Norges miljø- og biovitenskapelige universitet

Masteroppgave 2018 30 stp

Fakultet for realfag og teknologi
Ivar Maalen-Johansen

Estimering av biomasse for trær i urbane områder ved å bruke hyperspektrale flyfoto

Biomass estimation in urban areas with the use of hyperspectral airborne images

Simon Barane
Geomatikk
Fakultet for realfag og teknologi

Sammendrag

TerraTec og Plan og Bygningsetaten i Oslo kommune har i samarbeid skaffet et forsøksområde for HySpex hyperspektrale bilder. Terratec har stått for flyfotograferingen og Oslo kommune ønsker å finne ut hva og hvordan HySpex bilder kan brukes for miljøanalyser. I denne oppgaven blir det forsøkt å bruke spektralinformasjon fra HySpexs til å estimere biomasse for trær i urbane områder med stor variasjon i treslag, størrelser og tilstand. I tillegg til å estimere biomassen lages det også funksjonsuttrykk som beregner biomassen over bakken (AGB) til trær. Metoden er basert på Bernasconis artikkel som bruker flyfoto med synlig lys til å estimere biomasser ved hjelp av arealet til trekroner (Bernasconi et al., 2017).

For å lage uttrykkene og estimatene blir trær detektert og segmentert. Arealet til trekronene beregnes ut i antall piksler og vegetasjonsindekser og bånd multipliseres med antall piksler. Dette gjør at indeksene og båndene også representerer arealet av trær. Videre brukes dette til å lage funksjonsuttrykkene. Arealfunksjonen fikk R^2 = 0,656 og funksjonsutrykket $AGB = -0{,}11805 + 0{,}00106 \times \frac{Areal}{0{,}09}$. R^2 på 0,656 er en akseptabel verdi som viser at uttrykket korrelerer tilfredsstillende med biomasse. En bedre modell som krever mer data er vegetasjonsindeksmodellen som fikk R^2 = 0,826 og funksjonsuttrykket $AGB = \frac{1{,}74 \times GRVI + 4{,}68 \times LI_{ratio} - 1{,}88 \times SR_1 - 6{,}70 \times Vogelman_1}{1000}$, der vegetasjonsindeksene er multiplisert med antall 0,3 meters piksler for hvert tre. Analysene fant en korrelasjon mellom GRVI, SR1, LI_ratio, Vogelman1 og AGB. Dette vil si at 7 bånd gir svært nøyaktige estimat. Denne modellen klarte å gi bedre statistiske resultater enn laserdata klarte. Dette er en indikasjon på at hyperspektrale data gjøre bedre biomassetimeringer enn laserdata for urbane områder. Arealfunksjonene ender opp med enn lavere R^2 enn laseranalysene, men krever lite informasjon og kan kjøres uten å trenge annet enn et pankromatisk bilde eller et satellittbilde.

Estimater fra maskinlæring fikk lavere R^2 og noe større residual enn det regresjonsanalysene fikk for funksjonsuttrykkene. Det er fordi det brukes regularization og validering for å sikre at resultatet er stabilt, uten overfit, og korrekt. Arealberegningene fikk R^2 = 0,646 og vegetasjonsindeksene ga R^2 = 0,692. Maskinlæringen fant en korrelasjon mellom AGB og vegetasjonsindeksene GRVI, SR1, LI_ratio, Vogelman1, SAVI og TVI.

Abstract

TerraTec and Plan og Bygningsetaten has acquired a new hyperspectral dataset for Oslo. The dataset created by TerraTec and is an experimentation field for the HySpex hyperspectral data. Plan og Bygningsetaten and TerraTec wants to know what and how this new dataset can be used for environment analysis. This article tries to use the HySpex data for estimating biomass of trees in urban areas with a vast variation in trees size, species and health. Biomass functions is also created. These are calculated by linear regression and are based on the ideas from Bernasconi's article where visible light images are used to estimate aboveground biomass (AGB) from the area of the tree crowns (Bernasconi et al., 2017).

The images got segmented and everything except tree crowns got masked out. The number of pixels in each tree crown got calculated and the vegetation indices and bands got multiplied by the pixel amount. This created bands and vegetation indices with the spatial information. The first biomass function uses only the area of the tree crowns and has an R^2 of 0,656 and has the following function: $AGB = -0{,}11805 + 0{,}00106 \times \frac{Area}{0{,}09}$. The function has an acceptable R^2 that shows that the function does have some correlation with AGB. The function that uses vegetation indices gave better results and have a R^2 of 0,826 and the function is: $AGB = \frac{1{,}74 \times GRVI + 4{,}68 \times LI_{ratio} - 1{,}88 \times SR_1 - 6{,}70 \times Vogelman_1}{1000}$, where all the vegetation indicines is multiplied by the amount of 0,3 meter pixels in the tree crowns. The analysis finds a clear correlation with the indexes GRVI, SR1, LI_ratio, Vogelman1 and AGB. Using 7 bands to create the vegetation indices gave the best biomass function. This estimation gets a higher R^2 and lower residuals than the laser analysis in the same area. This means that HySpex data might be better than laser in urban areas for biomass estimation. The functions using only area of tree crowns for biomass estimation does not get better results than laser. The advantage with those functions is that they require little information and can be used together with panchromatic images or satellite images.

Estimations from machine learning gave a slightly lower R^2 than the biomass functions. This comes from the combination of validation data and regularization to ensure that the estimations do not suffer from large errors, overfit or instability. The estimation using only area as variable got an $R^2 = 0{,}646$ and the one with vegetation indices got $R^2 = 0{,}692$. There was a correlation between AGB and GRVI, SR1, LI_ratio, Vogelman1, SAVI og TVI.

Forord

Jeg har vært veldig heldig som har fått lov til å jobbe med hyperspektrale bilder. HySpex bildene er helt unike og det har vært lærerikt og interessant å få være blant de første som jobber med flyfotoene av Oslo. De har vært travle måneder og mye arbeid, skriving, prøving og feiling involvert i min masteroppgave. Det har lært meg hvordan det er å forske på noe nytt som svært få har kunnskap om. Da blir man nødt til å lære på egen hånd og teste det ingen har gjort før. Likevel har jeg fått mye hjelp, og vil takke mine veiledere for god hjelp.

Min hovedveileder, Ivar Maalen-Johansen, har vært til god hjelp og bidratt med råd og tips til både skriving og analyse. Jeg vil også takke Ingunn Burud for å ha hjulpet meg med teori, fagstoff og metoder for å bruke hyperspektral data. Dette har vært uvurderlig kunnskap som jeg har fått brukt for gjennom hele semesteret. Jeg vil også takke Plan og Bygningsetaten og TerraTec som begge har vært med på å skaffe disse bildene, og for å gi studenter muligheten til å jobbe med dem. Plan og bygningsetaten har også gitt meg en sitteplass, gode råd, motivasjon og kakao. Jeg vil også takke Erik Røstad, Kristoffer Cebaloss, Håkon Berg Lofthus og Åsmund Stemme som er de andre studentene som har jobbet med laserskanning og hyperspektrale bilder. Jeg har fått mye hjelp og informasjon fra dette samarbeidet. Til slutt vil jeg takke mor og far for gode skriveråd, korrekturlesing, motivasjon og støtte. Håper at oppgaven kan være interessant for alle som interesserer seg i fjernmåling, skog og statistikk.

Innholdsfortegnelse

1 Tabeller og formler

2 Figurer

3 Innledning

Innenfor fjernmåling har det ikke vært brukt mye hyperspektrale bilder. Det er ikke fordi en tviler på nytteverdien, men fordi den tidligere har vært nærmest utilgjengelig. Hyperspektral data har tidligere vært for dyrt og ikke hatt noe spesielt god romlig oppløsning, presisjon eller kvalitet. Det har omtrent ikke vært noen kameraer som har vært i stand til å skape gode nok bilder til å bruke de for fjernmåling. Ny teknologi har produsert bedre kameraer som kan ta bilder flere hundre meter over bakken og fortsatt gi god oppløsning og kvalitet.

Norsk Elektro Optikk AS (NEO) har produsert to nye kameraer som heter HySpex VNIR-1800 og SWIR-384. Kameraene gir oss hundrevis av bånd som kan brukes til å se objekters reflektans i det elektromagnetiske spekteret. VNIR og SWIR kameraene kan representere spekteret fra 400 nanometer og helt opp til 2500 nanometer fordelt på 476 bånd (Cerdeira, 2018; NEO, 2018). Dette gir oss muligheten til å få et kontinuerlig spekter i motsetning til multispektrale flyfoto og satellitter som har vanligvis mellom 3-12 bånd.

Informasjonen en får fra hyperspektrale kamera kan brukes til mye forskjellig. Hyperspektrale flyfoto kan brukes til klassifiseringer og analyser av vegetasjon og bygg. Siden HySpex har svært mange bånd, kan den brukes til omtrent alt som kan oppdages ved å se på objektets reflektans. I oppgaven er målet å se om biomassen til tre kan estimeres ved hjelp av treets spektralsignatur. I oppgaven prøver jeg å detektere tre, beregne statistikk for de og deretter bruke maskinlæring for å finne en estimert biomasse. Til slutt er ønsket å lage et funksjonsutrykk som bruker et sett av båndene til HySpex for å gi en estimert biomasse. Sammen med en deteksjon av tre skal den gi en estimert biomasse for trær i urbane områder.

3.1 Problemstilling

Min problemstilling i denne masteroppgaven handler om biomasseestimering ved hjelp av hyperspektrale flyfoto. Det jeg ønsker å se er å om det er mulig å estimere biomasse ved hjelp av 2-dimensjonale bilder der enn bruker arealet og spektralinformasjonen. Hvor bra kan en estimere biomasse for trær ved å bare bruke areal? Hvor mye bedre er estimatet dersom en legger til vegetasjonsindekser eller bånd? Hvor mange bånd er nødvendig for å få en akseptabel estimering av biomasse i urbane områder? Og til slutt, kan en metode som bruker kameraer konkurrere med resultater fra laserskanning?

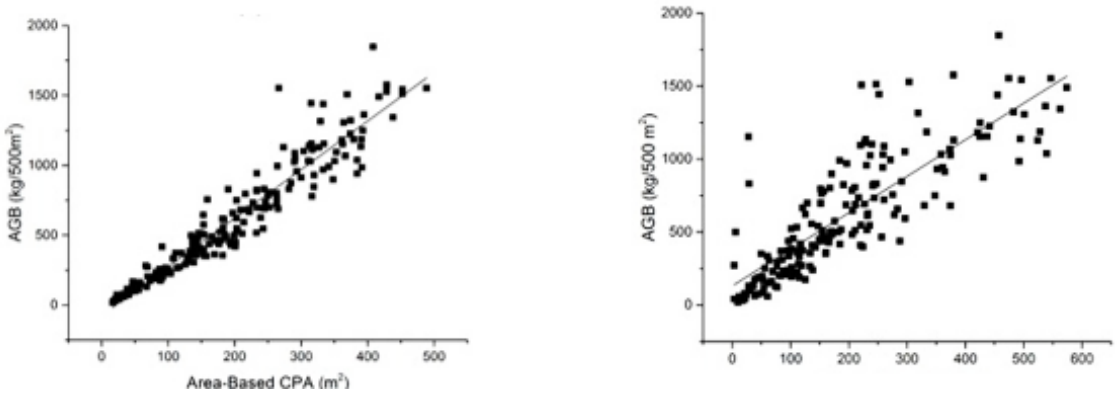
3.2 Bakgrunn

Det fins flere metoder å beregne biomasse for tre. Den tradisjonelle metoden er å gjøre feltarbeid og samle inn informasjon som er viktig for å beregne biomasse. Mer nøyaktige målinger og flere egenskaper som blir samlet for hvert tre, gir bedre biomasseestimeringer. I 1987 skrev Lars Gunnar Marklund en bok om hvordan en kan beregne biomasse for furu, gran og bjørk i Skandinavia. Den har vært en standard innenfor biomasseberegning basert på feltdata. Boka består av flere modeller som kan brukes for å estimere biomasse. Noen av modellene krever lite informasjon og krever bare brysthøydediameteren på stammen og høyden til treet. Andre funksjoner som gir mer nøyaktige resultat krever blant annet relativ barktykkelse, stubbediameter, diameter 5 meter, diameter 3 meter, kroneradius, grønn kronehøyde, høyde over vannet. Det blir laget 4 funksjoner for hver del av treet som en skal beregne biomasse og alle krever ulik mengde med egenskaper fra treet (Marklund, 1987). Det gjør at en kan bruke den som passer best basert på hvilke egenskaper en har samlet fra feltarbeid.

Forestry Commission har gjort mye av det samme i Storbritannia. Der ble målingene gjort for flere ulike treslag, og den inkluderer flere typer løvtre. Forestry commisions metoder bruker stammediameter ved brysthøyden og høyden i deres modeller. Videre utviklet de en metode som var laget for å unngå overestimeringer av biomasse for store tre (Randle et al., 2011). Modellene deres er også laget for å gi tilfredsstillende resultat dersom en ikke vet treslaget.

Svakheten med metoder som nevnt ovenfor er at de krever mye informasjon om hvert enkelt tre. Feltarbeid er nødt til å gjøres for alle tre en skal beregne biomasse for. Dette gjør at det er tidkrevende og nærmest umulig å gjennomføre for større områder. Å beregne biomasse for enkeltrær i et feltarbeid er ikke effektivt, men det gir svært gode resultater. Derfor brukes ofte funksjonene fra slikt feltarbeid fortsatt. Det gir en god fasit på hva biomassen skal være. Den moderne måten å beregne biomasse på er ved hjelp av laserskanning. Laserskanning kan gi 3-dimensjonal data som kan gi svar på hvor stor kronen er, høyden til treet og stammens diameter. Dette gjør at laserskanning kan finne volumene og beregne biomasse på nærmest nøyaktig samme måte som før. For validering kan det bruke små områder der en gjør både måling ved feltarbeid og laserskanning.

En annen metode som er brukt er å beregne biomassen ved hjelp av flyfoto. Det har blitt gjort en del forsøk med å bruke 3-dimensjonal informasjon fra fotogrammetri. Da har en tatt i bruk flyfoto av samme område, men i ulike vinkler som en kan skaffe 3-dimensjonal informasjon fra. Det eksister ikke mange forsøk uten noen som helst 3-dimensjonal informasjon, men det finnes noen få. Blant annet har Luca Bernasconi forsøkt å gjøre dette ved hjelp av vanlige ortofoto. Det ble funnet ut at vanlige ortofoto kan gi rimelig gode resultater for å estimere biomasse for tre under gode forhold. Resultatene ble gode fordi det var lite skygge i trekronene og lite overlapp mellom trær. Området som ble analysert var skogområde med stor avstand mellom tre. Det var også små trekroner med lite overlapp. Resultatene var svakere enn i lignede analyser som brukte multispektrale med høyere oppløsning. Biomassen var estimert ved hjelp av størrelsen på treet. Det vil si at bare mengden piksler trekronen bestod av bestemte biomassen (Bernasconi et al., 2017)



Figur 3.2 Grafene viser resultateten fra (Bernasconi et al., 2017). Y-aksen viser biomasse og x aksen viser trekronens areal. To ulike metoder ble brukt og ga litt ulike resultat.

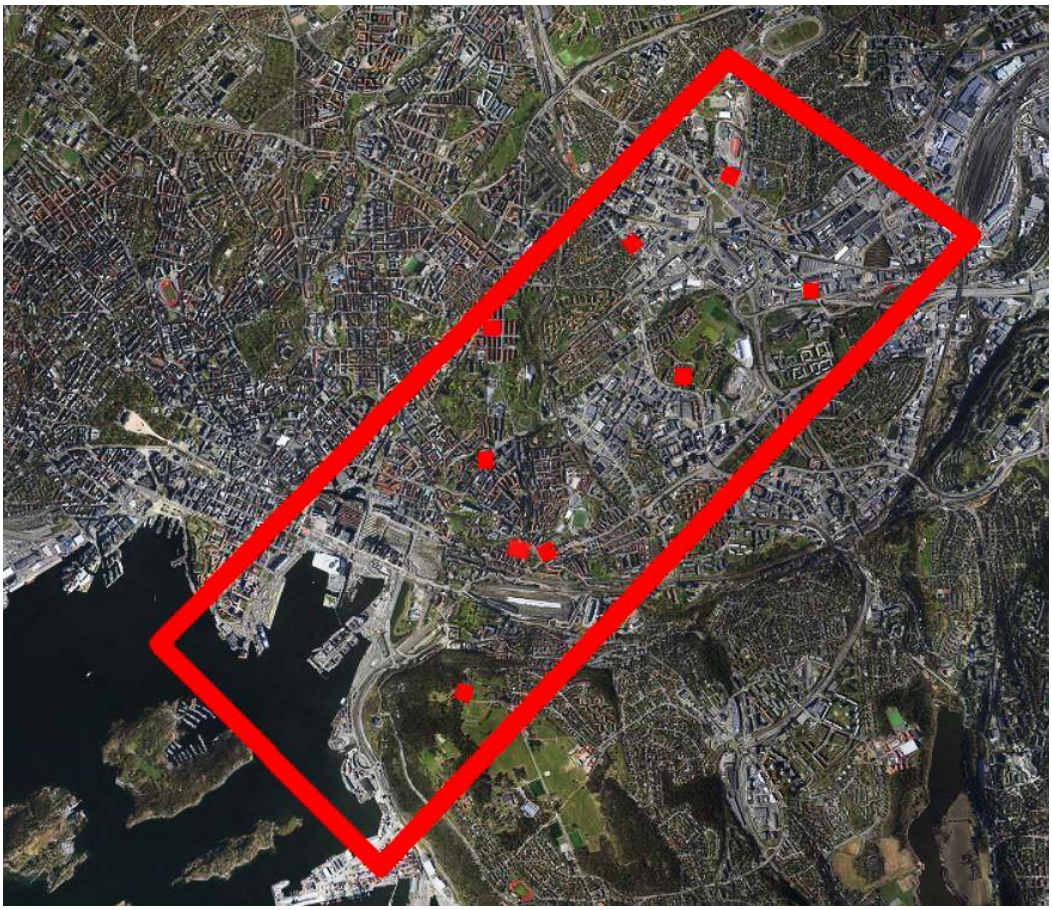
Ideen om å se på Leaf Area Index eller biomasse i urbane områder kommer fra artikkelen «Estimating Urban Leaf Area Index (LAI) of Individual Trees with Hyperspectral Data» skrevet av Ryan R. Jensen, Perry J. Hardin, og Andrew J. Hardin. De prøver å beregne LAI ved hjelp av hyperspektrale bilder i urbane områder. Dette gjør de ved hjelp av tre metoder. Hver av metodene bruker hver sin type informasjon. Den første bruker bånd, den andre bruker Prinsipalkomponenter fra PCA, og den tredje bruker vegetasjonsindekser. Det blir forsøkt å finne korrelasjon mellom «ground thruth» LAI fra feltarbeid og de ulike metodene. Korrelasjonen blir modellert med lineære regresjonsmodeller. De resulterte med at LAI kan beskrives ved hjelp av modellene, men ikke alle var like gode. Å bruke fire ulike bånd ga et estimat på LAI, men forklarte lite av variasjonen in LAI. PCA og vegetasjonsindekser ga litt bedre resultat. Videre ble det gjort forsøk med Neural Network. Neural Network ga en noe bedre forklaring i variansen av LAI, men fortsatt ikke like bra som håpet på forhånd. Både lineær regresjon og Neural Network ga best resultater med vegetasjonsindekser. Det fant og ut at det er ikke en enkel prinsipalkomponent som hadde en klar relasjon til LAI (Jensen et al., 2012).

I min oppgave vil jeg jobbe videre på resultatene fra artikkelen til Bernasconi. Jeg ønsker å bruke spektralinformasjonen i tillegg til arealet for å se om det kan forbedre biomasseestimatene. En annen forskjell blir at jeg kommer til å støte på skygge og overlappende trær og andre utfordringer som en urban by kan medføre. Jeg ønsker også å ta med noen av ideene som Jensen brukte for LAI. Det virker fornuftig å vurdere å bruke bånd,

PCA og vegetasjonsindekser ved hjelp av lineær regresjon for å se på hva som fungerer best. Forskjellen blir at jeg ønsker å summere sammen piksler istedenfor å jobbe med gjennomsnitt av piksler siden biomassen baserer seg på størrelsen til et tre.

3.3 Plan

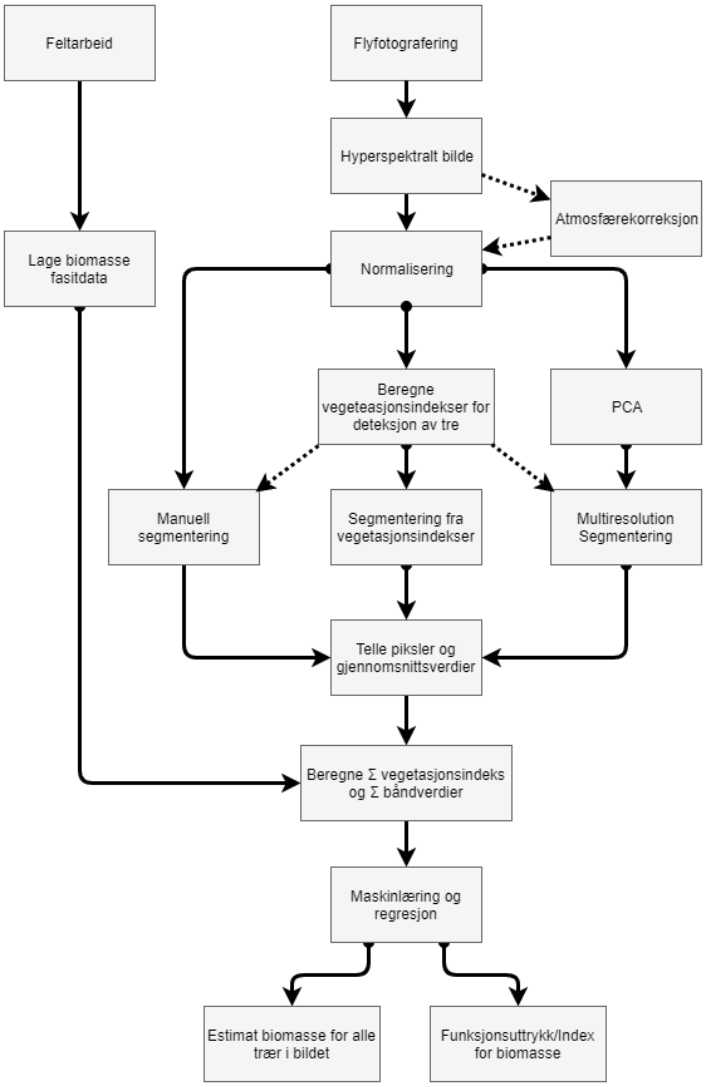
Ønsket fra Plan og Bygningsetaten var å finne ut hva HySpex bildene kan brukes til og å sammenligne dem med Optech Titan laserdata. Oslo kommune har vurdert å skaffe hyperspektrale data over hele kommunen og ønsker å vite om det er verdt å anskaffe. Siden bildene er dyre, avtalte Plan og Bygningsetaten å kjøpe hyperspektrale bilder for en del av Oslo. Dette er et samarbeid mellom TerraTec og Oslo kommune som gjør at prisen ble lavere. Begge parter ønsker å se hvilke analyser, metoder og resultater en kan få fra HySpex bildene. Jeg vurderte å skrive oppgaven på engelsk. Fordelen med engelsk er at da kan oppgaven bli lest av flere og kan brukes internasjonalt. Jeg valget å skrive oppgaven på norsk. Grunnen til dette er at det er en fordel for kommunen å ha oppgaven på norsk dersom metoden skal videreutvikles eller brukes. En annen grunn er at det også skal skrives en internasjonal publikasjon som inneholder de viktigste elementene fra oppgaven. Denne vil gi informasjonen som trengs fra oppgaven for de som ikke leser norsk.



Figur 3.3.A Ortofoto av Oslo, det røde rektangelet er området som er dekket av HySpex og de røde kvadratene er områder med feltarbeid fra masterstudenter

Masterstudenter som skrev for Oslo kommune fikk tilgang på bildene. Jeg ønsket å finne ut om en kunne estimere Leaf Area Index (LAI) ved hjelp av hyperspektral data og deretter se om det kan gi en god indikator på biomasse i urbane byområder. Etter hvert som jeg eksperimenterte fant jeg ut at oppgaven ikke var lett å gjennomføre. Dette var fordi jeg slet med å finne en måte å lage fasitdata for LAI. Feltarbeidet var ikke optimalt for å gi LAI. Istedenfor valgte jeg å vri på oppgaven. Hva med å se på biomasse istedenfor? Det var flere artikler på hvordan en estimerer biomasse fra feltdata. Planen for oppgaven ble da å se om spektralsignaturen fra et tre kan korreleres til biomassen fra feltdata. Videre ønsker jeg å se om jeg kan redusere båndmengden og likevel klare beholde et godt biomasseestimat. Om det er mulig ønsker jeg å redusere slik at bånd som brukes av flere typer sensorer kan gi et godt estimat. For eksempel ville det være ideelt om en kunne redusere båndmengden til synlig lys og et NIR bånd. I så fall vil flere satellitter og multispektrale sensorer gi gode estimat for

biomasse i urbane områder. Avslutningsvis ønsker jeg å sammenligne alle resultat med biomasseberegninger fra laserdata.



Figur 3.3.B Figur som viser fremgangsmåten og planen for masteroppgaven. Målet er å skaffe et presist biomasseestimat i tillegg til et funksjonsuttrykk som skal kunne estimere biomasse for trær i hvilket som helst område. Stiplede linjer er alternative operasjoner.

Figuren 3.2 er en generell visuell forklaring av hva jeg ønsker å gjøre i oppgaven. Alt her er lagt opp med de nødvendige stegene for å rekonstruere arbeidet eller dersom en ønsker å lage et eget funksjonsuttrykk for biomasse. Dersom en ønsker å bare bruke funksjonsuttrykket trenger en ikke gå gjennom alle stegene. I senere kapittel (8.9) blir det forklart hvordan en bruker funksjonsuttrykket og kan enkelt få biomassen for trær uten å trenge noe kraftig prosessering, regresjon eller maskinlæring. For å få de beste estimatene vil det likevel være nødvendig å gå gjennom hele metoden.

3.4 Hvorfor hyperspektralt?

Hyperspektral data blir brukt til denne oppgaven. Dette er ikke den vanligste måten å beregne biomasse på. Det eksisterer flere gode metoder for biomasseestimering ved laserdata. Vanligvis er laserdata brukt på skog, men det finnes også noen forsøk på å bruke den for enkelttrær (Popescu, 2007). Laserdata har den gode fordelen av å gi nøyaktige høyder og punktskyer som gjør det mulig å bruke volumer for å beregne biomasse (Nelson et al., 1988). Laserdata gir tilgang på 2,5 D og 3D informasjon og hyperspektral data gir 2D. Det er mulig å skaffe høyden ved hjelp av hyperspektral data også, men det vil kreve fotogrammetri og stereobilder. Det blir sett på mosaikkbilder uten høydeinformasjon i oppgaven. Likevel kan en anta at en har høyden for analyser siden den er notert i feltarbeid. Svakheten til laser er at den gir ingen eller lite spektral informasjon. Spektralinformasjon er hyperspektrale bilders beste egenskap.

Hyperspektralt har vært mindre tilgjengelig og mindre testet for biomassestimering. Fordelen med å eksperimenetere med hyperspektralt er en får alle mulige bånd. Med smale bånd kan vi teste hvilken som helst del av det elektromagnetiske spekteret og prøve å se om det er noen deler av dem som kan forklare biomassen i trær. Multispektrale kameraer har bare noen få bånd og har ikke muligheten til å finne ut hvilke bølgelengder som korreler best med biomasse. For senere analyser vil ikke hyperspektral data være like nødvendig. Med hyperspektral kan vi endre pikselstørrelser, slå sammen bånd, redusere bånd og skape nærmest hvilken som helst sensor, og da teste hvor bra de vil fungere. Kanskje viser de at med de riktige båndene kan en satellitt gi gode nok estimater for biomasse? Det ville vært optimalt om en kunne funnet en metode som gjør at laser ikke trengs for biomasseestimering. En annen fordel med hyperspektral er at den kan gjøre flere oppgaver. For eksempel kan det gjøres en klassifisering av tre ved hyperspektrale bilder kombinert med en biomasseestimering. Det er økonomisk gunstig å kunne bruke de samme datagrunnlaget til flere analyser.

Hyperspektral data er dyrt. Dette gjør at ikke alle har muligheten til å skaffe hyperspektral data over store områder. Om hyperspektral data i et lite område kan brukes til å lage en metode som fungerer ved å bruke billigere datasett, er det en stor fordel.

4 Teori

4.1 Biomasse

4.1.1 Vanlig definisjon

Biomasse er vanligvis definert som mengden organisk materiale som kommer fra planter og dyr (Forestry Commission, 2018). Begrepet blir ofte brukt når en snakker om energi og brensel. Det er det organiske materiale, biomassen som gjør at for eksempel ved brenner godt i peisen. I levende organismer og planter er tilnærmet alt regnet som organisk materiale bortsett fra vann. Biomasse blir definert som en masse og oppgis ofte i kilo eller tonn. For trær er det vanlig å oppgi biomassen i tonn tørrvekt.

4.1.2 Forestry Commissions definisjon

I oppgaven kommer jeg til å bruke definisjonen av biommase som Forestry Commission bruker. Forestry Commission er en britisk organisasjon som tar vare på skog i Storbritannia. De har blant annet flere artikler om biomasseestimering for skog. Forestry Commission definerer biomasse som en enhet for alt levende plantemateriell for et tre (Forestry Commission, 2014). Da blir både røtter, stubber, stamme, greiner og blader inkludert i definisjonen for biomasse. De tar ikke med tre som har en stamme med diameter under 7 cm i brysthøyde (Jenkins et al., 2011; Randle et al., 2011). Biomasse kan brukes til å fortelle mengden brensel og massen/størrelsen til et tre. Det kan og fortelle om hvor mye karbondioksid et tre har tatt opp. Andre ting biomasse kan brukes til er å si noe om verdien og nytten av et tre og det kan også korreleres med ulike nyttige vegetasjonsindekser som blant annet Leaf Area Index. Biomasse kan være verdifullt å ha for miljøanalyser.

4.1.3 Biomasse over og under bakken

For tre er det vanligvis stammen som inneholder den meste delen av biomassen, men også trekronen og røttene kan inneholde en god del biomasse. Ofte blir biomasse for et tre delt inn i

to hoveddeler. «Aboveground biomass» og «belowground biomass». Belowground vil si den biomassen som ligger under bakken og ikke synes. Det vil si røttene til et tre. En god del biomasseestimeringer tar ikke hensyn til det som er under bakken fordi det er vanskeligere å beregne eller se. Aboveground biomass (AGB) er trekronen, greiner og stammen. Siden stammen inneholder det meste av biomassen, blir AGB delt opp til biomasse fra stamme og biomasse fra resten. Ved å summere AGB og biomassen til røttene får vi den totale biomassen for et tre (Field Studies Council). I oppgaven blir det mest fokus på AGB siden det er den vanligste biomassen å bruke.

4.2 Elektromagnetiske spekteret

HySpex kameraet fotograferer synlig lys og infrarødt lys. Det vil si at den tar opp reflektert energi i det synlige og infrarøde lyset. Et menneske kan bare se reflektert lys i det vi kaller det synlige spekteret. Dette er fra ca. 380 nm og opp mot ca .700 nm. All reflektert energi over dette kan vi ikke se (GISGeography, 2018). Derimot kan HySpex se reflektert energi opp til 2523 nm (Cerdeira, 2018). Vi snakker ofte i fjernmåling om hvilke bånd og bølgelengder lys reflekterer i. Det er dette som er konseptet multi og hyper-spektrale bilder baserer seg på. For eksempel har trær en høy intensitet i reflektert infrarødt lys. Et menneske kan ikke se dette, men kameraet oppdager denne intensiteten. Ulike biokjemiske materialer som lignin og nitrogen reflekterer energi på ulike posisjoner. Ved å se på bølgelengden 1754 nm kan vi se sjekke om trær inneholder lignin, og ved 1680 nm reflekterer nitrogen (Serrano et al., 2002). Red edge er et annet område som ofte snakkes om i vegetasjonsanalyser. Dette er den store stigningen i reflektert energi for vegetasjon som er rett utenfor bølgelengdene mennesker kan se. Dersom denne energien har en bratt økning i intensitet kan dette være med på å identifisere helsen, klorofyll produksjonen og vekststadiet til vegetasjon (Vogelmann et al., 1993).

