista eli vesimuodostumista, jotka ovat vesienhoidon perusyksiköitä. Kaikkein pienimpiä pintavesiä ei ole määritelty vesimuodostumiksi, joten ne on rajattu aineiston ulkopuolelle.

Vesimuodostumat ovat aineistossa luokiteltu järvityyppien mukaan [1]. Tyyppi-muuttujassa luokittelu tehdään järven humuksisuuden, syvyyden, pinta-alan, sijainnin, kalkkisuuden ja viipymän mukaan. Järvityyppejä ovat keskikokoiset humusjärvet (Kh), lyhytviipymäiset järvet (Lv), matalat humusjärvet (Mh), matalat vähähumuksiset järvet (MVh), matalat runsashumuksiset järvet (Mrh), pienet humusjärvet (Ph), Pohjois-Lapin järvet (poLa), runsashumuksiset järvet (Rh), runsaskalkkiset järvet (Rk), runsasravinteiset järvet (Rr), runsasravinteiset ja -kalkkiset järvet (RrRk), suuret humusjärvet (Sh), suuret vähähumuksiset järvet (SVh), pienet ja keskikokoiset vähähumuksiset järvet (Vh).

Humuksisella järvellä tarkoitetaan eloperäistä hajonnutta tai osittain hajonnutta ainesta, jonka määrää vedessä mitataan veden värillä. Vesimuodostelman viipymä on teoreettinen luku, joka kuvaa veden vaihtuvuutta eli vesimuodostuman tilavuus jaettuna sillä, kuinka paljon vettä valuu puroa pitkin pois. Toisin sanoen se on teoreettinen aika, jossa koko muodostelman vesi vaihtuisi, jos vesi vaihtuisi tasanopeudella koko muodostumassa.

Alkuperäinen vaste eli vesimuodostuman ekologinen tila perustui arvioon siitä, kuinka paljon ihmisen toiminta on heikentänyt kyseisen pintaveden tilaa. Järvityypille ominaiset biologiset tilat eli tilat, joihin ihminen ei ole vaikuttanut, saatiin vertailuaineistoista tai mallintamalla. Jos kumpaakaan ei voitu toteuttaa, turvauduttiin asiantuntijoiden arvioihin. Näihin verrataan arvioitavia vesimuodostumia. Ekologisia tiloja oli alunperin viisi tasoa. Tilat olivat seuraavat raportin [1] mukaan:

- "**Erinomaisen** ekologisen tilan luokassa fysikaalis-kemiallisten, hydrologismorfologisten ja biologisten laatutekijöiden arvoissa on enintään hyvin vähän ihmistoiminnasta johtuvia muutoksia verrattuna niihin arvoihin (vertailuarvot), jotka tavallisesti liitetään kyseisen pintavesimuodostumatyypin häiriintymättömiin oloihin (vertailuolot). Ylläkuvatut vertailuolot siis määrittelevät erinomaisen tilan."

- "**Hyvän** ekologisen tilan luokassa biologisten laatutekijöiden arvoissa on merkkejä ihmistoiminnasta johtuvista vähäisistä muutoksista, mutta ne saavat erota ainoastaan vähän vertailuarvoista. Fysikaaliskemialliset laatutekijät eivät ylitä tasoja, jotka varmistavat biologisten laatutekijöiden hyvän tilan saavuttamisen. Hydrologis-morfologiset olot eivät haittaa biologisten laatutekijöiden hyvän tilan saavuttamista."

- "**Tyydyttävän** ekologisen tilan luokassa biologisten laatutekijöiden arvot eroavat kohtalaisesti vertailuarvoista. Arvot osoittavat kohtalaisesti ihmistoiminnasta johtuvia muutoksia, ja ovat muuttuneet selvästi enemmän kuin hyvää tilaa vastaavissa olosuhteissa. Fysikaaliskemialliset ja hydrologis-morfologiset olot eivät haittaa biologisten laatutekijöiden arvojen saavuttamista."

- "**Välttävän** ekologisen tilan luokassa ilmenee suurehkoja muutoksia kyseisen pintavesimuodostumatyypin biologisten laatutekijöiden arvoissa. Eliöyhteisöt eroavat merkittävästi häiriintymättömissä olosuhteissa olevista eliöyhteisöistä ko. pintavesimuodostumatyypissä."

- "**Huonon** ekologisen tilan luokassa ilmenee vakavia muutoksia biologisten laatutekijöiden arvoissa ja luokassa puuttuu suuri osa eliöyhteisöistä, jotka tavallisesti liitetään ko. pintavesimuodostumatyyppiin häiriintymättömissä olosuhteissa."

Ekologista tilaa määrättäessä käytetään seuraavia biologisia laatutekijöitä: kasviplaktonin määrä, haitallisen sinilevän osuus kasviplanktonissa, päällyslevät, vesikasvit, pohjaeläimet syvänteissä, pohjaeläimet rantavyöhykkeellä ja kalat. Eri vesimuodostumatyypeille oli eri rajat kaikille muuttujille [1].

Kasviplanktonin määrää mitattiin lehtivihreän eli a-klorofyllipitoisuudella, jonka yksikkö oli mikrogrammaa litrassa: mitä vähemmän a-klorofylliä, sitä parempi. Haitallisen sinilevänkin pitoisuus oli luonnollisesti oltava mahdollisimman pieni.

Päällyslevälaatutekijä perustui kahteen osaan, tyyppiominaisten taksonien esiintymiseen ja prosenttiseen mallinkaltaisuuteen. Tyyppiomaisten taksonomien esiintyminen vertaa vertailuaineistosta laskettujen tyypillisten lajien määrän osuutta arvioitavan vesimuodostuman koko lajimäärään. Prosenttisen mallinkaltaisuuden laskennassa verrataan tarkasteltavan vesikasvilajien suhteellisia osuuksia vertailuvesimuodostumien lajien runsausosuuksiin. Molemmille lasketaan vesimuodostumakohtaiset arvot ja verrataan niitä samantyyppisten vesimuodostumien arvoihin.

Vesikasvillisuuslaatutekijä oli kolmiosainen, se perustui tyyppiominaisten taksonien esiintymiseen, prosenttiseen mallinkaltaisuuteen ja referenssi-

indeksiin. Tyyppiominaisten taksonien esiintyminen ja prosentuaalinen mallinkaltaisuus olivat samoin määriteltyjä kuin päällyslevillä. Referenssi-indeksi perustui vesikasvien ravinnekuormituksen sietokykyyn.

Pohjaeläimet syvänteissä -laatutekijä perustui kahteen arvoon: PICM-arvoon ja prosenttiseen mallinkaltaisuuteen. PICM-arvo perustui lajien kuormituksen sietokyvyn pisteyttämiseen ja niillä painotettuihin 10-kantaisten logaritmien keskiarvoihin. [1]

Kalat-laatutekijä perustui verkkokoekalastukseen. Se jakautui neljään osaan: saaliin biomassaan, saaliskalojen lukumäärään, särkikalojen osuuteen biomassassa ja indikaattorilajien esiintymiseen. Biomassa on kaksisuuntainen muuttuja: sekä luonnontilaa suuremmat että pienemmät muuttujan arvot voivat ilmaista ihmistoiminnan rehevöittävää vaikutusta.

