Heikki Salo

Regressiomenetelmiä viljapellon biomassan estimointiin ortokuvista ja digitaalisesta korkeusmallista

Tietotekniikan
kandidaatintutkielma
11. kesäkuuta 2012

JYVÄSKYLÄN YLIOPISTO
TIETOTEKNIIKAN LAITOS

Jyväskylä

Tekijä: Heikki Salo
Yhteystiedot: heikki.ao.salo@iki.fi
Työn nimi: Regressiomenetelmiä viljapellon biomassan estimointiin ortokuvista ja digitaalisesta korkeusmallista
Title in English: Regression methods for estimating crop field biomass using orthophotographs and digital surface model
Työ: Tietotekniikan kandidaatintutkielma
Sivumäärä: 30
Tiivistelmä: Tutkielmassa esitellään käyttötarkoitus biomassan estimoinnille ja vertaillaan kolmea regressiomenetelmää, lineaarista regressiota, k:n lähimmän naapurin menetelmää sekä tukivektoriregressiota. Tutkielmassa esitellään myös aineisto ja aineistoon suoritetut muunnokset.
Abstract: Study presents purpose for biomass estimation and compares three learning methods: linear regression, k-nearest neighbours and support vector regression. Study describes also the used data with the needed transformations.
Avainsanat: biomassa, konenäkö, lineaarinen regressio, k-NN, tukivektoriregressio, estimointimenetelmät
Keywords: biomass, computer vision, linear regression, k-NN, support vector regression, estimation methods

Esipuhe

Tämä opinnäyte on tehty osana Tekes-rahoitteista UASI (Unmanned Aerial System Innovations) -hanketta, joka tutkii hyperspektrikameraa hyödyntäviä kaukokartoitussovelluksia maa- ja metsätalousympäristöihin. Tutkielma keskittyy hankkeen yhteen osakokonaisuuteen, viljapellon biomassan mittaukseen ja vertailee kolmea käytettävissä olevaa regressiomenetelmää. Kandidaatin tutkielman lisäksi hankkeesta on syntynyt myös useita konferenssijulkaisuja, joihin on viitteitä tutkielman lopussa.

Kiitokset hankkeen kaikille osapuolille mahdollisuudesta oppia ja olla mukana monipuolisessa tutkimushankkeessa. Erityiset kiitokset ohjaajalleni tutkijatohtori Ville Tirroselle, jolta olen hankkeen aikana 2010-2012 saanut paljon johdatusta konenäköön, oppimismenetelmiin ja Haskell-ohjelmointikieleen. Kiitokset myös Sami Äyrämölle opettavaisista keskusteluista tutkielman aiheisiin liittyen.

Sisältö

1 Johdanto

Biomassojen estimointi viljapellolta ilmakuvien avulla on yksi miehittämättömiä lentolaitteita hyödyntäviin innovaatioihin keskittyvän UASI-hankkeen (*Unmanned Aerial System Innovations*) tutkimuskohteista [18]. Tutkielma keskittyy kaukokartoituksessa ja konenäössä käytettyihin regressiomenetelmiin ja esittelee kolme estimointimenetelmää sekä niillä biomassan estimoinnissa saatuja tuloksia.

Tutkimuskysymyksenä on selvittää käytännössä sekä ominaisuuksia vertailemalla, mikä testatuista menetelmistä soveltuu soveltuu estimointitehtävään parhaiten. Tutkimuksen pitämiseksi kevyenä menetelmiin valitaan kirjallisuuden perusteella järkevät esikäsittelyt (kuten aineiston skaalaus), ja tutkielman lopussa vain pohditaan erilaisten variaatoiden vaikutuksia.

Tarkkuusviljely tarkoittaa mittaustiedon hyödyntämistä maanviljelyn hoitopäätösten tukena. Ennen lannoitusta pellosta muodostetun mittaustiedon avulla voidaan esimerkiksi muodostaa paikkakohtaisesti optimoituja hoitokarttoja lannoituksen tai kasvinsuojeluaineen jakeluun [4]. Lentolaitejärjestelmään pohjautuvassa mittaustiedon keruujärjestelmässä on useita hyötyjä vaihtoehtoihin nähden [3]. Tutkielmassa käytettyä ilmakuviin ja niistä laskettuun maaston korkeusmalliin pohjautuvaa lähestymistapaa verrataan vaihtoehtoihin luvussa 3.

Luku 4 esittelee estimointimenetelmien toimintaa ja kuvailee tutkielmassa käytetyt menetelmät, lineaarisen regression, k:n lähimmän naapurin regression ja tukivektoriregression. Menetelmistä k-NN on käytössä UASI-hankkeen osapuolilla, mikä tekee sen mukaanottamista vertailuun mielenkiintoista.

Tutkimusaineistona ovat Vihdissä sijaitsevasta MTT:n Hovin koepellosta otetut lähi-infrakanavan ilmakuvat, joista Geodeettinen laitos on muodostanut georeferoitavissa olevat kuvakaistat sekä korkeusmallin. Tutkimusaineisto esitellään luvussa 5. Luvussa 6 kuvaillaan tarkemmin tutkimuksessa käytetyt menetelmät, aineistolle tarvitut muunnokset sekä menetelmien implementaatiot.

Luku 7 esittelee tukimuksen tulokset ja vertaa niitä kirjallisuuteen. Tutkimuksessa havaittiin yllättäen, että lineaarisella kuvauksella saavutettiin tässä regressiotehtävässä parhaimman tulokset. Estimointimenetelmien tulokset poikkeavat kuitenkin kohtuullisen vähän toisistaan.

Luku 8 sisältää yhteenvedon ja pohdintaa. Pohdinnassa otetaan erityisesti kantaa tutkimuksessa suoritetun vertailun reiluuteen. Pohdinnassa nostetaan esille tässä tutkielmassa huomioimatta jääneitä ja tuloksiin mahdollisesti vaikuttavia aiheita, kuten piirteiden valintaa ja mahdollisia esikäsittelyitä.

2 Termejä

Luvussa kuvataan tutkielmassa käytetyt termit. Termeistä kerrotaan myös englanninkieliset vastineet ja mahdoliset lyhenteet, jotta tutkielma on asetettavissa helpommin englanninkielisen materiaalin kontekstiin.

estimaatti, ***estimate*** valistunut arvaus totuudesta, tässä yhteydessä piirteiden pohjalta ennustettu biomassan määrä (grammoja)

hyperspektrikuva, ***hyperspektral image*** lukuisia (esim. yli 20) säännöllisin etäisyyksin kuvattuja aallonpituuskaistoja sisältävä kuva (vrt. multispektrikuva)

kaukokartoitus, ***remote sensing*** ei-tuhoavaa mittausta, tässä yhteydessä ilmasta käsin tehtyä

korkeusmalli, ***Digital Elevation Model*** **(DEM)** kasvuston korkeutta maanpinnasta kuvaava pintamalli

miehittämätön lentolaite, ***Unmanned Aerial Vehicle*** **(UAV)** miehittämätön ilma-alus

miehittämätön lentolaitejärjestelmä, ***Unmanned Aerial System*** **(UAS)** UAV:n rinnalla käytössä oleva, laajemmin lentolaitejärjestelmää kuvaava termi

multispektrikuva, ***multispectral image*** useita aallonpituuskaistoja (värikanavia) sisältävä kuva

näytetiheys maastossa, ***Ground Sample Distance*** **(GSD)** pikselien koko maastossa (metrejä)

ortokuva, ***ortho image*** kartan projektioon oikaistu ilmakuva, johon maaston korkeuseroista johtuvat mittakaavaerot eivät vaikuta [6]

piirre, ***feature*** mielenkiinnon kohdetta kuvaileva suure

pintamalli, ***Digital Surface Model*** **(DSM)** korkeustietoa kuvaava kartta

RMSE, ***root mean square error*** keskineliövirheen neliöjuuri on tilastollinen tunnusluku, jota käytetään kuvaamaan estimaattien laatua, esitelty luvussa 6.8

tarkkkuusviljely, ***precision agriculture*** **(PA)** mittaustiedon hyödyntäminen maanviljelyn hoitopäätösten tukena

tekstuuri, ***texture*** kuvan sävyjen rakenteita kuvaava piirre

3 Tarkkuusviljelystä ja biomassan estimoinnista

Luvussa esitellään aihealue, jossa laskennallisia menetelmiä myöhemmin sovelletaan. Luvussa käydään läpi käytettyjä menetelmiä biomassan estimoimiseksi ja verrataan niiden lähtökohtia aineiston lähtökohtiin. Luvun tavoitteena on kuvata tutkimuksessa käytetyn menetelmän suhde tutkimuksissa ja teollisuudessa käytettyihin menetelmiin.