4.3 Hyperspektral data

Hysperspektral data er data produsert av et eller flere hyperspektrale kameraer. Hyperspektrale kameraer er en passiv sensor. Passive sensorer trenger lys fra solen for å

kunne fungere. Passive sensorer måler reflektert energi som den mottar, men den sender ikke ut noen bølger selv som den skal motta (Jonassen & Aarsten, 2017). I motsetning er radarinstrumenter slik som laser aktive sensorer som selv sender bølger til jorden som reflekterer tilbake (Thuy, 2012). En vanlig definisjon på hyperspektral data er at den består av flere hundre bånd (GISGeography, 2018). Hyperspektrale kameraer har muligheten til å finne endringer i små spesifikke områder av det elektromagnetiske spekteret. Dette er fordi den bruker smale bånd og ikke brede bånd.

Bredbånd er når båndene til kameraet er svært store. Når båndet dekker store deler av det elektromagnetiske spekteret får en mye større intensitet og tar opp alt i området. Svakheten er at en ikke klarer få detaljer som er i spesifikke bølgelengder. For eksempel nitrogen er vanskelig å oppdage ved bredbånd. Nitrogen absorberes på blant annet 1020 nm og vann absorberes ved 970 nm. Dersom bredbåndet er 900-1100 nm kommer nitrogen ikke til å være synlig fordi andre egenskaper som for eksempel vann vil skjule det. Alt blir samlet opp i dette området og intensiteten blir høy, men det betyr ikke at nitrogeninnholdet er stort. Bredbånd kan brukes til å definere høy intensitet eller lav intensitet i et område, men kan ikke brukes for egenskaper som en finner i små deler av det elektromagnetiske spekteret. Smale bånd derimot er optimale for slikt. Smale bånd er på noen få nanometer og kan oppdage absorbsjon av nitrogen, ligning, vann etc. Smale bånd gir lave intensiteter siden de tar opp små områder.

Forskjellen mellom multispektrale og hyperspektrale bilder kan være vanskelig å definere. Det er to egenskaper som skiller bildene. Hyperspektrale bilder har hundrevis av smale bånd og multispektrale har færre og bredere bånd. For eksempel en sensor med 90 smale bånd er ikke like lett å definere hvor passer best. Hvor smalt et bånd må være for at det defineres som hyperspektralt varierer. Artikkelen til Rodarmel og Shan definerer hyperspektrale sensorer med at de har kontinuerlige bånd med 5-10 nm bredde (Rodarmel & Shan, 2002). For HySex er bredden på båndene 5 nm og ca. 2,5 nm. For sammenligning har den hyperspektrale Hyperion satellitten til NASA en bredde på ca. 9,5 nm (USGS, 2011).

Figur 4.3 Spektralsignaturen til et tre i Haslevangen. Bildet er atmosfærekorrigert. Ved flere smale bånd får vi en kontinuerlig spektralkurve

Hyperspektrale bilder har den beste spektrale oppløsningen men ofte på bekostning av romlig oppløsning. Pikselstørrelsen ender ofte opp med å være dårligere. Hyperspektral data har flere praktiske bruksområder. Klassifiseringer, miljøanalyser, permeabilitet er noen av de viktige bruksområdene for hyperspektral data.

4.4 Atmosfærekorreksjoner

Satellittbilder og flyfoto blir ofte atmosfærekorrigert. Det vil si at det er forsøkt å fjerne all støy og reflektert strålingsenergi fra atmosfæren. Atmosfæren består av flere ulike gasser som reflekterer strålingsenergi. Denne fører til at sensoren får med seg mye mer enn bare objektets reflektans. Sensoren tar med reflektans til gasser som blant annet nitrogen, oksygen, argon og karbondioksid. Verdien når atmosfæren er inkludert i bildet kaller vi radians. Det vi egentlig ønsker er objektets reflektans (Richter & Schläpfer, 2016). For å gjøre om radians til reflektans, gjøres det en atmosfærekorreksjon. Det fins flere varianter av atmosfærekorreksjoner og de har ulike styrker og svakheter. Noen blir for enkle og klarer ikke fjerne alt av atmosfærens bidrag. Andre atmosfærekorreksjoner kan fjerne for mye og en blir sittende med et bilde som har fjernet deler av objektets reflektans (Jonassen, 2018). En annen

utfordring er at atmosfærekorreksjonene kan bli veldig komplekse og tidkrevende i tillegg til å kreve veldig mange ulike parametere som en ikke alltid har tilgang til.

Atmosfærekorreksjon er viktigere for satellittbilder enn for flybilder. Når høyden blir større, blir det mer atmosfære som vises i bildet. For et flyfoto, er det 500-1500 meter med atmosfære. For nærpolare satellitter er det ca. 800 km med gasser fra atmosfæren. Atmosfærekorreksjon kan gjøre at bilder gir et mer «true color» bilde. Det er fordi atmosfæren kan inneholde gasser som fører til at en får en misfarge på bildet og i noen tilfeller kan også atmosfærekorreksjon fjerne tynne skyer og gjøre bildet mindre «tåkete». For eksempel Atcor4 forsøker å fjerne blant annet vann, snø, cirrusskyer og tåke (Richter & Schläpfer, 2016).Dersom bildet er tatt fra lav høyde får en mindre av problemene og atmosfærekorreksjon blir mindre viktig.

4.5 Normalisering av data

Normalisering av data er en bildeprosessering. Normaliseringen som gjøres i denne oppgaven har som mål å fjerne skygge og gjøre hele bildet mer likt. Skyggepikslene skal ligne mest mulig på resten av pikslene og alle områder i bildet skal være lette å sammenligne med hverandre. Terratec har gjort normaliseringen (Jonassen, 2018). Formelen de har brukt er denne:

$$x_{ij}\ erstattes\ med\ uttrykket\colon \frac{x_{ij}}{\left(\frac{1}{K}\sum_j x_{ij}\right)}$$

Formel 4.5 Normalisering gjort med metoden fra (Yu et al., 1999)

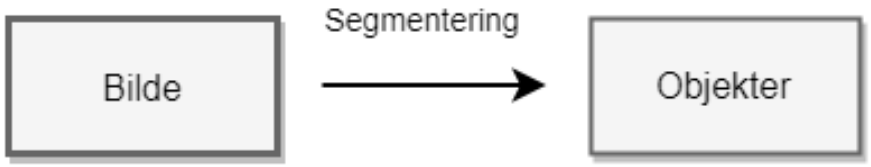
x_{ij} erstattes med uttrykket og x er piksel, j er radiansverdi til hver pixel x_i. i er indeksen til pikselen og K er antall bånd. Det vil si at en dividerer pikslene på summen av pikslene som er multiplisert med invers av antall bånd (Jonassen, 2018).

I de normaliserte bildene er det tilnærmet ingen skygge og derfor kan jeg velg å se bort fra hele skyggeproblematikken. Om jeg hadde brukt ikke-normaliserte bilder og ikke-normaliserte indekser ville jeg hatt problemer i skyggeområder fordi spektralsignaturen er helt ulik. Svakheten med normaliseringen er at den gjør en del objekt vanskeligere å skille i bildet. Gress og trær får en noe mer lik spektralsignatur og små forskjeller i spektralsignatur blir vanskeligere å oppdage.

4.6 Segmentering

4.6.1 Definisjon og bruksområde

Segmentering brukes når vi ønsker å slå sammen piksler til et større objekt. Segmenteringen vil slå sammen alle piksler som er i samme område og har samme egenskaper til et større objekt. Segmentering brukes når en ønsker å finne noen objekt fra et bilde, som for eksempel når en ønsker å bare se på en vei fra et flyfoto. Eller som det er i denne oppgaven, som skal segmentere slik en bare tar i bruk piksler som er fra tre. En segmentering gjøres vanligvis på intensitetsverdier fra bildet, men en kan også segmentere basert på egenskaper som pikslene har (Shapiro & Stockman, 2001). For eksempel om pikslene har en medfølgende tabell med info som masse, farge og materiale.



Figur 4.6.1 Input i en segmentering er et bilde. Output vil bli et sett med objekter.

4.6.2 Sammenheng mellom klassifisering og segmentering

Segmentering og klassifisering har mye til felles. Det er flere typer klassifiseringsmetoder som baserer seg på arealbaserte objekt istedenfor piksler. Det som blir gjort da er en segmentering før selve klassifiseringen. En segmentering i seg selv kan bli sett på som en enkel klassifisering. En segmentering kan også ha flere klasser eller den kan være binær.

4.6.3 Segmenteringsmetoder

Segmenteringsalgoritmer prøver som oftest å finne grensen/kanten som skiller objektet og resten av bildet. Dette kalles kantbasert segmentering (Darwish et al., 2003). For en vei vil dette være veikanten og for et tre blir det kanten av trekronen. Alt som ligger innenfor grensen blir satt som en verdi og det utenfor blir satt som en annen. Det finnes også to andre teknikker som kalles punktbasert og regionbasert segmentering. Noen av de vanligste formene for segmentering er K-means klustring, region growing, kvadtrebasert segmentering, sjakkbrett segmentering og, spektralforskjell segmentering og multiresolution segmentering.

4.6.4 K-means klustring

K-means klustring bruker minste kvadraters metode og prøver å minimere feilen. Noen varianter velger automatisk antall klasser (K) slik at minste kvadraters feil (D) blir minst. D forklarer hvor godt data passer i klassene sine. Andre krever at K er gitt på forhånd. Her er x verdien klustringen blir gjort på som i hyperspektrale bilder blir intensitetsverdien i et gitt bånd, og m er gjennomsnittet til klassen. K-means gjøres vanligvis med iterasjon og forsøker å finne globalt optimum for D (Shapiro & Stockman, 2001). K-means er ikke spesielt vanlig for å segmentere objekt i flyfoto og satellittbilder, men den brukes en del for å gjøre enkle klassifiseringer.

$$D = \sum_{k=1}^{K} \sum_{x_i \, \epsilon \, C_k} \|x_i - m_k\|^2$$

Formel 4.6.1 K-means klustring (Shapiro & Stockman, 2001)

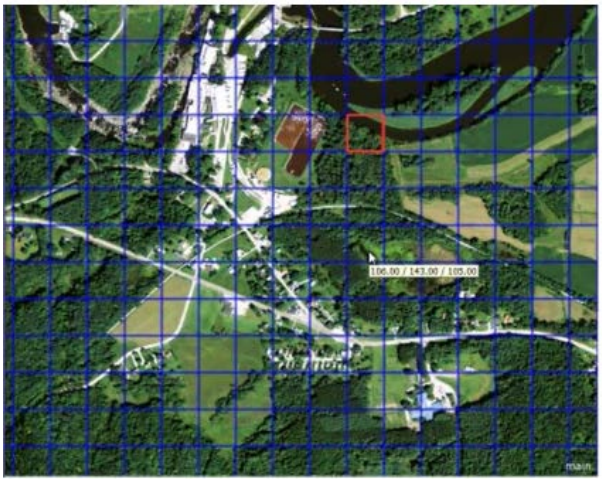
4.6.5 Region grow

Region grow velger en piksel i bildet og ser på alle naboer om de har likheter i intensitet. Dersom pikslene er innenfor en gitt terskel blir de slått sammen til et objekt. Videre sjekker den på nytt naboene for å se om de igjen ligner på det nye objektet. Den vil fortsette å vokse frem til segmentet ikke har noen nabopiksler som er innenfor terskelen. Algoritmen vil gå gjennom bildet piksel for piksel. Dersom en piksel ikke har likheter med noen naboer blir den

stående alene som et lite segment (Shapiro & Stockman, 2001). Det kan være en fordel å ha en regel som gir en min/max på hvor store klasser kan være. Da vil enkeltstående piksler bli tvunget inn i klasser.

4.6.6 Sjakkbrett segmentering og kvadtre segmentering

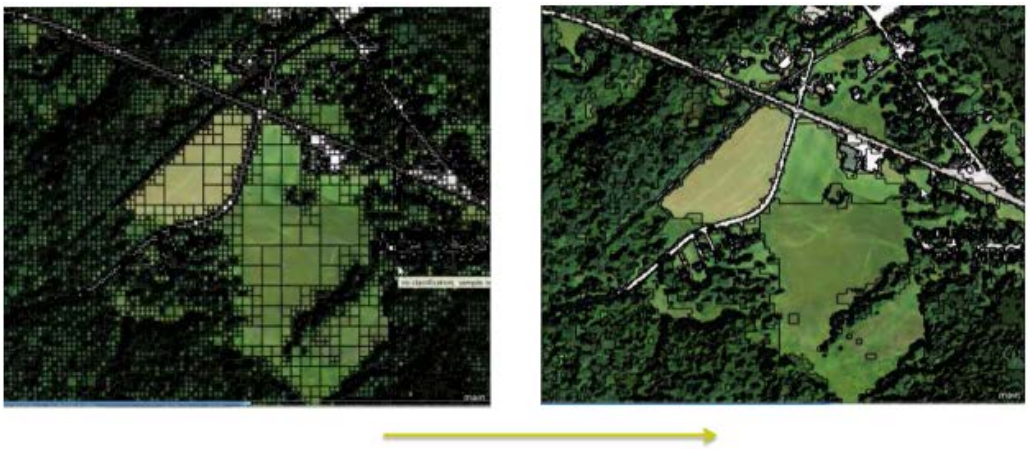
Sjakkbrettsegmentering er enkel å utføre, men er som oftest lite nyttig. Den går ut på at en lager et rutenett med like store ruter (Ouyang, 2015). Hver av rutene blir et objekt. Det kan prøves å velge en fornuftig rutestørrelse som gjør at hver rute inneholder mest mulig piksler som ligner hverandre. Som oftest er dette vanskelig. De fleste naturlige fenomen er ikke kvadratiske og passer dårlig med en rute. For en trekrone som er rund vil den enten måtte ta med mye område rundt treet eller så må det brukes mindre ruter slik den tar med treet. Da blir treet gjort til mer enn et objekt.



Figur 4.6.6.A Bildet viser en sjakkbrettsegmentering. Zutao Ouyang mener metoden er ubrukelig fordi objektene ikke gir noe meningsfull informasjon (Ouyang, 2015)

Kvadtremetoden er en bedre metode enn sjakkbrett segmenteringen. Kvadtre tar i bruk ulik rutestørrelse (Ouyang, 2015). For homogene områder med lite forskjeller blir det brukt store ruter. For områder med mye endring vil det brukes små ruter. Hver rute blir et objekt. For eksempel kan et stort tre nå være en stor rute i midten av treet og noen mindre ruter langs kanten av treet. Det som kan gjøres videre med kvadtreet er å bruke en spektralforskjell segmentering av rutene. Den sjekker spektralsignaturen til en rute og ser om noen naboruter

har en tilnærmet lik spektralsignatur. Denne fungerer slik som region grow metoden og går gjennom alle ruter og sjekker naboer. Er likheten innenfor en gitt terskel blir rutene slått sammen til et nytt objekt.



Figur 4.6.6.B Til venstre er det gjort en kvadtre segmentering og til høyre er kvadtreet segmentert på nytt basert på spektralsignaturen til rutene. (Ouyang, 2015)

4.6.7 Multiresolution segmentering

Multiresolution segmentering er en moderne segmenteringsmetode som eCognition speisliserer seg på. Dette er en regionsbasert segmenteringsalgoritme (Darwish et al., 2003). Den ser på hver piksel som et objekt i starten. Den sjekker lokal homogenitet i pikslene. To kriteria bestemmer om objekter er homogene. Det første er «color» og det andre er «shape». De to naboobjektene som er mest like blir slått sammen til et objekt. Videre vil det nye objektet bli sjekket på nytt for å finne naboen som er mest lik. Når ingen naboer av objektet er innenfor en gitt terskel for likhet, blir objektet lukket. Denne terskelen heter «Scale parameter». Når skaleringsparameteren har en høy verdi, lager segmenteringen større klasser som kan ha mindre homogene piksler. Med lav skaleringsparameter blir objektene mindre og en kan være mer sikker på at alle piksler innenfor objektet er rimelig like (Darwish et al., 2003). Det krevers ofte en del testing for å finne den rette skaleringsparameteren. Er den for stor får vi med piksler som ikke passer sammen med objektet, og er den for lav kan det føre til at hvert objekt ender opp med å bli sett på som flere objekt. Det vil si at et tre blir sett på som flere ulike tre. Ofte er det slik at skyggen kan bli et objekt og resten av treet blir et annet objekt.

«Color» funksjonen ser på spektralinformasjonen til objekter og sammenligner de. Den bruker da alle båndene i bildet til å bestemme om objekt er ulike. Funksjonen sjekker heterogenitet mellom objektene. Mye av formelverket bruker heterogenitet, som er det motsatte av homogenitet. Høy heterogenitet vil si at objekter er helt ulike og har lite til felles. Heterogenitet i et bildeobjekt er definert slik:

$$h = \sum_{k=1}^{m} w_k \cdot \sigma_k$$

Formel 4.6.7.A Formel for heterogenitet mellom to objekt (Ouyang, 2015)

Spektral heterogenitet (h) for et objekt er summen av standardavvikene til hvert enkelt bånd (σ_k) multiplisert med et vektall for hvert bånd (w_k). Vekten er lik for alle i vanlige tilfeller (Ouyang, 2015). Dersom en vet at noen bånd er spesielt viktige for å segmentere korrekt kan de få en høyere vekt.

$$h_{color} = \sum_{k=1}^{m} w_k \left[n_{mg} \cdot {\sigma_k}^{mg} - (n_{ob1} \cdot {\sigma_k}^{ob1} + n_{ob2} \cdot {\sigma_k}^{ob2}) \right]$$

Formel 4.6.7.B «color criterion» (Ouyang, 2015)

«color criterion» formelen gir oss heterogenitetsforskjellen (h_{color}) mellom de nye sammenslåtte objektet og de to objektene hver for seg. mg er det sammenslåtte objektet og ob1 og ob2 er objektene hver for seg. heterogeniteten for det sammenslåtte objekt vil være større enn heterogeniteten til ob1 og ob2. Dette gjør at vi får en positiv verdi. Er verdien nærme null er objektene lite heterogene og er verdien høy er områdene heterogene.

«Shape criterion» er den andre delen som trengs for å regne ut segmenteringsfunksjonen. Denne er todelt i heterogenitet for kompakthet og heterogenitet for glatthet.

$$cpt = \frac{l}{\sqrt{n}} \qquad smooth = \frac{l}{b}$$

Formel 4.6.7.C og D Kompakthet og glatthet (Ouyang, 2015)

Kompakthet (cpt) regnes ut ved å dele perimeteren på kvadratroten av antall piksler. Perimeter er omkretsen til objektet. Glatthet får en ved å dele perimeteren på omkretsen til objektets minimum bounding box. Etterpå beregner en heterogenitet mellom sammenslått objekt og objektene som blir vurdert å slå sammen. Dette gjøres hver for seg for glatthet og kompakthet. Høyt tall betyr heterogent, og lavt tall betyr det er homogent.

$$h_{cpt} = n_{mg} \cdot \frac{l_{mg}}{\sqrt{n_{mg}}} - \left(n_{ob1} \cdot \frac{l_{ob1}}{\sqrt{n_{ob1}}} + n_{ob2} \cdot \frac{l_{ob2}}{\sqrt{n_{ob2}}} \right)$$

Formel 4.6.7.E Heterogenitet kompakthet (Ouyang, 2015)

$$h_{smooth} = n_{mg} \cdot \frac{l_{mg}}{b_{mg}} - \left(n_{ob1} \cdot \frac{l_{ob1}}{b_{ob1}} + n_{ob2} \cdot \frac{l_{ob2}}{b_{ob2}} \right)$$

Formel 4.6.7.F Heterogenitet glatthet (Ouyang, 2015)

Heterogeniteten til shape (h_{shape}) beregnes ved å kombinere h_{cpt} og h_{smooth}. Det blir også brukt et vektall. w_{cpt} bestemmer hvor mye kompaktheten skal vektlegges. Når vekten er en 0,5 er glatthet og kompakthet vektet likt. Ved 0 er glatthet eneste som gir verdi. Når vekten er 1 blir glatthet ignorert.

$$h_{shape} = w_{cpt} \cdot h_{cpt} + (1 - w_{cpt}) \cdot h_{smooth}$$

Formel 4.6.7.G Heterogenitet shape (Ouyang, 2015)

Når en har beregnet både «shape criterion» og «color criterion» kan en lage «segmentation function» (S_f). Det er denne segmenteringsfunksjonen som bestemmer om et objekt blir slått sammen til et nytt segment eller ikke. Er S_f er de to objektene sett på som homogene og blir

om til et nytt større objekt. Er S_f stor vil ikke objektene kunne slås sammen fordi de ikke er homogene. Det vil være en terskel som bestemmer om objekt blir slått sammen. S_f må være en lavere verdi enn terskelen. Denne terskelen er scale og velges ofte av brukeren. I S_f funksjonen blir det og gitt en vekt. Vekten her bestemmer om spektralinformasjon eller form skal vektlegges mest. Ved 0,5 er begge vektlagt likt. Dersom form har lite å si for objektet bør vekten være nærme 1. Da blir det båndene som bestemmer verdien til S_f.

$$S_f = w_{color} \cdot h_{color} + (1 - w_{color}) \cdot h_{shape}$$

Formel 4.6.7.H Segmenteringsfunksjonen (Ouyang, 2015)

Etter å ha gjort en multiresolution segmentering er det ikke alltid slik at segmenteringen blir perfekt. Som oftest handler det om å få best mulig resultat i flest mulig tilfeller, men det er sjeldent alt blir slik en ønsker. Dersom en ikke er fornøyd med resultatet fra en multiresolution segmentering kan en kjøre en «spectral difference segmentation» Spektralforskjell ser på om ulike objekt har tilnærmet lik spektralinformasjon. Dersom objekter har lik spektralsignatur blir dem slått sammen. Denne brukes på objekter, og ikke piksler.

4.7 Vegetasjonsindekser

4.7.1 Hva er en vegetasjonsindeks?

Vegetasjonsindekser har som mål å gi en kvantifiserbar verdi for en egenskap eller intensitet til vegetasjon. Den skal for eksempel gi en verdi som kan vise hvor sunt et tre er, eller hvor mye nitrogen vegetasjonen inneholder. Vegetasjonsindekser minimerer eksterne faktorer som kan gjøre bånd vanskelige å analysere (Baret et al., 1989). Kjente vegetasjonsindekser som NDVI, EVI og SAVI brukes ofte for å vite tilstand og mengden vegetasjon i flyfoto og satellittbilder.

Vegetasjonsindekser lages ved hjelp av spektralinformasjon. To eller flere bånd brukes for å lage en vegetasjonsindeks (Bannari et al., 1995). Ved å multiplisere, dividere, addere og subtrahere bånd kan en lage vegetasjonsindekser. Målet er at de skal klare å fremheve noe som er viktig i vegetasjonen. Ved å bruke enkeltbånd kan en ha deler av informasjonen som trengs for å finne noe, men ofte klarer ikke et enkelt bånd forklare egenskaper som nitrogen, lignin, stress, etc. Det er ofte forskjeller i lysforhold, skygge og atmosfære som gjør at å bruke enkeltbånd kan bli ekstra krevende. Vegetasjonsindekser kan redusere problemer som dette.

4.7.2 Normaliserte og ikke-normaliserte vegetasjonsindekser

Normaliserte vegetasjonsindekser er laget slik at de fungerer bra når lysforholdene er varierende. Normaliserte vegetasjonsindekser gjør en normalisering og gir samme fordeler som en full normalisering av bildet slik som blir forklart i kapittel 4.5. Svakheten med normaliserte indekser er at de ikke er optimale for allerede normalisert data. Å normalisere data som allerede er normalisert kan gjøres, men kan endre resultatet. Ikke-normaliserte vegetasjonsindekser får problemer i skyggeområder og når en har flere lyskilder. På grunn av dette kan det være lurt å bruke de på normaliserte bilder (Jonassen, 2018).

4.7.3 Vegetasjonsindekser som kan brukes for biomasseestimering

Alle vegetasjonsindekser som er med i oppgaven eller som var vurdert å ha med forklares i dette delkapittelet. Mye av indeksene som brukes er indeksene som nettsiden Index DataBase har kategorisert for å kunne brukes til biomasseestimering.

Nr.	Name	Formula	Variables	Comment
1	Simple Ratio NIR/RED	$\frac{NIR}{RED}$		
2	Enhanced Vegetation Index	$2.5\frac{NIR-RED}{(NIR+6RED-7.5\,BLUE)+1}$		
3	Normalized Difference NIR/Red	$\frac{NIR-RED}{NIR+RED}$	RED=[670;50;30],NIR=[800;10;10]	
4	Soil Adjusted Vegetation Index	$\frac{800nm-670nm}{800nm+670nm+L}(1+L)$	L = 0,5	
5	Transformed Soil Adjusted Vegetation Index	$\frac{B(NIR-B\cdot R-A)}{RED+B(NIR-A)+X(1+B^2)}$	B=B	
6	Simple Ratio 800/600	$\frac{800nm}{600nm}$		
7	Simple Ratio 800/550	$\frac{800nm}{550nm}$		
8	Wide Dynamic Range Vegetation Index	$\frac{0.1NIR-RED}{0.1NIR+RED}$		
9	Normalized Difference 2160/1540	$\frac{2160nm-1540nm}{2160nm+1540nm}$		

Tabell 4.7.3.A Ulike vegetasjonsindekser som er tidligere brukt til biomasseestimering (Index DataBase, 2018)

NDVI:

NDVI står for normalized difference vegetation index. Dette er en gammel indeks som har vært i bruk for satellittbilder siden 70-tallet (Bannari et al., 1995). Veldig ofte er den med også i multispektrale og hyperspektrale analyser. NDVI bruker tradisjonelt to bredbånd, men det finnes flere moderniserte varianter. Den kan da bruke smalere bånd eller modifiseres slik at den tar i bruk ekstra bånd. Normalt bruker den nærinfrarødt og rødt synlig lys. Noen modifiserte varianter tar i bruk ekstra bånd i SWIR eller grønt i det synlige lyset. TerraTec testet noen NDVI indekser for å finne en god for HySpex bildene (Aarsten, 2018). Den anbefalte hadde 681 nm for rødt lys, og 780 nm for NIR (infrarødt lys). ENVIs standardverdier er 649 nm og 859 nm.

$$NDVI = \frac{NIR - RED}{NIR + RED} = \frac{780\ nm - 681\ nm}{780\ nm + 681\ nm}$$

Formel.4.7.3.B NDVI

NDVI gir en høy verdi for vegetasjon. Alt av planter og grønt får en høy verdi. Bygninger og vei får veldig lav NDVI. NDVI varierer fra -1 til +1. Når en skal maskere bort alt som ikke er vegetasjon brukes NDVI ofte. Verdier rundt 0,4 er vanlig for å skille mellom vegetasjon og ikke vegetasjon. Siden NDVI har vært i bruk lenge har det blitt gjort flere analyser og forskning med NDVI. Den er lett for å sammenligne resultater og er kjent i de fleste miljø som jobber med vegetasjon og fjernmåling. Det er funnet flere korrelasjoner mellom NDVI og ulike egenskaper for vegetasjon. Blant annet har Liu funnet en korrelasjon mellom NDVI og biomasse. En tredjegradsfunksjon kan brukes for å finne biomassen av vegetasjon (Liu et al., 2006). Tredjegradspolynomet er ikke laget for å passe for norsk vegetasjon, men burde passe delvis for de fleste typer vegetasjon. Dersom en ønsker best mulig resultat med NDVI bør en bruke den på atmosfærekorrigert data.

$$\mathrm{Y} = -5593.3\mathrm{NDVI3} + 7509.7\mathrm{NDVI2} - 1268.9\mathrm{NDVI} + 191$$

Formel 4.7.3.C Funksjonsuttrykk for estimering av biomasse med NDVI (Liu et al., 2006)

SAVI:

SAVI (Soil Adjusted Vegetation Index) er en annen vegetasjonsindeks som er mye brukt. Den har en del likheter med NDVI. SAVI har som mål å være en forbedret utgave av NDVI som forsøker å gi mer korrekte verdier for ulik jord og tett vegetasjon. NDVI har tendenser til å gi varierende resultater når der er jord som med ulik lyshet, fuktighet, skygger og organisk materiale. For satellitter er dette problematisk fordi de ofte har store piksler som fanger opp mye jord og får miksede piksler av tre og jord. Det fører til større verdier enn de egentlig skal ha. SAVI forsøker å løse dette problemet og likevel beholde fordelene som NDVI har som blant annet dens korrelasjon med biomasse og leaf area index. SAVI og NDVI har svært like formler. Forskjellen er SAVI bruker konstanten L som er korreksjonsfaktor for jordens lys, farge og fuktighet. Dersom L = 0 er SAVI og NDVI like. L anbefales å være høy for tett vegetasjon (L = 1), og L bør være lav for områder med lite vegetasjon (L = 0,25). For normale mengder vegetasjon anbefales L å være 0,5 (Huete, 1988). For oppgaven er det varierende mengder med vegetasjon og setter L som 0,5. I de mer urbane områdene passer en lav L best, men denne vil være mindre egnet for parker.

$$SAVI = \frac{NIR - RED}{NIR + RED + L} \times (1 + L) \qquad SAVI = \frac{800nm - 670nm}{800nm + 670nm + L} \times (1 + L)$$

Formel 4.7.3.E og F SAVI (Huete, 1988)

GRVI:

GRVI står for Green Ratio Vegetation Index. GRVI er en ratio indeks og er da ikke normalisert. GRVI er ifølge ENVI sensitiv til effektiviteten og hastigheten i fotosyntesen for trær. Dette gjør at den er god for å oppdage trekroner og tett vegetasjon. GRVI får med seg pigmenter i blader. Det gjør at den oppdager klorofyll.

En god fordel med GRVI er at den gir ulike verdier for trær og gress. Det gjør at den kan separere tre og busker fra gress nokså enkelt. For ulike trær er den derimot nokså lik i intensitet og det er bare små forskjeller i verdier. Normalt brukes GRVI for bredbånd, men den er også god for smale bånd.

$$GRVI = \frac{NIR}{Grønn} = \frac{780\ nm}{550\ nm}$$

Formel 4.7.3.F GRVI (Sripada et al., 2006)

Eventuelt kan en bruke GRVIhyper som bruker 560 nm for grønn og 658 nm for NIR. Denne er spesiallaget for hyperspektrale bilder (Shibayama et al., 1999). I oppgaven ble det brukt vanlig GRVI og ikke GRVIhyper, men jeg har valgt å ikke bruke denne for min oppgave. Verdiene blir annerledes når NIR er en lavere bølgelengde slik som i GRVIhyper. 658 nm er rett før red edge stiger og det fører til at NIR verdien kan være lavere enn grønn verdien. Dette fører til verdier som er veldig ulike de som en får fra bredbåndsutgaven av GRVI gir. I oppgaven ønsker jeg verdier som har likhet med bredbåndsverdier der NIR er et høyere tall enn grønn og derfor brukes NIR på 780 nm. Noen ganger brukes GRVI inversest av hva formelen viser. Det har lite å si, men en bør være klar over hvilken som er brukt.