Fysikaalis-kemialliset ja hydrologis-morfologiset laatutekijät ovat biologisia laatutekijöitä tukevia suureita. Idea luokittelussa on, että ne mahdollistavat hyvät biologiset laatutekijät. Fysikaalis-kemialliset tekijät olivat vesiympäristölle haitallisten aineiden määrä. Haitallisia aineita oli 15 kappaletta. Hydrologis-morfologisia laatutekijöitä olivat vesistön esteellisyys ja se, kuinka paljon ihminen on sitä muuttanut. [1]

Viisiluokkaisesta ekologinen tila -muuttujasta saadaan vaste eli kunnostustarve siten, että jos ekologinen tila on tyydyttävä, välttävä tai huono on, vaste saa arvon 0 eli kunnostustarvetta on, ja muutoin 1 eli ei kunnostustarvetta. [1]

Tämän tutkielman rekisterimuuttujat on koottu seuraavista rekistereistä: Maanmittauslaitos, Luonnonvarakeskus, Tilastokeskus ja Suomen ympäristökeskus [2]. Tarkasteltavat rekisterimuuttujat ovat vesimuodostuman nimi, järven nimi, vesimuodostuman koordinaatit, vesiala, rantaviivan pituus, keskisyvyys, suurin syvyys, tilavuus, kunnan pinta-ala, suuralue, maatalousmaan määrä kunnassa, kunnan väkiluku, kesämökkien määrä kunnassa, korkeus merenpinnasta, valuma-alueen kokonaisala, valuma-alueen metsäala, valuma-alueen peltoala ja valuma-alueen kesämökkien määrä. Järven ja vesimuodostuman nimet ja vesimuodostuman koordinaatit saatiin Maanmittauslaitokselta [3]. Kartoista saatiin laskettua vesiala ja rantaviivan pituus sekä piiri eli sellaisen ympyrän kehän pituus jonka pinta-ala on sama kuin vesimuodostuman pinta-ala ja pyöreys eli piiri jaettuna rantaviivan pituudella eli vesimuodostuman todellisen rantaviivan pituuden ja samankokoisen, mutta täysin pyöreän vesimuodostuman rantaviivan pituuden suhde. Myös syvyydet saatiin kartoista, jolloin tilavuus voitiin laskea. Tieto kunnan pinta-alasta on saatu niin ikään Maanmittauslaitokselta. Suuralue on määritelty kunnittain. Niitä oli neljä; Etelä-Suomi, Helsinki-Uusimaa, Länsi-Suomi, Pohjois- ja Itä-Suomi. Maatalousmaan määrä kunnassa on tieto Luonnonvarakeskukselta [4]. Väkiluku ja kesämökkien määrä kunnassa on peräisin Tilastokeskuk-

selta [5]. Korkeustaso saatiin Maanmittauslaitokselta. Valuma-alueen tiedot sen kokonaisala, metsäala, peltoala ja kesämökkien määrä saatiin Suomen ympäristökeskuksen valuma-aluemallista [6].

Kaikki muuttujat, joilla vesimuodostuman tilaa olisi mahdollista ennustaa, ovat taulukossa 1.

Taulukko 1: Kaikki muuttujat, joilla kunnostustarvetta voitaisiin ennustaa.

1	tyyppi
2	pituusaste
3	leveysaste
4	vesiala
5	rantaviivan pituus
6	keskisyvyys
7	suurin syvyys
8	tilavuus
9	kunnan pinta-ala
10	kunnassa oleva maatalousmaa
11	kunnan väkiluku
12	kunnassa olevien kesämökkien määrä
13	piiri
14	pyöreys
15	vesimuodostuman korkeus merenpinnasta
16	kunnan väkiluvun suhde kunnan pinta-alaan
17	kunnan kesämökkien määrän suhde kunnan pinta-alaan
18	kunnassa olevan maatalousmaan suhden kunnan pinta-alaan
19	vesimuodostuman pinta-ala
20	suuralue
21	valuma-alueen metsäala
22	valuma-alueen peltoala
23	valuma-alueen pinta-ala = A
24	valuma-alueen kesämökkien lukumäärä
25	valuma-alueen kesämökkien lukumäärän suhde A:han
26	valuma-alueen metsäalan suhde A:han
27	valuma-alueen peltoalan suhde A:han

3 Logistinen regressiomalli

Logistisella regressiolla mallinnetaan todennäköisyyksiä tai dikotomisia vasteita eli vasteita, jotka voivat saada arvot 1 tai 0. Tässä työssä 0 vastaa vesimuodostumen kunnostustarvetta ja 1 kunnossa olevaa vesimuodostumaa. Jotta logistisen regression voisi ymmärtää, pitää ensin ymmärtää vetokertoimen (odds) ja vetokerrointen osamäärän eli ristisuhteen (odds ratio) sekä sen logaritmin ja logit-funktion käsitteet. Symbolilla π tarkoitetaan tässä tutkielmassa jatkossa tapahtuman y todennäköisyyttä $P(y = 1) = \pi$ eikä matemaattista vakiota, jos ei muuta mainita. Vetokerroin on oleellinen asia logistisen regression ymmärtämisen kannalta, se määritellään seuraavasti [7]

$$\text{vetokerroin} = \frac{\pi}{1-\pi}.$$

Esimerkiksi nopan heitossa todennäköisyys saada jokin määrätty luku on (1/6), jolloin

$$\text{vetokerroin} = \frac{1/6}{5/6} = 1/5.$$

Vetokertoimien osamäärää sanotaan ristisuhteeksi

$$\text{ristisuhde} = \frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)} = \frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)}.$$

Luonnollinen logaritmi vetokertoimesta on niin hyödyllinen, että sille on oma merkintä ”logit”

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right), \quad 0 < \pi < 1.$$

Sen käänteisfunktio on

$$\text{invlogit}(x) = \log\left(\frac{e^x}{1+e^x}\right), \quad -\infty < x < \infty.$$

Logistisen regressiomallin kaava on

$$P(y_i = 1) = \pi_i = \text{invlogit}(\beta_0 + x_i^\top \beta),$$

missä β_0 on vakiotermi, β regressiokerrointen vektori ja x_i selittävien muuttujien vektori [7].

3.1 Logistisen mallin kertoimien tulkinta

Logistisessa regressiossa selittävät muuttujat linkitetään vasteen odotusarvoon logit-muunnoksen avulla. Vakiolle β_0 saadaan tulkinta, kun asetetaan $x = 0$. Tällöin invlogit$(\beta_0) = \pi$ ja edelleen $\beta_0 = \text{logit}(\pi)$. Usein β_0 tulkitaan todennäköisyyden avulla.

Yksinkertainen tapa tulkita regressiokertoimet on verrata ennustettuja todennäköisyyksiä erilaisten oletettujen havaintojen arvoilla. Kertoimien $\hat{\beta}$ estimointi tehdään yleensä suurimman uskottavuuden menetelmällä. Asetetaan $x_1 = x$ ja $x_2 = x + 1$, jolloin

$$\text{logit}(\hat{\pi_2}) - \text{logit}(\hat{\pi_1}) = \hat{\beta}_1(x + 1 - x)$$

eli

$$\log\frac{\hat{\pi_2}}{1-\hat{\pi_2}} - \log\frac{\hat{\pi_1}}{1-\hat{\pi_1}} = \log\left(\frac{\hat{\pi_2}/(1-\hat{\pi_2})}{\hat{\pi_1}/(1-\hat{\pi_1})}\right) = \hat{\beta}_1.$$

Kun otetaan eksponentti edellisestä,

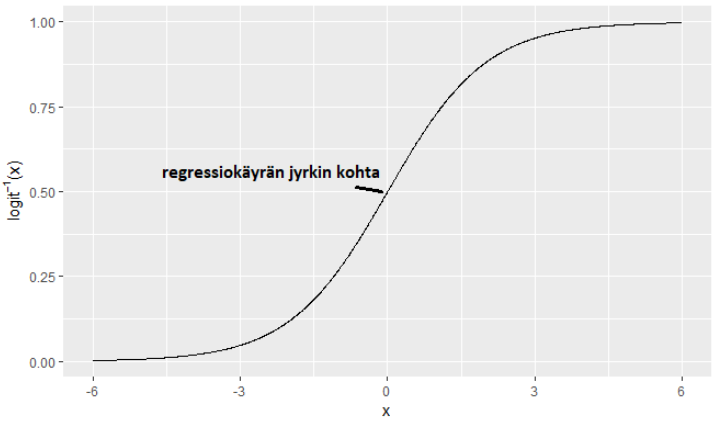
$$\left(\frac{\hat{\pi_2}}{1-\hat{\pi_2}}\right) \Big/ \left(\frac{\hat{\pi_1}}{1-\hat{\pi_1}}\right) = e^{\hat{\beta_1}}.$$

Termille $\hat{\beta}_1$ ei ole mielekästä tulkintaa, mutta $e^{\hat{\beta_1}}$ on selittävistä muuttujista x ja $x + 1$ laskettu ristisuhde.