Kaukokartoitusmenetelmissä sovelletaan erilaisia estimointimenetelmiä, erilaisia laitteistoja, erilaisissa olosuhteissa, erilaisten kaukokartoistustehtävien ratkaisemiseksi. Tämä tutkielma keskittyy vertailemaan estimointimenetelmiä ja kuvaa lyhyesti kaukokartoitustehtävässä käytetyt menetelmät seuraavissa luvuissa.

3.1 Taustaa

Maanviljelyssä hoitopäätökset tehdään perinteisesti viljapellolla kävelyjen ja satunnaisten näytteiden pohjalta. Näytteistä voidaan mitata esimerkiksi maaperän tyyppipitoisuutta tai kosteutta. Tarkkuusviljelyn yleistyessä maatalousteollisuudessa on tapahtunut suuria muutoksia. [8]

Erityistä tarkkuusmaanviljelyä alettiin kehittää 1980-luvulla maanviljelyn tuottavuuden parantamiseksi, minkä jälkeen sitä on omaksuttu käyttöön vaihtelevissa muodoissa useissa eri maissa. Esimerkiksi Yhdysvalloissa tarkkuusviljelyn menetelmien hyödyntäminen lannoituksen annostelussa vaihteli 2002 alueittain muutaman prosentin ja neljänkymmenen prosentin välillä. [15]

Viljapellon satoon merkittävästi vaikuttavista tekijöistä voidaan erotella muun muassa maaperän vaihtelu, maan muodon vaihtelu ja kylvössä tehdyt valinnat esimerkiksi siemen- ja lannoitemäärissä. Kaukokartoituksen ohella mittaustiedon keräämiseen on käytetty erilaisia maahan tai liikkuviin laitteisiin sijoitettuja sensoreita. [12]

Hinta on suurin tarkkuusviljelyn käyttöönottoa vähentävä tekijä [15]. Lannoitussuunnitelmien tekemiseksi koko peltoa ei ole kustannus- ja aikataulusyistä mahdollista tutkituttaa laboratoriossa. Kaukokartoitusmenetelmät tarjoavat suuria katettuja pinta-aloja halvemmalla, mutta vähemmän tarkasti.

Hoitopäätösten tueksi tarvittavan kasvustotiedon keräämiseksi kaukokartoitusmenetelmillä on potentiaalia muuttua tutkimustoiminnasta liiketoiminnaksi. Esimerkiksi Ranskassa viinitilat hyödyntävät palvelua, joka suosittelee satelliittikuva-aineistosta estimoidun tiedon perusteella viinipensaille optimaalisen poimimisajan

sekä varoittaa mahdollisista kasvitaudeista. [20]

3.2 Lähestymistapoja biomassan kaukokartoitukseen

Tutkielmassa käytetään estimointimointitehtävän opetus- ja vertailuaineistona tietoja koepellolta kerätyistä näytteistä (engl. *ground truth*). Pellolta otettujen näytteiden biomassa on määritetty laboratorio-olosuhteissa (tarkemmin luvussa 5).

Kaukokartoituksessa käytettävissä on erilaisia kuvantamisvälineitä sekä keilausmenetelmiä. Kuvantamisvälineiden lisäksi niiden sijainnilla on suuri merkitys. Esimerkiksi LandSat 7 -satelliitin monikanavaisen ETM+-kameran spatiaalinen resoluutio (erottelukyky maanpinnalla) on 30 metriä seitsemälle ja 15 metriä yhdelle kaistalle [13].

Laserkeilaus ei ole vielä yleistynyt maataloussovelluksissa [4]. Laserkeilausvälineistön lennättäminen vaatii tällä hetkellä painonsa takia miehittämättömän lentolaitteen sijaan pienen lentokoneen, ja on siten huomattavasti kalliimpaa. Kasvukaudella myös ajamista pellolla halutaan välttää. Miehittämättömän lentolentolaitteella otettujen ilmakuvien perusteella muodostettu pintamalli on siten kiinnostava tapa kaukokartoitusaineiston saamiseksi.

Vertailun vuoksi tässä tutkielmassa miehittämättömällä lentolaitteella lennätetyllä VTT:n kehittämällä hyperspektrikameralla saavutetun kuva-aineiston spatiaalinen resoluutio on 20 *cm*. Hyperspektrikameraa suuriresoluutioisemmalla, mutta vain NIR-kanavaa tallentaneella kameralla saatiin spatiaaliseksi resoluutioksi kolme senttimetriä.

Kasvillisuutta voidaan aineistosta riippuen tutkia muun muassa erilaisialla heijastumilla (kasvillisuusindeksit), kuvien tekstuuripiirteillä, maaston pintamallin tai edellisistä edelleen johdettujen piirteiden avulla. Yleisiä valintoja kasvillisuusindekseiksi ovat tässä tutkielmassa kokeillut NDVI *Normalized Difference Vegetation Index* [8] sekä SR *Simple Ratio* [8], jotka voidaan laskea näytteelle kaavoilla

$$NDVI = \frac{NIR - VIS}{NIR + VIS},$$

sekä

$$SR = \frac{NIR}{VIS},$$

joissa NIR ja VIS ovat infrapunan ja punaisen spektrialueiden mittausarvot. NDVI:n sekä lehtien osuutta pinta-alasta kuvaavan LAI *Leaf Area Index*:n arvojen vaih-

telusta voi päätellä maaston klorofyllipitoisuutta. Tässä tutkielmassa estimoitavana suureena on biomassa, joka riippuu indeksien arvoista vain välillisesti [8]. Kasvillisuusindeksien käytöstä tässä tutkimuksessa lisää luvussa 6.6.

3.3 Vastaavia kaukokartoitustuloksia

Saksalaisessa tutkimuksessa [4] tutkittiin biomassan estimoimista erilaisista peltotyypeistä **laserkeilauksen** avulla. Aineistoltaan 2008 suoritettu laserkeilaus antoi samankaltaiset lähtökohdat pintamallin tutkimiseen kuin tässä tutkimuksessa on käytettävissä. Instrumentti liikkui pellolla ajoneuvoon kiinnitettynä ja pystyi näin mittaamaan pulssien koordinaatit erittäin tarkasti. Testiaineistona oleviin kuivapainonäytteisiin verrattuna erot ja estimoiduissa kg/m^2 -luvuissa tutkimusryhmä pääsi alle 200 gramman keskivirheisiin.

Tutkimuksessa ilmakuvin toteutettu kaukokartoitus on ei-tuhoava tiedonkeruumenetelmä. Tuhoavista menetelmistä esimerkkinä on tämän tutkielman opetusmateriaali, joka on muodostettu kaivamalla maasta lyhdenäytteitä ja selvittämällä niiden kuivapainot. Samaa peltoa on myös tutkittu tutkimuksissa [14][19] erilaisin menetelmin.

4 Regressiomenetelmistä

Luvussa annetaan yleiskuva tilastollisista oppimismenetelmistä. Tämän luvun tavoitteena on antaa lukijalle yleiskuva oppimismenetelmistä ja lähtökohdat erilaisten menetelmien vertailuun. Seuraavassa luvussa 4 esitellään yksityiskohtaisemmin tutkimuksessa käytetyt menetelmät. Käytettyjen menetelmien lisäksi tämä luku kuvailee lyhyesti myös muita mahdollisia menetelmiä.

4.1 Mitä on estimointi?

Estimoinnissa on kyse aineistosta oppimisesta. Tavoitteena on muodostaa oppija (engl. *learner*), joka oppii tunnetuista koealoista kvantitatiivisen suureen (esimerkiksi grammaa/pinta-alayksikkö) laskettujen piirteiden (engl. *features*) avulla [5]. Tällaisessa estimoinnissa on kyse valvotusta oppimisesta (engl. *supervised learning*), sillä tieto estimoitavasta suureesta ohjaa oppimisprosessia [5]. Tällaista ongelmaa sanotaan myös regressio-ongelmaksi (engl. *regression problem*), koska ohjatun oppimisen tuloksena on kvantitatiivinen suure [5].