VARI:

VARI er en mindre brukt vegetasjonsindeks som bare bruker synlig lys. Indeksen bruker tre bånd og det er grønn, rød og blå. Dette kan være bredbånd eller smale bånd. VARI står for Visible Athospherically Resistant Index (Bernasconi et al., 2017). Bernasconi bruker denne indeksen for å oppdage og segmentere trekroner. Indeksen har fordelen av å bruke bånd som er veldig ofte tilgjengelig og indeksen er best når det ikke er atmosfærekorrigert data. Indeksen normaliserer og er ikke like god for normalisert data. Har ikke sett at denne indeksen blir brukt for annet enn å segmentere trær, men ønsker å teste om den kan også brukes til biomasseestimeringer.

$$VARI = \frac{Grønn - Rødt}{Grønn + Rødt - Blått}$$

Formel 4.7.3.G VARI (Bernasconi et al., 2017)

TVI:

TVI står for triangular vegetation index eller transformed vegetation index. Her er det snakk om triangular vegetation index. TVI ble utviklet for å karakterisere strålingen av bladpigmentene i området infrarødt/rødt lys og dets forbindelse med grønt lys. TVI bruker den høyeste intensitetstoppen i grønt lys, slutten av red edge og den laveste intensiteten i rødt lys (Driss Haboudane, 2004). Intensitetstoppen i grønt lys og intensitetsbunnen i rødt lys varierer for ulik vegetasjon, men vanligvis brukes en fast verdi istedenfor. Standarden er å bruke 550 nm for grønt lys, 670 nm for rødt lys og 750 nm før infrarødt lys. TVI er originalt laget for atmosfærekorrigerte bilder. TVI gir veldig varierende verdier og de kan være svært store. For de fleste vegetasjonsindekser får enn verdier mellom -1 og 1, eller 0 og 1. TVI kan gi verdier på flere hundre. Det eksisterer modifiserte utgaver av TVI som gjør at den ikke får slike høye verdier. MTV1 og MTVI2 er slike. MTVI1 og MTVI2 brukes omtrent likt og har stor korrelasjon med TVI, men bruker 800 nm istedenfor 750 nm. I oppgaven valgte jeg å bruke for TVI, men det kan være MTVI1 hadde passet like godt.

$$TVI = 0.5 \times \big(120(750nm - 550nm) - 200(670nm - 550nm)\big)$$

$$MTVI1 = \ 1.5 \times \big(1.2(800nm - 550nm) - 2.5(670nm - 550nm)\big)$$

$$MTVI2 = \ 1.5 \times \frac{\big(1.2(800nm - 550nm) - 2.5(670nm - 550nm)\big)}{\sqrt{(2 \times 800nm + 1)^2 - \big(6 \times 800nm - 5\sqrt{670nm}\big) - 0.5}}$$

Formel 4.7.3.H Formler for TVI, MTVI1 og MTVI2 (Driss Haboudane, 2004)

Simple Ratio:

Definisjonen av simple ratio er å dividere et bånd på et annet. Hvilke bånd som helst kan brukes. Dette er en måte å lage enkle indekser for hyperspektrale bilder. Dersom vi vet at vegetasjonen har en spesielle egenskaper eller absorbsjoner ved noen bølgelengde, kan vi lage simple ratios med dem. I oppgaven brukes to simple ratioer. Simple Ratio 1 (SR1) bruker 800 nm og 550 nm. Dette er de samme båndene som MTVI bruker. Det er også nokså likt det GRVI bruker. Simple Ratio 2 (SR2) bruker 800 nm og 600 nm. 550 nm er toppverdien når

vegetasjonen er svært grønn. Dersom den er mindre grønn og mer gulaktig er det rundt 600 nm som er toppen. Det forventes at et tre har høy SR1 og lav SR2 dersom det er har god helse og er lite stresset. Derimot er SR1 lav og SR2 høy er treet stresset. Uttrykk for SR1 og SR2 og generell formel for Simple ratio:

$$SR1 = \frac{800nm}{550nm} \qquad SR2 = \frac{800nm}{600\ nm} \qquad Simle\ Ratio\ n = \frac{\lambda_{abs}}{\lambda_{ref}}$$

Formel 4.7.3.I Simple ratio formler (Serrano et al., 2002)

Vogelman 1:

Vogelman 1 er en ratio vegetasjonsindeks som bruker to bånd i red edge. Den bruker båndene 720 nm og 740 nm og er gir store verdier når mye klorofyll, mye vanninnhold og stort bladareal. Vogelman1 er en god indeks når en ser på vekststadier, helsetilstand og klorofyll i trær (Vogelmann et al., 1993).

$$Vogelman\ 1 = \frac{740nm}{720nm}$$

Formel 4.7.3.J Vogelam1

NDNI:

NDNI står for Normalized Difference Nitrogen Index og er designet for å estimere nitrogenmengden i vegetasjon og trekroner. Det forventes at en kan oppdage nitrogeninnhold i bølgelengden 1510 nm. Denne blir sammenlignet med verdien i bølgelengden 1680 nm som er forventet å ikke absorbere nitrogen, men være en brukbar indikator på biomasse (Serrano et al., 2002). Normalt brukes indeksen for reflektans og den normaliserer bildet. Siden planen er å bruke normaliserte bilder lages det en ratio utgave av indeksen på samme måte som simple ratio:

$$NDNI = \frac{log\left(\frac{1}{1510nm}\right) - log\left(\frac{1}{1680nm}\right)}{log\left(\frac{1}{1510nm}\right) + log\left(\frac{1}{1680nm}\right)} \qquad NI_{ratio} = \frac{1680nm}{1510nm}$$

Formel 4.7.3.K NDNI og NI_ratio formler (Serrano et al., 2002)

NDLI:

NDLI står for Normalized Difference Lignin Index. Den er designert for å estimere ligninninnholdet i trekroner. Bølgelengden som en bruker for å oppdage lignin er 1754 nm, og referansen er 1680 nm som er det samme som NDNI bruker (Serrano et al., 2002). Begge indeksene er bygd opp likt og brukes på samme måte. Her lages det også en ratioutgave:

$$NDLI = \frac{log\left(\frac{1}{1510nm}\right) - log\left(\frac{1}{1754nm}\right)}{log\left(\frac{1}{1510nm}\right) + log\left(\frac{1}{1754nm}\right)} \qquad LI_{ratio} = \frac{1754nm}{1510nm}$$

Formel 4.7.3.L NDLI og LI_ratio formler (Serrano et al., 2002)

NDBleaf:

NDBleaf ble nevnt som en biomasse vegetasjonsindeks av Index DataBase, men jeg har ikke sett noen teori eller informasjon om den. Jeg tar med indeksen som en «wild card» for å se om den fungerer. Jeg forventer ikke at denne er spesielt god, men fra noen tester i ENVI viste det seg at den har kanskje en korrelasjon med biomasse.

$$NDBIeaf = \frac{log\left(\frac{1}{1540nm}\right) - log\left(\frac{1}{2160nm}\right)}{log\left(\frac{1}{1540nm}\right) + log\left(\frac{1}{2160nm}\right)} \qquad Bleaf_{ratio} = \frac{2160nm}{1540nm}$$

Formel 4.7.3.M NDLI og LI_ratio (Index DataBase, 2018)

4.8 PCA

PCA står for prinsipalkomponent analyse. PCA er en vanlig analysemetode når en har store datasett med flere variabler og det kan være vanskelig å finne korrelasjoner og sammenhenger ved å se på det originale datasettet. En PCA transformerer datasettet til komponenter. Det er en lineær transformasjon som reorganiserer variansen til båndene i et nytt datasett (Harris Geospatial solutions, 2014). Den første komponenten forklarer mest varians, og den siste komponenten forklarer minst varians. PCA gjør data lettere å håndtere og analysere. Komponentene i en PCA bygges opp av egenvektorene til kovariansmatrisen til bildet.

Artikkelen (Rodarmel & Shan, 2002) forklarer hvordan PCA er oppbygd i multi og hyper-spektrale bilder.

X_i er bilde piksel vektor og vi får en slik for hvert bånd bildet har.

$$X_i = [X_1, X_2, \dots X_N]_i^T$$

Formel 4.8.A Bilde pixel vektor. N er antall bånd.

Etterpå regnes det ut gjennomsnittsvektor for alle bildevektorene. M = kolonner multiplisert med rader.

$$m = \frac{1}{M}\sum_{i=1}^{M}[x_1 \;\; x_2 \;\; .. \;\; .. \;\; x_n]_i^T$$

Formel 4.8.B Gjennomsnittsvektor av bildevektorer

Kovariansmatrisen (C_x) til x må beregnes. D er diagonalen med egenvektorer og A er ortonormal matrise for C_x. Hvert bånd får en egenvektor og en egen otronormal vektor.

$$C_x = \frac{1}{M}\sum_{i=1}^{M}(x_i - m)\,(x_i - m)^T = ADA^T$$

$$A = a_1, a_2 \dots a_N \qquad D = diag(\lambda_1\,, \lambda_2 \dots \lambda_{N)}$$

Formel 4.8.C Kovariansmatrisen til x

A og x kan brukes til å finne PCA pikselvektorene.

$$A_i = A^T x_i\,, \;\; i = (1,2, \dots M)$$

Formel 4.8.D PCA piksel vektor

Dette her gir hver PCA verdi til hver piksel og de kan settes opp slik at vi får PCA båndene til det originale bildet (Rodarmel & Shan, 2002).

4.9 Maskinlæring og validering

4.9.1 Maskinlæring

Maskinlæring er en teknikk der en gir datamaskinen muligheten til å finne ut av resultatet selv. Den skal teste ut ulike metoder og beregninger, og lære av sine feil. Ved å kjøre flere iterasjoner og beregninger skal den klare å estimere eller klassifisere datasettet som vi mater datamaskinen med. For oppgaven skal maskinlæringsalgoritmene klare å gjøre regresjonsanalyser for å estimere AGB. Datamaskinen får datasett med arealer og spektralinformasjon i tillegg til AGB. Den bruker deler av datasettet til å lage en modell for estimering av biomasse, og tester den på en annen del av datasettet hvor den skjuler biomassen for seg selv. Denne metoden kalles for «supervised machine learning» der spektralinformasjon er input og AGB er output (MathWorks, 2018). Datamaskinen er god på å finne mønster i store kompliserte datasett. Datasettene kan være omtrent umulige å analysere for mennesker og likevel være en enkel prosess for datamaskiner.

Fire metoder for maskinlæringen brukes i denne oppgaven. SVM, Random Forest, Nevrale nettverk og lineær regresjon. Hver av dem blir gjennomgått kort om hva de gjør og hvordan de viktigste parameterne virker. Jeg forsøker å forenkle forklaringene mest mulig av algoritmene ettersom de er nokså kompliserte.

4.9.2 SVM

SVM står for Support Vector Machine. Dette er en utbredt metode som lenge har vært i bruk for klassifisering. I 1996 kom det en ny SVM som gjør regresjonsanalyser. Denne typen SVM kalles for SVR (Support Vector Reression) (Drucker et al., 1996). Metoden bruker hyperplan i flerdimensjonale rom som har som mål å separere verdier som ikke passer sammen., og den kan bruke ulike kerneltyper for dette som lineær eller polynomial kernel. SVM forsøker å minimere feil og holde seg innenfor regresionstoleransen for feil (epsilon) (Sayad). Epsilon og C må velges for SVM. C er cost penalty. Høy C gi større grad av overfit, og tillater

modellen å være fleksibel og endre seg mye. Lav C kan gi grad av underfit, hvor den ikke estimerer godt. Epsilon er området hvor feilen blir tolerert av modellen og cost penalty C ikke vil gjelde. Epsilon definerer en «true value» og det vil si at den her er den faktiske feilen i AGB som vi tolererer (Orange Documentation, 2015c). Det er da logisk at den må være en del lavere enn AGB, men den kan ikke være for lav. Er den for lav lager modellen overfit. For å kunne finne best mulig epsilon, anbefales det å finne ut hvor mye støy datasettet har (Cherkassky & Ma, 2002).

4.9.3 Random Forest

Random forest lager et sett med «decision trees» som inneholder deler av treningsdataen. Random forest deler opp attributtene i flere små sett i hvert av «decision trees». Attributtene er spektralinformasjonen, areal, vegetasjonsindekser eller andre egenskaper som algoritmen blir gitt. Hvert sett inneholder noen få av attributtene. Maskinlæringen tester de ulike attributtene i hvert sett og tar med den viktigste fra hvert sett. Deretter gjør algoritmen det samme om igjen flere ganger, men nå bare med de viktigste attributtene fra forrige forsøk. Slik fortsetter algoritmen til den finner de viktige attributtene og bruker de for regresjonsanalysen (Orange Documentation, 2015b). Viktige parameterne som velges for Random Forest er hvor mange attributter som skal være i hver oppdeling, og antall decision trees.

4.9.4 Nevrale nettverk

Nevrale nettverk forsøker å tenke på samme måte som nerveceller i hjernen gjør (Dvergsdal, 2017a). Nevrale nettverk forsøker å vektlegge ulike egenskaper forskjellig og teste seg frem til de som fungerer best. For å få gode resultater med nevrale nettverk er det veldig viktig med korrekte parametere. Ved for få iterasjoner klarer den ikke finne gode resultater. Regularization alfa må også velges. Dersom denne velges for høy vil modellen ende opp med overfit, men er den for lav vil modellen tilpasse seg for lite og føre til underfit. (Scikit-learn, 2017).

4.9.5 Lineær regresjon

Lineær regresjonsalgoritmer prøver å teste og lære hvilken lineær regresjon som passer best for et datasett. Algoritmen kan bruke flere variabler og gjøre multippel lineær regresjon. For lineære regresjon bør vi ha regularization (Orange Documentation, 2015a). Her kan vanlig ridge regularization brukes eller lasso regularization. Som oftest er rigde regularization med en lav alfa et godt alternativ. Når vi har svært mange variabler er lasso interessant. Denne prøver å bruke minst mulig variabler og reduserer mengden variabler som trengs (Aarshay, 2016). Den er god når vi har flere variabler som antageligvis ikke er spesielt gode.

4.9.6 Validering

Datasett bør valideres for å være sikre på at resultatene som en får er solide. Uten valideringen kan vi ikke se om resultatene er gode eller ikke. Kjører vi all data for å trene algoritmene og deretter bruker den samme dataen til å lage statistiske resultater for vi trolig gode verdier, men ikke realistiske resultater. Det er opplagt at en modell som lages og valideres på nøyaktig samme data gir et bra resultat. Resultatene sier ingenting om data overtilpasser eller om de fungerer for andre datasett. For å få til dette må vi dele opp datasettet slik at algoritmene trenes opp på deler av datasettet og valideres på en annen del. I oppgaven gjør jeg to former for valideringen. Den ene er «random sampling» og den andre er «leave one out».

Random sampling fungerer slik at vi deler datasettet opp i to. En treningsdel og en testdel. Treningsdelen brukes for å lære algoritmen opp, og testdelen brukes for å sjekke hvor bra resultatene blir. Datasettet deles opp flere ganger med tilfeldige deler av datasettet i hvert av settene (Orange Documentation, 2015d). Ved å gjøre dette flere ganger får vi en sikrere validering.

Leave one out trener modellen på hele datasettet bortsett fra en enkel instans (tre i mitt tilfelle). Deretter sjekker en hvor godt modellen passert for instansen. Nå velger den en annen instans som validering og trener opp en ny modell med resten av datasettet. Slik fortsetter valideringen frem til den har testet modellene for alle instanser. Dette er den sikreste valideringsmetoden som fungerer svært godt. Det som er problemet er at den må lage like

mange modeller som en har a instanser. Dette tar utrolig lang tid for store datasett (Orange Documentation, 2015d).

4.10Regresjon

Regresjonsmodellene som ikke kjøres i maskinlæringen i denne oppgaven er lineære med enten en eller flere variabler. Med flere variabler kalles det multippel lineær regresjon. Målet med regresjonen er å få linjen som passert best for biomasse. Vi klarer ikke lage en linje som passer perfekt for datasettet, men vi ønsker minst mulig feil. Den generelle formen til multippel regresjon er:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Formel 4.10.A Generell form av multippel regresjon (Mendenhall & Sincich, 1997)

I denne formelen er y den avhengige variabelen, i vårt tilfelle er det biomasse. x er de uavhengige variablene som i dette tilfellet blir spektralbånd eller vegetasjonsindekser. B viser bidraget til hver av variablene. Dette er koeffisientene til variabelen. Er den 0 har ikke variabelen noen som helst påvirkning av biomasse. ε er tilfeldig feil (Mendenhall & Sincich, 1997). Den lineære regresjonen vi bruker er minste kvadraters regresjon. Da er målet å gjøre kvadratsummen av vertikale avstander minst mulig. Det vil si at vi gjør slik at $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \cdots + \widehat{\beta_k} x_k$ gir lavest mulig SSE (sum of squared errors of prediction)

$$SSE = \sum (y_i - \hat{y}_i)^2.$$

Formel 4.10.B SSE (også kjent som RSS) (Mendenhall & Sincich, 1997)

Dette gir oss en lineær regresjonsmodell som har minst mulig feil. For å vite om denne modellen gir gode resultat og passer godt må en regne ut RMSE og R^2.

$$s^2 = MSE = \frac{SSE}{n - number\ of\ estimated\ \beta\ parameters} = \frac{SSE}{n - (k+1)}$$

Formel 4.10.C (Mendenhall & Sincich, 1997)

$$RMSE = \sqrt{MSE}$$

$$R^2 = \frac{Forklart\ varians}{Total\ varians} = 1 - \frac{SSE}{TSS}\ , der\ TSS = \sum (y_i - \bar{y})^2$$

Formel 4.10.E R^2 (Originlab Documentation)

I regresjonsmodellene vil vi få ulike mengder frihetsgrader basert på hvor mange variabler som brukes. Når det er mange variabler er det mindre frihetsgrader. R^2 bør korrigeres for dette. Det gjøres med denne formelen:

$$adjusted\ R^2 = 1 - \frac{\frac{RSS}{df_{error}}}{\frac{TSS}{df_{total}}} = 1 - \left[\frac{(n-1)}{n-(k+1)}\right] \times (1 - R^2)$$

Formel 4.10.F R^2 som, tar hensyn til antall frihetsgrader (Originlab Documentation)

Parameterne adjusted R^2, RMSE, MSE og SSE er de jeg ser på for å bestemme om en modell er god. Spesielt viktig er den justerte R^2 som bestemmer hvor godt en modell passer. I tillegg ser jeg også tabeller med alle residualer og faktiske estimat for å se at tallene er logiske. For eksempel dersom flere av estimatene blir negative er dette ikke logisk.

4.11 PLS analyse

PLS står for Partial Least Squares. Dette er en analysemetode som er ofte brukt når en skal gjøre regresjon. I denne oppgaven bruker jeg ikke selve regresjonen til PLS, men bare VIP (Variable Importance in Projection) verdiene for å finne ut hvilke variabler som er viktige å ha med for en vanlig minste kvadraters regresjon. Det vil si at PLS skal brukes til å redusere datamengden mest mulig. Normalt brukes VIP slik at en bare bruker variabler som har en VIP over en gitt terskel (ofte rundt 0,8 og 1,0). Jeg ønsker å bytte denne flere ganger og prøve å finne en verdi som reduserer denne maksimalt manuelt uten å svekke analysen til mine minste kvadraters regresjoner. VIP kan regnes ut med denne formelen:

$$VIP_j = \sqrt{\frac{p}{\sum_{m=1}^{M} SS(b_m.t_m)} \cdot \sum_{m=1}^{M} w_{mj}^2 \cdot SS(b_m \cdot t_m)}$$

Formel 4.11 VIP formel (Cassotti & Grisoni)

Cnonical Powered PLS antar at all relevant informasjon for å estimere y eksisterer i M. p er antallet variabler totalt, M er antallet gjenværende latente variabler. Algoritmen bruker bare de gjenværende latente variablene om gangen (antall skjulte variabler varierer med summasjonstegnet). Variabelen vil få et lavere resultat når flere variabler skjules og et bedre resultat når flere variabler er med. w_{mj} er PLS vekten for variablene. $_j$ for de vanlige variablene og $_m$ for de skjulte variablene. $SS(b_m \cdot t_m)$ er prosenten av y (i mitt tilfelle biomasse) som er forklart av de skjulte variablene (Cassotti & Grisoni). Formelen gir en VIP_j verdi for hver variabel. Er denne høy betyr det at variabelen inneholder mye nyttig informasjon.

5 Utsyr, programvare og metode

5.1 HySpex

HySpex består av to hyperspektrale kameraer og er produsert av Norsk Elektro Optikk AS (NEO). Det ene er et VNIR-1800 kamera og det andre er et SWIR-384 kamera. Totalt har dem 474 bånd, der 186 kommer fra VNIR og 288 kommer fra SWIR. VNIR gir oss synlig lys i tilegg til nærinfrarødt lys. Synlig lys er det et menneske kan se med øyet og er lys fra ca. 380 nm og opp til ca. 750 nm. VNIR starter på 407 nm og opp till 997 nm. Det vil se at den ikke tar opp fiolett lys som er 380-390 nm. SWIR starter på 955nm og slutter på 2523 nm. Det vil si at SWIR tar opp infrarødt lys. Det er noe overlapp mellom VNIR og SWIR siden begge tar området 955-997 nm. Begge sensorene har de som blir kalt «narrowband». Det vil si at båndene er svært små og tar bare opp det som er i et lite område av spekteret. VNIR har veldig smale bånd på 3,19 nm og SWIR har 5,46 nm. Pikselstørrelsen er også bedre på VNIR som har en oppløsning på 0,3 meter. SWIR har 0,7 meter (Jonassen & Aarsten, 2017).

Description	VNIR-1800	SWIR-384	Total
Number of bands	186	288	474
Spectral range	407 - 997 nm	955 - 2523 nm	407 - 2523 nm
Spectral resolution	3.19 nm	5.46 nm	
Spatial resolution (GSD)	30 cm	70 cm	
Field of view	17°	16°	
Normal operating altitude	1 300 m AGL	1 300 m AGL	
Normal swath width	390 m	370 m	

Tabell 5.1 Spesifikasjonene til HySpex (Jonassen & Aarsten, 2017)

Når vi ser på spesifikasjonene til HySpex ser vi at dette er et unikt sett med kameraer. Det er sjeldent at hyperspektrale flyfoto kan oppnå god romlig oppløsning i tillegg til å ha god spektral oppløsning. At kameraet dekker hele spekteret fra 400-2500 nm er også positivt. HySpex har et utrolig stort potensial og kan brukes til flere ulike analyser. VNIR-1800 er perfekt for vegetasjonsanalyser og for klassifisering og analyser av vei, bygg og materialer, er SWIR-384 veldig bra. Det er gjort lite forskning på hva SWIR kan bidra med innenfor vegetasjonsanalyse, og det kan være mulig at SWIR kan være nyttig for vegetasjonsanalyser

også. Det kan være interessant å se om SWIR kan brukes til å estimere biomasse eller detektere tre.

TerraTec har flydd med HySpex kameraene over Oslo. Området som har blitt flydd over inneholder vann, skog, parker og urbane områder av Oslo. Dette gjør at bildene kan brukes til flere analyser og forsøk. Oslo kommune har i samarbeid med TerraTec vært med på å betale for anskaffelsen av bildene og sitter på et svært verdifullt datasett. Området kan bli sett i figur 3.3.A. TerraTec har i tillegg også levert et normalisert radiansbildet og atmosfærekorrigerte flylinjer.

5.2 Feltarbeid

5.2.1 Datainnsamling av studenter og NINA

I september ble det gjort feltarbeid i Oslo. Det ble valgt ut ni ulike områder i Oslo som var innenfor de hyperspektrale flybildene. I tillegg hadde NINA (Norsk institutt for naturforskning) tre områder innenfor området til flybildene. Områdene til NINA er plassert noe tilfeldig og inneholder ikke nødvendigvis de beste områdene til å analysere i masteroppgaven. Feltarbeidet som ble gjort av studenter for å brukes for vegetasjonsanalyse, er bedre egnet. Områdene er 100x100 meter store og er valgt slik at de alle inneholder en god mengde med trær, og står i ulike områder som er typiske for Oslo. Områdene er bygårder, parker, lekeplasser, industriområder og trær langs veier. Dette er områder som en kan forvente å finne i de fleste urbane områder. Feltarbeidet gikk ut på å finne alle trær, og notere viktig informasjon som trengs for å lage en fasit til flyfotoene. Alle trær er fotografert med et håndholdt kamera, og alle trær over 5 meter har en tilhørende egenskapstabell. I tabellen ligger informasjonen: Treslag, område, høyde, stamme omkrets, antall stammer, trekrone lengde, trekrone bredde. I tilfeller der treslag var ukjent, ble det fotografert nærbilder av treets stamme og blader for å kunne klassifisere treet senere. Informasjon om stammen kan være nyttig for å si en del om treet, og den kan ikke oppdages fra flyfoto.

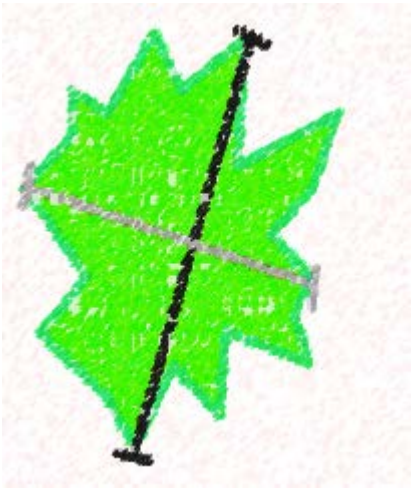
5.2.2 Egenskaper, målemetoder og nøyaktighet

Egenskapene som er lagret er valgt på grunnlag av det NINA bruker for sitt Urban EEA feltarbeid (Urban EEA vegetation survey sample summer 2017, 2017). For feltområdene fra studentene er egenskapene noe forenklet og inneholder ikke alle egenskapene som URBAN EEA bruker.

Det er varierende nøyaktighet på målingene. Målingene for stamme omkrets og kroneutbredelse er gjort med målebånd. Å måle stammeomkrets med målebånd gir nøyaktige målinger.

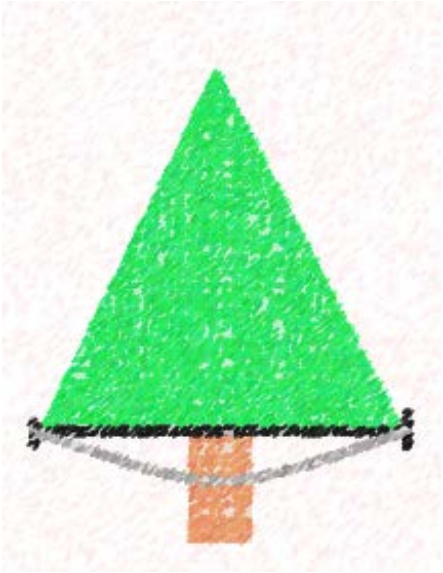
Figur 5.2.2.A Måling av omkretsen til tre gjort med målebånd.

Derimot er kroneutbredelsen mer krevende. Kronen er målt ved å først måle den lengste diameteren på kronen, og deretter måle lengden normalt på den. Dette er en forenkling av metoden i Urban EEA som måler kronen i Nord-Sør og deretter Øst-Vest. Et problem med metoden er at vi kan ofte bomme og ikke måle normalt. Dette gjør at metoden ikke er like konsekvent. Å bruke kompass og sikre at målene alltid er i nord-sør, øst-vest sikrer konsekvente målinger, men krever mer tid. På grunn av begrenset tid ble det antatt at å måle slik vi gjorde var godt nok for masteroppgavene.



Figur 5.2.2.B som viser hvordan trekronen er målt. Lengste diameter er målt (sort linje) og deretter er grå linje målt som står normalt på den sorte linjen)

Et annet problem med målingene av kronen er at de blir for store. For det første tar vi alltid den lengste diameteren og det kan representere et tre dårlig slik som vises i figuren. Dette kan gi for store bladareal om vi beregner bladarealet til kronen ved hjelp av arealet til en ellipse. Det andre problemet er at målebåndet går i en bue. Dette kan til en viss grad korrigeres for, men skaper noe usikkerhet i feltarbeidets målinger.



Figur 5.2.2C. (venstre) og Figur 5.2.2.D (høyre) Sort linje viser lengden på kronen til treet, den grå linjen viser avstanden vi får fra målebåndet. På bildet til høyre ser en at det er en bue på målebåndet.

Der er også en god del usikkerhet i høydemålingene. Ingen som var med på feltarbeidet hadde erfaring med å måle høyden på trær. Dette gjorde at målingene i høyde er noe upresise. Spesielt i starten var høydene unøyaktige, men etter hvert ble målingene bedre. Høydene ble sammenlignet med høyder fra bytre-datasettet til Bymiljøetaten og høyder fra laser. I Ekeberg og Lilleberg er høydene mest unøyaktige. Dette var de to første områdene som ble målt i feltarbeidet. Her er høydene konsekvent for store.

5.2.3 Oppdatering og korreksjon

Tabellene fra feltarbeidet ble lagret som en PostGIS database som ble brukt for å oppdatere ukjent informasjon og rette feil. Når databasen var ferdig, ble den konvertert til shapefiler i QGIS som har blitt brukt i analysen. En ny oppdatering av ukjente treslag ble senere gjort med hjelp fra eksperter i Plan og Bygningsetaten og Bymiljøetaten. For å validere at treslag var korrekt, ble det også tredatabasen til Oslo kommune sjekket. En del tre var feilklassifisert eller manglet klassifisering. Treslag er en fordel å vite for å beregne biomasse. Tre som ikke var mulig å klassifisere ble satt som enten løvtre eller bartre. Metoden kan håndtere ukjent treslag dersom det er bestemt om det er løvtre eller bartre (Field Studies Council).

5.3 Testområder

5.3.1 Vurdering og forkasting av områder

Til sammen var det ni områder fra studenter, tre områder fra NINA og et sett med bytrær i Oslo. Alle områdene fra feltarbeidet ble beholdt, men ikke alle trærne var brukbare. Trær som manglet informasjon om stamme eller høyde ble fjernet. Noen tre manglet informasjon fordi dem ikke var tilgjengelige eller glemt.

Områdene til NINA hadde svært få tre. I tillegg var det informasjon som manglet. På grunn av dette var det vanskelig å ta i bruk informasjonen far områdene og blir ikke brukt videre.

Datasettet til bymiljøetaten besto av flere tusen tre og hadde treslag korrekt i tillegg til en delvis korrekt posisjon. For klassifisering virker dette som et svært godt datasett. Dessverre var det stor usikkerhet på målingene i høyde og stammediameter. Flere av trærne hadde stammeomkrets, men den var målt omtrentlig og satt i grupper. For å beregne biomasse fasitdata var det nødvendig med nøyaktige stammemålinger og da holdt det ikke å vite at stammen var mellom 15-30 cm i diameter. For eksempel vil et tre med 30 cm stamme ha ca. 1,8 ganger større volum enn om det samme treet hadde 22,5 cm diameter. Dette gjør at en ikke kan bruke gjennomsnittet for gruppenes diameter for stammen fordi den har en svært stor betydning for volumet, som er hovedfaktoren for stammens biomasse.

5.3.2 Informasjon om gjenstående områder

Områdene som blir brukt er feltarbeidet fra studenter. Områdene er: Ekeberg, Lilleberg, Tøyen, Ulven, Helgesensgate, Haslevangen, Galgeberg, Gamle Oslo og Økern. Områdene er spredt ut i ulike områder av Oslo og alle er fullstendig innenfor de hyperspektrale bildene.

Figur 5.3.2A Oversiktskart over Oslo med alle testområdene som brukes. Den røde linjen viser området som er fotografert med HySpex. Bakgrunnskartet viser vegetasjon i Oslo. Grå farge er vegetasjon. Sirklene viser hvor områdene ligger.

Ekeberg er det første området som ble målt i feltarbeidet. Det er ikke et typisk byområde og ligger utenfor den urbane delen av Oslo. Ekeberg har mange løv og bartre. Noen står i klynger og noen tre står alene. Alle de røde og blå objektene en ser på bildet er telt som er satt opp grunn av Norway Cup. Området er et skog/parkområde. Høydene er nokså unøyaktig oppmålt.

Figur 5.3.2.B Ekeberg, bilde tatt fra HySpex. Under fotograferingen var Ekeberg full av telt og personer.

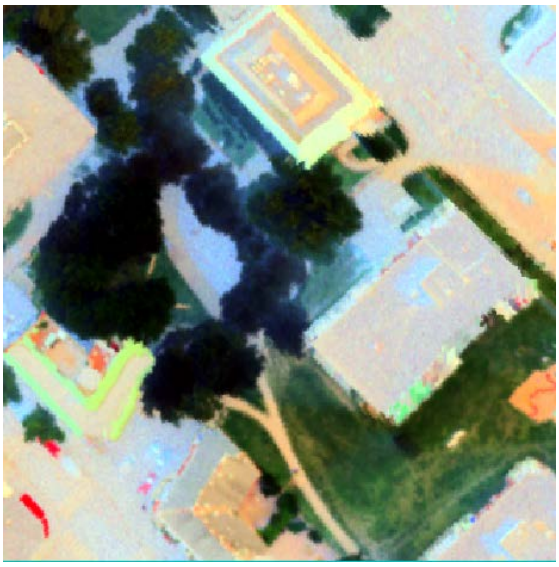
Lilleberg er et boligområde med en god del gress og vegetasjon. På fremsiden er det noen få trær, men på baksiden er det en stor klynge av tre. Trærne står i en bratt bakke med en bekk i bunnen av den. Noen av trærne er svært store. Alt er løvtrær. Vanskelig å finne treslag på alt, og høydene er lite nøyaktige.