Ristisuhteiden etu todennäköisyyksiin verrattuna on muun muassa se, että β-kertoimia voi suurentaa niin paljon kuin on tarvetta ilman, että rajat 0 ja 1 tulevat vastaan.

Joskus syötemuuttujat kannattaa keskistää, niin β-kertoimet on helpompi tulkita.

Neljällä jakamisen sääntö on kätevä, kun arvioidaan selittävän muuttujan maksimivaikutusta logistiseen regressioon. Regressiokäyrä on jyrkimmillään keskikohdassaan, kun invlogit$(\beta_0 + \beta_1 x) = 0.5$ (kuvio 1). Tässä kohdassa myös logistisen funktion derivaatta $\beta_1 \exp(\beta_0 + \beta_1 x)/(1 + \exp(\beta_0 + \beta_{1x}))^2$ on suurimmillaan eli selittävän muuttujan vaikutus on isoimmillaan. Lähellä



Kuvio 1. Regressiokäyrä jyrkimmillään.

regressiokäyrän keskikohtaa voi estimaattia $\beta_1/4$ käyttää karkeana arviona selittävän muuttujan β_1 vaikutuksesta [7].

3.2 Logistisen mallin interaktion tulkinta

Logistisessa mallissa voidaan kahden syötemuuttujan yhteisvaikutusta mallintaa interaktiolla. Esimerkiksi jonkun muuttujan vaikutus voi olla erilainen eri tyyppisillä vesimuodostumilla, jolloin vesimuodostumatyypeille pitää olla omat logistisen regression käyrät. Yhteisvaikutusta voi olla myös jatkuvilla syötemuuttujilla. Esimerkiksi valuma-alueen maatalousmaan vaikutus saattaa olla sitä isompi, mitä pohjoisempi on vesimuodostuman koordinaatti. Toisin sanoen leveysasteen koordinaatin vaikutus kasvaa, kun valuma-alueella on maatalousmaata paljon. Tällöin logistisen regression kaava on

$$P(y_i = 1) = \pi_i = \text{invlogit}(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i1} \times x_{i2}\beta_3),$$

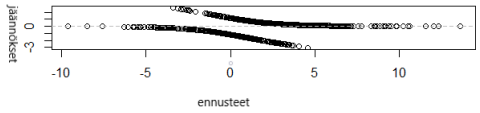
missä x_{i1} ja x_{i2} ovat kaksi eri syötemuuttujaa, ja ne sekä $x_{i1} \times x_{i2}$ muodostavat ennustimen. Kuten edellä mainittiin, on joskus järkevää keskittää syötemuuttujat ennen kuin muodostetaan interaktioennustin. Jos syötemuuttujat vain keskitetään eikä skaalata, niin tuloksia voi tulkita alkuperäisten muuttujien mittayksiköissä. Käyttäen neljällä jakamissääntöä saadaan β-kerroin muutettua likimääräiseksi maksimimuutokseksi todennäköisyydessä sille, että vesimuodostelma vaatii kunnostusta [7].

3.3 Logistisen mallin jäännökset

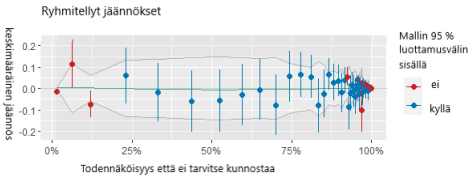
Logistisen mallin jäännökset määritellään samoin kuin lineaarisen mallin jäännökset, eli havaittu arvo vähennetään ennusteesta seuraavasti

$$\text{jäännös}_i = y_i - \mathrm{E}(y_i|x_i) = y_i - \text{logit}^{-1}(\beta_0 + x_i^\top \beta).$$

Toisin kuin lineaarisessa regressiossa nämä ennusteet eivät ole sellaisenaan hyödyllisiä, koska vaste on dikotominen, jolloin jäännökset muodostavat vain kaksi käyrää (kuvio 2). Tavallisia jäännöksiä hyödyllisempää on tarkastella ryhmiteltyjä jäännöksiä (binned residuals). Tällöin data jaetaan ennustearvojen mukaan ryhmiin, joissa on kaikissa yhtä monta datapistettä. Sitten piirretään jäännöskuvio, jossa on keskimääräinen jäännös keskimääräistä sovitetta vasten (kuvio 3) [7].



Kuvio 2. Tavalliset jäännökset.



Kuvio 3. Ryhmitellyt jäännökset.

3.4 Logistisen mallin devianssi

Devianssia käytetään eri mallien vertailuun. Devianssi yksin ei kerro mallin hyvyydestä mitään, vaan sen käyttö vaatii jonkin vertailumallin. Devianssi D on määritelty seuraavasti [8]

$$D = -2 \log \left(\frac{\text{sovitetun mallin uskottavuus}}{\text{saturoidun mallin uskottavuus}} \right). \tag{1}$$

Koska logaritmisessa mallissa saturoidun mallin uskottavuus on yksi, niin kaava sievistyy muotoon

$$D = -2\log(\text{sovitetun mallin uskottavuus}). \tag{2}$$

Saturoidulla mallilla tarkoitetaan mallia, jossa on yhtä paljon parametrejä kuin havaintoja, jolloin malli sopii aineistoon täydellisesti. Nolladevianssilla tarkoitetaan sellaisen mallin devianssia, jossa on vain vakio, eli mallia joka ennustaa kaikille havainnoille havaintojen keskiarvon. Devianssi saa arvoja nollasta äärettömään ja mitä pienempi devianssi on, sitä paremmin malli sopii aineistoon. Mutta jos malli sopii liian hyvin aineistoon, on vaarana ylisovittuminen, jolloin mallin käyttö ennustamiseen on epäluotettavaa. Myös kahta erilaista sisäkkäistä mallia voidaan vertailla deviansseilla kaavalla [8]

$$D = -2\log\left(\frac{\text{pienemmän mallin uskottavuus}}{\text{suuremman mallin uskottavuus}}\right). \tag{3}$$

Tällöin mallien välinen devianssi noudattaa χ^2-jakaumaa vapausasteella mallien vapausasteiden erotus.

3.5 AIC

AIC (Akaike informatin criterion) [9] on devianssin tavoin menetelmä, jolla eri malleja verrataan keskenään. Devianssista poiketen AIC huomioi mallin parametrien määrän, ja se antaa pienempiä arvoja yksinkertaisemmilla malleille eli malleille, joissa on vähemmän parametrejä. AIC soveltuu myös sellaisten mallien vertailuun, jotka eivät ole sisäkkäisiä. AIC ja devianssi poikkeavat toisistaan kaavan viimeisen termin osalta. Siinä k on parametrien määrä:

$$\text{AIC} = -2\log(\text{sovitetun mallin uskottavuus}) + 2k.$$

3.6 Luokitteluvirheet

Luokitteluvirhe on yksinkertaisesti väärä luokittelu. Tässä työssä se tarkoittaa, että joko vesimuodostuma luokitellaan kunnostustusta vaativaksi, vaikka

se ei kunnostusta tarvitse tai vesimuodostuma luokitellaan sellaiseksi, että se ei tarvitse kunnostusta, vaikka kunnostustarvetta on. Tutkielmassa luokitteluvirhe niin päin, että kunnostusta vaativa vesimuodostuma luokitellaan kunnossa olevaksi on vakavampi asia kuin se, että kunnossa oleva vesimuodostuma luokitellaan, että kunnostustarvetta on. Kun mallia käytetään ennustamiseen, tarkoittaa luokitteluvirhe todennäköisyyttä, että ennuste on virheellinen. Tämä on tilanne tässäkin tutkielmassa.