Tavoitteena on siis muodostaa lasketuista piirteistä mahdollisimman tarkka kuvaus kohdeattribuuttiin[1]. Tutkielmassa käytetään merkintätapaa, jossa piirrevektoriin viitataan käyttäen vahvennettua kirjainta. Yksittäiseen piirteeseen viitataan joko vahventamattomalla kirjaimella ja indeksinumerolla tai vahvennetulla kirjaimella yhdessä sulkuoperaattorin kanssa, jonka alaindeksissä on piirteen numero:

$$x_i = (\mathbf{x})_i \tag{1}$$

Näin voidaan erikseen viitata eri piirrevektoreihin käyttäen alaindeksejä, eikä merkintätapa sekoitu yksittäisiin piirteisiin (kuten kaavassa 2). Lisäksi, matriiseihin viitataan vahvennetulla isolla kirjaimella, kuten $\mathbf{A}$. Näitä merkintöjä käyttäen myöhemmin tässä luvussa esitellyt regressorit f ovat kuvauksia piirrevektoreista $\mathbf{x}$ estimaateiksi y ja virheiksi e:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{2}$$

jossa kuvauksella f piirteistä $\mathbf{x} \in \mathbb{R}^p$ saatu estimaatti poikkeaa todellisesta arvosta $y_i \in \mathbb{R}^1$ virheen $\epsilon_i \in \mathbb{R}^1$ verran. Varsinkaan kaukokartoituksessa kuvaukset eivät

[1] Tässä tutkielmassa kohdemuuttuja on biomassa ($g/0.1m^2$), jonka arvoja opitaan estimoimaan 33 $cm * 33$ cm koealoista laskettujen piirteiden avulla. Piirteistä on kerrottu lisää luvussa 6

ole täydellisiä. Tässäkin tutkielmassa kuvauksien hyvyyttä mitataan laskemalla oppijalle ennestään tuntemattomalle ns. testausaineistolle laskettujen estimaattien keskimääräisiä virheitä. Aineiston jakaminen tutkielmassa opetus- ja testausaineistoksi on kuvattu luvussa 6.1.

Luvussa 4 esitellään kolme toisistaan hyvin erilaista estimaattoria. Lineaarisessa regressiossa tavoitteena on löytää numeerisille piirteille optimaaliset kertoimet (luku 4.2), kun puolestaan k:n lähimmän naapurin (luku 4.3) menetelmässä tuntemattomat koealat estimoidaan etsimällä opetusaineistosta kullekin k lähintä naapuria ja päättelemällä suureet näiden ominaisuuksista. Tukivektoriregressiossa aineistosta etsitään korkeaulotteisen aineiston sisältä vaihtelun jakava taso, jota käyttäen regressiota tehdään.

Numeeristen piirteiden vaihteluvälit ja omainaisuudet vaihtelevat, minkä takia piirteiden käyttö edellyttää jonkinlaista esikäsittelyä. Esimerkiksi lähimpien naapureiden etsintä laskee euklidisia etäisyyksiä näytteiden välillä, jolloin vaihteluvälin $[0, 1]$ ja $[1, 10000]$ omaavat piirteet olisivat eri asemassa näytteiden läheisyyttä määrittämisessä. Näin kävisi myös tukivektoriregressiossa, sillä tukivektorikoneen opetus tarkoittaa vaatii optimointitehtävän ratkaisemista, jossa minimoidaan piirteistä riippuvaa muuttujaa [1]. Tutkielmassa käytetty skaalaus esitellään myöhemmin luvussa 6.2.

4.2 Lineaariset menetelmät

Lineaarimalli on ollut tilastotieteen tukipilari viimeiset 30 vuotta [5]. Siinä annetulle vektorille syötteitä (tuntemattomasta koealasta lasketuille piirteille)
$\mathbf{x} = (x_1, x_2, \ldots, x_p)$ ennustetaan haluttua arvoa (biomassaa) y mallilla

$$f(\mathbf{x}; \boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{p} x_i \beta_i, \tag{3}$$

jossa $\boldsymbol{\beta}$ on kerroinvektori, jonka β_0 on kuvauksen epäkeskisyys (engl. *bias*) [5]. Kerrointen $\boldsymbol{\beta}$ löytämiseksi on monia tapoja, mutta ylivoimaisesti yleisin on pienimpien neliösummien (engl. *least squares*) metodi. Siinä kertoimet $\boldsymbol{\beta}$ etsitään siten, että jäljelle jäävät neliösummat (RSS) minimoidaan N näytteen mittaisen opetusaineiston kesken:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{N}(y_i - f(\mathbf{x}_i; \boldsymbol{\beta}))^2 \tag{4}$$

Kertoimien etsimiseksi on tarjolla vaihtoehtoja. Tässä tutkielmassa käytetään algoritmia, joka ratkaisee kertoimet käyttäen pääakselihajotelmaa (engl. *Singular Value Decomposition*, SVD). Pääakselihajotelmassa piirrevektorit $\{\mathbf{x}_i\}_{i=1}^N$ ovat opetusaineiston näytemäärän N pituisessa matriisissa A, joka esitetään hajotelmana

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T, \tag{5}$$

jossa $\mathbf{U}$ (engl. *left singular vectors*) ja $\mathbf{V}$ (engl. *right singular vectors*) ovat ortonormaaleja matriiseja ja $\boldsymbol{\Sigma}$ diagonaalimatriisi, joka sisältää $\mathbf{A}$:n ominaisarvot, joita on piirteiden n määrän verran. Lineaarisen mallin ratkaisu pääakselihajotelmaa käyttäen määritellään suoraan hajotelman osien $\mathbf{U}$, $\boldsymbol{\Sigma}$ ja $\mathbf{V}$ avulla [23].

Lineaarisen ratkaisun löytyminen luonnollisestikin edellyttää, että tehtävään on lineaarinen ratkaisu. Lineaarimallin etuna on etenkin ratkaisun hyvä tulkittavuus, koska syötteiden vaikutus ennustettuun arvoon on selkeästi nähtävissä. Lisäksi pienillä näytemäärillä lineaariset menetelmät tehtävään soveltuessaan voivat tuoda etua kohinaisessa, harvassa tai vähänäytteisessä opetusaineistossa [5].

Tutkielman sovellusalueessa hyvän mallin löytämisessä ei ole erityisiä aikarajoja. Sen sijaan mallin soveltaminen eli sesonkiaikana kasvukauden alussa peltojen biomassojen estimointi on prosessi, jolta edellytetään nopeutta. Tässä suhteessa lineaarikuvaus olisi soveltuessaan menetelmistä ihanteellisin, sillä kuvauksen löytymisen jälkeen lineaarimallilla estimaattien laskeminen on erittäin nopeaa. Tämä on selkeä etu seuraavaksi luvussa 4.3 esiteltyyn k-NN-menetelmään nähden.

4.3 k-NN-regressio

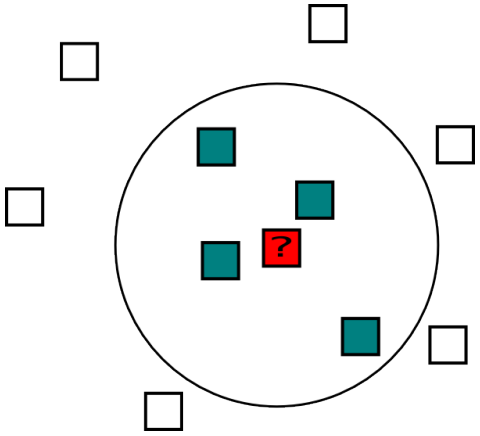
K:n lähimmän naapurin (engl. *k-nearest neighbours*) menetelmä on luokittelija, joka etsii tutkittavalle näytteelle k lähintä näytettä opetusaineistosta. K-NN-menetelmä edustaa näytepohjaista oppimista (engl. *instance-based learning*), jossa tietämys koostuu siis suoraan näytteistä itsestään [17].

Regressiota k-NN-menetelmällä tehdään keskiarvoistamalla tuntemattoman pisteen k:n lähimmän naapurin arvot tuntemattoman jatkuvan attribuutin selvittämiseksi [5]:

$$f(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i, \tag{6}$$

jossa $N_k(\mathbf{x})$ tarkoittaa näytteen $\mathbf{x}$ k:n lähimmän pisteen naapurustoa. Tuntemattoman pisteen $\mathbf{x}$ ominaisuudet päätellään siis lähimpien tiedettyjen pisteiden ominaisuuksien perusteella. [5]

Kaksiulotteisessa piirreavaruudessa näytteet voidaan esittää tasossa kuten kuvassa 1, jossa pisteiden etäisyys vastaa pisteiden etäisyyttä kuvassa. Tässä tutkielmassa käytetään seitsemää piirrettä, jolloin vastaavaa visualisointia ei voida suoraan tehdä, vaikka etäisyydet määritelläänkin samalla tavalla.



Kuva 1: Tuntemattoman pisteen (keskellä, kysymysmerkillä) neljä lähintä naapuria kaksiulotteisessa piirreavaruudessa.

K-NN-menetelmä ei tee tiukkoja oletuksia aineistosta, ja sopeutuu kaikkiin tilanteisiin [5]. Toisaalta, kaikkialla piirreavaruudessa uusien näytteiden ominaisuuksien arviointi riippuu vain tunnetuista pisteistä ja niiden tarkasta sijainnista, mikä tekee k-NN:stä epävakaan [5]. Mallin yleistäminen tuntemattomille alueille riippuu siis opetusaineistosta.