Figur 5.3.2.C Lilleberg. Veldig mye overlapp i rekken med trær. Flere små trær er ikke mulige å se fra bildet fordi de store trekronene ligger ovenfor.

Ulven er et industriområde med omtrent ingen vegetasjon. Trærne står langs vei. Noen av trærne i økern er hogd ned, trolig fordi greinene begynte å vokse ut mot veien og blokkerte sikt. Totalt hadde Ulven bare fire trær.

Figur 5.3.2.D Bilde av Ulven. Feltarbeidet er gjort etter fotograferingen og noen tre har blitt hogd ned.

I Tøyen ble det valgt ut en lekeplass/park som område. Dette området hadde veldig mange tre (41 tre på 100x100 meter). Området ligger mellom 4 boligblokker og i en sentral bydel av Oslo med stor befolkningstetthet, men likevel et område med mye vegetasjon. Flertallet av trær står ikke i klynger, men på rad. De er av samme treslag og er trolig plantet på samme tid. Ofte overlapper kronen på trær i rekkene. Tøyen har mye svenskeasal og lind. Svenskeasal er ikke et typisk tre å se i skog, men det blir plantet en del i Oslo. Lind er treslaget det er mest av i Oslos bytredatabase. I vest på bildet er det en stor klynge med trær, her er det ulike treslag og mye overlappende trær.



Figur 5.3.2.E utsnitt fra Tøyen. Urbant område som likevel har mye tre og annen vegetasjon.

Gamle Oslo området har to bygårder. Innenfor bygårder har kommunen mindre informasjon om trærne. Bytredatabasen tar ikke med private tre eller tre innenfor bygårder. Her er det en stor variasjon i vegetasjon og treslag. Her er det store tre som asp og or, men og flere ulike prydbusker og små tre plantet for sin estetikk. Noen av trærne har vært vanskelige å få nøyaktige mål på fordi de er inngjerdet. To trær i området var inngjerdet og har derfor bare en estimert diameter på stammen.

Figur 5.3.2.F Gamle Oslo. Fra utsiden kunne en ikke se mye vegetasjon, men på innsiden av bygården var det flere trær.

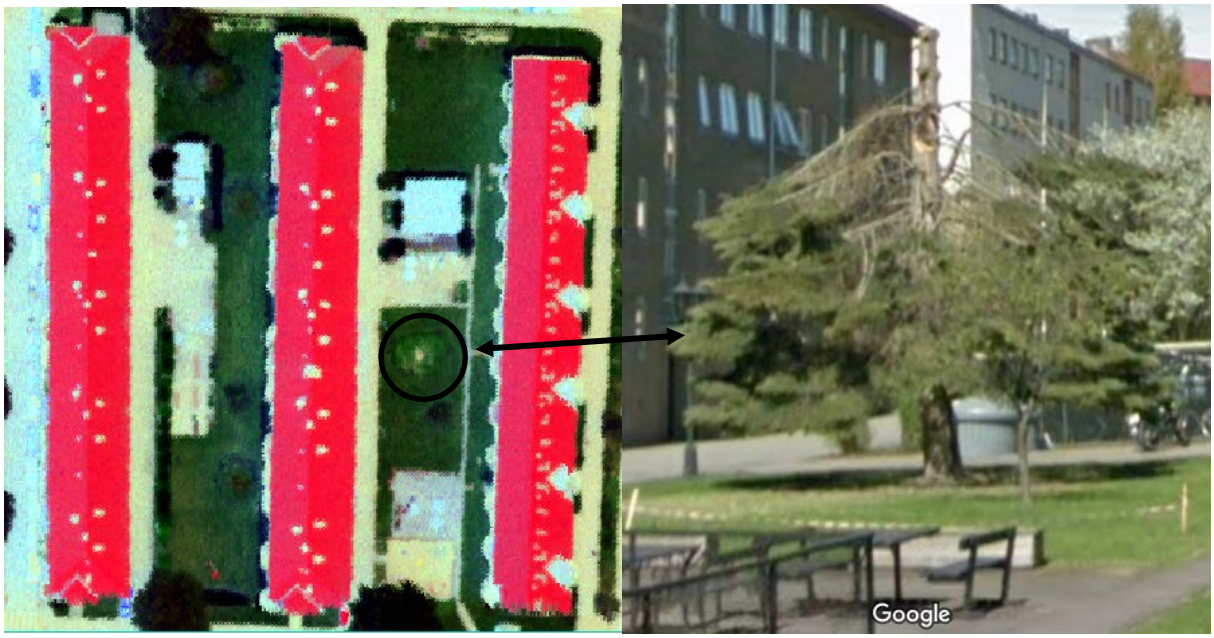
Økern er et industriområde med lite vegetasjon og har noen likheter med Ulvenområdet. Dette området er også nær trafikkert vei. Alle trær her var små trær på rundt 3-6 meter høyde. Det var også små trekroner og en del trær så ut til å være i noe dårlig helsetilstand. Trærne stod i klynge, men hadde små nok trekroner til å ikke overlappe.

Figur 5.3.2.G Økern. Industriområde med lite tre og stresset vegetasjon.

Haslevangen var et tidligere industriområde, men som i nyere tid har blitt bygd om til et boligområde. Her var det store bjørketrær langs veien. I midten av området var det en park rundt nyere boligblokker. Her var trærne fortsatt små, ca. 5 meter høye. Her var det piletrær, bøk, or og rogn. Noen tre var inngjerdet og noen stod i privat hage. De ble ikke målt i feltarbeidet.

Figur 5.3.2.H Haslevangen er et område der industrien er på vei bort. Nå er området fullt av nyplantede trær.

Det siste området som ble målt var Helgensgate. Helgesensgate er en del av Grünerløkka og består av mye boligblokker. Området har mye vegetasjon mellom byggene og store tre og mye busker og gress. Det var bjørketre og lindetre, grantre og syrin i området. Det største treet i området var hogd ned da vi gjorde feltarbeidet. Når flyfotoet er tatt kan en se at det ikke er hogd ned, men at det likevel var noe rart med treet. Dette ga en noe annerledes spektralsignatur og med en synlig stamme i midten av treet. Ved å bruke streetview viste det seg at halvparten av treet var hogd ned. Trolig ga bildet oss signaturen til stammen og greinene på toppen uten barnåler.



Figur 5.3.2.I Bilde av området i Helgesensgate og treet som skapte noe forvirring.

5.4 Filformater og programvare

5.4.1 PCI Geomatica

Det første programmet jeg brukte i masteroppgaven var PCI Geomatica. Dette er et program som er laget for fjernmåling. Hovedgrunnen til å bruke Geomatica var at den støttet Python. Python integreringen fungerte fint og denne var god for å lese inn filer og lage vegetasjonsindekser og eksportere til format som QGIS og ENVI kunne lese.

Dessverre oppsto det flere vanskeligheter med Geomatica. En av dem var mangelen på innstillinger for PCA. Den virket ikke spesielt god å bruke og ENVI hadde en PCA som viste bedre hvilke bånd som var relevante for PCA-komponentene. Det var også mer strevende å få bilder til å få korrekt kontraststrekking og å få de til å se visuelt gode. Alt dette gjorde at det eksplorative tidlige arbeidet og analysen gikk tregt. Etter hvert oppsto flere problemer som at Geomatica manglet lisens for en del vegetasjonsindekser og funksjoner. Dette gjorde at nærmest alt jeg ønsket å gjøre måtte jeg selv programmere i Python. På grunn av dette gikk jeg over til ENVI som hadde flere av de ønskende funksjonene innebygd.

5.4.2 ENVI

ENVI er et annet program som er laget for fjernmåling. ENVI har svært mange funksjoner og innebygde vegetasjonsindekser. Programmet er bra for eksplorativ analyse, beregning av indekser, klassifisering (Harris Geospatial solutions, 2018b). ENVI er ofte brukt for satellittbilder og har også innebygde atmosfærekorreksjoner.

ENVI hadde for meg flere fordeler over PCI Geomatica. Blant annet var det enklere å få rett kontraststrekking og lettere å lage visuelt gode bilder. Spesielt bra var samlingen av innebygde vegetasjonsindekser og PCA funksjonen. Utfordringer med å bruke ENVI var brukervennligheten. Flere ganger fant jeg ikke vegetasjonsindeksen eller funksjonen jeg trengte. Blant annet sammenslåing og resampling av bilder fant jeg ikke i ENVI. Da endte jeg opp med å selv programmere det i IDL, som er programmeringsspråket som er integrert i

ENVI. Når jeg hadde endelig laget koden viste det seg at funksjonene jeg trengte allerede var i programmet, men vanskelige å finne. I slike tilfeller har jeg forkastet egen kode fordi den var tregere enn innbygde funksjoner. Et annet problem som oppsto var mangelen av feature extraction lisens. Ved å ha denne kunne jeg ha spart mye tid når det skal samles statistikk for objekter. Dersom forsøket skal rekonstrueres eller metoden skal kjøres med ENVI, anbefaler jeg å få tak i denne lisensen.

5.4.3 Python

Python er et av de mest kjente programmeringsspråkene. Det er veldig mye brukt innefor forskning, dataanalyse og automatisering (Dvergsdal, 2017b). Python har flere utvidelsespakker som gjør det i stand til å kunne beregne vegetasjonsindekser, kjøre maskinlæring og regresjon. Python ble brukt sammen med PCI Geomatica for å lage vegetasjonsindekser i starten ved hjelp av PCI-Python utvidelsen som er god for å lese filer, beregne indekser og kjøre Geomaticas funksjoner i Python. Ved overgangen til ENVI kunne ikke Python brukes like enkelt. Det er ingen integrering mellom ENVI og Python. Senere ble også Python og utvidelsespakken scikit-learn pakken testet for å kjøre maskinlæring. Mer om dette er i kapittel 6.5.

5.4.4 Origin Pro

Origin og Origin Pro er statistikkprogram som inneholder svært mange regresjonsmodeller og fitting modeller. Programmet kan også kjøre PLS analyse og PCA for å finne viktige variabler og sammenhenger. Programmet har også flere funksjoner for å kjøre statistiske tester for å se om resultater er signifikante. I oppgaven brukes flere regresjonsmodeller sammen med PLS for å lage funksjonsuttrykkene for AGB. Alle variabler testes også med Student T-test for å se om variabler er signifikante. Origin er et godt alternativ dersom en ønsker en ferdig pakke som inneholder alt det viktigste av statistiske funksjoner og tester.

5.4.5 Orange

Orange er et gratis dataminingsprogram med åpen kildekode (Orange, 2018). Det er laget for å visualisere og kjøre maskinlæring på store datasett. Orange er lett å lære og bruke, og kan integreres med Python. Mye av funksjonene i Orange kommer fra Scikit-learn. Orange kan håndtere blant annet CSV filer og vi kan velge targets og variabler. Programmet kan selv ved hjelp av maskinlæring finne ut estimater for biomasse eller andre targets. Orange kan klassifisere og beregne estimater ved hjelp av blant annet lineær regresjon, SVM, Random Forest, nevrale nettverk og logistisk regresjon. Programmet kan også kjøre PCA, eksportere data og lage grafer/scatterplots. Alle resultater kan lagres i rapporter. Orange brukes i oppgaven for å kjøre maskinlæringsalgoritmer for å estimere AGB. Orange har noe mindre funksjonalitet enn Python og tillater ofte bare at de vanligste parameterne for maskinlæring kan justeres.

5.4.6 eCognition

eCognition er for denne oppgaven veldig aktuelt for å segmentere delvis automatisk. Cognition er en svært populær programvare innenfor arealbasert romlig data. eCongition brukes mye innenfor segmentering og klassifisering (Ouyang, 2015). eCognition har en veldig god segmentering som heter multiresolution segmentation i tillegg til flere andre segmenteringer som også er aktuelle (Darwish et al., 2003). Når en ønsker å bruke metoden 8.8 trenger en segmentert data, og da er trolig eCognition det beste programmet for å segmentere og lagre data.

5.4.7 QGIS

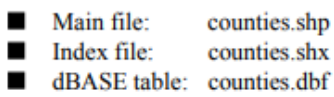
QGIS er et GIS program med åpen kildekode. QGIS er gratis og et av de mest populære produktene for analyse, prosessering og visualisering av romlig data (University of Pennsylvania, 2018). QGIS kan gjøre svært mange operasjoner for både raster og vektor datasett. Det kan også brukes sammen med flere ulike tilleggspakker. I forhold til denne oppgaven er QGIS aktuelt for å laste inn databaser og eksportere til vektorformat, lagring av egenskaper for trær, redigeringer og oppdatering av data og manuell tegning av trekroner.

QGIS var effektivt for oppdatering av trær og rasterkalkulatoren er mye brukt i masteroppgaven. QGIS er programmet jeg bruker for å binde sammen vektordata og rasterdata siden den kan lese PostGIS databasen og lese inn HySpex bildene. Omtrent alt oppdatering og justering av punkter og egenskapstabeller skjer i QGIS. Tegning av trær var noe tungvint i QGIS, og defor gjøres dette i ENVI.

5.4.8 Filformater

Flere ulike typer filformater er brukt i metoden. De viktigste formatene er ESRI Shapefile, ENVI hdr/image file, ENVI roi og CSV. De eksiterer flere alternativer til filformatene som nevnes her. Filformatene er valgt fordi de passer til programvaren som jeg har valgt bruke i kapittel 7.

ESRI shapefiles er et filformat som er produsert av ESRI. Det brukes for å lagre attributt-tabeller med egenskaper for objekter i tillegg til objektenes geometri for romlig vektordata (ESRI, 1998). Shapefiler er laget for å være raske å lese og enkle å redigere. For å gjøre dette er filen separert i flere deler. Attributter lagres i dBASE format (mye brukt databaselagringsformat) istedenfor å være i samme del som geometrien. Det gjør at punkter kan leses inn og redigeres uten å laste attributter. Samme gjelder for attributtene som kan redigeres uten å lese punktene. Alle attributter har en en-til-en kobling med den geometriske delen av filen. Geometrien som shapefiler kan lagre er punkter, linjer og arealer (ESRI, 1998). Shapefiler kan leses av de fleste GIS programmer. GIS programmer som QGIS, ArcGIS og ENVI støtter Shapefiler. Siden shapefiler er enkle å redigere og kan leses av både QGIS og ENVI er det et fornuftig filformat å bruke for lagring av egenskaper til trær og overføre romlig data mellom ENVI og QGIS.



Figur 5.4.8.A Hovedstrukturen til en shapefil. Attributter er separert fra geometrien og lagres i dBASE. (ESRI, 1998)

ENVI hdr og ENVI image file er formatet som ENVI bruker som default files. ENVI hdr er headerfilen og ENVI image file er selve bildet. Bildefilen består av en binær rasterfil med bildedata. Filen inneholder ingen metadata. Rekkefølgen på bytes i den binære bildefilen kan variere. Det er tre ulike format som brukes og kalles ofte for «interleave type». De ulike metodene er BSQ (band sequential), BIP (Band interleaved by pixel) og BIL (Band interlave by line). (Harris Geospatial solutions, 2018d). Dette har med rekkefølgen bytes leses inn i. I oppgaven har filene blitt lest inn som BSQ. BSQ er regnet for å være den mest effektive metoden for å få tak i den romlige posisjonen til et enkelt spektralbånd. BSQ leser inn et og et bånd om gangen for hele bildet. Det gjør at den er litt tregere når den skal prosessere med hele spektralsignaturen i flere piksler.

Derimot BIP som leser inn alle bånd for en og en piksel er effektiv på å skaffe spektralsignaturen, men tregere romlig posisjon av enkeltbånd. Siden det er mye spektralsignaturer vi ønsker å se på, kan denne metoden være mer effektiv. Den siste metoden BIL, leser inn en linje om gangen istedenfor en piksel. Metoden er noe lik BIP, men litt tregere for spektralsignaturer, men raskere for romlig prosessering. ENVI anbefaler å bruke BIP siden det er den mest balanserte måten å lese og prosessere filen (Harris Geospatial solutions, 2018d). For analysene som er gjort i ENVI ville nok både BIP og BIL være raskere enn BSQ og kunne nok spart litt tid i prosessering. Siden BIL var anbefalt av ENVI trodde jeg dette var default for bildene, men det viste seg at de fleste var BSQ.

ENVI hdr lagrer alt metadata som trengs til bildefilen. Her vil det stå om filen må leses som BSQ, BIP eller BIP. Det vil også stå tidspunkt bildet er produsert, navn på bånd og bølgelengde, datatype, klassifiseringer, projeksjoner, datum, pikselstørrelse etc. All informasjon som lagres om et bilde ligger i headerfilen siden den binære filen inneholder bare rasteret. ENVI header file lagres som en ACII fil og en kan enkelt redigere filen ved hjelp av enkle programmer som notepad. ENVI har også en funksjon i programmet for å lese og redigere headerfiler (ENVI Help, 2007; Harris Geospatial solutions, 2018c). De fleste operasjoner og filer som lages og brukes i ENVI er i image file/header file formatet. Unntaket er ROI filer som lagres som ENVI ROI filer og eksporteres til CSV.

```
ENVI
description = {
  Create Layer File Result [Fri Mar 09 10:04:47 2018]}
samples = 334
lines   = 334
bands   = 369
header offset = 0
file type = ENVI Standard
data type = 4
interleave = bsq
sensor type = Unknown
byte order = 0
map info = {UTM, 1.000, 1.000, 600498.150, 6643782.800, 3.0000000000e-001, 3.0000000000e-001, 32, North, WGS-84, units=Meters}
coordinate system string = {PROJCS["UTM_Zone_32N",GEOGCS["GCS_WGS_1984",DATUM["D_WGS_1984",SPHEROID["WGS_1984",6378137.0,298.2
wavelength units = Nanometers
band names = {
 Layer (Resize (Band Math (Band 1:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 2:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 3:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 4:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 5:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 6:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 7:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 8:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 9:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 10:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
 Layer (Resize (Band Math (Band 11:VNIR_mosaic.bsq):VNIR_normalized_mosaic.dat):VNIR_3_Lilleberg),
```

Figur 5.4.8.B En del av en headerfil. Filen er fra en sammenslått SWIR og VRIR mosaikkfil

ROI står for region of interest. Den brukes når en ønsker å tegne opp objekter som en ønsker å se på. I oppgavens tilfelle er det trær som er interresant. Da kan alle trær tegnes opp og lagres som en ROI fil. ROI filen vil bestå av et raster med bare pikslene til objektene som er tegnet inn. ROI kan brukes for å filtrere bort det som vi ikke ønsker å jobbe med i tillegg til at ROI også kan brukes til å generere en statistikk fil med båndverdier og gjennomsnitt for objekter. ENVI kan eksportere ROI filene til vektorformat som ESRI shape eller som CSV tabell (Harris Geospatial solutions, 2018e).

CSV er et tabellformat. CSV står for comma-seperated value files. CSV bruker ASCII tekst og separerer hver kolonne med et komma (noen program bruker semikolon) og hver linje er en ny rad i tabellen (Shafranovich, 2005). Fordelen med CSV er at omtrent alle programmer kan importere og eksportere filformatet. Det er et enkelt format å bruke og QGIS, ENVI, Excel, Origin, og Orange kan tolke CSV filer. Statistikk, spektralinformasjon fra bånd, pikselantall og metadata lagres i CSV. Vegetasjonsindekser for gjennomsnittsverdier genereres i CSV filen direkte ved hjelp av Excel. Det er CSV filen jeg gir til Origin og Orange for å gjøre regresjon, maskinlæring og analyse. For at Origin og Orange skal tolke filen korrekt må det brukes semikolon og ikke komma for å separere kolonner.

5.5 Metode

Metoden er gjennomgått i dette kapittelet. Det er laget slik at metoden og forsøkene skal kunne rekonstrueres, forbedres og testes. Metoden min er inspirert av Bernasconis artikkel og «single tree» metoden fra artikkelen. Metoden har blitt brukt for nokså få trær og i små områder i denne oppgaven. Det har ikke vært nok tid eller feltarbeid til å gjøre analysen i storskala. I kapittelet fremgangsmåte (kapittel 7) blir metoden gjennomgått i programvaren som jeg har valgt å bruke. Da blir den gått gjennom mer detaljert og konkret. Resultater og problemer som oppstår blir og tatt med der. Dersom det er et ønske om å bruke en hyperspektral biomasseestimering bør en vurdere hvilken nøyaktighet enn trenger. Dersom en trenger bedre nøyaktighet bør en gjøre slik som det står forklart i 5.7, men med flere trær. Dersom en mener nøyaktigheten fra min biomasseestimering er god nok, er det ikke nødvendig å rekonstruere hele metoden for å få resultater. Da kan en bruke funksjonsuttrykket og den forenklede metoden i kapittel 8.8

Metoden her antar at SWIR og VNIR bildet er ferdig bearbeidet og klar for bruk. HySpex bildene fra SWIR og VNIR er tatt fra samme flygning og er slått sammen til et bilde før en gjør metodeprosessen.

Metoden som jeg har laget i oppgaven er basert på og inspirert av single tree segmentation og arbeidet fra (Bernasconi et al., 2017)

5.5.1 Fasitdata

Fasitdata er basert på Forestry Commisions metoder og (Field Studies Council), og bruker data samlet inn fra feltarbeid. I selve fremgangsmåten er beregningene gjort for AGB, men også total biomasse kan estimeres. Eneste forskjellen mellom AGB og total biomasse er biomassen til røttene. Biomassen for røtter vises også i 5.7.1.

Biomassen er delt opp i tre deler. Den første er stammen, den andre kronen og den tredje er røttene. Starter med å finne biomassen til stammen. For å finne biomassen til stammen må en vite stammeradius,høyde og nominal specific gravity. Høyden gis i meter når en beregner stammebiomasse. Nominal specific gravity kan en finne i tabeller og varierer for hvert treslag. Dersom treslag er ukjent kan en forenklet verdi for nominal specific gravity brukes. Da må en bare vite om treet er bartre eller løvtre. Starter med å skaffe radiusen til treet i brysthøyde om en ikke har den allerede og deretter volumet:

$$Radius = \frac{omkrets}{2\pi} \qquad Radius = \frac{DBH}{2}$$

Formel 5.5.1.A Formler for å finne radius basert på om en har omkretsen eller diameteren

$$Volum = \pi r^2 \times \left(\frac{h}{3}\right)$$

Formel 5.5.1B Volum av stammen der r = radius og h = høyde. (Field Studies Council)

Stammens biomasse er enkel å finne når en har volumet. Alt en trenger å gjøre da er å multiplisere nominal specific gravity med volumet. Nominal specific gravity er vanligvis rundt 0,5 for løvtrær og mellom 0,3 og 0,4 for bartrær. Se tabell 5.7.1.A for tabell over nominal specific gravity for ulike treslag. Dersom en ikke har treslaget kan verdien 0,53 brukes for løvtrær, og 0,39 for bartrær.

$$Biomasse\ stamme = Nominal\ Specific\ Gravity \times \text{Volum}$$

Formel 5.5.1.C Stammens biomasse

Species	Allocated Species	Nominal Specific Gravity (NSG)
Scots pine (SP)	XP,	0.42
Corsican pine (CP)	AUP, BIP, RAP, PDP	0.40
lodgepole pine (LP)	MOP, MCP	0.39
maritime pine (MAP)		0.41
Weymouth pine (WEP)		0.29
Sitka spruce (SS)		0.33
Norway spruce (NS)	XS, XC, MC	0.33
Omorika spruce (OMS)		0.33
European larch (EL)		0.45
Japanese larch (JL)		0.41
hybrid larch (HL)		0.38
Douglas fir (DF)		0.41
western hemlock (WH)		0.36
western red cedar (RC)	JCR	0.31
Lawson cypress (LC)		0.33
Leyland cypress (LEC)		0.38
grand fir (GF)	RSQ, WSQ	0.30
noble fir (NF)	XF	0.31
silver fir (ESF)		0.38
oak (OK)	POK, SOK	0.56
red oak (ROK)		0.57
beech (BE)		0.55
sycamore (SY)	NOM, RON, MB, XB	0.49
ash (AH)		0.53
birch (BI)		0.53
poplar (PO)		0.35
sweet chestnut (SC)		0.44
horse chestnut (HCH)		0.44
alder (AR)	CAR, GAR, RAR, SAR, VAR	0.42
lime (LI)	CLI, SLI, LLI	0.44
elm (EM)	EEM, SEM	0.43
wych elm		0.50
wild cherry, gean (WCH)	BCH	0.50
hornbeam (HBM)		0.57
raoul		0.37

Tabell 5.5.1.D tabell med nominal specific gravity for ulike treslag. (Jenkins et al., 2011)

Det neste er å finne biomassen til trekronen og greiner. Dette gjøres ved å bruke fire konstanter i tillegg til stammediameteren. To av konstantene er for tre under 50 cm i stammediameter og to er for trær over 50 cm. Det er ingen konstanter for trær som er under 7 cm i stammediameter. Dersom en har tre med stammediameter mindre enn 7 cm kan de fjernes fordi metoden ikke fungerer for dem. Dersom et tre har en stamme under 50 cm blir kronens biomasse beregnet med formel 5.5.1.C og dersom den er over 50 cm brukes formel 5.5.1.D. Grunnen til at det brukes to ulike formler er fordi estimeringer som bruker samme formel for både store og små trær ender opp med å overestimere eller underestimere trær. Det er vanskelig å finne et uttrykk som passer for alle trær. Det er viktig å huske at formelen bruker centimeter og ikke meter.

$$Kronebiomasse = a \times DBH^{b}$$

Formel 5.5.1.E Brukes for trær med stamme under 50 cm (Field Studies Council)

$$Kronebiomasse = c + (d \times DBH)$$

Formel 5.5.1.F Brukes for trær med stamme over 50 cm (Field Studies Council)

For å vite hva a, b, c og d skal være trenger en treslaget. For trær med ukjent treslag trenger en bare vite om det er bartre eller løvtre. Løvtrær bruker konstantene for eik dersom en ikke har treslag, og Nobeledelgrans konstanter brukes om det er et bartre uten nøyaktig treslag.

Estimating crown biomass				
Species	**a**	**b**	**c**	**d**
Larch	0.000044	2.0291	-0.129047	0.005039
Corsican pine	0.000012	2.4767	-0.299529	0.009949
Lodgepole pine	0.000018	2.4767	-0.430537	0.014300
Scots pine	0.000016	2.4767	-0.394206	0.013094
Douglas fir	0.000017	2.4767	-0.411768	0.013677
Grand Fir	0.000015	2.4767	-0.353198	0.011732
Noble Fir & other conifers	0.000015	2.4767	-0.353198	0.011732
Hemlock	0.000015	2.4767	-0.353198	0.011732
Norwegian Spruce	0.000015	2.4767	-0.353198	0.011732
Cedar	0.000015	2.4767	-0.353198	0.011732
Sitka Spruce	0.000015	2.4767	-0.353198	0.011732
Beech, Sycamore & Maple	0.000019	2.4767	-0.459519	0.015263
Oak and all other broadleaved trees	0.000017	2.4767	-0.411551	0.013670

Tabell 5.5.1.G Tabell som viser konstantene for å beregne trekronenes biomasse (Field Studies Council)

Når biomassen er beregnet for både trekronen og stammen kan en addere dem for å få AGB:

$$AGB = Kronebiomasse + Stammebiomasse$$

Formel 5.5.1.H Formel for AGB når en har stammen og trekronens biomasse

For å finne den totale biomassen må også røttene tas med. Dette gjøres på samme måte som trekronen. Da må en vite treslaget og for de ukjente må en vite om det er bartre eller løvtre.

Tre konstanter i tillegg til DBH trengs for å finne biomassen til røttene. Den første formelen er for trær under 50 cm diameter og den andre er over 50 cm diameter.

$$Biomasse\ røtter = e \times DBH^{2.5}$$

Formel 5.5.1.H Biomassen til røttene for trær med stammediameter under 50 cm (Field Studies Council)

$$Biomasse\ røtter = f + (g \times DBH)$$

Formel 5.5.1.I Biomassen til røttene for trær med stammediameter over 50 cm (Field Studies Council)

Species	e	f	g
Larch	0.000017	-0.133480	0.007296
Corsican pine	0.000011	-0.082603	0.004515
Lodgepole pine	0.000017	-0.133480	0.007296
Scots pine	0.000015	-0.118673	0.006487
Douglas fir	0.000017	-0.133480	0.007296
Grand Fir	0.000015	-0.118673	0.006487
Noble Fir	0.000011	-0.082603	0.004515
Hemlock	0.000015	-0.118673	0.006487
Norwegian Spruce & other conifers	0.000012	-0.091547	0.005004
Cedar	0.000011	-0.082603	0.004515
Sitka Spruce	0.000021	-0.157579	0.008614
Beech	0.000023	-0.174882	0.009559
Oak and all other broadleaved trees	0.000023	-0.174882	0.009559

Tabell 5.5.1.J Tabell med konstantene som brukes for å beregne biomassen til røttene (Field Studies Council)

For å finne den totale biomassen adderes biomasse røtter med AGB:

$$Total\ Biomasse = Biomasse\ røtter + AGB$$

Formel 5.5.1.K Total biomasse for et tre

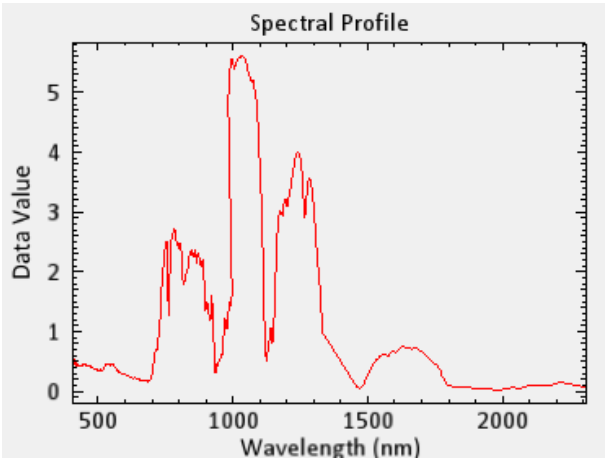
Det som ikke er nevnt her er problemet når stammen er todelt. Dersom treet splitter over brysthøyde har det vært antatt å være en stamme og beregnet slik som vist ovenfor. Dersom et tre har to eller flere stamme før brysthøyde, regnes biomassen for alle stammene. For trekronen og røttene endres ingenting.

5.5.2 Fjerning av bånd med støy

I hyperspektrale bilder forventer en å finne noen bånd som inneholder støy. Bånd med støy vil ikke gi noen brukbar informasjon og kan føre til å gi uklare inkonsistente resultater. Støy kan komme fra at noen bånd ikke har fungert og da inneholder ingenting eller bare tilfeldige intensiteter. Et annet problem er områdene for vannabsorbsjon. Båndene som viser hvor vann blir absorbert inneholder ofte mye støy og det er lite informasjon å hente her. Atmosfæren er lite gjennomsiktig der den absorberer vann. Flere av båndene med støy viser bare vannet i atmosfæren og trenger ikke gjennom. Informasjonen i de båndene forklarer lite eller ingenting annet enn atmosfærens vannabsorbsjon. Båndene med støy er best å fjerne tidlig slik at en ikke trenger bruker harddiskplass, tid og prosesseringskraft på data som ikke kan brukes til noe. Det er lurt å se gjennom alle bånd og lete etter støy. Det er helt normalt å ha flere bånd med vannabsorbsjon i hyperspektrale bilder og som regel blir de fjernet eller ikke brukt. Det er og bånd der vegetasjon absorberer vannet og de kan være nyttige. Problemet er at de er ofte på samme plass som atmosfæren absorberer vann.

Absorbing biochemical	Wavelength (nm)
Water	970, 1200, 1400, 1450, 1940

Tabell 5.5.2.A Utsnitt fra tabell i artikkelen (Serrano et al., 2002). De forventes at noen av båndene i områdene kan inneholde veldig mye støy.