4 Puuttuvan tiedon käsittely

Tässä luvussa kerrotaan havaintojen puuttuvuusmekanismeista eli puuttumisen ja puuttuvien arvojen riippuvuuksista ja siitä, miten työssä on selvitetty minkälaista puuttuvuutta aineistossa on ja miten se vaikuttaa tuloksiin. Lisäksi käydään läpi, miten tietojen imputointi käytännössä tehtiin.

4.1 Puuttuvuusmekanismit

Seuraava puuttuvan tiedon teoria perustuu lähteisiin [10] ja [11]. Puuttuvuusmekanismeilla tarkoitetaan erilaisia puuttuvuuden ja puuttuvien arvojen riippuvuuksia toisistaan tai vasteesta. Puuttuvuusmekanismeja on useita MCAR, MAR ja MNAR. Näistä MCAR on helpoin käsitellä. Nimitys tulee sanoista "Missing Completely At Random" eli puuttuvuus ei riipu vasteen arvosta eikä muuttujasta itsestään. Koska imputoituja aineistoja voidaan käyttää mallintamaan monia eri muuttujia, niin tämä periaatteessa tarkoittaa, että puuttuminen ei riipu muistakaan muuttujista. MAR tulee sanoista "Missing At Random". Nimestä huolimatta se tarkoittaa ehdollista riippumattomuutta eli puuttuvuus ei riipu vasteesta ehdolla muut muuttujat. MNAR tulee sanoista "Missing Not At Random". Se tarkoittaa, että puuttuvuus riippuu myös itse puuttuvan muuttujan arvoista. Nämä puuttuvuuden tyypit määritellään seuraavasti. Merkinnöissä ψ on tuntematon parametri, M on puuttuvuusmatriisi ja Y sisältää kaikki muuttujat.

MCAR eli Missing Completely At Random

$$P(M|Y,\psi) = P(M|\psi) \quad \text{kaikilla} \quad Y,\psi,$$

MAR eli Missing At Random

$$P(M|Y,\psi) = P(M|Y^{obs},\psi) \quad \text{kaikilla} \quad Y^{mis},\psi$$

ja MNAR, kun MAR ehto ei päde, eli puuttuvuus riippuu muuttujien Y arvoista.

4.2 Sensitiivisyysanalyysi

Kun muuttujien puuttuvuusmekanismeja ei varmuudella tiedetä, niin tätä tutkitaan sensitiivisyysanalyysilla, jotta saataisiin saataisiin selville erilaisten puuttuvuusmekanismien vaikutus estimaatteihin [10]. Sensitiivisyysanalyysi toteutetaan tekemällä samanlaisia analyyseja olettaen, että puuttuvat arvot ovat joko systemaattisesti suurempia tai pienempiä kuin havaitut arvot ja mallintamalla vastetta tällaisella koeaineistolla.

4.3 Ketjuimputointi

Tämän luvun teksti perustuu kirjaan [10]. Tutkielmassa käytettiin ketjuimputointia, eli imputoitiin monimuuttujaista dataa ketjutettujen yhtälöiden avulla. Siitä tulee nimikin `MICE` (Multivariate Imputation by Chained Equations.) Tämä moni-imputointi huomioi vain muuttujien suorat vaikutukset vasteeseen, joten mahdolliset interaktiot pitää erikseen koodata malliin. Periaatteessa se ei erottele lopullisen mallin vastetta ja selittäviä muuttujia.

Imputointi tapahtuu yksi muuttuja kerrallaan ehdolla kaikki muut muuttujat. Tämä imputointi on tehtävä useamman kerran jo senkin takia, että ensimmäisellä kerralla on muiden muuttujien arvot arvottava satunnaisesti datasta ja vasta ensimmäisen kierroksen jälkeen on ehtona olevina arvoina jo imputoidut arvot. Imputoinnin vaiheet ovat seuraavat:

1. Suoritetaan algoritmi m kertaa.
2. Arvotaan jokaisen muuttujan kaikkiin puuttuviin kohtiin olemassa olevista arvoista jotkin arvot käyttäen jokaiselle muuttujalle sille valittua imputointimenetelmää. Määritellään kaikille muuttujille, joissa on puuttuvaa tietoa, jokin malli, joilla niitä imputoidaan ja mallin parametrit ϕ.
3. Toistetaan T kertaa:
   (a) Imputoidaan puuttuvat arvot muuttujaan Y_1 ehdollisesta jakaumasta $Y_1|Y_1^{obs}, Y_{-1}, X, \phi$. Toisin sanoen siis käytetään imputointiin muuttujan Y_1 havaittuja arvoja ja kaikkien muiden muuttujien havaittuja ja imputoituja arvoja.
   (b) Toistetaan sama kaikkien Y-muuttujien kohdalla. Toisin sanoen imputoidaan kaikkiin muuttujiin kaikkiin puuttuviin kohtiin arvot käyttäen havaittuja muuttujia ja uusimpia imputointeja.

Mallissa m on imputoimalla tuotettavien datasettien määrä. T on imputointikierrosten lukumäärä, siis jokaisen m:n vaiheen alussa arvotut täysin

satunnaiset arvot korvataan arvoilla, jotka imputoidaan käyttämällä muita muuttujia ja saman muuttujan havaittuja arvoja. Seuraavalla kierroksella nämä korvataan uusilla arvoilla, jotka on imputoitu käyttäen edellisiä arvoja ja muita muuttujia. Tätä toistetaan T kertaa. Muuttujat Y ovat muuttujia, joissa on puuttuvaa tietoa. Muuttujat X ovat täysin havaittuja muuttujia.

Passiivinen imputointi tarkoittaa arvojen suoraa laskemista muista muuttujista. Se on kuvattu luvussa 5.4

Imputointimenetelmät

Yleisimmät imputointimenetelmät, joita myös tässä tutkielmassa harkittiin käytettävän, on lueteltu taulukossa 2.

Taulukko 2: Imputointimenetelmät.

menetelmä	muuttujan tyyppi
ennustekeskiarvokaltaistus(pmm)	mikä vaan
normaalijakauma perusteinen imputointi (norm)	numeerinen
passiivinen imputointi	johdetut muuttujat

Ennustekeskiarvokaltaistus käyttää imputointiin aineistossa esiintyviä arvoja, joiden muut muuttujan arvot ovat lähellä imputoitavan rivin muiden muuttujien arvoja. Algoritmi arpoo viidestä lähimmästä arvosta yhden puuttuvan arvon tilalle. Normaalijakaumaperusteinen imputointi arpoo imputoitavat arvot normaalijakaumasta, jonka parametrit ovat uskottavimpia havaituille muuttujan arvoille [10](7.6.1).

5 Aineiston analyysi

Tässä luvussa kuvataan, kuinka analyysi toteutettiin ja millaisia ongelmia sitä toteutettaessa tuli vastaan.