Esimerkiksi tutkielman $\mathbb{R}^7$-piirreavaruus piirteiden arvojen ollessa välillä $[0, 1]$ on miellettävissä 7-ulotteiseksi kuutioksi. Hyvä lähtökohta olisi olla ainakin yksi piste kutakin hyperkuution $2^7 = 128$ särmää kohden. Aineistona olevat 61 havaintopistettä eivät kuitenkaan välttämättä ole jakautuneet havainnoitavan arvon suhteen edullisesti. Tästä johtuen vähämmän pisteitä sisältävien alueiden ominaisuudet riippuvat kussakin sijainnissa niistä pisteistä, jotka 7 ulottuvuudessa sattuvat

olemaan lähimpinä. K-NN yleistyy siis tuntemattomillekin alueille, mutta sen ominaisuudet riippuvat täysin aineistosta. Huomionarvoista on myöskin, että k-NN:n estimaattien minimi- ja maksimiarvot ovat ääripäiden k:n reuna-arvon keskiarvot, eikä mitään piirrevektoria vastaava estimaatti voi olla yli tai alle näiden arvojen.

Laskennalliselta vaativuudeltaan k-NN:n kustannukset ovat yksinomaan estimointia tehdessä, sillä tuntemattomia alkioita verrataan kaikkiin opetusaineiston pisteisiin. Tämä ominaisuus voi olla sovelluskohteesta riippuen etu tai haitta. Tutkimuskäytössä k-NN:llä on nopea kokeilla erilaisia aineistoja, mutta esimerkiksi muuttumatonta koulutusaineistoa käyttäen nopeiden vasteaikojen saavuttaminen on k-NN:llä haastavaa. K-NN-menetelmän hyötyjä ovat sen implementoinnin helppous sekä tulosten tulkinnan selkeys.

4.4 Tukivektoriregressio

90-luvulla esitelty tukivektoriregressio [25] on tähän tutkielmaan valituista menetelmistä selkeästi uusin. Tukivektorikoneita (engl. *support vector machines*, SVM) voidaan käyttää sekä luokitteluun (engl. *support vector classification*, SVC) [5] että yleistettynä myös regressiotehtävien ratkaisuun (engl. *support vector regression*, SVR) [25].

Tukivektorimenetelmistä on muutamia erilaisia variaatioita [21]. Tutkielmassa käytetty `libsvm`-implementaatio [2] tarjoaa jatkuva-arvoisten funktioiden approksimointiin Vapnikin [25] esittelemää *Epsilon SVR*:ää.

Aineiston oletetaan olevan riippumatonta sekä identtisesti jakautunutta [21]. Karkeasti voidaan sanoa, että tukivektorimenetelmä etsii opetusjoukon piirreavaruudesta ne pisteet, jotka eivät ole täysin määritellyllä jakotasolla mutteivat toisaalta ole täysin vieraita havaintoja (engl. *outlier*) [5]. Tässä tutkielmassa aineisto koostuu biomassoista $y \in \mathbb{R}^1$, jota pyritään estimoimaan piirreavaruuden $\mathbf{x} \in \mathbb{R}^p$ avulla.

Tarkistellaan ensin lineaarimallia

$$f(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{x}^T \boldsymbol{\beta} + \beta_0, \tag{7}$$

jossa vektorin $\boldsymbol{\beta}$:n ratkaisemiseksi muotoillaan minimointitehtävä

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{N} V(y_i - f(\mathbf{x}_i; \boldsymbol{\beta})) + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2, \tag{8}$$

jossa V on alla kaavassa 9 esitelty häviöfunktio. Tyypillisesti tukivektoriregressiossa aineiston informaatio esitetään pienen siitä etsityn osajoukon avulla [21]. Os-

ajoukon etsimiseksi Vapnik kehitti epsilonin kokoisen virheen sallivan häviöfunktion (*ϵ-insensitive loss function*):

$$V_\epsilon(r) = max\left\{0, r - \epsilon\right\}, \tag{9}$$

joka ei rankaise etukäteen valittua epsilonia $\epsilon > 0$ pienemmistä virheistä [25]. Kun optimointitehtävä on ratkaistu, ratkaisussa piirreavaruuden vektoreihin on liitetty kertoimia, joista tyypillisesti vain osa poikkeaa nollasta [5]. Nämä vektorit ovat tukivektoreita, jotka kuvailevat aineiston optimaalisella tavalla annettuihin parametreihin nähden [5].

Tukivektoreilla esitetyn ratkaisun kompleksisuus datan erottelutehtävässä riippuu estimoidun funktion kompleksisuudesta ja estimoinnin tarkkuudesta. Ratkaisun kompleksisuus ei siis riipu suoraan ongelma-avaruuden dimensiosta.[25]

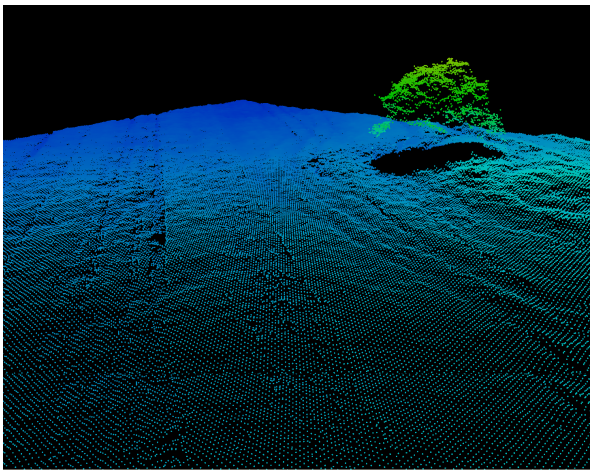
Tukivektoriregression yleistämiskyky riippuu hyvästä parametrien C, ϵ, kerneli-tyyppi ja kerneliin liittyvien parametrien valinnasta. C-parametrin valinta vaikuttaa trade-offiin tukivektorimallin kompleksisuuden ja ϵ:tä suurempien virheiden sallimisen välillä. Esimerkiksi liian suuri C-arvo aiheuttaa koulutusaineistoon ylisovittumista, liian pieni ei hyödynnä sitä tarpeeksi. Optimaalinen epsilon-arvo puolestaan riippuu aineistossa esiintyvän kohinan määrästä. [26]

Tukivektoriregressiossakin opetuksesta tulokseksi jää malli, joka on laskennallisesti kevyt kuvaus piirrevektoreista estimaateiksi. Heikkoutena mallissa suhteessa lineaarikuvaukseen ja k-NN:ään nähden on mallin kahteen muuhun menetelmään verrattuna huomattavasti hankalampi tulkittavuus. Kuten luvun 4.2 lineaarinen kuvaus, myös tukivektoriregressio olisi laskennallisen vaativuuden jaksottumiseltaan houkutteleva regressiomenetelmä sovellusalueeseen. Tukivektorikoneen ja muiden menetelmien implementaatioista lisää luvussa 6.

5 Tutkimusaineisto

Tutkimusaineisto on kerätty Vihdissä sijaitseva Maa- ja elintarviketeollisuuden tutkimuskeskuksen (MTT) Hovin peltolohkolta. Peltolohko on kylvetty järjestämällä kasvustoon vaihtelua varioimalla lannoite-, kylvö- ja ruiskutusmääriä. Kuvausalue peltolohkolla oli kooltaan 3 hehtaaria.

Kuva-aineisto hankittiin miehittämättömillä lentojärjestelmillä (engl. *unmanned aerial vehicle*, UAV), jotka kuvasivat alueen Panasonic Lumix NIR-kameralla sekä VTT:n kehittämällä hyperspektrikameralla [9].