Figur5.5.2.B Den første store økningen er red edge. Den andre og tredje økningen som er markert med blå sirkel er vannabsorbsjon. Legg merke til at dem starter på 970 nm og 1200 nm. Vannabsorbsjonen rundt 1400 nm er fjernet fullstendig her og er bare en rett linje siden den inneholder ingen punkter. 1400 nm inneholdt ingen informasjon og var ren støy.

Støy kan oppdages på flere måter. En kan bruke PCA og fjerne bånd som forklarer minst varians. Dersom noen av båndene ikke gir noe informasjon, er de trolig full av støy. Dersom de ikke inneholder støy er det likevel ikke farlige å fjerne de siden PCA viser til at båndet har lite informasjon. Å se visuelt gjennom hvert bånd og lete etter støy er mulig. Dette tar noe tid og metoden er vanskelig å automatisere. Fordelen med å se gjennom hvert bånd for støy er at da kan en være sikker på at en fjerner de rette båndene. Andre aktuelle metoder kan være å gjøre en partial least square (PLS) analyse. Denne vil vise hvilke bånd som er viktige for estimering av biomasse. Svakheten med denne er at dette ikke kan gjøres før en har både definert tre, laget fasitdata og lagret statistikk. Med denne metoden må en ha med alle båndene med frem til analysedelen.

5.5.3 Valg av trær

For at metoden som blir forklart i oppgaven skal fungere, må en ha hele trekronen til et tre. Siden den viktigste variabelen for å estimere biomasse er pikselstørrelse, er det viktig at arealet på trekronen er riktig. Det vil si at både for store og for små trekroner fører til upresise resultater. Alle trekroner som har overlapp med andre tre eller deler av trekronen ikke vises i bildet fører til problemer. Alternativ en er å fjerne alle trær som ikke har hel trekrone synlig i bildet. Dette reduserer størrelsen på datasettet fort, spesielt dersom mange trær står i klynger eller skog. Derimot har tre langs vei sjelden overlapp. Det andre alternativet er å sjekke hvor mye av arealet som mangler. Det gjøres med den enkle formelen:

$$A = a_s + \left(\frac{100}{a_p} - 1\right) \times a_s$$

Formel 5.5.3.A Formel for å finne ut hvor stor en overlappet krone er der A = Arealet til hele trekronen i piksler. a_s er piksler som er synlig i bildet, og a_p er arealet som er synlig i prosent.

Fordelen med å bruke er formelen at det blir mulig å beholde langt flere trær og kan få et større datasett å gjøre analyse på. Svakheten er at det skaper usikkerhet. Det er vanskelig å si om en ser 30% eller 40% av en trekrone. Det er også vanskelig å automatisere enn metode som sjekker om en trekrone er hel eller ikke.

Figur 5.5.3.B Ved Galgeberg står det flere trengs langs veien. Trær som dette er enkle siden dem ikke har noen overlapp.

Figur 5.5.3.C Ekeberg har tett skog. Trekronene overlapper hverandre og her er det mange trekroner som ikke kan brukes. De største kan tas med siden de ikke har noen overlapp.

Når en skal velge trær og testområder som skal brukes bør en tenke gjennom hvilke trær en ønsker. Fordelingen av treslag og høyden på trær innenfor treslag har mye å si for resultatene. For at resultatet skal bli best bør en ha omtrent like mange av hvert treslag og like mange store og små trær innenfor hvert treslag. Dersom en ikke tenker gjennom treslagene kan en få en dårlig fordeling som skaper dårlige resultat. For eksempel kan vi ha et datasett med 50 trær. 25 trær er små løvtrær med lav biomasse og 25 trær er svært store bartre med høy

biomasse. Regresjonsmodeller og maskinlæring vil tolke dette som om alle løvtrær er små og alle bartrær er store. Da vil spektralsignaturen til bartre indikere høy biomasse og spektralsignaturen til løvtrær indikere lav biomasse. Dette vil være en svært dårlig fordeling siden det ikke er slik i virkeligheten. Trær bør være fordelt med varierende størrelse innenfor alle treslag.

5.5.4 Deteksjon og segmentering av trær

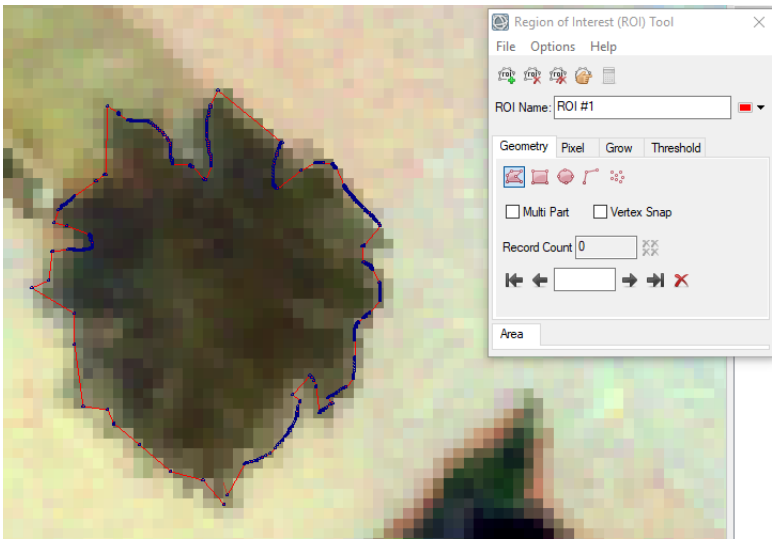
Før en kan begynne å se på biomasse må det lages et sett med objekter. Det er ikke ønskelig å bruke hele bildet for operasjoner. Bildene kan bli store og det gir ingen mening i å bruke metoden for annet en trær. Med å lage objekter for trær får vi fjernet resten av bildet og redusert data som må gjennom prosessering senere. Når hvert tre blir gjort om til et objekt går vi også fra pikselbasert måling over til arealbasert. Det blir ikke brukt enkeltverdier for piksel, men istedenfor gjennomsnitt for hvert objekt. Piksler av tre har stor grad av romlig autokorrelasjon. Det vil si at om det blir oppdaget en piksel som er tre i et bilde, er det sannsynligvis flere piksler rundt den som er tre. Å bruke objekter istedenfor piksler passer bra når det er stor grad av romlig autokorrelasjon. Dette vil også være med på å fjerne «salt and pepper». «salt and pepper» er et fenomen som skjer for pikselbaserte metoder. Det er enkeltpiksler som har blitt klassifisert/segmentert for seg selv. Eksempelvis et blad som ligger på toppen av et tak vil bli sett på som en piksel med tre. Salt and pepper blir en type støy som har ingen nytte for biomasseestimering.

Det er flere måter å detektere og segmentere trær. Det kan gjøres manuelt ved å tegne opp hver enkelt trekrone som en kan se i et bilde. En annen metode er å bruke en automatisk segmenteringsformel. Det er også mulig å bruke en vegetasjonsindeks eller et sett av vegetasjonsindekser som klarer å skille mellom trær. Uansett hvilken metode som blir brukt, er det viktig at hele trekronen blir definert som et enkelt tre. Det er ødeleggende for resultatet om et tre blir segmentert til flere mindre trær. Dette er et problem som oppstår i artikkelen «Biomass Estimation of Xerophytic Forests Using Visible Aerial Imagery». Her er multiresolution segmentering med eCognition brukt og den strevde med å håndtere store trær. Noen ganger ble de delt opp i flere småtre som hadde omtrent ingen biomasse. Antageligvis

vil multiresolution segmentation gi bedre resultater i det hyperspektrale bildet. Årsaken er at hyperspektrale bilder har flere bånd å bruke for å finne heterogenitet i «color» funksjonen. I forsøket var GSD på 0,4 meter, det vil si pikselstørrelsen var 0,4 meter. Dette er dårligere enn hva VNIR bruker, men bedre enn SWIR. Bedre romlig oppløsning vil gjøre at «shape» funksjonen fungerer bedre. Det er sannsynlig at multiresolution gir bedre resultater for deteksjon av trær for bildene brukt for denne oppgaven, spesielt i VNIR.

Ved manuell opptegning av trekroner blir resultatene svært gode. Da tegner man opp et tre om gangen. Dette gjør en ved å bruke et program som tillater å tegne et vektorsett eller rastersett over bildet. Program som for eksempel QGIS, ENVI og PCI Geomatica kan gjøre denne jobben. Når en skal tegne kan det være greit å starte med å bruke naturlige farger og tegne opp en nokså nøyaktig trekrone. Etterpå kan en bruke et par vegetasjonsindekser som kan hjelpe med å skille gress og kanten av trekronen. En må huske at SWIR og VNIR har forskjellige pikselstørrelser. En kan risikere at SWIR tar med mye gress/asfalt i kantpisklene siden de er større. En måte å unngå dette på er å tegne grensen en piksel på innsiden. Eventuelt kan en tegne opp alle trærne og lage en buffer på -1 piksel. Gjør vi noe av dette ender vi opp med å redusere arealet og det fører til noe mer unøyaktige resultater.

Ved å bruke manuell segmentering får vi nøyaktige resultater, men på bekostning av tid. Det tar lang tid å tegne manuelt alle trekroner, spesielt dersom det er snakk om flere hundre trær. For små områder eller områder med få trær fungerer denne metoden veldig bra. Siden metoden baserer seg på at en skal bruke noen få testområder til å produsere et funksjonsuttrykk som kan kjøres på et større område, er den manuelle segmenteringsmetoden god. For det store området med testområder vil det ta lang tid.



Figur 5.5.4 Manuell tegning av trekrone. Tegner rundt kanten til trekronen og prøver å få med mest mulig av den. Ønsker å unngå å ta med piksler utenfor. Når en trekrone er tegnet kan en redigere for å få det helt korrekt. Dette er den mest nøyaktige metoden.

For større områder er multiresolution et godt alternativ. For å gjøre denne typen segmentering er eCognition et godt alternativ. Multiresolution har fordelen av å være delvis automatisk. Den krever en del parametre og litt prøving og feiling for å gi gode resultater, men selve segmenteringen og oppdeling blir gjort uten tegning. For å få gode resultater med multiresolution må en bruke de riktige vektene og parametere og de kan variere for hvert område. Gjøres dette på et stort område må en akseptere at de ikke blir perfekt. Det beste er å sjekke at segmentene tar med bare tre og ikke noe annet. Om et tre blir delt opp i 2 eller 3 objekter er det mindre problematisk. Det er to grunner til dette. Den første er at mye av metoden baserer seg på lineær data. Det gjør at eksempelvis to trær på 100 piksler hver vil ha ca. like stor biomasse som et tre på 200 piksler. Det blir ikke helt slik, men nokså nøyaktig. Feilen en får av å ha med vei eller gress i objekter er mye mer alvorlig enn at et tre blir delt i to objekt. Den andre grunnen er at en kan slå sammen objekter etterpå. Dette kan gjøres manuelt eller ved å bruke spectral difference segmentation.

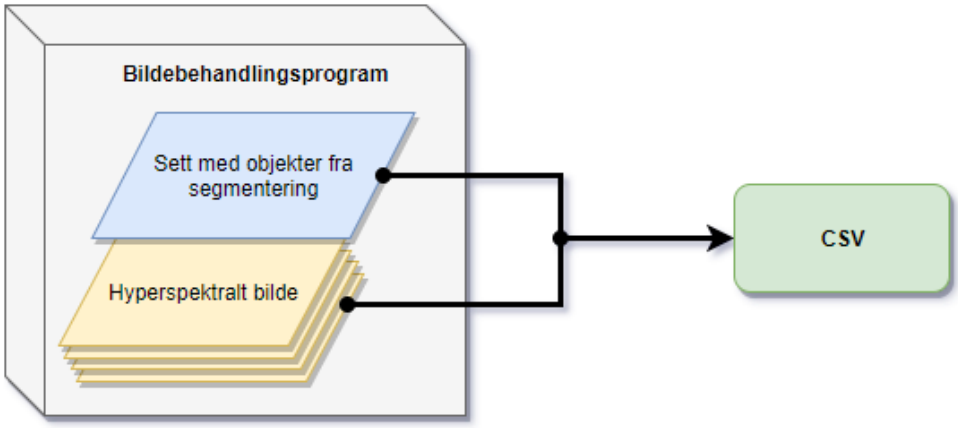
PCA kan være en god måte å forenkle data før en gjør en multiresolution segmentering. Det kan være vanskelig å finne de riktige vektene for enkeltbånd, men med PCA får en redusert mengden. Istedenfor å ha 369 bånd som skal vektlegges korrekt, får en istedenfor 5-7 PCA bånd. Mer om dette vil bli forklart i eksperimentering i kapittel 6.3.

Den tredje metoden er å bruke vegetasjonsindekser. En kan finne vegetasjonsindekser eller sett av vegetasjonsindekser som kan hjelpe til med å terskle bort alt som ikke er trær. Denne metoden krever noe manuelt arbeid ettersom en må se gjennom bildet etterpå og sjekke om alt en har er trær. En må også slå sammen pikslene til objekt. Dette går rimelig fort ettersom alt bortsett fra tre skal være maskert bort fra bildet. Vegetasjonsindekser som har vist seg å være gode for dette er: SR1, Vog1hyper, GRVI, NDBleaf og Bleaf_ratio. Spesielt GRVI er god for å oppdage trekroner. GRVI skiller greit mellom gress og trekroner og gjør det lett å filtrere bort gresset. Når en bruker en terskelverdi og en vegetasjonsindeks får en alltid litt «salt and pepper» støy. Dette kan fjernes ved å bruke morfologiske operatorer som «dilate» og «erode» (Burger & Burge, 2016). Ved å først erodere fjerner en alle enkeltpiksler. Dette fjerner støy. Eroderingen fjerner også kantene på alle objekter. Dilate vil gjøre dem større igjen. Kombinert vil dem gjøre at bildet ser omtrent likt ut, bortsett fra at alt støy er fjernet. Kombinasjonen av en erodering og en dilate kalles for «opening». I for eksempel ENVI kan dette gjøres med å bruke funksjonen «clump classes» (Harris Geospatial solutions, 2018a).

5.5.5 Samle inn nødvendig statistikk

Når alle trær som skal være med i analysen er valgt, kan en starte å samle inn statistikk. Det er ikke spesielt mye statistikk som trengs for å gjøre analysen. Det som trengs er gjennomsnittsverdier for hvert bånd i hvert tre i tillegg til antall piksler hvert tre består av. I tillegg blir ID-nummeret, treslaget, høyden, stammediameteren tatt med som ekstra informasjon. Selve analysen bruker ikke nødvendigvis noe av informasjonen, men den er grei for å se hvilket tre vi har beregnet biomasse for. Det kan være at noen trær ikke passer med modellen og må fjernes.

I program som for eksempel eCognition og ENVI kan beregne gjennomsnittsverdier og antall piksler i hvert segment, og deretter konvertere dem til en CSV tabell. For videre arbeid er det ikke nødvendig å jobbe med bildebehandlingsprogram. Det er lettere å jobbe videre i CSV, xlsx, eller andre enkle tabellformater.



Figur 5.5.5.A Segmenteringen og spektralinformasjonen blir slått sammen og eksportert ut sammen fra bildebehandlingsprogrammet.

Vi ønsker at hver rad skal være et tre. Hver kolonne skal bestå av en egenskap eller båndverdi. Egenskaper som må være med er trenummer, antall piksler, og alle bånd som skal med i analysen. Resten av egenskapene som er samlet fra feltdata kan være greit å ta med, men er ikke nødvendig. Tabellen skal da se slik ut:

Trenummer	*Treslag*	*pikselantall*	B_1	…	B_n
ID_1	Bjørk	250	1,1	…	1,1
ID2	Lerk	400	1,3	…	1,2
ID3	Gran	350	0,5	…	1
…	…	…	…	…	…
ID_m	Lind	2400	1,7	…	0,7

Tabell 5.5.5.B Eksempeltabell som viser hvordan innholdet i CSV/XSLX filen skal se ut

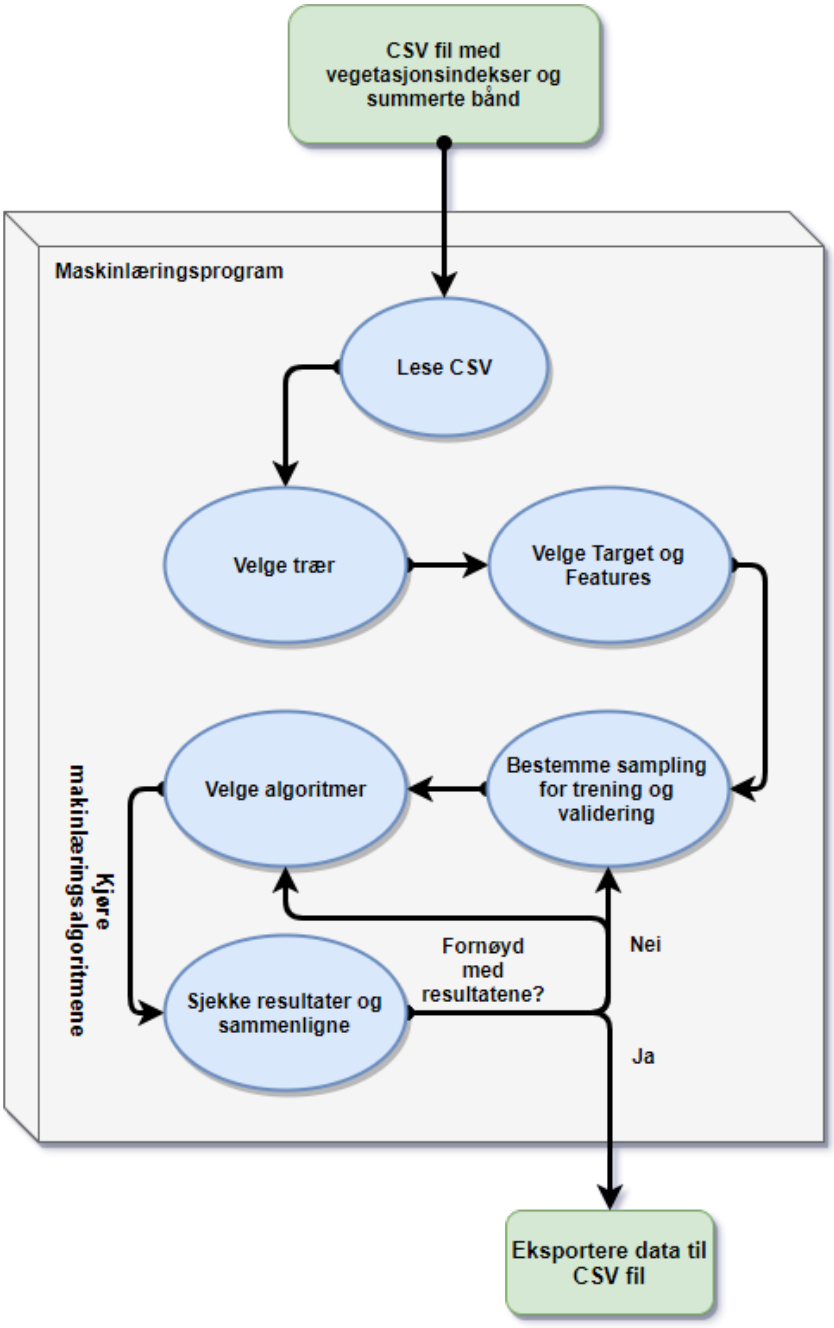
5.5.6 Beregning av vegetasjonsindekser og summeringer

Det neste steget er å bruke informasjonen fra tabellen til å lage summer av alle båndgjennomsnittene. Dette gjør vi ved å lage en ny kolonne for hvert bånd. I denne kolonnen multipliseres båndgjennomsnittet med pikslene. Dette gir verdien til summen av alle pikslene. Siden alle båndverdiene er positive vil verdiene blir store. For et stort tre, blir verdien mest sannsynlig stor siden den multipliseres da med et større pikseltall.

Etterpå skal det beregnes vegetasjonsindekser og summer av dem. Det må da velges hvilke vegetasjonsindekser som er relevante. Siden det jobbes med summeringer vil vegetasjonsindekser som gir både negative og positive verdier kunne risikere å havne rundt null. For eksempel NDVI kan få veldig lave verdier dersom deler av treet er dødt. Denne delen kan få verdier som er nærmere null eller til og med negative verdier dersom det ikke er tett nok med blader. Da kan det være at det som reflekteres er spektralsignatur fra asfalt eller betong som er negativt med NDVI. Når en bruker denne metoden er det enklest om indeksene gir positive verdier. Vegetasjonsindeksene bør også kunne representere noe som kan være viktig for biomasse. For eksempel er det å vite om lignin eller nitrogen kan indikere biomassen til et tre. For å finne dette ut må en ha med indekser for dette. Mer om hver enkelt vegetasjonsindeks kan en se i teorikapittelet. Indeksene blir også gjennomgått mer i fremgangsmåten. Indekser som blir testet i oppgaven er: NDVI, SAVI, TVI, VARI, GRVI, SR1, SR2, Vogelman1, NDNI, NI_ratio, NDLI, LI_ratio, NDBleaf, Bleaf_ratio. Det er gunstig å ta med mange vegetasjonsindekser i regresjonen. Da får en testet flere indekser og de som ikke passer kan fjernes. For vegetasjonsindekser som er normaliserte er det best å bruke ikke-normaliserte bilder. Anbefaler å bruke et normalisert bilde for ikke-normaliserte vegetasjonsindekser og ikke-normalisert bilde for normaliserte vegetasjonsindekser. Skygge er et problem når bildene ikke er normalisert. Dersom det er mye skygge i bildene, bør en ikke bruke områder med skygge, eventuelt korrigere eller fjerne skyggen.

5.5.7 Maskinlæring

Maskinlæring kan gjøres i flere ulike programmer og programmeringsspråk. Jeg synes de mest aktuelle er Python, R og Orange. Maskinlæring er den første delen av analyse i metoden. Resten har hovedsakelig vært bearbeiding og forberedelse av data frem til nå. I maskinlæringen handler det om å bygge opp en modell som kan ta inn en CSV-fil, velge ut trær og attributter. Deretter skal den kjøre ulike typer algoritmer og selv iterere og teste seg frem til det beste resultatet. Fire ulike metoder testes ut. Metodene er SVM, Random Forest, Neural Network og Linear Regression..



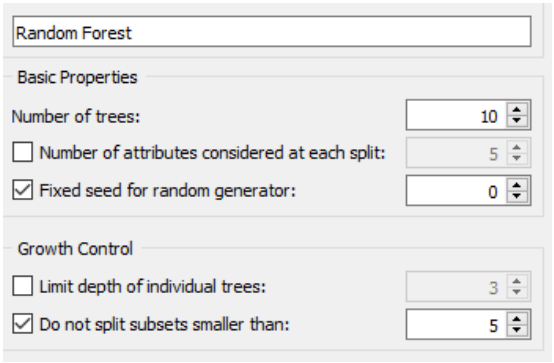
Figur 5.5.7.A Tegning av alt som trengs å gjøre for maskinlæringsprosessen

Det første blir å lese inn filen og sjekke at den leser den korrekt. Etterpå må det velges hvilke trær som skal være med. Dersom en planlegger å bruke alle trær bør en likevel ha muligheten for å kunne fjerne trær. Det kan vise seg at en må fjerne noen treslag fordi dem har feil eller at enkelte tre skal fjernes fra modellen fordi de er outliers.

Neste steget er å kunne velge kolonner som skal brukes. Metadata må bli satt som metadata, og fasitdata blir valgt som target. Enten sum vegetasjonsindekser eller sum bånd settes som variabler.

Siste delen av bearbeiding av data er å velge valideringsmetode. Dette kan gjøres på flere måter. De vanligste metodene er kryssvalidering, «leave one out» og dele opp sett i treningsett og valideringssett. Det anbefales å bruke treningssett/valideringssett eller kryssvalidering for store datasett. For mindre datasett er leave one out en aktuell metode også. Denne er som oftest mer tidkrevende. Dersom en ikke vet hva en bør bruke kan det være lurt å gå for treningssett og valideringssett. Treningssettet bør være stort nok til å klare lage en god modell, men bør ikke være såpass stort at det ikke er plass til et valideringssett. En grei fordeling er 66% treningssett og resten validering.

Nå er datasettet klar for å kjøres av algoritmene. Den første metoden som er aktuell er Random Forest. Random Forest har gitt gode resultater i mine klassifiseringstilfeller. Random Forest kan få stor grad av overfit. For å unngå dette er en nødt til å passe på at parameterne er satt korrekt. Dersom number of trees blir for stor ender den opp med potensiell overfit. Det er også lurt å bestemme hvor små subsets Random Forest skal få lov til å lage.



Figur5.5.7.B Random Forest parametere

En annen algoritme er SVM. Den har fordelen av å kunne bruke ulike kerneltyper. RBF brukes vanlig ofte, men også lineær kernel kan være interessant. Den viktigste parameteren for SVM er C verdien. Når justeres opp reduseres muligheten for overfit, men kan gjøre at

modellen passer dårligere. Iterasjonsmengden kan også justeres. Dersom den er høyere tar prossesen mer tid, men gir nøyaktigere resultat.

Nevrale nettverk er et annet alternativ for maskinlæring. Algoritmen fungere fint for lineær data. Nevrale nett krever også rett parameter for å unngå overfit. Dersom en bruker nevrale nett uten å se på parametere kan den fort gi store overtilpasninger som absolutt ikke ønskes. Verdien alfa må bestemmes korrekt for å sikre modellen fra overfitting. Lav alfa hjelper mot dette, men det fører til at C blir lav og tillater da ikke algoritmen å gjøre like store endringer og tilpasninger. Når datamengden økes vil nevrale nett bruke mye tid siden den kreve mange iterasjoner for å gi gode resultat.

Den siste algoritmen er lineær regresjon. Lineær regresjon er den som jeg antar kommer til å passe best for estimering av biomasse siden tidligere analyser har som oftest endt opp med å bruke lineær regresjon. I tillegg er det tydelig at data blir nokså lineært når bånd multipliseres med antall piksler. Ved lineær regresjon er det også noen parametere å velge mellom. Den viktigste er å velge om en vil bruke regularization eller ikke. Det er ikke nødvendig å bruke det, men det kan være lurt for å unngå både underfit og overfit fra minste kvadraters metode. For eksempel om en ønsker å bruke minst mulig variabler kan en bruke lasso regularization. Dette gjør at regresjonsmodellen kun tar i bruk de viktigste båndene og redusere bort resten. Rigde regresjon kan også brukes. Den prøver å kontrolere at koeffisienter ikke blir for ekstreme fra minste kvadraters metode. Gir som oftest noe lavere fit enn ingen regularization, men ikke spesielt mye. Alfa for regularization må velges. Er den for lav ender modellen opp med å passe dårlig og er den for høy øker muligheten for overfit.

Når algoritmene er valgt og parametere er innstilt kan en kjøre maskinlæringen. Ved første forsøk har en som oftest ikke optimale resultat. Overfit er sannsynlig, men også underfit og dårlige estimeringer er vanlig. Da må en endre parametere og prøve å tillate algoritmene enten å tilpasse seg mer eller mindre ved å endre alfa. Eventuelt er beregningsmetoden svak eller regularization. Valideringsdata kan også gi problemer. Dersom for stor andel av dataen brukes til validering får ikke algoritmene nok trær til å lage en god modell. Da må valideringsdataen

være mindre og treningssettet være større. Er det lite data til både treningssett og valideringssett må en prøve en annen metode som kryssvalidering eller leave one out validering.

Når en har kjørt algoritmene og mener resultatet er godt nok eksporterers tabellen med de estimerte biomasseverdiene til en CSV fil. Lurt å ha med trenummer i modellen som metadata slik at estimert biomasse kan kobles opp til trenummeret. I tillegg til å eksportere CSV filen bør en også huske å lagre en rapport med statistiske verdier. De vanligste å ha med er R^2, MSE og RMSE, men RSS og kovariansmatrise er også greit å ha.

5.5.8 Regresjon

Regresjonsmetoden som jeg anbefaler å bruke er multippel lineær regresjon. Veldig mye av informasjonen er nær lineært og har sjeldent store kurver slik som andregradsfunksjoner og tredjegradsfunksjoner modeller bra. For å lage et uttrykk velges biomassen som en ønsker å estimere som avhengig variabel og enten summerte vegetasjonsindekser eller summerte bånd som uavhengige variabler. Metoden er helt lik om en kjører regresjon for bånd eller indekser. Eneste forskjellen er at når en bruker bånd kan en fort risikere at en har flere variabler enn observasjoner. Da får en ingen frihetsgrader og kan ikke kjøre regresjonen. For å ordne dette problemet blir en PLS kjørt først og reduserer bort bånd med lav VIP score.

Fra regresjonen er det flere resultat som ønskes. Det ene er selve den estimerte biomassen og residualene som viser avviket til den faktiske biomassen. I tillegg må også statistiske verdier tas med. Det statistiske resultatet som viser hvor bra regresjonsmodellen passer. Verdiene som en må få fra det statistiske resultatet er R2 og RMSE. I tillegg kan det være greit å få med RSS, adjusted R2, kji-kvadrat, kovariansmatrise og korrelasjonsmatrise. Det siste som en også må få med er koeffisientene til hver uavhengig variabel. Det er koeffisientene som skal bygge funksjonsuttrykket. Det er også ønskelig å få med T-test verdier eller P-verdien til hver koeffisient der Nullhypotesen blir at variabelens koeffisient er null.

For å skaffe verdiene som trengs her må en enten lage et program som gjør jobben eller bruke et statistikkprogram som inneholder det nødvendige. For eksempel kan Python med scikit learn og R være kode programeringsprogram for å kjøre multippel lineær regresjon. Dersom en ønsker å bruke et program kan Origin og Origin Pro anbefales. Programmet gir i resultatrapporten for regresjonen all informasjon som trengs uten å trenge å kode. Programmet har også muligheten til å kjøre og integreres med blant annet Python.

Når en multippel regresjon er kjørt starter en med å se på koeffisientene og følgende P verdi. til hver variabel. Dersom P-verdien er lavere enn en terskel vi har bestemt oss for, blir variabelen forkastet. Det er fordi den ikke gir noe signifikant informasjon og kunne like godt vært null. P verdien viser sannsynligheten for at den faktisk skal være null i modellen. For å være sikker på at bare gode variabler er med settes maksimalt tillatt P lavt. 0,01 eller 0,05 er greie verdier. 0,05 vil si at det er 5% sannsynlighet for at koeffisienten skal være 0.

$$H_0 : \beta_j = 0 \qquad\qquad H_\alpha : \beta_j \neq 0$$

Formel 5.5.8.A H_0 er nullhypotesen. Den sier at $B_j = 0$ og H_α sier at den ikke er null. Dersom P-verdien er lav forkastes H_0 og da beholdes variabelen. Dersom P-verdien er høy settes koeffisienten til null og det er det samme som å forkaste variabelen. β er koeffisienten og j er båndnummeret/indeksnummeret

Variabler reduseres stegvis. En og en variabel fjernes om gangen. Når en variabel fjernes vil alle andre variabler få nye T-verdier og P-verdier. Koeffisientene endres også. Noen ganger kan alle andre variabler ende opp med å være signifikante etter at en variabel fjernes. Da er den fjernede variabelen en typisk grov feil som ødela modellen og fjerning av den retter alt.

Neste steget er å sjekke hvor bra de statistiske resultatene er. Er R^2 høy nok? Høy R^2 betyr at modellen passer bra. For eksempel har Bernascoris resultat en R^2 på 0,7 ved hjelp av Single Tree multiresolution segmentering og en lineær modell. Ønskelig da er å oppnå et bedre resultat en dette når det brukes hyperspektral data. Det er også mulig å kjøre en lineær regresjon hvor bare arealet til et tre er en variabel. Da er det mulig å se hvor mye bedre R^2

ender opp med å være når hyperspektral data brukes. MSE og RMSE bør også sammenlignes for å se hvilke modeller som gir minst residualer. Her er det og greit å sammenligne hyperspektral mot areal. Etterpå kan en sjekke hvor mye bedre modellen hadde blitt om høyden var med.