5.1 Imputointimallit

Tutkielmassa käytettiin R-ohjelmointikielen `mice` -kirjastoa ja sen vastaavaa funktiota. Se on tarkoitettu juuri tämän tutkielman tyyppisiin ongelmiin. Oleellisia puuttuvia tietoja oli muuttujissa keski- ja suurin syvyys, joista puuttui havainto 1871 ja 1776 kohteesta, keskisyvyydestä laskettu tilavuus, josta puuttui 1892 havaintoa, valuma-alueen metsäala, valuma-alueen peltoala, valuma-alueen kesämökkien määrä ja valuma-alueen kokonaisala. Kaikkiaan havaintoyksiköitä, joista puuttui ainakin yksi tieto, oli 1080. Korkeustason tieto puuttui 250 tapauksessa. Näitä arvoja imputoitiin eli ennustettiin muiden muuttujien avulla, tässä työssä.

Kaikkia muita paitsi tilavuutta ja skaalattuja muuttujia, eli muuttujia valuma-alueen kesämökkien lukumäärän suhde valuma-alueen pinta-alaan, valuma-alueen metsäalan suhde valuma-alueen pinta-alaan ja valuma-alueen peltoalan suhde valuma-alueen pinta-alaan, imputoitaessa päädyttiin käyttämään ennustekeskiarvokaltaistusmenetelmää. Toinen vaihtoehto eli normaalijakaumalla imputointi tuotti hyvin vaihtelevia arvoja. Tilavuus imputoitiin suoraan kertomalla imputoiduista tai havaituista arvoista keskisyvyys pinta-alalla, niin kuin se oli tehty silloinkin, kun molemmat arvot olivat käytettävissä. Skaalatut muuttujat niin ikään imputoitiin samoin kuin jos käytettävissä olisi ollut havaitut arvot käyttäen imputoituja arvoja eli $havaittuarvo/10 * valuma - alueenala$. Menetelmä on passiivinen imputointi [10].

R-kirjaston `mice`-funktio tarjosi tosin myös johdetuille muuttujille ennustekeskiarvokaltaistusmenetelmää, mutta se vaihdettiin ennen imputointia käsin passiiviseksi imputoinniksi.

5.2 Ennustinmatriisi

Tämän luvun teksti perustuu kirjaan [10](9). Ennustinmatriisi on matriisi, jossa on tieto, miten muuttujia ja vastetta käytetään toistensa imputoimiseen. Vielä imputointivaiheessa ei aina välttämättä tiedetä, mikä muuttuja on vasteena varsinaisessa analyysissä, vaikka tässä tutkielmassa se tiedettiin. Ennustinmatriisi tehtiin siten, että alkuperäiseen datamatriisiin lisättiin jokaiselle muuttujalle puuttuvuussarake, jossa oli 1, jos muuttuja oli havaittu,

ja muuten 0. Tämän puuttuvuussarakkeen yhteyttä muiden muuttujien arvoihin samoilla tilastoyksiköillä eli vesimuodostumilla tarkasteltiin. Toinen asia, mitä ennustinmatriisia tehdessä tarkasteltiin, oli imputoitavan muuttujan imputoinnissa käytettävien muuttujien välinen korrelaatio. Jos imputoitavan muuttujan ja toisen muuttujan arvojen välinen korrelaatio oli yli 0.05, niin muuttujalla imputoitiin toista. Koska puuttuvuus muuttujissa on erilaista, on mahdollista, että matriisi ei ole symmetrinen. Toisin sanoen jos muuttujalla y_1 imputoidaan muuttujaa y_2, niin välttämättä ei muuttujalla y_2 imputoida muuttujaa y_1.

Myös sillä, kuinka paljon havaintoyksikköjä on käytettävissä imputointiin, on luonnollisesti suuri merkitys imputoinnin tarkkuudelle. Tähän liittyvät sisä- ja ulkohyötysuhteiden käsitteet [10](4.1.2). Sisä- ja ulkohyötysuhde käytettäville muuttujille lasketaan vain puuttuvuustiedosta, niiden laskemiseen ei siis käytetä muuttujien arvoja. Sisä- ja ulkohyötysuhde antaa kuitenkin jonkinlaista tietoa siitä, kannattaako jotakin määrättyä muuttujaa käyttää imputointiin. Sisähyötysuhde muuttujalle on sellaisten puuttuvien arvojen lukumäärä, joista on toisilla tilastoyksiköillä tietoa, jaettuna kaikkien tilastoyksiköiden lukumäärällä. Ulkohyötysuhde on vastaavasti sellaisten arvojen lukumäärä, joita puuttuu toisilta muuttujilta, eli sellaisten arvojen lukumäärä, joita voi käyttää imputointiin. Sisä- ja ulkohyötysuhteet selittävät, miksi ennustinmatriisi ei ole välttämättä symmetrinen toisin kuin korrelaatiomatriisi. Kun prediktorimatriisi oli tehty luvussa 5.2 mainitulla funktiolla, niin imputoidessa havaittiin, että valuma-alueen tiedot eivät konvergoidu kunnolla, eli todennäköisesti niiden välillä oli riippuvuutta. Matriisiin tehtiin vielä käsin muutos, jolla määriteltiin, että valuma-alueen tiedoilla ei imputoida toisia muuttujia.

Ennustinmatriisi quickpred-funkfiolla

Ennustinmatriisin tekoon käytettiin `quickpred`-funktiota, jolla voi toteuttaa edellisen luvun asiat kerralla, kun ensin määrittelee minimikorrelaation ja käytetäänkö Pearsonin, Spearmanin vai Kendallin korrelaatiota. Lisäksi funktiossa voi määrittää itse joiksikin ennustinmatriisin arvoiksi joko 0 tai 1, eli voi määrittää muuttujapareille y_i ja y_j, käytetäänkö niitä toistensa imputointiin. Valuma-alueen tiedoilla ei imputoitu toisia muuttujia, koska tällöin imputoinnit eivät konvergoineet.

5.3 Muuttujien valinta

Muuttujat, joilla kunnostustarvetta ennustettiin, valittiin imputoimalla 50 aineistoa ja sen jälkeen tekemällä jokaiselle aineistolle logistinen regressio si-

ten, että lisättiin yksi kerrallaan muuttujia ja katsottiin jokaisen muuttujan vaikutus logistisen regression lopputulokseen suurimman uskottavuuden menetelmällä. Jos muuttujalla oli merkitystä, se jätettiin, muutoin poistettiin. Menettely toistettiin kaikille aineistoille. Sitten katsottiin, kuinka moneen malliin yksittäiset muuttujat olivat jääneet. Näin saatiin taulukko, jossa oli jokaiselle muuttujalle arvo nollan ja viidenkymmenen välillä. Ne muuttujat, jotka jäivät malliin yli 25 kertaa testattiin vielä Waldin testillä siten, että kyseinen muuttuja jätettiin pois mallista ja verrattiin tällaista mallia malliin, jossa muuttuja on mukana. Jos p-arvo jäi alle 0.05:n niin muuttuja otettiin mukaan lopulliseen malliin. Osa muuttujista oli joko kokonaan toistensa muunnoksia kuten muuttujat pinta-ala ja piiri tai osittain päällekkäisiä kuten tyyppi ja keskisyvyys. Näiden annettiin olla mallissa, kun muuttujien valintaa tehtiin. Pinta-alaa kuvaavista muuttujista piiri jäi malliin. Erilaisten mallien luokitteluvirheitä tarkasteltiin myös ennen kuin lopullinen malli valittiin. Kaikilla imputoiduilla malleilla luokitteluvirhe jäi yli 10 prosentin ja vielä niin päin, että malli ei löytänyt vesimuodostumia, joilla oli kunnostustarvetta. Näin päin luokitteluvirhe oli 52.4 % ja toisin päin, ts. malli ehdotti turhaan kunnostustarvetta vesimuodostumalle, vain 3.2 %.