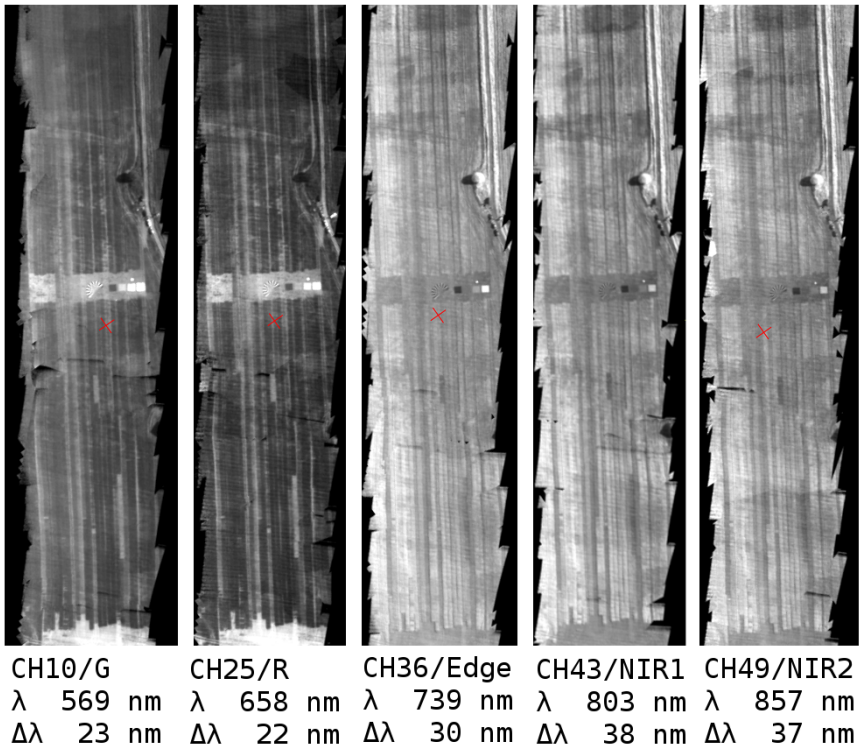
Kuva 2: Näkymä geodeettisen laitoksen NIR-kuvista muodostamaan 3D-pistepilveen, josta on erotettavissa mm. puu sekä traktorinjälkiä.

Piirteytysaineistona ovat kirjoitushetkellä käytettävissä olleet geodeettisen laitoksen Hovin pellosta muodostamat

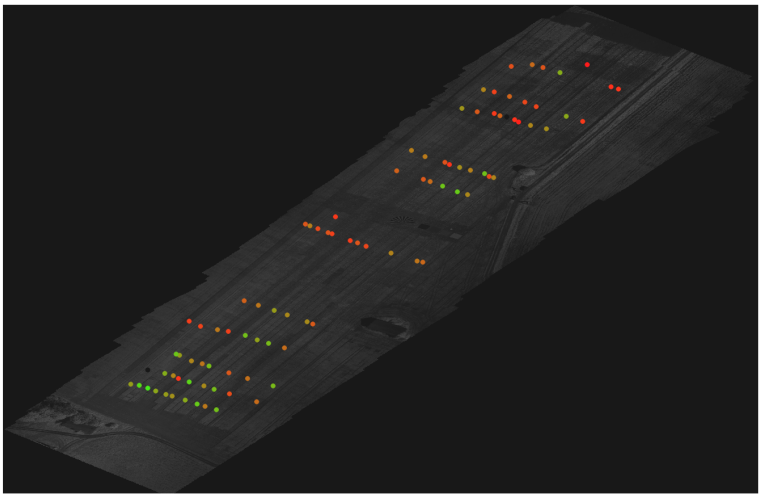
- lähi-infrakanava (GSD 3 *cm*) (kuva 4)
- ilmakehäkorjattu multispektriortokuva (GSD 20 *cm*) (kuva 3) sekä
- digitaalinen pintamalli pistepilvenä (GSD 3 *cm*, RMSE korkeudelle 10-20 *cm*) (kuva 2).

Opetus- ja testausaineistona on lisäksi MTT:n peltolohkolta teettämät kuivabiomassamittaukset, joissa neliönmuotoiselta 33 *cm* x 33 *cm* -koealalta (0.1 m^2) on kaivettu kaikki biomassa. Näytteet oli kuivattu uunissa ennen punnitusta.

Menetelmien testausta varten 91 koealan aineisto jaettiin opetus- ja testausaineistoksi, 61 opetuskoealaan ja 30 testauskoealaan. Perusteet jaolle ja jaon toteutus on kuvattu luvussa 6.1.



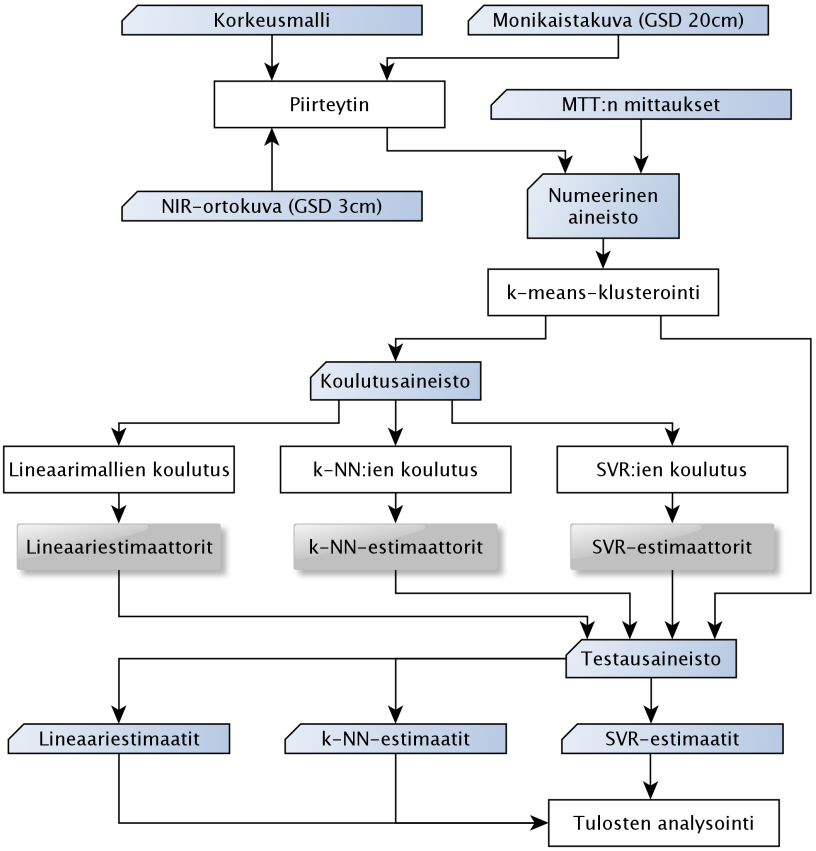
Kuva 3: Hovin hyperspektriaineistosta luodun monikanavakuvan viisi kaistaa 10,25,36,43 ja 49 vierekkäin, alla spektrikaistojen keskustat (λ) ja kaistojen leveydet ($\Delta\lambda$).



Kuva 4: Geodeettisen laitoksen muodostama NIR-ortokuva peltolohkosta, jonka yl-le on lisätty 91 biomassamittausarvoa (punainen vähän biomassaa, vihreä paljon). Kuva on pohjois-itäsuunnassa, toisin kuin kuva 3.

6 Metodit

Luvussa kuvaillaan käytetyt menetelmät ja tutkimuksessa tehdyt valinnat sekä niiden implementaatiot.



Kuva 5: Mukailtu tietovuokuva aineiston käsittelystä

6.1 Opetus- ja testausaineiston valinta

Käytössä olleet 91 koealaa jaettiin 61 koealan kolutusaineistoksi ja 30 koealan testausaineistoksi käyttäen *k-means*-klusterointialgoritmia [7]. Klusterointi jakaa aineiston piirreavaruudessa ryhmiin, joiden jäsenet muistuttavat toisiaan. Kun nämä ryhmät edelleen puretaan tasapuolisesti kahdeksi ryhmäksi, opetus- ja testausaineistoksi, voidaan toivoa, ettei jokin ominaisuusryhmä ylikorostu toisessa aineistossa. K-means-toteutuksena käytössä oli Haskell-kielelle tarjolla ollut `kmeans-vector`-kirjasto [11].

Vaikka jakoperuste on järkevä, olisi mielenkiintoista kokeilla opetus- ja testausaineiston jakoperusteen vaikutusta lopputuloksiin. Tätä voisi tutkia generoimalla sa-

tunnaisesti erilaisia opetus- ja testausaineistojakoja.

6.2 Skaalaus

Tässä tutkielmassa skaalataan kaikki piirteet i vuorotellen välille $[0, 1]$ vähentämällä niistä minimi ja jakamalla opetusaineiston X vaihteluvälillä

$$x_i' = \frac{x_i - \min_{\mathbf{x} \in X}(\mathbf{x})_i}{\max_{\mathbf{x} \in X}(\mathbf{x})_i - \min_{\mathbf{x} \in X}(\mathbf{x})_i} \tag{10}$$

Testausaineiston skaalaus tehdään opetusaineiston vaihteluväliä käyttäen. Tämä on sovellusaluetta ajatellen tehty valinta, joka mahdollistaa skaalauksen tekemisen yksittäisellekin testausarvolle ja kuvaus alkuperäisistä piirrevektoreista $\mathbf{x}$ skaalatuiksi vektoreiksi regressiomallin käyttöä varten on varmasti sama kuin opetusaineistolla, jolla malli koulutettiin. Lopputuloksena opetusaineiston kukin piirre sisältää koealoissa minimit ja maksimit (0 ja 1). Opetusaineiston parametreillä skaalatun testausaineiston vaihteluväli ei kuitenkaan välttämättä pysy täysin samalla välillä.