Istedenfor for å starte med lineær regresjon kan en starte med en PLS analyse. Denne vil finne ut hvilke variabler som er relevante. Da kan variabler som ikke er relevante reduseres bort før en begynner regresjonen. For vegetasjonsindekser kan en trolig finne noen som har lav VIP score. Sannsynligvis er det mye av de samme som reduseres bort ved lineær regresjon, men det er ikke nødvendigvis helt likt.

For summerte vegetajsonsindekser er det mer viktig med PLS analysen. For det første er det mangel på frihetsgrader om det er få observasjoner når en bruker lineær regresjon. Det andre er at det tar veldig lang tid å kjøre multippel lineær regresjon for flere hundre variabler. Å kjøre det en gang går rimelig greit, men når reduseringen begynner blir det vanskelig. Siden reduseringen må gjøres stegvis må det gjøres veldig mange ganger. For eksempel har en 300 summerte bånd og det viser seg at 15 er signifikante for lineær regresjon. Da må en kjøre 285 lineære regresjoner for å fjerne alle bånd med høy P-verdi.

Jeg anbefaler å ta med ca. 15-20 viktigste summerte bånd fra en PLS med mindre det er tydelig at det er flere eller færre som trengs. Etterpå kjører en vanlig multippel lineær regresjon og fjerner de som får lav P-verdi. En kan også sjekke korrelasjonsmatrise og se om noen har veldig stor korrelasjon. Da kan dem også fjernes. De som har høy korrelasjon får som oftest at en av variablene har en høy P-verdi siden dens koeffisient kunne likeså godt være 0 i modellen.

Når alle regresjonsmodellene er ferdig kan en sammenligne resultatene og hvilke som er aktuelle. Hvor mange bånd trengs, og hvor nøyaktige blir modellene? Noen av modellene

skiller seg kanskje ut med å være mye bedre eller dårligere. Det kan også være at det er stor forskjell mellom modeller som bruker SWIR og VNIR.

For modellene som har gode nok resultater kan en lage et funksjonsuttrykk. Koeffisientene til variablene brukes for å gjøre dette. Ved å multiplisere hver koeffisient med tilhørende variabel får vi stigningstallet til det lineære funksjonsuttrykket. I tillegg kan det være et konstantledd β_0.

$$y = \beta_0 + \beta_1 \times v_1 + \beta_2 \times v_2 + .. + \beta_n \times v_n$$

Formel 5.5.8.B Funksjonsuttrykket for den multiple lineære regresjonen. B_0 er konstantledd, $\beta_1 \ldots \beta_n$ er koeffisient og v_n er variablene som gjenstår etter redusering.

Funksjonsuttrykket kan brukes på tabeller der trær er segmentert og har fått summerte båndverdier eller summerte vegetasjonsindekser. Kan også brukes på treklynger, men da vil nøyaktigheten være dårligere og biomasseverdien ender trolig opp med å være underestimert.

6 Eksperimentering

I kapittel 6 går jeg gjennom forsøk, og tester som ikke direkte blir brukt i fremgangsmåten min, men som likevel kan være gode å bruke dersom en skal rekonstruere metoden. Jeg går også gjennom metoder og forsøk som ikke har fungert.

6.1 Klassifiseringer

I starten ble det kjørt en del klassifiseringer. Før jeg hadde bestemt meg for å jobbe med LAI eller biomasse ble klassifiseringstester gjort. Det var hovedsakelig noen klassifiseringer av treslag. Metodene som jeg testet var support vector machine, maximum likelihood og spectral angle mapper. Det var tydelig at HySpex er svært gunstig data å bruke for treslagsklassifisering. Ved å kjøre slike algoritmer kan en klassifisere treslag og gress presist. Siden klassifisering ikke er målet med oppgaven la jeg bort dette teamet, men det kan ha en nytte for biomasseestimering. Dersom en allerede har kjørt en treslagsklassifisering kan denne brukes som grunnlag for biomasseestimeringen. Dette blir diskutert videre i kapittel 8.10.

6.2 Atmosfærekorrigert data

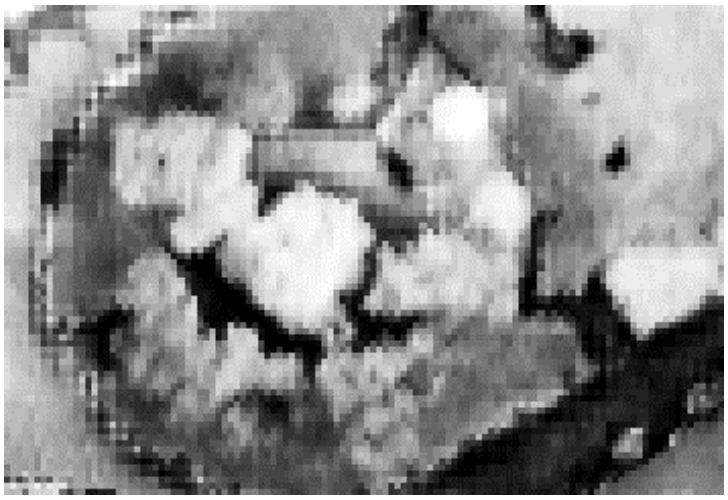
Atmosfærekorrigert data ble vurdert å brukes for biomassestimeringen. Fordeler med dette er at støy fra atmosfæren forsvinner og noen vegetasjonsindekser som er laget for reflektansverdier gir mer korrekte svar. Å bruke atmosfærekorrigerte unormaliserte data ville vært det beste for normaliserte vegetasjonsindekser. Det er tre årsaker til at atmosfærekorrigert data ikke brukt i selve analysen. Den viktigste var harddiskkapasitet. Oppgaven ble gjort på en bærbar pc i tillegg til en ekstern harddisk og de hadde ikke plass til de atmosfærekorrigerte flystripene. Filene var svært store og en flystripe var flere hundre gigabytes. En annen viktig årsak var usikkerheten med å bruke atmosfærekorrigert data. Flere operasjoner som gjøres på det datasett gir økt usikkerhet. Er korreksjonen korrekt og er alt atmosfære fjernet, og kan det være at mer enn bare atmosfæren er fjernet? Dersom mer enn atmosfæren er fjernet har vi mistet data som kan være nyttig. Det er ikke garantert at atmosfærekorreksjoner gjør bildene bedre. Det siste problemet var alle verdier endte opp som 0 i atmosfærekorrigerte bilder. Dette er ikke et stort problem, men noen vegetasjonsindekser

får rare tall i flere piksler fordi det blir delt på verdien 0. Dette skjer når det er bare atmosfære i en piksel for et bånd. Dette skjedde nokså ofte.

6.3 PCA

PCA var en viktig del av den eksplorative analysen i starten. Det ble brukt for å se hvilken informasjon HySpex bildene inneholder. Det ble sjekket hvilke bånd som forklarte mest varians og hvilke som var viktigst for hver komponent. Dette viste omtrentlig hvor støy befant seg, og hvor den viktigste informasjonen var. PCA var og effektiv for å se hvor skygger var og hvor trekronens grenser ligger. Metoden og fremgangsmåten for forsøket bruker ikke PCA til i noen av hovedleddene, men likevel har PCA vært mye brukt gjennom prosessen.

Den viktigste oppgaven til PCA var å være med som segmentering. Det er vanskelig å vektlegge flere hundre bånd korrekt for å få best mulig segmentering av trær, men det ble lettere med noen få PCA komponenter istedenfor. Ved å bruke 7 PCA komponenter fikk en et godt bilde å bruke for segmentering. Det var ikke slik at det er fast hvilken PCA komponent som skal vektlegges høyt for segmenteringen. Dette varierer litt i hvert bilde. I mine tilfeller var alltid komponent 1 god å vektlegge høyt, men det varierte i område til område om komponent 4, 5 og 7 var viktige for segmentering av trær.

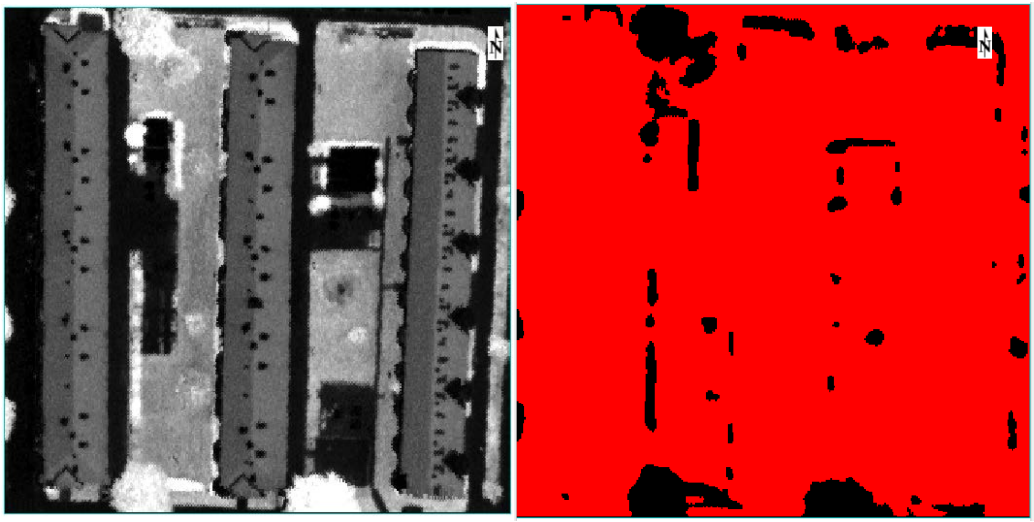


Figur 6.3 I en del tilfeller var PCA komponent 5 viktig for å segmentere trær. Her klarer den å skille gress og trær fra hverandre i et vanskelig område. Hvite områder er trær og grått er gress. Det sorte er en stil som går i en sirkel rundt treet i midten.

6.4 Segmenteringer

Flere segmenteringsmetoder har vært testet i løpet av arbeidet med masteroppgaven. De som har vært testet er manuell segmentering, vegetasjonsindeks terskel og deretter manuell segmentering eller automatisk segmentering, PCI Geomatica objectbased segmentation og segmenteringer i eCognition. Manuell segmentering slik som er nevnt i 5.5.4 fungerer bra og er forklart i kapittelet. De andre segmenteringsmetodene som for eksempel K-means klustring og sjakkbrettsegmentering blir for simple til å klare segmentere trær nøyaktig. Disse simple metodene følte jeg ikke det var verdt å bruke tid på fordi det tidlig var klart at de ikke kom til å gi akseptable segmenteringer. Kvadtremetoden fikk jeg ikke tid til å teste, men med de rette parametre og litt bearbeiding kan nok denne metoden fungere.

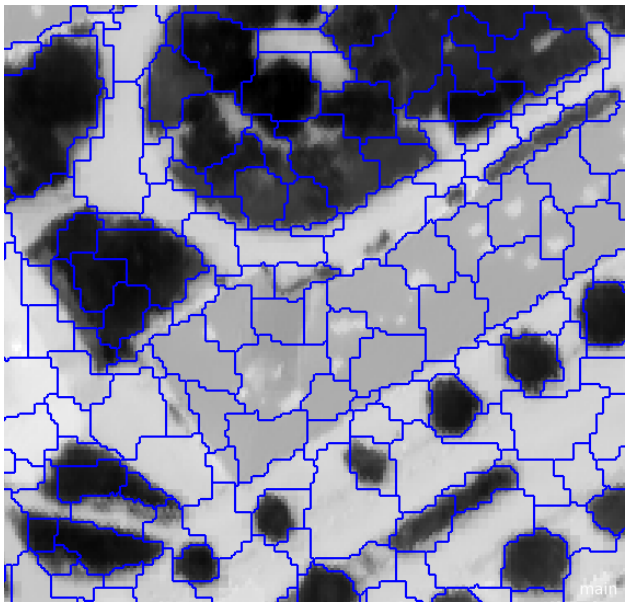
Å segmentere manuelt ved hjelp av thershold fra vegetasjonsindekser viste seg å fungere. Ved hjelp av vegetasjonsindeksen GRVI var det rimelig enkelt å fjerne alt som ikke var vegetasjon i tillegg til gress. Når dette er gjort er det enklere å segmentere trær inn til objekt. Dette ble testet for ikke-normaliserte bilder. For at denne metoden skal fungere bra må morfologiske operasjoner som dilate og erode brukes for å fjerne støy siden vi ikke ønsker «salt and pepper støy» i bildet.



Figur 6.4 Til venstre er GRVI hvor hvit farge er høy GRVI. Til høyre er samme området med en binær threshold. Bildet er også gjort den morfologiske operasjonen erode og deretter dilate med en kernel på 5.

PCI Geomatica har en object analyst pakke som kan segmentere. Metoden kalles OBIA (Object Based Image Analysis). Denne ga parametere som lignet på det eCognitions multiresolution segmentation ga. Dessverre ga ikke OBIA noen brukbare resultater i dette tilfellet. Grunnen for dette var trolig at den ikke ga vekter på båndene. Alle bånd hadde lik vekt som førte til at segmenteringen ikke klarte skille trær til objekter. Enten endte trær opp i flere deler (5-10 mindre objekt) eller så endte dem opp med å ta med mye gress og annet rundt treet. Noen av segmentene så omtrent tilfeldige ut.

eCognition ga de beste segmenteringene. Multiresolution segmentation med vektet PCA ga svært solide resultat. Objektene treffer godt på trær og ofte ender enkelttrær opp som et eget objekt. I de tilfellene hvor et tre har endt opp som flere objekt var det enkelt i eCognition å slå dem sammen manuelt. En annen mulighet er også å bruke en ny segmentering for å slå de sammen. Da fungerte spectral differences segmentation. Det var også lite annet enn trær som endte opp i segmentene. To metoder fungerete bra for multiresolution segmentation. Den ene var å sette veldig høy vekt for båndene i red edge og grønt lys området, eventuelt kjøre en PLS for å finne de viktigste båndene for å definere tre og bruke de. Den andre metoden er å kjøre PCA og sjekke hvilke bånd som skiller gress og tre best. Båndene som klarer dette best får høyest vekt. Bånd 1 vil omtrent alltid skille mellom vegetasjon og ikke-vegetasjon og kan også få høy vekt.



Figur 6.4.B Segmentering med vektet PCA i Galgeberg. Segmenteringen traff godt på de fleste trærne. I toppen til venstre av bildet er det et tre som ikke treffer spesielt godt. Hele plenen og treet har endt opp som et objekt.

6.5 Python

Scikit-learn ble testet for å kjøre maskinlæring for biomasseestimering. Filen ble lest inn som en SCV i Python ved hjelp av funksjoner i Pandas biblioteket og deretter ble data som var irrelevant filtrert bort. AGB ble satt som target og vegetasjonsindekser og bånd ble satt som variabler. Koden som ble brukt var hovedsakelig laget for kategorisk klassifisering. Dette gjorde at den måtte redigeres en del for å brukes til biomasseestimeringer. Biomassen er ikke kategorisk og må bli estimert som en numerisk verdi. Det så ut som om Python kan fungere fint for biomasseestimeringer og maskinlæring. Grunnen til jeg ikke valgte å jobbe videre med koden var fordi Orange allerede gjorde det samme. Orange baserer seg også på scikit-learn. Forskjellen mellom Python og Orange er at Orange tvinger store deler av innstillingene til å være default. Det er bare de viktigste parameterne for maskinlæringsalgoritmen en kan justere. Ved å bruke Python kan en justere alle innstillinger. Siden jeg ikke bruker maskinlæring for å lage funksjonsuttrykkene bruker jeg ikke mye tid på å finne alle optimale parametere for maskinlæringen. Det eneste som jeg ser på som viktig er å unngå overfit og de klarer også parametervalgene i Orange. Det er verdt å vurdere å bruke Python for biomasseestimering. Med god koding kan en gjøre både segmentering, regresjon og maskinlæring i Python.

7 Fremgangsmåte

7.1 Beregning av biomasse fra feltdata

For å beregne den faktiske biomassen for testområdene brukte jeg QGIS. Alt av feltdata som var lagret i PostGIS kobles opp til QGIS. Da fikk jeg importert alle trær som vektordata. Hvert tre har alle egenskaper som trengs for å beregne biomasse. Det som trengs er høyde, stammeomkrets og treslag. Ut ifra dette regnes det ut stammediameter og stammevolum med formlene som forklares i kapittel 5.5.1. Beregninger gjøres ved hjelp av raster kalkulatoren i QGIS. Denne lager en ny kolonne for hver ny egenskap som legges til. Rasterkalkulatoren kan gjøre operasjoner og fungerer på omtrent samme måte som en database med et enkelt SQL lignende språk. Rasterkalkulatoroperasjonene som brukes for å lage fasitdata kan en finne i vedlegg A. Noen få eksempler fra operasjonene vises i dette kapittelet.

Stamme_radius:

CASE

WHEN «stamme» IS NOT NULL THEN «stamme»/(2*pi())

END

Formel 7.1.A Beregning av radiusen til stammen ved hjelp av rasterkalkulator

For å finne biomassen holder det ikke med stammevolum. For å finne den trengs nominal specific gravity som varierer med treslagene. Operasjonen fungerer slik at alle treslag som er bestemt får riktig verdi, og alle trær som er klassifisert som bartre uten kjent treslag får verdien 0,39 som brukes når en er usikker. For trær som ikke er bartre og ikke har en bestemt nominal specific gravity brukes verdien 0,49. Dette er verdien som brukes for ukjente treslag som er løvtre. Dersom treslagskolonnen er null får ikke treet biomasse og brukes ikke videre.

Biomasse_stamme:

```
CASE
    WHEN "treslag" = 'Bjork' THEN "stamme_vol"*0.53
    WHEN "treslag" = 'Bartre' THEN "stamme_vol"*0.39
    WHEN "treslag" = 'Furu' THEN "stamme_vol"*0.42
    WHEN "treslag" = 'Lind' THEN "stamme_vol"*0.44
    WHEN "treslag" IS NULL THEN 0
    ELSE "stamme_vol" * 0.49
END
```

Formel 7.1.B Beregning av stammens biomasse. Dette er bare et utsnitt av operasjonen som bruker bare bjørk, bartre furu og lind. Ukjente løvtrær får nominal specific gravity på 0,49.

For trekronen og røttene trengs det 7 konstanter som er ulike for treslagene. For hver konstant gjøres det er operasjon med rasterkalkulator. Etterpå brukes konstantene til å beregne biomassen av trekronen og røttene. For å bestemme hvilken formel som skal brukes for å beregne biomassen til trekronen og røttene må en sjekke stammediameteren. Dette gjøres også i rasterkalkulator.

```
CASE
    WHEN "stamme_dia" > 0.5 THEN "c" + ("d" * "stamme_dia" * 100)
    ELSE "a" * ("stamme_dia"*100)^"b"
END
```

Formel 7.1.C Beregning av kronens biomasse. Sjekke om diameteren er over eller under 0,5meter for å bestemme om formel 5.5.1.C eller formel 5.5.1.D skal brukes

Når biomassen før stammen, røttene og trekronen er regnet ut, kan dem adderes sammen til total biomasse og AGB slik som formlene 5.5.1.E og 5.5.1.H.

Fremgangsmåten var forskjellig for noen få trær. Noen tre var splittet og hadde to eller flere stammer. De er regnet ut i et excelark. Det var få av dem og det var raskere å beregne de manuelt enn å finne på en QGIS operasjon for å gjøre jobben.

7.2 Bestemme datasett

For å gjøre analysen hadde jeg flere tilgjengelige datasett. Fra TerraTecs leveranse hadde jeg tre ulike datasett over samme område. Det ene var mosaikkbildet fra Hyspex, det andre var normalisert mosaikk, og det tredje var atmosfærekorrigerte flystriper. Jeg bestemte meg for å fokusere på å bruke normaliserte bilde fordi skygge er mindre problematisk i normaliserte bilder. Skyggen er ikke synlig og jeg kan da se bort fra skyggen. Et annet alternativ ville vært å bruke atmosfærekorrigerte normaliserte bilder. Problemet da er at atmosfære korrigert data fører også til en del usikkerhet som nevnt i kapittel 4.4. I tillegg måtte også flystripene gjøres om til mosaikkbilder fordi to flystriper trengs for å dekke noen av områdene. Dette krever mye tid og harddiskkapasitet. En siste grunn for at jeg også valgte å ikke bruke atmosfærekorrigert data var fordi bildet ble lest slik at piksler uten informasjon etter atmosfærekorreksjon ble lest som 0. Dette ga feilmeldinger i noen vegetasjonsindekser som da ender opp med å dividere på null i noen tilfeller.

Både VNIR og SWIR velges å brukes i oppgaven. Det er ønskelig å kunne teste alle bånd og vegetasjonsindekser for å se hva som fungerer best for biomassestimering. For å skaffe den filen ble den normaliserte filen av SWIR og VNIR slått sammen til en fil. Da brukes den romlige oppløsningen til VNIR siden den er minst. Da resamples SWIR til 0,3 meter pikselstørrelse før den slås sammen med VNIR. Resamplingsmetoden som brukes er Nærmeste nabo. Denne velges fordi den ikke endrer mye på bildet. For eksempel cubic convolution tar i bruker flere piksler for å lage de nye pikslene og interpolerer mer. Det gir ofte mykere overganger, men fører til at kanter i bildet ikke blir slik de opprinnelig var. Uansett hvordan vi gjør det får vi en interpolering i pikslene når 0,7 meter piksler skal gjøres om til 0,3 meter piksler. Både VNIR og SWIR fotografiene er tatt på likt og mosaikken bruker de samme flystripene. Dette gjør at bildene vi får passer godt sammen.

7.3 Valg av trær

Deteksjon og segmentering av tre er mulig på flere ulike måter som nevnt tidligere i kapittel 4.6 og 5.5.4. Siden det totalt er 161 trær, og flere av dem ikke kan brukes ender vi opp med et lite datasett. Da er det viktig at kvaliteten på datasettet er god. For å gjøre den best mulig tegnes alle trær for hånd i ENVI ved hjelp av ROI tool. Det er enkelt å finne trærne i bildet ved hjelp av å importere vektorpunktene fra QGIS. De inneholder koordinatene til trær. Alle trær med hel trekrone i tillegg til feltdata tegnes inn. Noen trær som hadde skjulte deler av trekronen ble også tatt med dersom det var lett å se hvor stor det egentlig var. For eksempel var halvparten av et tre innenfor testområdet og halve trekronen var klippet bort. Det var lett å se at kronen var en halvsirkel. I slike tilfeller brukes formel 5.5.3. For å gjøre det lettere å tegne inn trekronene korrekt ble PCA og noen vegetasjonsindekser brukt. Spesielt SR1 og GRVI var hjelpsomme. I tillegg ble noen bånd i SWIR sjekket for å se at de ikke ble feil. De store pikslene i SWIR gjør det noe vanskeligere å tegne korrekt for SWIR. Bedre romlig oppløsning gjør det lettere å tegne kroner manuelt som gir mer korrekte areal. Areal er forventet å være en svært viktig faktor for biomasseestimeringen.

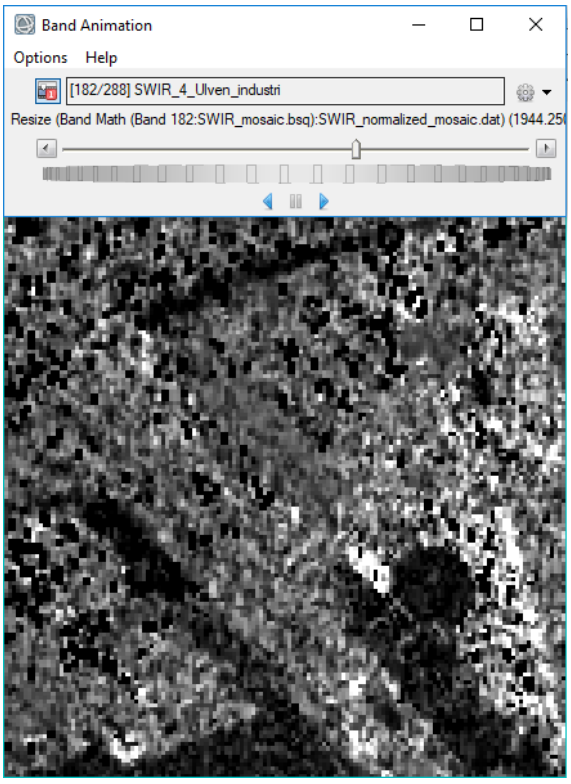
Figur 7.3 ROI tegnet manuelt i en bygård i Gamle Oslo. Fargene er kronene til trærne som er med

Når en har tegnet fjernes alle trær som ikke har stor nok stammediameter. Siden kravet var at stammens diameter måtte være 7 cm eller større ved brysthøyde, kan ikke trær som har mindre stamme brukes. Ved å se på biomassen til de små trærne ser vi også grunnen til dette. Fasitberegningene gir tvilsomme resultat for de minste trærne. De med svært små stammer for en tørrvekt biomasse på 4-7 kilo. Dette er urealistisk lavt. I feltarbeidet er bare trær over 5 meter tatt med og de har garantert ikke en biomasse som er rundt 5 kilo. Når de er fjernet sammen med alle trær som har overlappende trekroner er det 91 trær igjen. Det er et lite datasett, men det har i det minste rimelig god kvalitet.

Tre nummer 128 er et spesielt tre. Det har mye større biomasse enn resten. Treet er det eneste treet som har en AGB over 3 tonn. Problemet er at den har rundt 7 tonn biomasse. Treet passer ikke inn med resten Om treet brukes i modellene vil dens spektralsignatur ende opp som definisjonen på hva som gjør at biomassen blir høy. Treet er en ask og dette treet er det få av. Dette treet kan ødelegge modeller fordi den må tilpasse seg mye til treet. Alle lineære regresjonsmodeller prøver å gi minst mulig residual, og må derfor passe med treet. Tidlige regresjonstester viser at modellene passer bra, men ender opp med å overestimere alle andre trær når denne er med. Vi har ingen trær mellom 3 og 7 tonn biomasse og det gir ingen mening å prøve å lage en modell som skal passe for dette området. Tre nummer 128 fjernes fra modellen i de fleste forsøk. Dette gjør at modellen gir mer realistiske estimat, men den vil fungere dårlig for trær over 3 tonn biomasse. En måte å definere hvilke trær modellen fungerer på er å sjekke stammediameteren. Det største treet i diameter bortsett fra tre nummer 128 er tre nummer 9. Den har en stammediameter på 0,81 meter. Det betyr at modellen som lages er best å bruke for trær med en stammediameter mellom 0 og 0,81 meter. Modellen kan gi fornuftige svar for større trær, men vil mest sannsynlig gi mer avvik når trær blir større.

7.4 Fjerne bånd med støy

For å fjerne støy ble ENVI brukt. Ved å bruke funksjonen «view band animation» viser ENVI hvert enkelt bånd i en liten videospiller. Der kan en stoppe og se gjennom hvert bånd og notere ned de som ikke er verdt å ta med.



Figur 7.4 Her er et bilde av animasjonen som ENVI viser. Bånd 182 i SWIR (1944,25 nm) er et bånd med støy. Dette er et vannabsorbsjonsstøybilde.

Det var ingen bånd i VNIR som inneholdt betydelige mengder med støy. Derimot SWIR hadde en del bånd som nærmest bare bestod av tilfeldige piksler som ble definert som støy. Båndene som ble fjernet i SWIR er: 1-5, 72-92, 156-188, 191-197 og 250-288. Totalt ble 98 bånd fjernet. Det er mulig at å ha med båndene kunne tilført noe informasjon, men i de fleste tilfeller ville de bare gjort prosessering og maskinlæring tregere. Det ville og potensielt gitt en del uforklarlig informasjon som kunne svekke maskinlæring og regresjonsmodellene.

7.5 Summering av bånd og vegetasjonsindekser

Nå er det klart for å eksportere data ut av ENVI ved hjelp av CSV. Det som eksporteres er antall piksler til hvert tre og alle gjennomsnittsverdier fra båndene. Det var også mulig å eksportere hver eneste pikselverdi. Begge deler kan brukes for å lage summeringene. CSV filen blir videre redigert i Excel. Her lages summeringene og vegetasjonsindeksene. Data ryddes opp og gjøres slik det kan leses av Origin Pro, Python og Orange. Informasjon som lagres er trenummer, treslag, biomasse (AGB, total og hver enkel del av biomassen), høyde, stammediameter, volum og antall piksler.

Ved å multiplisere antall piksler med gjennomsnittsverdien til hvert bånd, får en summerte verdier for alle trær. For å beregne vegetasjonsindeksene beregnes de ved å bruke gjennomsnittsbånd og deretter multiplisere med pikselmengden. Dette gir summerte verdier for bånd. Når filen er ryddet og ser bra ut kan den lagres og importeres videre til statistikkprogram og datamining programvare. Origin Pro og Orange leser CSV filer som er delt med semikolon.

$$Summert\ bånd = SB_n = \sigma_n \times p_n$$

Formel 7.5.A Summering av bånd. σ er gjennomsnittsverdi for et bånd til tre n, og p_n er pikselmengden til tre n. Dette gjøres for alle bånd.

For vegetasjonsindekser er det akkurat det samme

$$Summert\ vegetasjonsindeks = SVI_n = VI_n \times p_n$$

Formel 7.5.B Summering av vegetasjonsindeks. VI_n er gjennomsnittsverdi for en vegetasjonsindeks til tre n, og p_n er pikselmengden til tre n. Gjøres for alle vegetasjonsindekser som ønskes.