5.4 Interaktiot

Interaktiotermejä valittaessa huomioitiin, miten muuttujien vaikutukset saattavat riippua toisistaan sekä ovatko muuttujat osittain päällekkäisiä, kuten esimerkiksi tyyppi ja keskisyvyys: joidenkin vesimuodostuma tyyppien määritelmissä on mukana keskisyvyys. Myös interaktion p-arvon käytetyssä mallissa piti luonnollisesti olla alle 0.05, jotta se kannattaisi malliin ottaa. Malliin valituista 14 muuttujasta saisi 91 kahden muuttujan interaktiotermiä. Jos niistä valittaisiin käytettävät interaktiot, niin keskimäärin 4.55 tulisi valittua, vaikka mikään interaktio ei olisi tilastollisesti merkitsevä viiden prosentin merkitsevyystasolla. Lisäksi voisi määritellä kolmen tai neljän tai useamman muuttujan interaktioita. Pelkästään merkitsevyyden perusteella interaktioita ei siis voinut valita. Yksittäisistä muuttujista tilastollisesti merkitsevimpiä olivat seuraavat piiri $4.4 \cdot 10^{-10}$, kunnan maatalousmaan suhde kunnanpinta-alaan $6.6 \cdot 10^{-8}$ ja leveysaste $1.3 \cdot 10^{-7}$. Näiden interaktioista vain leveysasteen ja skaalatun maatalousmaan interaktio oli merkitsevä, joten se oli ainoa interaktio, joka jäi malliin. Lisäksi interaktio on järkevä, sillä on mahdollista, että pohjoisen luonto on herkempi ravinteille ja muille päästöille. Tälläkin interaktiolla saatiin luokitteluvirhettä pienennettyä vain noin prosenttiyksikön verran. Useampien kuin yhden interaktion lisääminen ei pienentänyt luokitteluvirhettä.

Interaktio mice-funktiolla

Aineistoja imputoitiin 50 kappaletta ja imputointialgoritmia toistettiin 50 kertaa jokaista aineistoa imputoitaessa. Luvut ovat kohtalaisen suuria, mutta imputoitava aineisto oli sen verran pieni, että aikaa imputointeihin ei kulunut liikaa. `MICE`-imputoinnin perusversio imputoi ja mallintaa vain muuttujien suorat vaikutukset [10]. Käytettäessä `mice`-functiota interaktion imputointi tehdään niin sanotulla passiivisella imputoinnilla. Siinä datamatriisiin lisätään uusi sarake, johon imputoituja interaktion arvoja sijoitetaan. Interaktioita kokeiltaessa jokaiselle sellaisen kategorisen muuttujen arvolle, jonka interaktiota mallinnettiin, tehtiin oma dikotominen sarake, joka sai arvon 1, jos kyseisen muuttujan arvo oli tämä, ja muuten 0. Lisäksi menetelmätvektoriin kirjattiin menetelmä, joilla kyseiset sarakkeet imputoitiin.

5.5 Logistinen regressiomalli

Lopulliseen vesimuodostuman kunnostustarvetta ennustavaan logistiseen malliin otettiin luvussa 5.3 kerrotut ennustavat muuttujat. Niiden tilastolliset merkitsevyydet on esitetty taulukossa 3.

Taulukko 3: Regressiokerrointen estimaatit ja niiden 95 %:n luottamusvälit.

	Estimaatti	95 %:n luottamusväli
vakio	24	(17, 31)
leveysaste	-0.37	$(-0.48, -0.25$)
keskisyvyys	0.23	$(9.0 * 10^{-2}$,0.38)
suurin syvyys	$5.4 * 10^{-2}$	$(2.0 * 10^{-2}, 8.9 * 10^{-2})$
kunnan pinta-ala	$1.7 * 10^{-4}$	$(9.8 * 10^{-5}, 2.4 * 10^{-4}$)
kunnan väkiluku	$-5.3 * 10^{-6}$	$(-7.7 * 10^{-6}$, $-2.9 * 10^{-6})$
piiri	$-8.9 * 10^{-3}$	$(-1.1 * 10^{-2}$, $-6.4 * 10^{-3})$
korkeus merenpinnasta	$6.7 * 10^{-3}$	$(3.9 * 10^{-3}, 9.6 * 10^{-3})$
maatalousmaan osuus = M	129	(58, 200)
suuralue Helsinki	-1.4	$(-2.0, -0.83)$
suuralue pohjoinen	0.77	(0.46, 1.1)
suuralue etelä	-0.75	$(-1.1, -0.36)$
valuma-alueen peltoala = P	$-1.3 * 10^{-3}$	$(-2.4 * 10^{-3}, -2.5 * 10^{-4})$
P:n suhde valuma-alueen pinta-alaan	-0.17	$(-0.19, -0.15)$
leveysaste x M	-2.2	$(-3.3, -1.0)$

Koska muuttujien tyyppi ja kunnan kesämökkien määrä p-arvo jäi yli 0.05, niin ne poistettiin vielä mallista. Kun suppeampi logistinen malli so-

vitettiin, niin siinä jäi valuma-alueen metsäalan suhde valuma-alueen pinta-alaan muuttujan p-arvo suuremmaksi kuin 0.05, joten sekin poistettiin.

5.6 Mallin hyvyyden tarkastelu

Tässä luvussa kuvaillaan, miten hyvin malli sopi aineistoon ja arvioidaan, voiko sitä käyttää kunnostustarpeen ennustamiseen ja kuinka luotettava se on tällaiseen ennustamiseen. Lisäksi pohditaan, onko mahdollista, että malli on ylisovittunut.

Luokitteluvirhe

Sekaannusmatriisi laskettiin imputoitujen aineistojen yli siten, että kullakin iteraatiokierroksella laskettiin havainnolle ennuste, ennusteiden keskiarvo ja havainnon luokittelu keskiarvon perusteella ja tätä luokittelua verrattiin todelliseen luokkaan. Ehdollisesta jakaumasta laskettu luokitteluvirhe taulukon 3 mukaisella mallilla oli 52.4 % niin päin, että ennuste oli, että ei ole kunnostustarvetta, kun todellisuudessa kunnostustarvetta oli (taulukko 4). Toisin päin eli niin, että ennusteen mukaan vesimuodostumaa pitää kunnostaa, kun todellisuudessa kunnostustarvetta ei ole, oli 3.2 %. Mallilla, jossa oli lisäksi muuttujat tyyppi, kunnan kesämökkien määrä ja valuma-alueen metsäalan suhde valuma-alueen pinta-alaan, oli luokitteluvirhe niin päin, että ennuste oli ei kunnostustarvetta, vaikka todellisuudessa oli, 48.9 % ja virhe toisinpäin eli niin, että ennustettiin kunnostustarvetta, vaikka sitä ei ollut, oli 3.2 % (taulukko 5). Kun taulukon 3 malli sovitettiin täydellisten havaintorivien aineistoon saatiin luokitteluvirheeksi 75.1 % niin päin, että vesimuodostuma, joka oli kunnostustarpeessa, luokiteltiin sellaiseksi, jota ei tarvitse kunnostaa ja 19.8 % niin päin, että vesimuodostuma, joka ei tarvitse kunnostusta, luokiteltiin sellaiseksi, joka pitää kunnostaa (taulukko 6).