Ylläoleva tapa ei ole kuitenkaan ainoa mahdollisuus skaalaukseen. Esimerkiksi keskiarvon vähentäminen ja keskihajonnalla jakaminen on yleinen tapa samanarvoistaa piirteet [1]. Sen käyttäminen olisi mahdollista tässäkin tapauksessa.

6.3 Valitut piirteet

Piirreavaruuden olemuksella on mahdollisesti paljonkin merkitystä käytettyjen menetelmien onnistumiselle. Piirrevalinnasta seuraavia ominaisuuksia pohditaan luvussa 8.

Piirre	**Nimi**	**Kuvattu luvussa**
1	Kanava 10 (green)	6.4
2	Kanava 25 (red)	6.4
3	Kanava 36 (red edge)	6.4
4	Kanava 43 (nir1)	6.4
5	Kanava 49 (nir2)	6.4
6	DSM, varianssi	6.5
7	DSM, vaihtelumaksimi	6.5

6.4 Sävypiirteet

Spektrikaistoista tehtyjen ortokuvien asemoinnissa oli kokeiden perusteella virhettä 1-2 metrin verran. Vaikka ortokuvien pikselit ovat georeferoituja, niitä ei asemointivirheen takia voi käyttää ilman esikäsittelyä. Tässä tapauksessa asemointivirhe huomioitiin tavalla, joka vastaa resoluution pienentämistä.

Raakakuvia käsiteltiin resoluutiolla 1 m^2, mikä tarkoitti 20 cm leveiden pikseleiden ottamista $5 * 5$ pikselin kokoisena alueena ja keskiarvoistamalla niiden intensiteettiarvoista yksi keskiarvo. Huomioitavan alueen koko ja keskiarvoistustapa ovat valintoja, joiden kyseenalaistaminen olisi jatkotutkimuksissa mielenkiintoista.

6.5 Pintamallipiirteet

Piirteinä käytettiin lisäksi kahta pintamallista (DSM) johdettua piirrettä. Pintamalli on pistepilvi, jossa pisteitä on hilana kolmen senttimetrin välein. 33 cm leveälle neliön malliselle koealalle osuu keskimäärin $11 * 11 = 121$ pistettä.

Koealan ympäristöön osuvista pisteistä johdettiin useita piirteitä, joista tässä tutkimuksessa päädyttiin käyttämään kahta. Alla esitellyt kaksi piirrettä paransivat suppeassa testauksessa tuloksia. Ajatuksia piirteiden valinnasta ja optimoinnista on kuvattu lisää lopussa luvussa 8.

Piirre 6, "DSM, varianssi" on määritelty täsmälleen koealan päälle osuvien pisteiden korkeusarvojen populaatiovarianssina.

Piirteen 7, "DSM, vaihtelumaksimi" tavoitteena on kuvata koealan korkeusvaihtelun maksimiarvoa suhteessa ympäristöön. Ympäristön huomioimiseksi koealan ympäristöstä etsitään 4 m^2 kokoisesta ruudusta pienimpiä arvoja, johon verrataan täsmälleen koealan kohdalta löytynyttä maksimiarvoa. Ympäröivän alueen minimi lasketaan etsimällä korkeusarvo, jota suurempia 90 % ympäröivän alueen arvoista on.

6.6 Kasvillisuusindeksit

Monikanavakuva mahdollistaa kasvillisuusindeksien käytön [8]. Käytössä ollut maapikselikoon 3 cm lähi-infrakanavan ortokuva oli radiometrisesti korjaamaton, mistä johtuen sävyarvot mosaiikin eri puolella eivät olleet keskenään verrannollisia. Maapikselikoon 20 cm monikanavakuvat sen sijaan olivat korjattuja ja sävyt ovat keskenään verrannollisia. Kasvillisuusindeksejä voitaisiin siis hyödyntää tässäkin

tutkimuksessa.

Kasvillisuusindekseillä vain välillinen suhde kasvillisuuden biomassan kanssa, kuten edellä luvussa 3.2 kuvattiin. Edellä esitellyt indeksit ovat yksinkertaisia suhdelukuja, joiden informaatio vaikutti tässä tutkimuksessa sisältyvän kanavakohtaisiin mittausarvoihin. Käsin piirteitä valitessa asvillisuusindeksien lisääminen aiheutti kaikkien estimaattien tuloksien heikkenemisen, joten ne jätettiin tässä tutkimuksessa käyttämättä piirteinä. Lisää ajatuksia piirteisiin liittyen luvussa 8.

6.7 Regressiomenetelmien implementaatiot

Tutkimuksessa käytettiin luvussa 4.2 esiteltyä lineaarikuvausta, luvussa 4.3 esiteltyä k-NN-menetelmää sekä 4.4 esiteltyä tukivektoriregressiota.

Pienimmät neliösummat antavan lineaarimallin etsimiseen käytettiin `LAPACK`-kirjastoa [23] (*Linear Algebra PACKage*) Haskell-ohjelmointikielelle rajapinnan tarjoavan `hmatrix`-paketin [16] avulla. Luvussa 4.2 kuvattu pienimpien neliösummien menetelmä ratkaistiin käyttäen hmatrix-kirjaston tarjoamaa `linearSolveSVDR`-funktiota (toteutuksena LAPACK:n `dgelss`) ja ratkaisee lineaarisen yhtälöryhmän käyttäen pääakselihajotelmaa. Lineaarikuvauksen etsiminen tällä tavoin ei vaatinut parametreja tai muita optimoitavia valintoja.

K-NN-menetelmä toteutettiin tutkimusta varten luvussa 4.3 esitellyn määritelmän pohjalta. Tuntemattomalle havainnolle k lähintä naapuria etsittiin laskemalla euklidiset etäisyydet kaikkiin opetusaineiston pisteisiin [5]. Optimoidun k-arvon löytämiseksi opetusaineisto jaettiin edelleen kahtia, käyttäen uuden aineiston toista puolikasta opetusaineistona ja toista testausaineistona parametrien etsintää varten. Optimaalisinta k-arvoa etsittiin väliltä $[2, 20]$, ja paras arvo valittiin käytettäväksi varsinaisen testausaineiston kanssa.

Tukivektoriregressiosta käytettiin Chih-Chung Chaning ja Chih-Jen Linin kehittämää `libsvm`-implementaatiota [2] `svm-simple`-paketin [22] avulla. Regressiotehtävän tyypiksi valittiin *Epsilon SVR* ja kerneliksi Radial Basis Function -kerneli (RBF), sillä se on yleinen valinta regressiotehtäviin [26]. Optimoitavia parametrejä tätä regressiotehtävää varten oli kolme: mallin kompleksisuutta säätelevä C, Epsilon-SVR:n epsilon ϵ sekä RBF-ytimen gamma γ.

Hyviä valintoja parametrien arvoiksi voisi kirjallisuuden avulla [1][2][5] aineiston ominaisuuksia samalla tutkien. Tässä tutkimuksessa valinta tehdään jakamalla opetusaineisto edelleen kahtia kuten k-NN:n kanssa ja tekemällä hilahaku (engl. *grid search*), jossa käydään läpi parametriyhdistelmät kunkin parametrin arvovälin

kymmenestä kohdasta, jonka jälkeen valitaan sitten $20^3 = 8000$ lopputuloksesta paras yhdistelmä. Hakualueet eri parametreille olivat $\epsilon \in [2^{-11}, 2^8]$ (logaritminen), $C \in [1, 570]$ ja $\gamma \in [2^{-11}, 2^8]$ (logaritminen).

6.8 Mallin validointi

Käytössä olleet estimaattorit koulutettiin koulutusaineistolla, ja estimaattoreiden laatua mitattiin laskemalla estimaattorin virheitä erilliselle testausaineistolle. Tässä tutkimuksessa koulutusaineistosta opittujen mallien suorituskykyä vertaillaan muutaman yleisen tilastollisen tunnusluvun avulla.

Tunnusluvut lasketaan estimaateista, jotka ovat yksikössä grammaa neliöjalan kokoista koealaa kohti ($g/0.1m^2$). Estimaateille virheen keskiarvo (engl. *mean absolute error*, MAE) lasketaan käyttäen:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \mu_i| \tag{11}$$

jossa y_i on tässä yhteydessä koealalle i piirteistä opetusaineiston perusteella estimoitu biomassa ja μ_i samalle koealalla punnittu biomassa.