1	område	tre_nummer	hoyde	treslag	stamme_dia	stamme_bio	krone_bio	røtter_bio	AGB	Total_biomasse	Pixels	GRVI	VARI
17	1_Helgesens	19	19	Hengebjork	0,416985951	0,4584	0,17499	0,25824	0,63339	0,89163	1098	3949,6438	660,115606
18	6_toyen_lek	20	18	Lind	0,350140875	0,2542	0,11352	0,16685	0,36772	0,53457	476	1814,58955	322,564918
19	6_toyen_lek	21	18	Lind	0,416985951	0,36053	0,17499	0,25824	0,53552	0,79376	358	1358,07808	242,732233
20	6_toyen_lek	23	18	Lind	0,324676084	0,21857	0,09416	0,13815	0,31273	0,45088	346	1645,19196	213,826725
21	6_toyen_lek	26	15	Kastanje	0,423352149	0,30968	0,18168	0,26821	0,49136	0,75957	632	2473,96587	902,628748
22	6_toyen_lek	27	18	Lonn	0,426535247	0,4201	0,20685	0,27328	0,62695	0,90023	1058	5178,07904	764,054126
23	6_toyen_lek	28	15	Lonn	0,305577491	0,17968	0,09056	0,11872	0,27024	0,38896	670	3470,17393	484,657391
24	6_toyen_lek	39	23	Lind	0,630253575	1,0524	0,45001	0,42758	1,50241	1,92999	1038	5527,44874	733,659971
25	6_toyen_lek	49	15	svenskeasal	0,413802852	0,32949	0,1717	0,25334	0,50119	0,75453	529	2085,06472	360,141685
26	6_toyen_lek	54	15	Lind	0,356507073	0,21961	0,1187	0,17454	0,33831	0,51285	691	2561,5545	361,515592
27	6_toyen_lek	59	16	Lind	0,410619753	0,31076	0,16844	0,2485	0,47920	0,72770	741	2359,43925	617,239589
28	6_toyen_lek	60	20	Lind	0,416985951	0,40058	0,17499	0,25824	0,57557	0,83381	880	3086,74379	704,961443
29	9_Ekeberg_p	63	6,8	Furu	0,2801127	0,30147	0,39579	0,40546	0,69726	1,10272	472	1496,55644	248,501848
30	9_Ekeberg_p	64	8,5	Ask	0,264197206	0,08232	0,05651	0,08252	0,13883	0,22135	457	2098,80212	355,30724

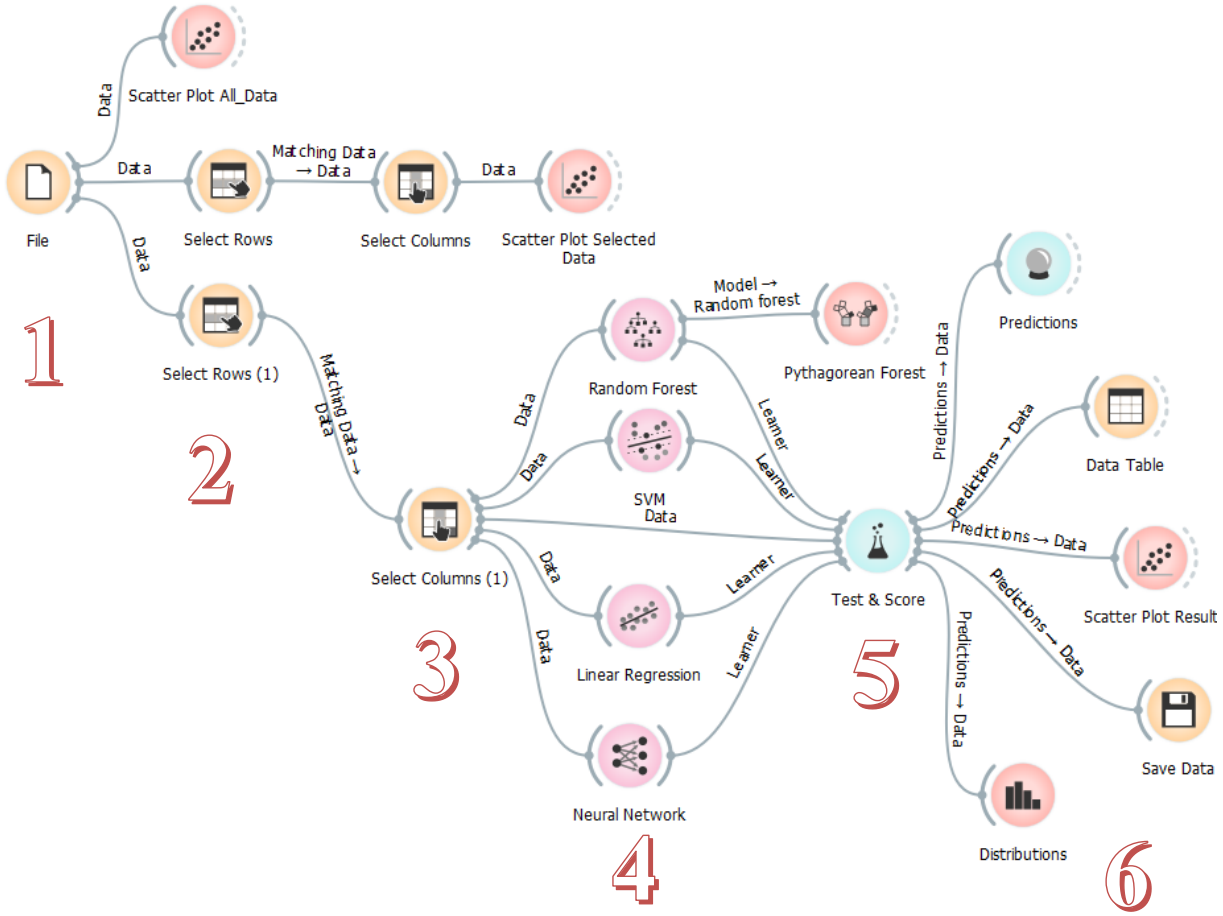
Figur 7.5.C CSV filen vist i Excel. Bare en liten del av filen. Hele filen består av 81 rader med trær og 750 kolonner med egenskaper

Vegetasjonsindekser som har blitt summert er:

GRVI, VARI, TVI, SAVI, SR1, SR2, Vogelman1, NDBleaf, Bleaf_ratio, NDNI, NI_ratio, NDLI og LI_ratio. I tillegg er NDVI funksjonsuttrykket fra (Liu et al., 2006) tatt med og summert.

7.6 Maskinlæring

Maskinlæringsdelen gjøres ved hjelp av programmet Orange. Lager en modell i Orange. Modellen består hovedsakelig as seks ulike deler: Import av CSV, valg av rader, valg av kolonner, maskinlæring, validering og statistiske resultat og til slutt lagring av predikert biomasse. Modellen ser slik ut:



Figur 7.6 Modellen som kjører maskinlæringen i Orange.

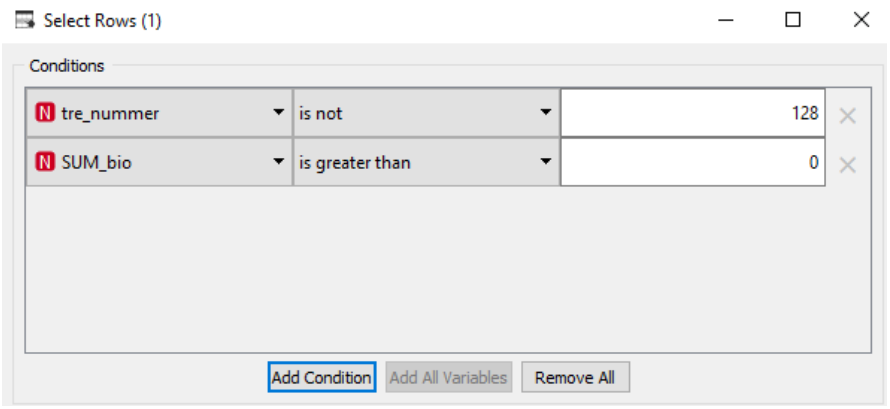
7.6.1 Importere CSV

Modellen starter med å lese inn CSV filen. Filen importeres inn i Orange ved å bruke file modulen. Denne leser filen og forstår CSV filer som er separert med semikolon. Her kan en velge om kolonner skal tolkes som kategoriske eller numeriske verdier. Alt bortsett fra treslag er numerisk. Videre velges det at treslag, trenummer, treets testområde skal være metadata. Biomasse total og AGB settes som target value. Det vil si at de brukes som fasitdata og det

som modellen skal estimere seg frem til ved hjelp av variablene. Variablene kalles for features og alle bånd og vegetasjonsindekser settes som features.

7.6.2 Valg av rader

Del 2 er å velge hvilke rader som skal brukes i maskinlæringen. Hver rad består av et tre. Trær som ikke skal være med må fjernes her. Vi vet at tre nummer 128 er problematisk. Dette treet er svært stort og passer ikke inn. Treet fjernes fra maskinlæringen. Fjerningen av treet gjør at modellen som lages kommer til å egne seg for trær mellom 0-3 tonn tørrvekt biomasse. Etter å ha fjernet tre nummer 128 velges det at alle andre trær som har en biomasseverdi skal tas med i modellen.



Figur 7.6.2 Velger alle trær som ikke er tre nummer 128. Alle trær med en biomasse over 0 er med i modellen.

7.6.3 Valg av kolonner

Det må velges hvilke kolonner som skal brukes. Slik filen er definert fra innlesingen er alle biomasseverdier targets. Maskinlæringen bruker bare en target om gangen. Det må velges om AGB, kronebiomasse eller total biomasse skal estimeres. Her velges AGB. Features må også velges. Summerte Vegetasjonsindekser eller de summerte båndene kan velges som features. Eventuelt kan også høyde og areal velges for å lage fiktive laserberegninger. Maskinlæringen må kjøres hver for seg for variablene ovenfor. Settes alt som features bruker den alle verdiene på en gang for å estimere biomasse. Først velges laseren, deretter vegetasjonsindeksene, og etterpå båndene. NDVI funksjonen kan også kjøres.

7.6.4 Maskinlæringen

Del 4 og 5 har en del sammenheng. De gjøres ofte parallelt og endringer i del 5 påvirker resultater i del 4. Validering og treningsfelt er en avgjørende del sammen med regresjonsparameterne for å få tilfredsstillende resultater. Det kan være enklere å lese 7.6.5 før enn leser 7.6.4. Resultatene fra maskinlæringen vil sammenlignes og diskuteres mer i kapittel 8.

Areal:

Jeg starter med å kjøre en maskinlæringsalgoritme der bare arealet brukes. Feautre blir da antall piksler. Prøver da å finne en sammenheng mellom arealet til trær og biomassen. Her forventes det at det er sammenheng. For random forest brukes parameterne fast gjennom hele analysen. Antallet trær settes til 10, og subests splittes ikke ned til mindre enn 5. SVM bruker RBF kernel med C = 0,5 og regression epsilon på 0,4. SVM får bedre resultater med lineær kernel i dette tilfellet, men SVM var også da lavere enn både Lineær regresjon og Nevrale nettverk. Nevrale nettverk og SVM kjøres begge to med 200 iterasjoner i alle tilfellene og alfa på 0,1. Dersom det er tydelig at nevrale nettverk ikke klarer å lage gode estimat må alfa endres. Er det overfit må alfa økes, og er det underfit må den senkes. Resten av innstillingene for nevrale nettverk er satt til standardverdiene i alle estimeringene. Standardverdiene i Orange er de samme som er standard i scikit-learn modulen til Pyhton. For lineære regresjon vil både rigde regression og brukes i alle beregninger. I tilfeller med flere variabler testes også lasso regression. I begge tilfeller settes alfa til 0,1.

Sampling type: Leave one out

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.151	0.389	0.330	0.346
Random Forest	0.124	0.352	0.223	0.465
Neural Network	0.084	0.290	0.202	0.637
Linear Regression	0.084	0.290	0.202	0.637

Tabell 7.6.4.A Statistiske resultater fra areal og leave one out validering

Sampling type: Stratified Shuffle split, 10 random samples with 66% data

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.154	0.392	0.333	0.320
Random Forest	0.116	0.341	0.219	0.485
Neural Network	0.080	0.283	0.200	0.647
Linear Regression	0.080	0.283	0.200	0.646

Tabell 7.6.4.B Statistiske resultater fra areal og random sample 66% treningsdata

I den første analysen er det tydelig at Nevrale nett og lineær regresjon har best tilpasning. Dette er en lineær modell med en parameter. Da er det ikke snakk om noe overfit problemer for den lineære modellen. SVM og random forest får ikke noe spesielt gode resultater her, men lineær regresjon og SVM er overraskende bra til å være med bare en parameter. Det som er også interessant er at nevrale nettverk og lineær regresjon gir tilnærmet det samme resultatet. Det vil si at nevrale nettverk ser lineær regresjon som den beste metoden. Resultatet som en får fra å bruke areal er en god referanse for alle andre modeller. Målet er å bruke modeller som gir bedre resultater enn denne. Etter en del prøving og feiling har jeg funnet ut hvilke indekser som ikke gir noe. SR1, SR2, VARI, NI_ratio og Bleaf_ratio.

Areal og høyde, fiktiv laser:

Legger til høyden og ser om det endrer noe i resultatet. Fra en enkel ratevisering av laserdata får en høyde og kronens areal. Dette forventes å gi et bedre resultat enn bare areal alene.

Sampling type: Leave one out

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.145	0.381	0.334	0.373
Random Forest	0.104	0.322	0.193	0.551
Neural Network	0.077	0.277	0.197	0.669
Linear Regression	0.077	0.277	0.198	0.669

Tabell 7.6.4.C Areal og høyde, leave one out

Settings

Sampling type: Stratified Shuffle split, 10 random samples with 66% data

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.142	0.377	0.328	0.370
Random Forest	0.096	0.309	0.192	0.577
Neural Network	0.073	0.269	0.194	0.679
Linear Regression	0.073	0.269	0.194	0.679

Tabell 7.6.4.D Areal og høyde, random sample 66% treningsdata

Resultatet ender opp med å bli bedre når høyde legges til. Det som er interessant er om resultatet er bedre eller dårligere enn det som kommer når spektral data brukes istedenfor høyde. Dersom høyde og areal ikke er noe bedre enn spektralinformasjonen fra de summerte båndene/vegetasjonsindeksene vil det være et tegn på at HySpex kan konkurrere med laser.

Summerte Vegetasjonsindekser:

Nå kjøres maskinlæringsalgoritmen med de summerte vegetasjonsindeksene. Features da blir: SAVI, SR1, SR2, Bleaf_ratio, Vogelman1, NI_ratio, GRVI, Li_ratio, VARI og TVI.

Sampling type: Stratified Shuffle split, 10 random samples with 66% data

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.157	0.396	0.337	0.305
Random Forest	0.100	0.316	0.202	0.559
Neural Network	0.096	0.309	0.208	0.577
Linear Regression	0.088	0.296	0.195	0.612

Tabell 7.6.4.E Summerte vegetasjonsindekser, random sample 66% treningsdata

Resultatet ble lavere enn forventet. Resultatet er dårligere enn når bare arealet ble brukt. Dette er et uforventet resultat. Arealet er inkludert med vegetasjonsindeksene og det var da ikke forventet å få et lavere resultat. Trolig er noen av indeksene nokså dårlige og ikke egnet for

biomasseestimering. Gir et forsøk med å fjerne vegetasjonsindekser som er dårlig egnet for å se om det gir et bedre resultat.

Settings

Sampling type: Stratified Shuffle split, 10 random samples with 66% data

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.157	0.396	0.337	0.307
Random Forest	0.113	0.336	0.216	0.501
Neural Network	0.083	0.288	0.196	0.634
Linear Regression	0.070	0.264	0.176	0.692

Li_ratio, SR1, GRVI, SAVI, TVI og Vogelman1

Tabell 7.6.4.F Gjenværende summerte vegetasjonsindekser, random sample 66% treningsdata

Resultatet ble en god del bedre når noen av indeksene ble fjernet. Spesielt lineær regresjon ble mye bedre. Nå er verdien høyere enn både areal og areal + høyde. Dette er et bra tegn for hyperspektral data. Grunnen til verdiene økte når indekser ble fjernet kan være fordi dem inneholdt tall som ikke hadde noen sammenheng med AGB. En annen grunn som kan være forbedre resultatet er at det blir flere frihetsgrader for regresjonen. Dette er en liten årsak og kan ikke forklare alt av forskjellen i resultatet.

SVM gir ofte dårlige resultater. Det ser ut som om RBF kernel ikke passer i for biomasseestimeringene. Ved å bytte til lineær kernel, ender den opp med R^2 verdier rundt 0,59. For tidligere modeller blir den også bedre når den er lineær, men fortsatt ikke like god som nevrale nett eller lineær regresjon. Dersom epsilon settes veldig lavt (ca. 0,1) forbedres de statistiske resultatene seg, men dette er antageligvis bare overfit. SVM klarer ikke å matche lineær regresjon uten å risikere enn fullstendig overtilpasset modell. Random forest ser ikke ut til å gi noen spesielt gode resultater til nå. Det er Lineær regresjon og nevrale nettverk som er mest interessant.

LI_ratio gir alene veldig gode resultater og er den beste enkeltindeksen for biomasseestimering. Det ser ut som at mengden lignin korrelerer sterkt med biomassen til trær. Det kan kanskje være aktuelt å bruke et enkeltbånd for estimering av biomasse. SVM med lineær kernel gir god R^2 og RMSE, men større MAE (mean absolute error). Lineær regresjon med R^2 på 0,656 er bedre enn å bare bruke arealet.

Settings

Sampling type: Stratified Shuffle split, 10 random samples with 66% data

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.089	0.298	0.229	0.607
Random Forest	0.110	0.332	0.211	0.513
Neural Network	0.078	0.279	0.197	0.656
Linear Regression	0.078	0.279	0.197	0.656

Tabell 7.6.4.G LI_ratio summert vegetasjonsindeks, random sample 66% treningsdata

Summerte bånd:

Dette er den største og mest tidkrevende maskinlæringen. Nå skal alle summerte bånd brukes. Det vil si 369 features. Det er trolig flere bånd som er helt unødvendige og skaper svekker resultatene når alle bånd brukes. Det er vanskelig å fjerne bånd manuelt. For å gjøre dette kan en lasso regresjon gjøres. Dette er en regularization som prøver å bruke minst mulig bånd for å lage regresjonsmodellen. Dersom denne blir god er det et tegn på at veldig mange bånd ikke trengs. Ved å øke alfa i en lasso regression vil flere av koeffisientene til variabler ende opp som 0.

Sampling type: Stratified Shuffle split, 10 random samples with 66% data

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.153	0.391	0.332	0.324
Random Forest	0.106	0.326	0.211	0.531
Neural Network	0.172	0.414	0.259	0.241
Linear Regression	0.091	0.302	0.196	0.597

Tabell 7.6.4.H Summerte bånd, random sample 66% treningsdata

Resultate i 7.6.4.H gir ikke like gode resultater som tidligere. Dette kan ha med at variablene må endres en del når mengden parametere økes. Nevrale nettverk har fått en veldig stor nedgang og har en svært stor RMSE. Random forest ser ut til å gi et nokså likt resultat som tidligere. Lineær regresjon har fått lavere R^2.

Forsøker med lasso regression. Da øker jeg også alfa opp til 0,8 for å fjerne mest mulig uviktige bånd.

Sampling type: Stratified Shuffle split, 10 random samples with 66% data

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.153	0.391	0.332	0.324
Random Forest	0.106	0.326	0.211	0.531
Neural Network	0.149	0.386	0.248	0.340
Linear Regression	0.082	0.286	0.196	0.638

Lasso, alfa = 0,8

Tabell 7.6.4.I Summerte bånd, random sample 66% treningsdata og lasso regression med alfa = 0,8

Ved å øke alfa og kjøre lasso regression forbedres resultatene betraketlig. Det er ikke slik at resultatene blir bedre enn de fra areal eller vegetasjonsindekser, men det er ikke mye dårligere. Dette er et resultat som kan være interessant å se videre på ved hjelp av PLS analyse i kapittel 7.7.

NDVI funksjon:

Det siste som forsøkes ved hjelp av maskinlæring er om NDVI funksjonen som er laget for å beregne biomasse estimerer bedre enn de andre forsøkene. Denne består av en parameter.

Scores

Method	MSE	RMSE	MAE	R2
SVM	0.153	0.391	0.332	0.324
Random Forest	0.114	0.338	0.218	0.496
Neural Network	0.079	0.281	0.198	0.652
Linear Regression	0.079	0.281	0.198	0.652

Tabell 7.6.4.J Summert NDVI funksjon, random sample 66% treningsdata

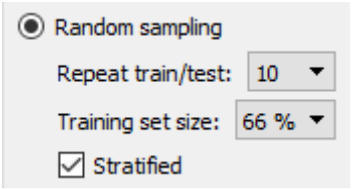
NDVI funksjonen gir et resultat med lineær regresjon som er så vidt lavere enn det lignin gir, men høyere enn det areal alene gir. Dette betyr at NDVI er bedre enn et estimat som bare bruker areal.

7.6.5 Validering og statistiske resultater

I del 5 velges valideringsmetoden. Valideringsmetoden som fungerer best varierer med hvor mange features som brukes og størrelsen på datasettet. Siden datasettet er lite passer det bra å bruke leave one out validering. Den er en god valideringsmetode, men tar for lang tid når det er mange features. Det gjør at den bare passer når bare noen få features brukes slik som når bare høyde og areal er variabler. Dersom alle båndene brukes vil det ta svært langt tid med leave one out validering. For å kunne bruke samme valideringsmetode på alle datasett brukes random sampling med treningssett og testsett. Orange skjuler biomassen for trær i testsettet og

bruker treningssettet til å trene opp algoritmene. I tillegg til å kjøre treningssett og testsett kjøres også leave one out der det er få variabler med.

Samplingen av testdata og treningsdata gjøres tilfeldig. Orange deler opp data selv og deler det med 66% trenings og resten som testdata. Denne oppdelingen gjør Orange 10 ganger. Oppdelingen gjøres forskjellig og kan variere og kan variere hver gang den kjøres. Siden Orange lager 10 slike tilfeldige datasett med trening og testsett. Det totale statistiske resultatet blir mindre tilfeldig når det repeteres mer.



Figur 7.6.5 Slik er innstillingene for validering

7.6.6 Lagring av estimat

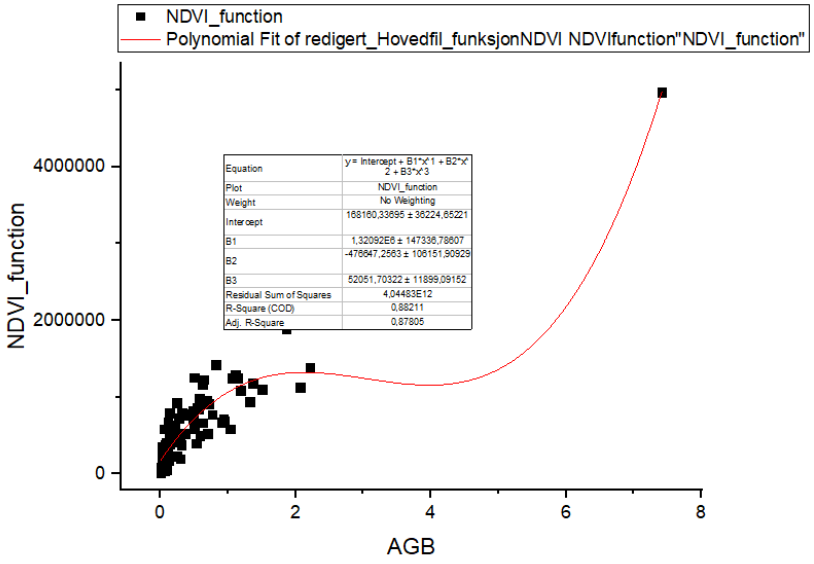
Den siste delen er å sjekke estimatet og deretter lagre resultatet. Selve resultatet kan en se «Predictions». Her kan en se hva den faktiske biomassen er for trær som har den beregnet, og hva Orange har estimert for hver av maskinlæringsmetodene. I scatterplot kan en se visuelt hvordan modellen passer med faktisk biomasse. Dersom en er fornøyd med resultatet kan en bruke knappen «save data» for å lagre resultatet som en CSV med estimerte biomasseverdier.

7.7 Regresjon

I dette kapittelet gjennomgås beregningene og forsøkene som brukes for å lage funksjonsuttrykkene ved hjelp av regresjon. I kapittel 8.1.2 vil de gjøres mer sammenligninger og videre analyse for å finne ut hvilke av dem som er best. Her er både de svake og gode modellene forklart og beregnet. Funksjonsuttrykkene som lages i kapittelet bruker alle antall piksler som en verdi for areal. Dette gjør at funksjonsuttrykkene bare

fungerer for sensorer med en GSD på 0,3 meter med mindre dette korrigeres for med dette uttrykket: $antall\ 0{,}3GSD\ pixels = \frac{Areal}{0{,}3^2}$

Regresjon valgte jeg å gjøre ved hjelp av Origin Pro. Målet er å klare å estimere biomasse mest mulig nøyaktig for trær i tillegg til å klare lage et funksjonsuttrykk som skal kunne brukes i videre biomasseanalyser. Det er viktig at jeg ikke lager modeller som har stor grad av overfit. Da vil modellen ikke fungere i alle tilfeller. I datasettet har trær en biomasse mellom null og tre tonn, bortsett fra et tre som har nærmere 7,5 tonn tørrvekt. Dette er tre nummer 128. Jeg tenker å lage en modell som bruker alle trær bortsett fra dette. Det samme ble gjort i maskinlæringen også. Dette er fordi modellen har lite grunnlag for hvordan biomassen skal være for større trær. Ut i fra feltarbeidet ser det ut som at trær i urbane miljø vanligvis ikke har svært stor biomasse. I analyser og fremgangsmåte vil det alltid brukes AGB. Det er regnet noen verdier for røtter og summen av biomasse for hele treet, men dem har noe mer usikkerhet. AGB er også den vanligste måten å gi biomasse for trær.



Figur 7.7 Tredjegradsfunksjon som tydelig overfitter data. Her er NDVI funksjonen og AGB plottet og prøvd å finne et tredjegradspolynom fit. Vi vet at biomassen ikke synker når NDVI funksjonen øker, men det viser ikke modellen. Modeller som dette kan ikke brukes. Det er for lite data til å estimere AGB for trær over 2,5 tonn.

7.7.1 Lineær regresjon areal:

Den første Regresjonen jeg kjører er den enkleste. Den tar inn en parameter og det er arealet til trekronene. Antall piksler beskriver arealet til trær. Regresjonen som kjøres er en vanlig lineær regresjon. I Origin bruker jeg ingen regularization for regresjonene. Det vil si at alle lineære regresjoner er minste kvadraters regresjon. Denne modellen vil være en referanse for de andre modellene. De som bruker spektralinformasjon bør være bedre enn den som bare bruker areal. Dersom det ikke er det, betyr det at spektralinformasjonen som brukes i modellen ikke har noen korrelasjon med biomasse.

		Value	Standard Error	t-Value	Prob>|t|
AGB	Intercept	-0,11805	0,04978	-2,37174	0,01989
	"Pixels"	0,00106	8,09135E-5	13,05209	0

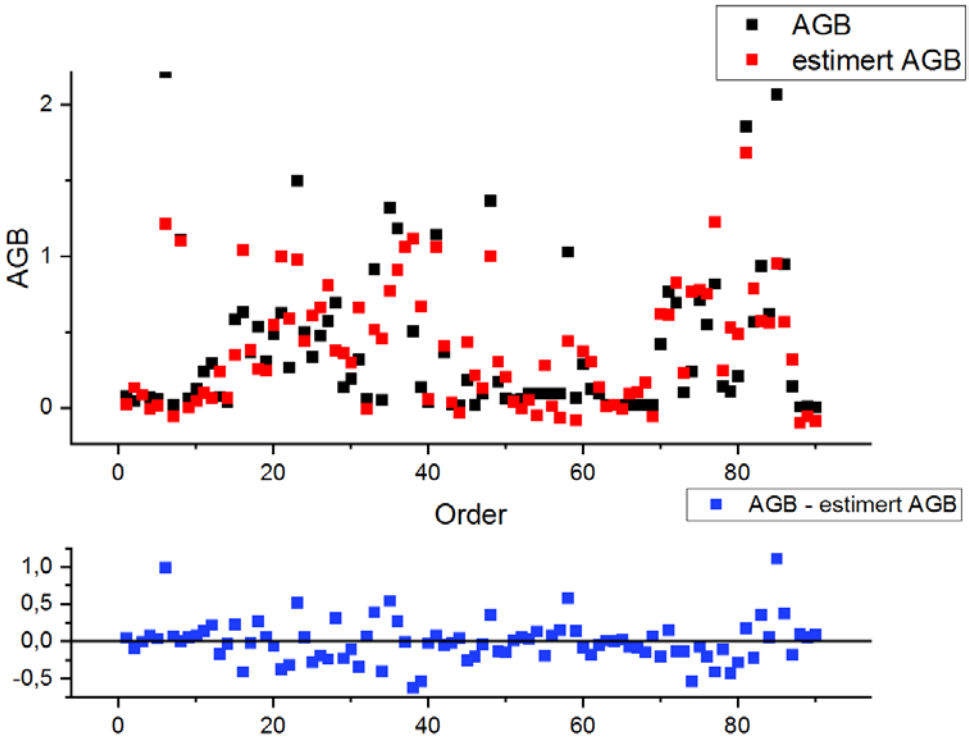
Standard Error was scaled with square root of reduced Chi-Sqr.

Statistics

	AGB
Number of Points	90
Degrees of Freedom	88
Reduced Chi-Sqr	0,08052
Residual Sum of Squares	7,08596
R Value	0,81203
R-Square (COD)	0,65939
Adj. R-Square	0,65552
Root-MSE (SD)	0,28376
Norm of Residuals	2,66195

Tabell 7.7.1.A Statistiske resultater for areal regresjonsmodellen

Modellen får en R^2 på ca. 0,655 som er litt høyrere enn Orange modellens lineære regresjon. Dette kommer av at denne ikke har separert treningssett og testsett. Metoden med å beregne areal er den samme som brukes i (Bernasconi et al., 2017). For å se hvor godt estimatet er kan en lage et plot med både estimert biomasse og faktisk biomasse. Da ser vi at modellen både over og underestimerer litt, men treffer noenlunde greit for de fleste punkter. To trær har fått et avvik i faktisk AGB på over 1 tonn. Det er de største trærne som får størst avvik. Trær som er under 11 piksler ender også opp med å få en negativ biomasse. For en storskala biomasseestimering har dette lite å si fordi dette er trær som har tilnærmet ingen biomasse. I mindre analyser kan dette påvirke resultatet mer og er ikke ønsket.



Figur 7.7.1.B Plot av estimert AGB og faktisk AGB

Funksjonsuttrykket lages av koeffisientene til pikselmengden. Det er viktig å huske at denne arealfunksjonen baserer seg på antall piksler. Dersom en bruke en annen romlig opplønsing enn 0,3 meter GSD må det lages en nye funksjon for dette.

$$AGB = -0{,}11805 + 0{,}00106 \times Pixels$$

Formel 7.7.1.C AGB funksjonsuttrykk der antallet piksler er eneste parameter

7.7.2 Lineær regresjon areal og høyde:

Den andre regresjonen bruker både høyde og areal. Det vil si det samme parameterne som laserdata bruker. Det forventes at resultatet blir en del bedre enn det som bare bruker areal. Resultatet som kommer fra denne regresjonen vil ligne på det en får dersom en direkte kjører laserdata på et 2,5 dimensjonalt datasett der en kan finne arealet til trær og høyden til pikslene. Regresjonen er en multippel lineær regresjon med to variabler.

		Value	Standard Error	t-Value	Prob>|t|
AGB	Intercept	-0,27596	0,06808	-4,0533	1,09568E-4
	"hoyde"	0,02402	0,00745	3,22536	0,00177
	"Pixels"	8,71025E-4	9,59542E-5	9,0775	3,10862E-14

Standard Error was scaled with square root of reduced Chi-Sqr.

Statistics

	AGB
Number of Points	90
Degrees of Freedom	87
Reduced Chi-Sqr	0,07275
Residual Sum of Squares	6,32915
R Value	0,83413
R-Square (COD)	0,69576
Adj. R-Square	0,68877
Root-MSE (SD)	0,26972
Norm of Residuals	2,51578

Tabell 7.7.2.A Statistiske resultater for areal og høyde regresjonsmodellen

Som forventet er R^2 høyere for denne modellen. Modellen blir noe bedre, men den blir mindre bedre enn forventet. Det ser ut til at arealet til trekronen er viktigere enn høyden når en skal beregne biomassen, eventuelt at arealet og høyden har en del korrelasjon som gjør at høyden ikke gir mye ny informasjon. Antageligvis øker nok både høyden og arealet lineært og tilnærmet like mye AGB. Både areal og høyde har en positiv koeffisient for AGB. Funksjonsuttrykket for regresjonen blir slik:

$$AGB = -0{,}27596 + 0{,}02402 \times høyde + 0{,}000871025 \times Pixels$$

Formel 7.7.2.B AGB funksjonsuttrykk der antallet piksler er eneste parameter

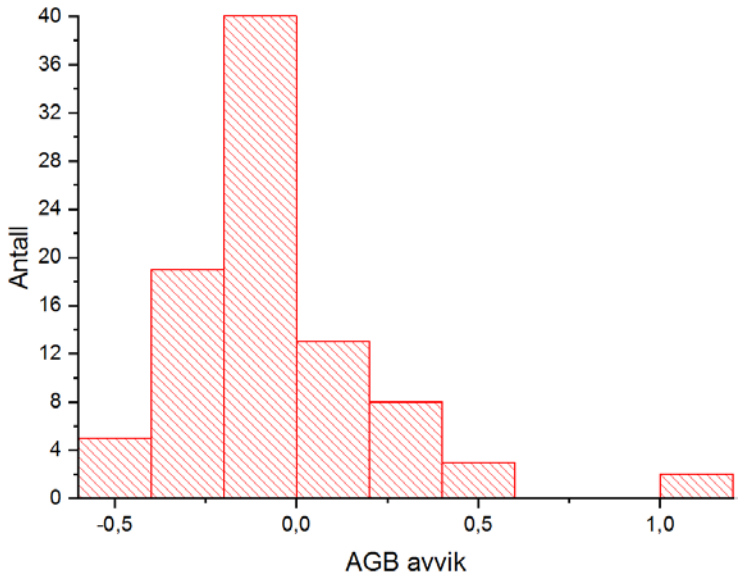
7.7.3 Lineære regresjon NDVI funksjon:

Denne regresjonen er for å se hvordan NDVI funksjonen fra (Liu et al., 2006) fungerer. Denne funksjonen forventer reflektansverdier og ikke-normaliserte radiansbilder. Likevel er det verdt å gi den et forsøk siden NDVI fungerer ok på normaliserte bilder også likevel om det ikke er det den er laget for. Denne funksjonen er laget for å fungere for mer enn bare trær, og skal fungere for gress og annen vegetasjon. Dette kan gjøre at den gir svake resultater i dette forsøket. Den er ikke laget for norsk klima.