Taulukko 4: Sekaannusmatriisi taulukon 3 mukaiselle mallille.

kunnostustarve	todellinen kyllä	todellinen ei
ennuste kyllä	354	114
ennuste ei	390	3502

Taulukko 5: Sekaannusmatriisi mallille, jossa on lisäksi muuttujat tyyppi, kunnan kesämökkien määrä ja valuma-alueen suhteellinen metsäpinta-ala.

kunnostus tarve	todellinen kyllä	todellinen ei
ennuste kyllä	380	115
ennuste ei	364	3501

Taulukko 6: Sekaannusmatriisi täydelle mallille.

kunnostus tarve	todellinen kyllä	todellinen ei
ennuste kyllä	66	349
ennuste ei	199	1417

Mallin devianssi

Käytetyllä mallilla, jossa oli taulukon 3 muuttujat, saatiin lähes yhtä pieni devianssi kuin mallilla, jossa käytettiin kaikkia muuttujia ja samaa interaktiota. Kun malli sovitettiin imputoituihin aineistoihin, niin keskimäärin devianssi oli 2515 vapausasteella 4345, kun nolladevianssi oli 3984 vapausasteella 4359. Täydellisten havaintorivien analyysissä mallin devianssi oli 1202 vapausasteella 2016 ja nolladevianssi oli 1987 vapausasteella 2030.

Mallilla, jossa oli näiden lisäksi muuttujat tyyppi, kunnan kesämökkien määrä ja valuma-alueen suhteellinen metsäala, oli keskimääräinen devianssi 2410 vapausasteella 4329. Täydellisten havaintorivien devianssi oli 1159 vapausasteella 2002.

Kaikkia muuttujia ja samaa leveysasteen ja suhteellisen maatalousmaan interaktiota käyttävässä mallissa imputoituja aineistoja mallinnettaessa keskimääräinen devianssi oli 2390 vapausasteella 4316, kun nolladevianssi oli 3984 vapausasteella 4359. Täydellisten havaintorivien tapauksessa laajemman mallin devianssi oli 1133 vapausasteella 1980 ja nolladevianssi oli 1980 vapausasteella 2021.

Näiden perusteella suurimmalla ja toiseksi suurimmalla mallilla ei ole suurta eroa. Kaikki mallit eroavat nollamallista, joten taulukon 3 malli on riittävä.

Mallin AIC

Kun suppeampi malli sovitettiin imputoituihin aineistoihin, niin keskimääräinen AIC oli 2545, kun taas täydellisten havaintorivien suppeamman mallin AIC oli 1232.

Kun imputoituihin aineistoihin sovitettiin malli, jossa oli näiden lisäksi muuttujat tyyppi, kunnan kesämökkien määrä ja valuma-alueen suhteellinen metsäala, keskimääräinen AIC oli 2472. Täysien havaintorivien AIC oli tällä mallilla 1213.

Laajemmalla mallilla imputoitujen aineistojen keskimääräinen AIC oli 2478 ja laajemman mallin täysien havaintorivien AIC oli 1217.

Myös AIC:n perusteella taulukon 3 malli on paras.

Sensitiivisyysanalyysi

Yksittäisten muuttujien vaikutus β-kertoimiin tai luokitteluvirheeseen oli hyvin pieni, kun käytettiin mallia, jossa oli mukana vähintään taulukon 3 muuttujat. Tämän vuoksi sensitiivisyysanalyysissä käytettiin mallia, jossa oli vain muuttuja, jonka sensitiivisyyttä tarkasteltiin. Kun kunnostustarvetta mallinnettiin niin, että keskisyvyyteen joko lisättiin tai vähennettiin 1 metri, niin mallin muuttujien β-kertoimet muuttuivat vai vähän (taulukko 7). Tässä mallissa kunnostustarvetta mallinnettiin vain keskisyvyydellä.

Taulukko 7: Sensitiivisyysanalyysi muuttujalle keskisyvyys.

keskisyvyyden arvo	β_0	β_1
4.39	0.91	0.19
5.39	0.86	0.18
3.39	1.03	0.18

Kun kunnostustarvetta mallinnettiin niin, että suurimpaan syvyyteen joko lisättiin tai vähennettiin 10 metriä, niin mallin muuttujien β-kertoimet muuttuivat vain vähän (taulukko 8). Tässä mallissa kunnostustarvetta mallinnettiin vain suurimmalla syvyydellä.

Taulukko 8: Sensitiivisyysanalyysi muuttujalle suurin syvyys.

suurin syvyys	β_0	β_1
13.72	0.89	0.059
23.72	0.85	0.047
3.72	1.29	0.037

Kun kunnostustarvetta mallinnettiin niin, että vesimuodostuman korkeu-

teen merenpinnasta joko lisättiin tai vähennettiin 150 metriä, niin mallin muuttujien β-kertoimet muuttuivat vain vähän (taulukko 9). Tässä mallissa kunnostustarvetta mallinnettiin vain korkeudella merenpinnasta.

Taulukko 9: Sensitiivisyysanalyysi muuttujalle korkeus merenpinnasta.

korkeus merenpinnasta	β_0	β_1
143.09	-0.502	0.017
293.09	0.001	0.012
-6.91	0.288	0.011

Kun kunnostustarvetta mallinnettiin niin, että valuma-alueen peltopinta-alaan joko lisättiin tai vähennettiin 100 hehtaaria, niin mallin muuttujien β-kertoimet muuttuivat vain vähän (taulukko 10). Tässä mallissa kunnostustarvetta mallinnettiin vain valuma-alueen peltopinta-alalla.

Taulukko 10: Sensitiivisyysanalyysi muuttujalle valuma-alueen peltopinta-ala.

valuma-alueen peltopinta-ala	β_0	β_1
84.51	1.63	-0.0026
259.06	1.63	-0.0020
32.54	1.62	-0.0027

Kun kunnostustarvetta mallinnettiin niin, että valuma-alueen peltopinta-alan suhteelliseen osuuteen joko lisättiin 0.5 tai vähennettiin 0.01, niin mallin muuttujien β-kertoimet muuttuivat vain vähän (taulukko 11). Tässä mallissa kunnostustarvetta mallinnettiin vain valuma-alueen suhteellisella peltopinta-alalla. Ja lisäys sekä vähennys eivät olleet yhtä suuria siksi, että näin saatiin mallinnettava osuus realistiseen väliin 0 - 1.

Taulukko 11: Sensitiivisyysanalyysi valuma-alueen suhteellinen pinta-ala.

valuma-alueen suhteellinen pinta-ale	β_0	β_1
0.019	2.54	-0.218
0.52	2.58	-0.222
0.0074	2.54	-0.218

Puuttuvuuden tarkastelu

Koska mitään prosessia, joka tuottaisi aineistoon puuttuvuutta siten, että se riippuisi vasteesta, ei keksitty, oletettiin, että puuttuvuus on joko MAR- tai jopa MCAR-tyyppistä. Koska vasteen kaikki arvot olivat havaittu, puuttuvuuden riippuvuutta muista arvoista testattiin sovittamalla logistinen malli, jossa eri muuttujien puuttuvuutta mallinnettiin muuttujan ekologinen tila arvoilla, joka on se muuttuja, mistä varsinainen dikotominen vaste saatiin. Ainoastaan valuma-alueen tietojen puuttuvuus riippui ekologisen tilan arvoista, ja siinäkin tilastollisesti merkitsevä logistisen mallin p-arvo oli vain, kun ennustettiin erinomaisella ekologisella tilalla puuttuvuutta valuma-alueen tiedoissa. Dikotominen vaste saatiin viisiarvoisesta ekologinen tila -muuttujasta siten että, jos ekologinen tila oli tyydyttävä, välttävä tai huono, vaste sai arvon 0 eli kunnostustarvetta on, ja muutoin 1 eli ei kunnostustarvetta. Siten puuttuvuus ei riippunut suoraan varsinaisen dikotomisen muuttujan arvosta. Tosin valuma-alueen tietojen puuttuvuus ei ollut täysin riippumatonta. Tämä luonnollisesti vaikuttaa niiden käytön vasteen ennustamiseen.