Toinen tunnusluku on keskineliövirheen neliöjuuri (engl. *root mean square error*, RMSE), joka esitellään tulosten vertailemisen helpottamiseksi toisen osin samaa aineistoa käyttävän tutkimuksen [14] kanssa. Keskineliövirheen neliöjuuri määritellään seuraavasti:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_i)^2} \tag{12}$$

Edelliset tunnusluvut ovat edelleen yksikössä $g/0.1m^2$, ja kuvaavat suoraan biomassaestimaattien keskimääräisiä virheitä testausaineistossa. Lopuksi esitellään vielä kaksi suhteellista tarkkuutta kuvaavaa tunnuslukua. Normalisoitu keskineliövirheen neliöjuuri (engl. *normalized root mean square error*, NRMSE), joka määritellään RMSE:n sekä aineiston keskiarvon avulla:

$$NRMSE = \frac{RMSE}{\frac{1}{n} \sum_{i=1}^{n} \mu_i} \tag{13}$$

NRMSE kuvaa siis estimaattien suhteellista virhettä testausaineistossa. Estimaattien ja mitattujen arvojen suhteesta annetaan lisäksi Pearsonin korrelaatiokerroin, joka määritellään seuraavasti:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\mu_i - \bar{\mu})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\mu_i - \bar{\mu})^2}}, \tag{14}$$

jossa $\bar{y}$ ja $\bar{\mu}$ ovat estimaattien ja mitattujen arvojen keskiarvot. Jälkimmäisiset tunnusluvut eivät ole määriteltyjä kaikissa tapauksissa, kuten esimerkiksi NRMSE aineiston keskiarvon ollessa 0 ja korrelaatio tapauksissa, joissa estimaatit tai mitatut arvot ovat vakioita. Tässä tutkimuksessa kaikki tunnusluvut olivat kuitenkin määriteltyjä.

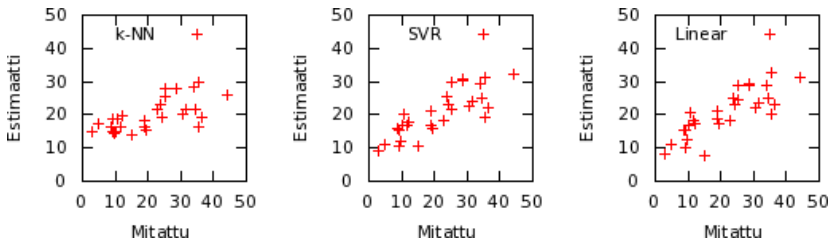
7 Tulokset

Luvussa esitellään tutkimustulokset luvussa 6 kuvattuja menetelmiä käyttäen. Luvussa kuvataan myös käytettyjen menetelmien ominaisuuksia optimaalisia malleja etsiessä.

Regressiomenetelmä	**MAE** $\frac{g}{0.1m^2}$	**RMSE** $\frac{g}{0.1m^2}$	**NRMSE** %	**r**
k-NN$_{k=7}$	7.64	10.44	46.3	0.59
SVR-RBF$_{\gamma=2^{-6},C=407,\epsilon=4.0}$	6.53	9.03	40.09	0.73
Lineaarikuvaus	**6.27**	**8.85**	**39.2**	**0.74**

Taulukko 1: Mallien testaustulokset luvussa 6.8 kuvatuilla mittareilla.

Koealoista irroitettiin 6 erilaista piirrettä luvussa 6.3 kuvatulla tavala. Koealat jaettiin menetelmien testausta varten opetus- ja testausaineistoihin käyttäen luvussa 6.1 kuvattua *k-means*-klusterointialgoritmia. Aineisto jaettiin piirreavaruutta jakaen 7 klusteriin, joiden koot olivat 8, 16, 21, 11, 8, 8 ja 19 koealaa klusterissa. Aineisto saatiin klustereita käyttäen jaettua tasapuolisesti 61 opetuskoealaan ja 30 testikoealaan.



Kuva 6: Hajontakuviot estimointimenetelmillä saaduista tuloksista.

Tuloksissa oli odotuksiin nähden monia yllätyksiä. Silmiinpistävintä on, että lineaarikuvaus suoriutuu tehtävästä yhdessä tukivektoriregression kanssa parhaiten. Seitsemälle piirteelle löydetyn luvussa 4.2 kuvatun lineaarikuvauksen optimaalinen kerroinvektori oli pyöristettynä $\beta = (2.97, 0.30, -4.82, 6.16, 3.82, 17.50, -1.97, 12.58)$. Vektorista nähdään, että NIR2-kanavan kerroin on suurin (17.50), mutta myös muilla piirteillä on selkeästi nollasta poikkeavia kertoimia. Lineaarikuvauksen hyvästä korrelaatiosta testausarvojen kanssa voisi päätellä, että estimointitehtävällä on käytettyjen piirteiden kanssa lineaarinen luonne. Piirteissä on myös paljon kohinaa, ja kuten luvussa 4.2 todettiin, lineaariset menetelmät voivat pärjätä hyvin kohinaisessa ja vähänäytteisessä opetusaineistossa.

K-NN-menetelmän k-arvoja kokeiltiin välillä $[1, 30]$. Arvoilla opetusaineiston kesken kokeillessa saadut virheen keskiarvot vaihtelivat välillä 6.55 ($k = 7$) ja 10.0 ($k = 30$). Virhearvot laskevat $k = 1$:n jälkeen tasaisesti 8.9:stä kohti optimaalista k-arvoa 7, ja nousevat sitten tasaisesti k:n lähestyessä arvoa 30.

Vastaavat SVR-RBF-tukivektoriregression parametrien γ, ϵ ja C arvojen kanssa saadut virheen keskiarvot ($g/0.1m^2$) vaihtelivat optimaalisen 5.01:n ja hakualueen huonoimman 15.23:n välillä. Optimaalinen ratkaisu oli $\gamma = 2^{-6}, C = 407, \epsilon = 4.0$. 8000 yhdistelmän kokoisen hakualueen huonoin ratkaisu valittua mittaria käyttäen oli $\gamma = 2.0$, $C = 552$, $\epsilon = 2^{-11}$.

Tukivektoriregression parametrien optimointia eri osajoukolla saman pellon aineistosta on tutkinut myös Salo [19], jolloin opetusaineiston sisäisesti optimoidut virheen mediaanit olivat hyvinkin pieniä (optimi 1.4). Lukua ei suoraan voi verrata tämän tutkimuksen virheen keskiarvoihin sillä ne kuvaavat eri lukuja. Huomattavaa on silti, että molemmissa tutkimuksissa testausaineistoa vasten tehdyt testit olivat hieman opetusaineiston kesken tehtyjä huonompia (MAE-arvot nousivat noin $1g/0.1m^2$). Parametrien optimointi johtaa ilman erillisiä toimenpiteitä opetusaineiston ylioppimiseen.

8 Yhteenveto ja pohdintaa

Tässä tutkielmassa keskityttiin kolmen eri estimointimenetelmän käyttöön. Kuten tutkimustuloksista 7 nähdään, menetelmillä saaduissa tuloksissa on pieniä eroja. Huomattavinta oli, miten hyvin lineaarikuvaus tässä tapauksessa toimi. Tutkimustulosten ja linaarikuvauksen luvussa 4.2 esiteltyjen ominaisuuksien perusteella lineaarikuvaus on soveltuu tutkimuksen regressiotehtävään parhaiten.

Lineaarikuvauksella saadut lupaavat tulokset viittaavat, että regressiotehtävällä on käytetyn piirreaineiston kanssa lineaarinen luonne. Piirteiden suuri määrä ja turhat piirteet aiheuttavat kohinaa, mikä vaikeuttaa estimoinnin onnistumista. Piirteitä lisäämällä ja niitä kehittämällä saavutetaan mahdollisesti merkittäviä parannuksia tuloksiin. Lisätyt piirteet saattavat yhdessä tässä tutkimuksessa käytettyjen piirteiden kanssa parantaa estimaatteja toisilla menetelmillä, mutta voivat regressiotehtävän muuttuessa heikentää lineaarikuvauksen suhteellista suorituskykyä.

Tällä hetkellä lineaarikuvaus on kuitenkin selkeytensä vuoksi pidettävä jatkotutkimuksissa mukana, vaikka se olisi pärjännyt vertailussa huonomminkin. Kuten luvussa 4.2 pohdittiin, riittävän hyviä tuloksia antava lineaarikuvaus olisi sovellusalueelle menetelmistä houkuttelevin. On myös mahdollista, että lineaarikuvauksen korrelaatio paranisi entisestään esimerkiksi RANSAC-menetelmää käyttämällä, jossa aineiston ulkopuoliset havainnot (engl. *outlier*) pyritään jättämään huomiotta.