6 Yhteenveto

Mallia sovitettaessa testattiin useita ennustavia muuttujia ja niiden parittaisia interaktioita. Moni-imputoituun malliin jäivät seuraavat muuttujat: leveysaste, keskisyvyys, suurin syvyys, kunnan pinta-ala, kunnan väkiluku, piiri, korkeus merenpinnasta, maatalousmaan osuus, suuralue Helsinki, suuralue pohjoinen, suuralue etelä, valuma-alueen peltoala, valuma-alueen peltoalan suhde valuma-alueen pinta-alaan sekä maatalousmaan osuuden ja leveysasteen interaktio. Tyyppimuuttujan poisjääntі tämän työn perusteella oli ehkä suurin yllätys, koska intuitiivisesti se on tärkeä muuttuja.

Malleja vertailtiin useilla kriteereillä. Mallissa käytettävät ennustavat muuttujat valittiin tilastollisen merkitsevyyden perusteella ja sen, kuinka paljon ne vaikuttivat luokitteluvirheisiin. Erilaisten mallien devianssit pienenivät, kun muuttujien määrää kasvatettiin. Mallien parametrien määrän huomioivan AIC:n mukaan taulukon 3 malli oli toisaalta riittävä. Tämän mukaan kaikkia muuttujia ei tarvitse käyttää kunnostustarpeen ennustamiseen.

Mallin käyttötarkoitus huomioiden tärkein hyödyllisyyttä mittaava arvo on luokitteluvirhe. Mallin β-kertoimet taas vaikuttavat suoraan luokitteluun ja luokitteluvirheeseen. Sekaannusmatriiseista laskettujen luokitteluvirheiden perusteella mallia ei suoraan kannata käyttää kunnostustarpeen mallintamiseen: Kaikilla malleilla luokitteluvirhe varsinkin niin päin, että malli ennustaa, ettei vesimuodostumaa tarvitse kunnostaa, kun kunnostustarvetta olisi, on liian suuri. Imputoimalla muuttujia saatiin luokitteluvirheitä jonkin verran pienennettyä. Muuttujien valinnassa olisi moni-imputoinnin vaikutuksen tarkastelu siten, että vaikuttaako valitut imputointimenetelmät siihen, mitä muuttujia ennustamiseen kannattaa käyttää, ollut vielä mielenkiintoista, mutta se rajattiin työn ulkopuolelle.

Taulukon 3 mukaisen mallin β-kertoimien muutokset sensitiivisyysanalyysissä mallinnettiin yksinkertaistetulla mallilla, koska isommilla malleilla ei kertoimiin tullut juurikaan eroja. Sensitiivisyysanalyysi tehtiin kaikille taulukon 3 muuttujille, joissa oli puuttuvaa tietoa. Jos imputoidut arvot ovat selvästi suurempia tai pienempiä kuin oikeat, niin sillä olisi vaikutusta mallin β-kertoimiin. Sensitiivisyysanalyysin perusteella malli toimii, ja jos puuttuvat tiedot imputoidaan hyvin, niin luokitteluvirhe on pienempi.

Tämän työn tavoite oli etsiä muuttujat, joilla voitaisiin helposti ja edullisesti mallintaa vesimuodostumien kunnostustarvetta logistista mallia käyttäen. Logistinen malli oli luonnollinen valinta, kun vaste on dikotominen. Jos haluttaisiin ennustaa myös, mihin suuntaa kunkin vesimuodostuman kunto kehittyy tulevaisuudessa, voisi logistinen sekamalli olla sopiva.

Muutkin luokittelumenetelmät olisivat voineet ehkä sopia tarkoitukseen. Niistä mahdollisia esimerkkejä ovat Bayes-luokittelijat. Nämä ovat luokit-

telumenetelmiksi yksinkertaisia ja nopeita. Päätöspuut olisivat helposti visualisoitavissa ja myös tilastotieteeseen vain vähän perehtyneiden nopeasti ymmärrettävissä. Päätöspuut kärsivät korrelaatiosta. Satunnainen metsä on luokittelualgoritmi, joka perustuu päätöspuiden kokoelmaan ja vähentää korrelaatio-ongelmaa. Neuroverkot on nimensä mukaan kehitetty aivojen mallintamiseen ja niitä on ”hypetetty”. Ne ovat kuitenkin epälineaarisia tilastollisia malleja, jotka sopivat luokitteluun [12].

Tämän työn tutkimusongelman ratkaisuun valittiin logistinen regressio, koska se mahdollisti moni-imputoinnin, sillä puuttuvan tiedon imputointi oli myös tärkeä osa työtä. Puuttuvan tiedon käsittely on tarpeen, koska tutkielman aineistosta puuttui havaintoja. Lisäksi täydellisesti havaitun aineiston hankkiminen olisi todennäköisesti hankalaa ja kallista myös jatkossa vastaavanlaisia aineistoja hankittaessa. Tosin jos osa kohteista pystyttäisiin muilla keinoin rajaamaan kunnostettavien kohteiden ulkopuolelle, niin puuttuvan tiedon määrää pystyttäisiin vähentämään.

Logistinen malli saattaa olla riittävä kunnostustarpeen mallintamiseen. Jotta saataisiin malli, jonka käyttö on järkevää, olisi tiedettävä, kuinka kallista ja aikaa vievää jokaisen käytettävän muuttujan tietojen hankinta on. Silloin tiedettäisiin, miten mallilla voidaan säästää aikaa ja rahaa. Hyvä olisi, jos malliin löydettäisiin joitakin uusia hyvin ennustavia muuttujia, luokitteluvirhe saattaisi pienentyä. Nykyiselläkin mallilla voitaisiin ehkä joissakin olosuhteissa päästä rahan ja ajan säästöön, mutta eräs vaihtoehto olisi valita mallin muuttujat suoraan niiden hinnan perusteella ja laajentaa mallia kalliimpiin muuttujiin vain, jos edullisempi malli ei toimi.

Viitteet

[1] Aroviita, J., Mitikka, S. & Vienonen, S: *Suomen ympäristökeskuksen raportteja 37, Suomen ympäristökeskus, 2019.*

[2] Koski, V., Kärkkäinen, S. & Karvanen, J: *Subsample selection methods in the lake management. Käsikirjoitus, 2024.*

[3] *Maanmittauslaitos 2020*

`https://www.maanmittauslaitos.fi/sites/maanmittauslaitos.fi/files/attachments/2020/01/Vuoden_2020_pinta-alatilasto_kunnat_maakunnat.xlsx`

[4] *Luonnonvarakeskus 2017*

`http://statdb.luke.fi/PXWeb/pxweb/fi/LUKE/LUKE__02%20Maatalous__04%20Tuotanto__22%20Kaytossa%20oleva%20maatalousmaa/02_Kaytossa_oleva_maatalousmaa_kunta.px/`

[5] *Tilastokeskus 2017*

`http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__asu__rakke/statfin_rakke_pxt_116j.px/table/tableViewLayout1/`

[6] *Suomen ympäristökeskus 2017*

`http://rajapinnat.ymparisto.fi/api/jarvirajapinta/1.0/`

[7] Gelman, A. & Hill, J: *Data Analysis Using Regression and Multilevel/Hierarchical Models,* Cambridge University, 2007.

[8] Dobson, A. J: *An Introduction to Generalized Linear Models,* Chapman I& Hall/CRC, 2002.

[9] Aho, K. Derryberry, D. & Peterson, T: Model selection for ecologists: the worldviews of AIC and BIC, *Ecological Society of America*, 95(3), 631-636, 2014.

[10] van Buuren, S: *Flexible Imputation of Missing Data.* Chapman & Hall/CRC, 2018.

[11] Seaman, M, Galati, J, Jackon, D. & Garlin, J: What is meant by "missing at random"? *Statistical Science*, 28(2), 257-268, 2013.

[12] Hastie, T., Tibshirani, D. & Friedman, J. H: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2. painos), Springer, 2009.