Ominaisuuksiensa puolesta k-NN-menetelmä todettiin luvussa 4.3 sovellusaluetta ajatellen vähiten kiinnostavaksi. Sen antamat tulokset olivat myös selkeästi muita menetelmiä huonompia. On kuitenkin huomattava, että k-NN:n käyttäytyminen riippuu täysin aineistosta, ja esimerkiksi piirteet olisi syytä valita erityisesti k-NN:ää käyttäen. Tässä tutkielmassa piirteitä ei optimoitu toisaalta erityisesti mitään menetelmää ajatellen.

Tukivektoriregressio selviytyi verrattaen hyvin, ja on ominaisuuksiensa puolesta kiinnostava menetelmä sovellusaluetta ajatellen. Tukivektoriregression kouluttamisessa tehdyillä valinnoilla on paljon vaikutusta, ja koulutusaineiston ylioppimista on huomattavissa tässäkin tutkielmassa. Kun parametrien hakualuetta laajennettiin niin, etteivät optimaaliset parametrit ole hakualueen reunalla, koulutusaineiston kesken löydetyt optimaaliset parametrit toimivat lopulta testausaineistoa vasten huonommin. Vapaammin haetuista parametreista seurasi, että aiemmin lineaarikuvauksen kanssa tasavertaisen tukivektoriregression suorituskyky putosi tuloksissa hieman lineaarikuvauksen alle. Luvussa 6.7 kuvattu parametrien hakutapa ei siis ole täydellinen. Ylioppimisen välttämiseksi esimerkiksi Salo [19] käytti vastaa-

via parametrejä optimoidessa *bootstrap aggregation* -menetelmää, jossa parametrien hakemisessa käytetystä koulutusaineistosta jätetään osa käyttämättä nimenomaan parempaa yleistymiskykyä tavoitellen.

Tutkielmassa oli käytössä vain osajoukko kaikista mahdollisista aineistosta irroitettavissa olevista piirteistä, jotka valittiin tutkimusta varten käsin. Käytetyissäkin piirteissä saattoi olla mukana estimointitehtävän kannalta tarpeettomia piirteitä.

Piirteiden määrän kasvattaminen ei kuitenkaan yksin auta. Jatkotutkimuskohteita olisivat itse piirteiden etsiminen ja kehittäminen (*feature extraction*), piirteiden valinta (*feature selection*) esimerkiksi geneettisiä algoritmeja käyttäen [24] sekä dimensionpienennyksen (*dimension reduction*) soveltaminen [10].

Kuten luvussa 6.1 mainittiin, olisi mielenkiintoista kokeilla erilaisten opetus- ja testausaineistojakojen vaikutusta. Vaikka jako tehtiin erityisesti aineiston jakamiseksi tasapuolisesti, eri yhdistelmiä generoimalla ja estimoinnin kokeilemista niillä saataisiin lisää tietoa siitä, kuinka herkkiä menetelmät sekä toisaalta aineisto ovat jakotavalle.

Pölönen [14] sai vastaavaa aineistoa käyttäen hieman parempia tuloksia soveltamalla diffuusiokarttoja ja k-NN-regressiota. Näin saatu RMSE oli noin gramman pienempi kuin nyt saadun lineaarikuvauksen RMSE ja oli usean ajon keskiarvo, kun tämän tutkimuksen luvut ovat yksittäisen ajon arvoja. Kyseinen tutkimus käytti myös hieman erilaista k-means-jakoperustetta opetus- ja testausaineiston välillä tehden jaon koealojen kylvöattribuuttien (lajike, lannoitemäärä ja niin edelleen) perusteella, mikä olisi myös perusteltu tapa jakaa koealat. Ylipäänsä menetelmien herkkyys opetus- ja testausaineiston valinnalle olisi kuitenkin syytä selvittää seuraavaksi.

Lähteet

[1] Lorenzo Bruzzone. Support vector machines in remote sensing: the tricks of the trade. Kirjassa *Proceedings Vol. 8180, 81800B, Image and Signal Processing for Remote Sensing XVII*. SPIE, 2011.

[2] Chih-Jen Chang, Chih-Chung and Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1—-27:27, 2011.

[3] R.B. Brown D.W. Lamb. Remote-Sensing and Mapping of Weeds in Crops. *Journal of Agricultural Engineering Research*, 78(2):117–125, helmikuu 2001.

[4] Detlef Ehlert, Hans-Jürgen Horn, ja Rolf Adamek. Measuring crop biomass density by laser triangulation. *Computers and Electronics in Agriculture*, 61(2):117–125, toukokuu 2008.

[5] Trevor Hastie, Robert Tibshirani, ja Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag, viides laitos, 2011.

[6] Fotogrammetrian ja kaukokartoituksen Seura. Ohjeita ortokuvien tuotannolle ja käytölle Suomessa. Tekninen raportti, 2005.

[7] Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, kesäkuu 2010.

[8] Hamlyn G. Jones ja Robin Antony Vaughan. *Remote sensing of vegetation*. Oxford University Press, illustrate laitos, 2010.

[9] Tapani Antila Jussi Mäkynen, Christer Holmlund, Heikki Saari, Kai Ojala. Unmanned aerial vehicle (UAV) operated megapixel spectral camera. Kirjassa *Proceedings Vol. 8186,Electro-Optical Remote Sensing, Photonic Technologies, and Applications V*. SPIE, 2011.

[10] F. Melgani ja L. Bruzzone. Support vector machines for classification of hyperspectral remote-sensing images. Kirjassa *IEEE International Geoscience and Remote Sensing Symposium*, osa 1, ss. 506–508. IEEE, 2002.

[11] Alp Mestanogullari. kmeans-vector, 2011. <URL: `http://hackage.haskell.org/package/kmeans-vector`>.

[12] Ning Wang Naiqian Zhang, Maohua Wang. Precision agriculture—a worldwide overview. *Computers and Electronics in Agriculture*, 36(2-3):113–132, marraskuu 2002.

[13] NASA. The Enhanced Thematic Mapper Plus, 2012. <URL: `http://landsat.gsfc.nasa.gov/about/etm+.html`>.

[14] Ilkka Pölönen, Ismo Pellikka, Heikki Salo, Heikki Saari, Jere Kaivosoja, Sakari Tuominen, ja Eija Honkavaara. Biomass estimator for CIR-image with few additional spectral band images taken from light UAS. Kirjassa *SPIE 8369*, 2012.

[15] P. C. Robert. Precision agriculture: a challenge for crop nutrition management. *Plant and Soil*, 247(1):143–149, marraskuu 2002.

[16] Alberto Ruiz. hmatrix - Linear algebra and numerical computation, 2011. <URL: `http://hackage.haskell.org/package/hmatrix`>.

[17] Stuart J. Russell ja Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.

[18] Heikki Saari, Ismo Pellikka, Liisa Pesonen, Sakari Tuominen, Jan Heikkilä, Christer Holmlund, Jussi Mäkynen, Kai Ojala, ja Tapani Antila. Unmanned Aerial Vehicle (UAV) operated spectral camera system for forest and agriculture applications. Kirjassa *Proceedings Vol. 8174, Remote Sensing for Agriculture, Ecosystems, and Hydrology XIII*. SPIE, 2011.

[19] Heikki Salo, Ville Tirronen, ja Ferrante Neri. Evolutionary Regression Machines for Precision Agriculture. sarjan *Lecture Notes in Computer Science* osa 7248, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[20] Henry Samuel. satellites help french winemakers pick a perfect harvest, 2011. <URL: `http://www.telegraph.co.uk/news/worldnews/europe/france/8673485/Satellites-help-French-winemakers-pick-a-perfect-harvest.html`>.

[21] Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, ja Peter L. Bartlett. New Support Vector Algorithms. *Neural Computation*, 12(5):1207–1245, toukokuu 2000.

[22] Ville Tirronen ja Paulo Tanimoto. svm-simple - Medium level, simplified, bindings to libsvm, 2011. <URL: http://hackage.haskell.org/package/svm-simple>.

[23] Berkeley; Univ. of Colorado Denver; Univ. of Tennessee; Univ. of California ja NAG Ltd. LAPACK — Linear Algebra PACKage, 2011. <URL: http://www.netlib.org/lapack/>.

[24] F Vancoillie, L Verbeke, ja R Dewulf. Feature selection by genetic algorithms in object-based classification of IKONOS imagery for forest mapping in Flanders, Belgium. *Remote Sensing of Environment*, 110(4):476–487, lokakuu 2007.

[25] Steven E. Golowich Vladimir Vapnik. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. Kirjassa *Advances in Neural Information Processing Systems*, osa 9, ss. 281–287. MIT Press, 1997.

[26] X J Yao, A Panaye, J P Doucet, R S Zhang, H F Chen, M C Liu, Z D Hu, ja B T Fan. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *Journal of chemical information and computer sciences*, 44(4):1257–66, tammikuu 2004.