	AGB
Number of Points	90
Degrees of Freedom	89
Reduced Chi-Sqr	0,08381
Residual Sum of Squares	7,45885
R Value	0,88806
R-Square (COD)	0,78865
Adj. R-Square	0,78628
Root-MSE (SD)	0,28949
Norm of Residuals	2,73109

Tabell 7.7.3.A Statistiske resultater for lineær regresjon med NDVI-funksjonen

Ved å bruke NDVI funksjonen fikk jeg bedre resultater enn forventet. Modellen gir en R^2 på 0,786 der er langt mye bedre enn det areal-regresjonen gjorde. Det er også bedre enn resultatet som areal og høyde gir. Det som er overraskende er at R^2 blir mye bedre, men RSS (Residual Sum of Squares) økes en god del. Ønsket er at denne skal være nærmest mulig 0 og denne kan også beskrive hvor godt en modell passer til datasettet. Høy RSS gir også høy RMSE. Alt i alt virker det som om NDVI kan være et greit utgangspunkt for å estimere biomasse. Modellen gir små avvik for de fleste trær, men svært store avvik for to trær. Det er de to største trærne som får store avvik. Den har små underestimeringer for små trær og store avvik i de største. Modellen underestimerer oftere enn den overestimerer. Dette vises godt i et histogram:



Figur 7.7.3.B Histogram som viser avvikene til trær. 2 trær har feilestimert AGB på mer enn 1 tonn

Resultatene i NDVI funksjonen gir trolig svært ulike verdier fra de tidligere modellene fordi intercept (verdien til AGB når NDVI-funksjon = 0) er satt til 0. Intercept bør ikke settes til 0 med mindre verdien som regresjonen foreslår har en høy P-verdi. Jeg kjørte en ny NDVI funksjon hvor jeg lot modellen endre intercept. Denne fikk en lav P-verdi. De nye tallene ser slik ut:

		Value	Standard Error	t-Value	Prob>|t|
AGB	Intercept	-0,11854	0,04941	-2,39917	0,01854
	"NDVI_function"	9,63966E-7	7,31843E-8	13,17176	1,59946E-22

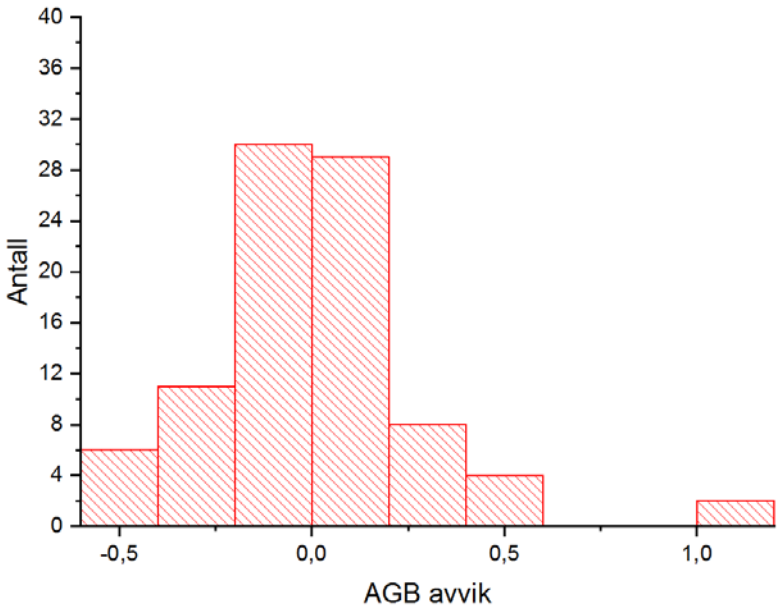
Standard Error was scaled with square root of reduced Chi-Sqr.

Statistics

	AGB
Number of Points	90
Degrees of Freedom	88
Reduced Chi-Sqr	0,07956
Residual Sum of Squares	7,00092
R Value	0,81454
R-Square (COD)	0,66347
Adj. R-Square	0,65965
Root-MSE (SD)	0,28206
Norm of Residuals	2,64592

Tabell 7.7.3.C Statistiske resultater for lineær regresjon med NDVI-funksjonen med intercept

Det som skjer nå er at R^2 synker og RSS synker. Dette er nok den mer korrekte regresjonen å bruke. Nå er resultatene omtrent like med det resultatet som kommer fra å bare bruke areal. Det en liten forbedring i R^2. Det vil si at NDVI inneholder noe informasjon som hjelper for å estimere biomasse, men det er svært lite. NDVI er en indeks som omtrent alle sensorer har og den er bedre enn å bare bruke areal, men forskjellen er omtrent ubetydelig. Ved å legge til intercept endret også histogrammet seg. Nå er det ikke slik at modellen har små underestimeringer for svært mange trær.



Figur 7.7.3.D Nytt histogram som viser avvikene til trær. De fleste trær har et avvik innenfor 0,2 tonn.

7.7.4 Lineær regresjon vegetasjonsindekser:

I denne regresjonen bruker jeg mer enn bare et bånd. Nå kjøres en multippel lineær regresjon med alle ikke-normaliserte idekser. Fra eksperimentering og testing har det vist seg at normaliserte indekser ikke virker like bra på normalisert data, spesielt NDNI, NDLI og NDBleaf. Vegetasjonsindeksene SR1, SR2, TVI, GRVI, Vogelman1, NI_ratio, LI_ratio, Bleaf_ratio beholdes. Den lineære regresjonen gir oss dette resultatet:

		Value	Standard Error	t-Value	Prob>|t|
AGB	Intercept	-0,07029	0,04825	-1,45668	0,14907
	"GRVI"	0,00194	0,00116	1,67294	0,0982
	"TVI"	7,2821E-6	1,08269E-5	0,67259	0,50312
	"SR1"	-0,002	0,00124	-1,62156	0,10878
	"SR2"	-1,72725E-5	2,29151E-4	-0,07538	0,9401
	"Bleaf_ratio_sum"	-2,49676E-4	2,25192E-4	-1,10873	0,27083
	"Vog1hyper"	-0,0067	0,00254	-2,64044	0,00993
	"NI_ratio"	-6,33296E-4	8,83736E-4	-0,71661	0,47567
	"LI_ratio"	0,00545	0,0022	2,47671	0,01534

Tabell 7.7.4.A Parameterresultatet fra den lineære regresjonen med ikke-normaliserte indekser.

SR2 har lavest absolutt T-verdi og høyest P-verdi. P-verdien gir sannsynligheten for at nullhypotesen er sann. I lineær regresjon er nullhypotesen at koeffisienten (value) er lik 0.

$$H_0 : \beta_j = 0 \qquad\qquad H_\alpha : \beta_j \neq 0$$

Formel 7.7.4.B Til venstre er nullhypotesen som sier koeffisienten er 0.

Dersom nullhypotesen er sann vil det si at indeksen har ingen betydning på estimering av biomasse. Siden P = 0,94 for SR2 betyr det at denne har en 94% sannsynlighet for burde være 0. SR2 har ikke noe bidrag som forbedrer biomasseestimeringen. Dette er veldig høyt, og det er ønskelig å ha P < 0,05 for alle variabler. Vi kan ikke forkaste H0 når denne er svært høy. Vi ser at også NI_ratio, TVI og Bleaf_ratio har høy P-verdi, men gjør ikke noe med de enda. Vi fjerner en og en variabel om gangen fordi når SR2 fjernes kan alle andre P-verdier endres betraktelig. Noen ganger er det bare en feil som gir utslag for hele modellen. Setter koeffisienten til SR2 til 0 og får da et nytt resultat. I tillegg til å fjerne SR2 forsøkte jeg å fjerne GRVI istedenfor. Når denne fjernes ender SR2 opp med å være svært signifikant. Fra min eksperimentering ser det ut som at informasjonen i SR2 overlapper med GRVI som baserer seg på bølgelengder i omtrent det samme området. Da er det ikke nødvendig å ha med begge to. Dette vil si at SR2 ikke nødvendigvis er en dårlig parameter å bruke for biomasse, men at den ikke trengs om en har GRVI. Etter fjerningen av SR2 ser resultatet slik ut:

		Value	Standard Error	t-Value	Prob>|t|
AGB	Intercept	-0,07058	0,04781	-1,47627	0,1437
	"GRVI"	0,00195	0,00115	1,7023	0,09249
	"TVI"	6,98552E-6	1,00253E-5	0,69679	0,48791
	"SR1"	-0,00204	0,00112	-1,83064	0,07079
	"Bleaf_ratio_sum"	-2,49697E-4	2,23822E-4	-1,11561	0,26785
	"Vog1hyper"	-0,00676	0,00242	-2,79458	0,00647
	"NI_ratio"	-6,68113E-4	7,48826E-4	-0,89221	0,37489
	"LI_ratio"	0,00555	0,00171	3,25001	0,00167

Tabell 7.7.4.C Parameterresultatet etter fjerning av SR2

Etter reduseringen av SR2 er det fortsatt flere variabler som ikke er gode. Da må igjen variabelen med høyest P-verdi fjernes. Dette gjøres helt til alle har en lav P-verdi. Ved å gjøre dette får vi sett hvor mye vi kan maksimalt redusere en vegetasjonsindeksene ifølge multippel lineær regresjon. Håpet er at svært få vegetasjonsindekser trengs slik at enklere multispektrale sensorer kan brukes for å estimere biomasse med vegetasjonsindekser. Etter stegvis fjerning av alle indekser som har høy P-verdi blir resultatet slik:

		Value	Standard Error	t-Value	Prob>|t|
AGB	Intercept	0	--	--	--
	"GRVI"	0,00174	6,66238E-4	2,61097	0,01065
	"SR1"	-0,00188	6,17047E-4	-3,04186	0,00312
	"Vog1hyper"	-0,0067	0,002	-3,35325	0,00119
	"LI_ratio"	0,00468	0,00121	3,88305	2,01881E-4

Tabell 7.7.4.D Parameterresultatet etter stegvis fjerning av SR2, TVI, NI_ratio, Bleaf_ratio og intercept

Regresjonsanalysen kommer frem til at 4 vegetasjonsindekser er signifikante for å estimere biomasse. Indeksene er GRVI, SR1, Vogelman1 og LI_ratio. GRVI, Vogelman1 og LI_ratio var indekser som jeg forventet at kom til å være signifikante, men jeg trodde at NI_ratio også kom til å være signifikant. Fra resultatene kan en se at LI_ratio har høyest positiv innvirkning på biomassen. Det vil si at ligninmengden i trær har mye å si for biomassen. GRVI er også positiv og dette virker fornuftig ettersom den var svært god til å identifisere trekroner og busker med høy biomasse fra gress og tynn vegetasjon i kapittel 6.4. Vogelman1 gir som vanlig høye verdier for trær som har god helse eller er i vekstfasen. Trær i vekstfasen har liten stamme og det gir liten biomasse. Det er logisk at denne har en negativ innvirkning på biomassen. Det samme gjelder for SR1 som også baserer seg red edge. Min forventning var at NI_ratio ville ha en negativ koeffisient fordi nitrogeninnholdet synker når trær vokser og blir eldre. Denne variabelen var ikke signifikant nok ble fjernet fra regresjonen. Nå som, alle variabler er signifikante kan vi se på det statistiske resultatet:

	AGB
Number of Points	90
Degrees of Freedom	86
Reduced Chi-Sqr	0,06837
Residual Sum of Squares	5,88006
R Value	0,9129
R-Square (COD)	0,83339
Adj. R-Square	0,82564
Root-MSE (SD)	0,26148
Norm of Residuals	2,42488

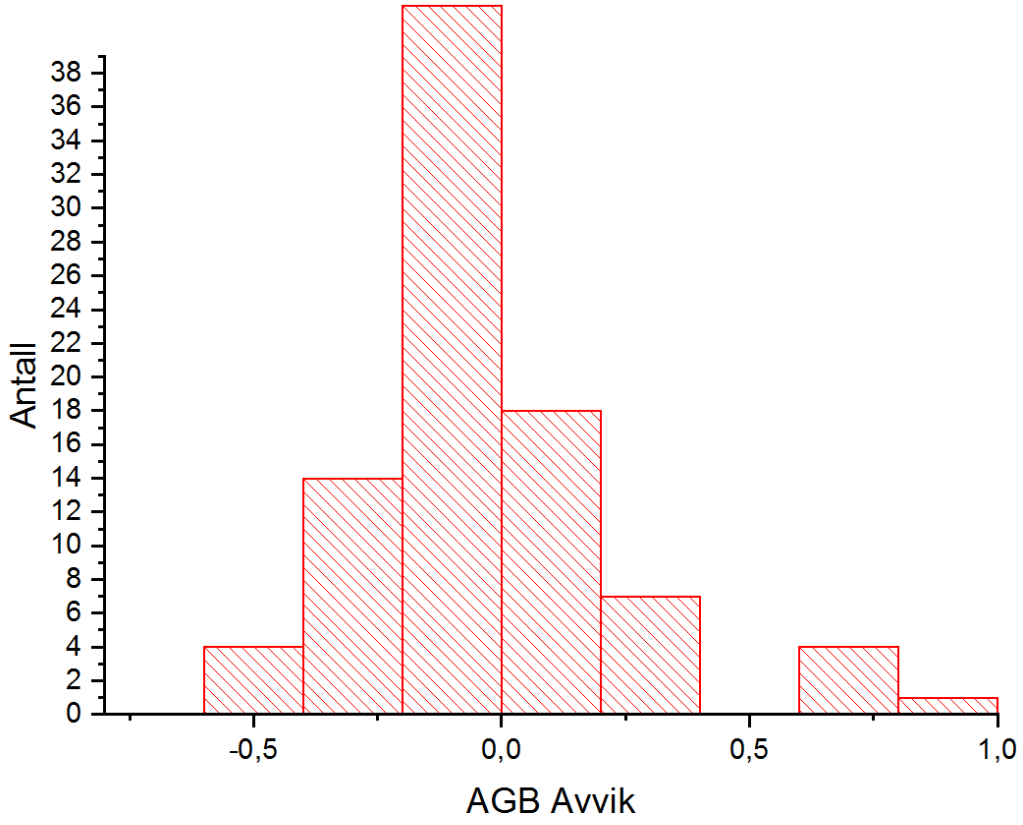
Tabell 7.7.4.E Statistikk for regresjon med vegetasjonsindekser

For denne regresjonen blir resultatene svært gode. RSS er lavere enn for de tidligere modellene og R^2 er mye høyere. GRVI, SR1, Vogelman1 og LI_ratio lager et godt funksjonsuttrykk for å estimere biomasse:

$$AGB = \frac{1{,}74 \times GRVI + 4{,}68 \times LI_{ratio} - 1{,}88 \times SR_1 - 6{,}70 \times Vogelman_1}{1000}$$

Formel 7.7.4.F Funksjonsuttrykk laget fra multippel lineær regresjon av vegetasjonsindekser

Avvikene for store trær har også blitt mindre med denne modellen. Fra histogrammet kan vi se at ingen trær har et avvik over 1 tonn. Ved å sammenligne de estimerte biomassene med de faktiske biomassene, er det tydelig at små trær har små avvik, og store trær har store avvik. Avviket i prosent av faktisk biomasse ser ut til å være mer likt. Mer sammenligning og sjekk av resultater finner en i kapittel 8.1.2.



Figur 7.7.4.G Histogram som viser avvikene til trær

7.7.5 PLS analyse av vegetasjonsindekser:

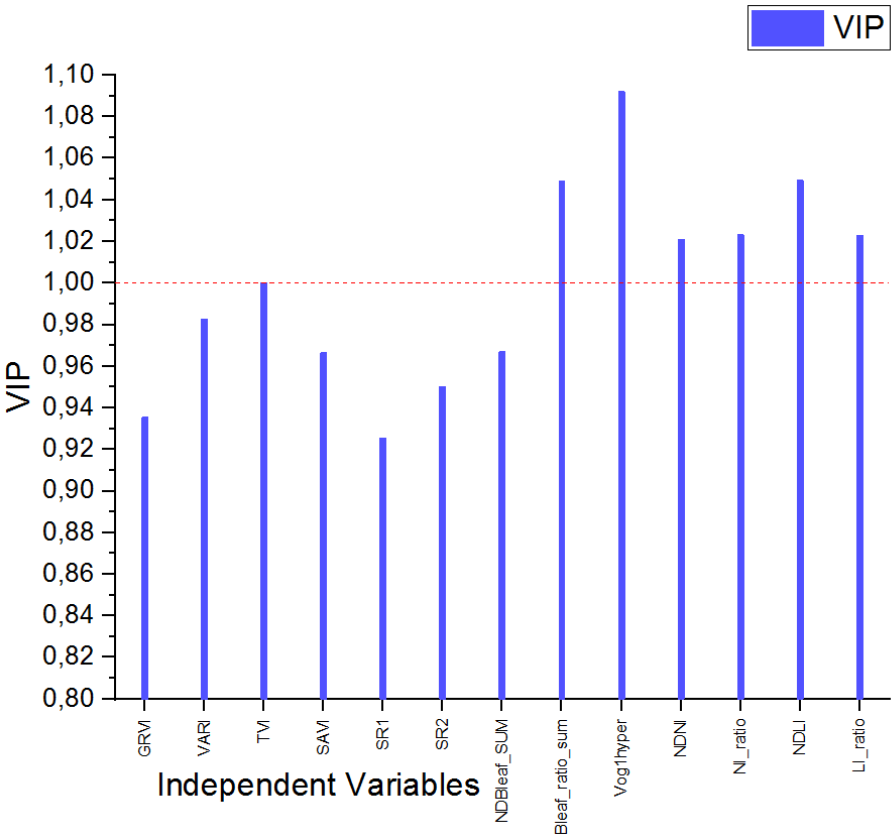
Istedenfor å kjøre lineær regresjon og redusere bort vegetasjonsindekser manuelt kan vi bruke en PLS analyse. PLS analysen vil lage et VIP plot som viser hvilke indekser som er verdt å ta med og hvilke som ikke gir noen nødvendig informasjon.

Det første som blir gitt når en kjører en PLS i Origin er antall faktorer som PLS bruker. Ved hjelp av kryssvalidering og Wold's iteration finner origin selv ut hvor mange faktorer som er optimalt. For mange faktorer vil føre til overfit og for få vil gi en dårlig modell. I dette tilfellet var den optimale mengden faktorer 8 med en Root Mean PRESS på 0,6122. For dette bruket er det resultatet som Origin gir automatisk godt nok. Flere faktorer gir mer forklart varians.

Number of Factors	Variance Explained for X Effects(%)	Cumulative X Variance(%)	Variance Explained for Y Responses(%)	Cumulative Y Variance(%)
1	95,75688	95,75688	63,65431	63,65431
2	2,84618	98,60306	3,14124	66,79554
3	0,86967	99,47273	1,24694	68,04249
4	0,29857	99,7713	1,72002	69,76251
5	0,1406	99,9119	1,85539	71,6179
6	0,05011	99,96201	1,64974	73,26765
7	0,01305	99,97505	2,6575	75,92515
8	0,00531	99,98037	1,53686	77,46201

Tabell 7.7.5.A antall faktorer og varians forklart

Det neste som gjøres i analysen er å se på VIP plottet. Det forteller oss hvilke variabler som er mest signifikante. Det første vi kan legge merke til er at alle variablene har nokså like verdier i plottet. Det er ingen verdier som er nær 0 fordi alle av dem er summerte indekser som inneholder informasjon om arealet. VIP scoren er likevel forskjellig for variablene, og vi kan velge bort de som er lavest. I dette tilfellet kan vi sette en grense på 1,0 og ta med alle vegetasjonsindekser over 1,0.



Figur 7.7.5.B VIP plott for vegetasjonsindekser

VIP plottet gir interessante resultater. Det er bare en indeks fra VNIR som har et høyt resultat her. For SWIR har alle høy verdi bortsett fra NDBleaf. Det er overraskende at SWIR indeksene blir sett på som svært viktige. Jeg tar med alle indekser over 1,0 VIP score videre til en multippel lineær regresjon. Indeksene som er med er Vogelman1, LI_ratio, NI_ratio, NDNI, NDLI og Bleaf_ratio. Regresjonen gir dette resultatet:

		Value	Standard Error	t-Value	Prob>|t|
AGB	Intercept	-0,08959	0,04899	-1,82853	0,07106
	"Bleaf_ratio_sum"	-3,77329E-4	1,99945E-4	-1,88717	0,06263
	"Vog1hyper"	-0,00363	0,00201	-1,80839	0,07417
	"NDNI"	-0,01443	0,00968	-1,49083	0,1398
	"NI_ratio"	0,00213	0,00238	0,89655	0,37255
	"NDLI"	0,01373	0,00958	1,43422	0,15527
	"LI_ratio"	0,00202	0,00286	0,70615	0,48207

Figur 7.7.5.C Regresjon av indeksene til PLS

Fra regresjonen får vi at NI_ratio og LI_ratio har en høy P-verdi. Dette betyr at de ikke er signifikante i modellen. Dette kommer nok trolig av at de inneholder mye av det samme som NDNI og NDLI. Jeg velger da å prøve å redusere bort indekser og starter med de som har høyest P-verdi, og da fjernes LI_ratio. Vogelman 1 ble også fjernet etterpå fordi den hadde en høy P-verdi. Etter reduseringen av de to indeksene fikk jeg dette resultatet:

		Value	Standard Error	t-Value	Prob>|t|
AGB	Intercept	-0,10407	0,04843	-2,14886	0,03449
	"Bleaf_ratio_sum"	-4,05707E-4	1,92739E-4	-2,10496	0,03825
	"NDNI"	-0,01594	0,00456	-3,49279	7,6115E-4
	"NI_ratio"	0,00266	7,70723E-4	3,45638	8,57065E-4
	"NDLI"	0,01414	0,00377	3,74569	3,26737E-4

Standard Error was scaled with square root of reduced Chi-Sqr.

Statistics

	AGB
Number of Points	90
Degrees of Freedom	85
Reduced Chi-Sqr	0,07226
Residual Sum of Squares	6,14181
R Value	0,83951
R-Square (COD)	0,70477
Adj. R-Square	0,69088
Root-MSE (SD)	0,26881
Norm of Residuals	2,47827

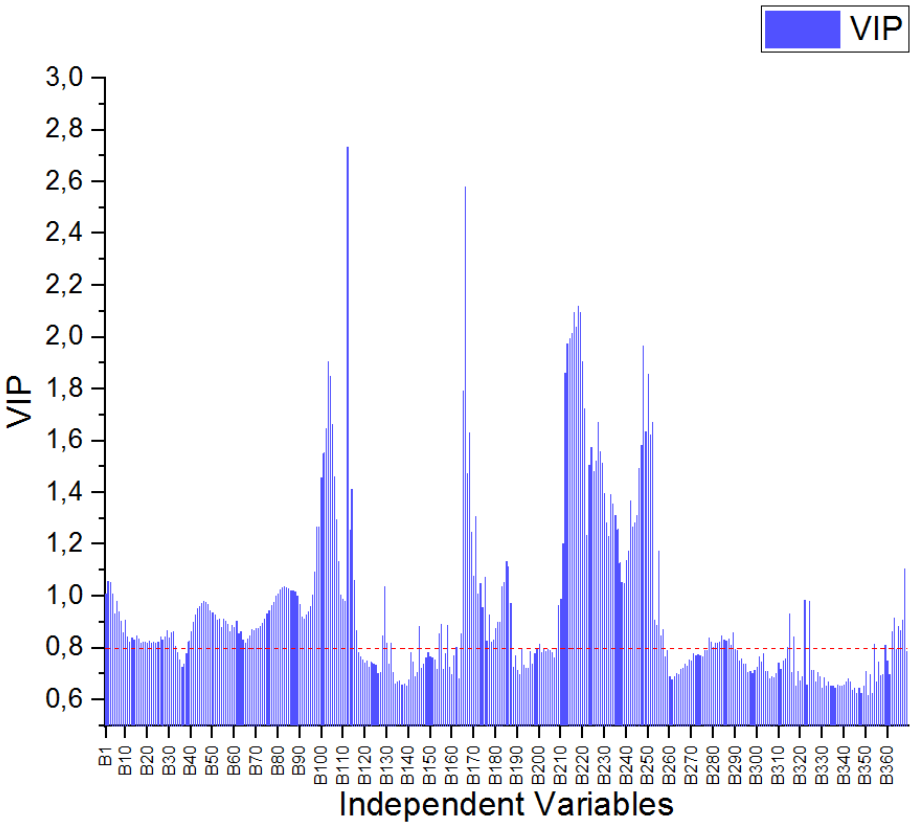
Figur 7.7.5.D Regresjonsresultater etter redusering av indekser

Dette resultatet er vanskelig å tolke. R^2 er ikke spesielt god og er den svakeste bortsett fra å bare bruke areal. Indeksene har også unormale koeffisienter. For eksempel har NDNI negativ koeffisient og NI_ratio har positiv koeffisient. Det er forventet at begge burde være negative. Bleaf_ratio burde ikke ha en negativ innvirkning på biomasse. Den har fått en svært liten koeffisient som er negativ. Det vil si at den har liten betydning i modellen, men den er likevel rar. Det eneste som gir mening her er at NDLI er positiv. Lignin er trolig en viktig faktor for biomasse. Jeg tror ikke denne modellen har noe som helst nytte og er svært tvilsom til å bruke den for biomasseestimeringer. Årsaken for de merkelige verdiene kan komme av kraftig overfit og tilfeldighet. Det kan være at verdiene her fungerer fint for mitt datasett, men at det bare er tilfeldige grunner for det. Modellen vil nok estimere unøyaktig biomasse for alle andre datasett.

7.7.6 PLS analyse av bånd:

Det første jeg ønsker å gjøre før jeg kjører selve analysen er å finne ut hvor mye av informasjon arealet har i forhold til spektralinformasjonen i båndene. Ved å kjøre en PLS med alle 369 bånd og arealet får jeg en VIP score som viser dette. Areal variabelen får en VIP score på 8,6. Båndene har en VIP score mellom 0 og 2,1. Her er det ikke brukt summerte bånd. Båndene inneholder gjennomsnittsverdier for hvert tre. Dette viser at arealet er helt klart den viktigste faktoren for biomasseestimeringene, men det er også informasjon i en del bånd. Båndene som har VIP score rundt 2 kan inneholde en del spektralinformasjon som kan forbedre regresjonene.

Jeg fjerner nå arealvariabelen fra PLS analysen og kjører en ny PLS med bare bånd. Jeg ønsker å bruke denne til å finne de 20 viktigste båndene for å estimere biomasse. Plottet jeg får da ser slik ut:



Figur 7.7.6.A VIP plott for bånd

Siden båndene ikke er summerte i denne analysen får vi store utslag på VIP resultatene. Noen bånd får svært lave verdier og noen andre har veldig høye verdier. Områdene der VIP er lavest ser ut til å treffe godt med hvor det er mye spektralinformasjon fra atmosfæren og vannabsorbsjon (men de med mest vannabsorbsjon er fjernet i 7.4).

Jeg velger de 20 båndene med høyest VIP verdi og gjør en multippel lineær regresjon med båndene. De 20 viktigste båndene her var:

Nummer	Variable	VIP
1	B112	2,7338
2	B166	2,57773
3	B218	2,11971
4	B219	2,09692
5	B216	2,09638
6	B217	2,03809
7	B215	2,01489
8	B214	1,99521
9	B213	1,97711
10	B248	1,9681
11	B103	1,9073
12	B220	1,90663
13	B212	1,8622
14	B250	1,85742
15	B104	1,84887
16	B165	1,79102
17	B221	1,72565
18	B227	1,67321
19	B252	1,66935
20	B105	1,66177

Tabell 7.7.6.B 20 båndene med høyest VIP score

De fleste av båndene er SWIR bånd, men 5 av båndene er VNIR bånd. Den lineære regresjonen kjøres på samme måte som tidligere og båndene reduserer bort på samme måte som tidligere. Dette gjøres helt til alle bånd har en P-verdi under 0,05. Når dette er gjort står det 6 bånd igjen. 3 fra SWIR og 3 fra VNIR:

		Value	Standard Error	t-Value	Prob>|t|
AGB	Intercept	0	--	--	--
	"SUM_B103"	0,01384	0,00312	4,43977	2,72061E-5
	"SUM_B104"	-0,01452	0,00316	-4,59207	1,52647E-5
	"SUM_B112"	0,00414	8,50958E-4	4,86581	5,26917E-6
	"SUM_B214"	0,02155	0,00388	5,54989	3,26085E-7
	"SUM_B218"	-0,04516	0,00583	-7,7434	1,97436E-11
	"SUM_B219"	0,0197	0,0025	7,87783	1,06455E-11

Standard Error was scaled with square root of reduced Chi-Sqr.

Statistics

	AGB
Number of Points	90
Degrees of Freedom	84
Reduced Chi-Sqr	0,0488
Residual Sum of Squares	4,09914
R Value	0,94013
R-Square (COD)	0,88385
Adj. R-Square	0,87556
Root-MSE (SD)	0,22091
Norm of Residuals	2,02463

Tabell 7.7.6.C Resultater fra multippel lineær regresjon med bånd fra PLS analysen

Dette resultatet ble svært bra og bedre enn forventet. Her er R^2 svært høy og RSS er svært lav. Dette gjøt at også RMSE ender opp med å være vesentlig lavere enn for de tidligere modellene. Fra statistikken er denne modellen overlegen ovenfor de tidligere modellene. Det som gjør meg litt usikker er båndene som brukes. Båndene i VNIR er som forventet i red edge området. B112 (761,1 nm) og B103 (732,4 nm) er tilnærmet de sammen båndene som Vogelman1 bruker. Derimot SWIR båndene er mer unormale. 1130nm, 1152nm, 1157nm er ikke bølgelengder som jeg kjenner til at skal være relevante for biomasse. Jeg vet ikke hvilke biokjemiske absorberinger som skjer i de bølgelengdene. Det nærmeste jeg vet om er at 1120 nm absorberer lignin (Serrano et al., 2002), og burde ha en positiv koeffisient for biomasse. Det jeg også synes er merkelig er at bånd som er svært nær hverandre har helt ulik påvirkning på biomasse. 1152 nm har en negativ koeefisient og 1157 nm har positiv koeffisient. Jeg har ingen forklaring på hvorfor det er slik og vet ikke om dette er normalt. Det jeg frykter er at resultatet kommer av overfit eller tilfeldigheter. Funksjonsuttrrykket blir slik:

$$AGB = \frac{1{,}384 \times B103 - 1{,}452 \times B104 + 0{,}414 \times B112 + 2{,}155 \times B214 - 4{,}516 \times B218 + 1{,}97 \times B219}{100}$$

Formel7.7.6.D Funksjonsuttrykk fra PLS analyse med bånd og lineær regresjon

Absorbing biochemical	Wavelength (nm)
Water	970, 1200, 1400, 1450, 1940
Nitrogen	1020, **1510**, **1730**, **1980**, **2060**, 2130, **2180**, **2240**, **2300**
Lignin	1120, 1200, 1420, 1450, 1690, **1754**, 1940, **2262**, **2380**

Tabell 7.7.6.E biokjemiske absorberingsområder fra artikkel (Serrano et al., 2002)

Båndnummer	Bølgelengde
B103	732,4 nm
B104	736,5 nm
B112	761,1 nm
B214	1130,2 nm
B218	1152,1 nm
B219	1157,5 nm

Tabell 7.7.6.F Bølgelengdene til båndene som brukes i regresjonen

7.8 Kontroll og manuell testing

Det siste som gjøres i fremgangsmåten er å gjøre noen tester for å se at estimatene virker. Funksjonsuttrykkene er laget basert på et «best case» tilfelle, der alle trær står alene og hele trekronen kan indentifiseres. Teste blir gjort på to ulike klynger av trær som er realistiske utfordrerne områder. Det første området er en klynge bestående av 10 trær i en klynge. De største trærne er svært store. Under de store trærne er det mindre trær som ikke er synlige i bildene. Dette er en vanskelig klynge å estimere og det forventes en stor feil her. Det andre området er en rekke med 10 trær. Her er det vanskelig å skille nøyaktig når neste tre starter og slutter, men arealet som overlapper er lite. Her forventes det at biomassen blir estimert noe

bedre. Jeg velger å teste 3 ulike funksjonsuttrykk siden jeg mener de er mest aktuelle for estimeringer: Areal 7.7.1, vegetasjonsindekser 7.7.4, bånd 7.7.6.

Figur 7.8.A Klynge forsøk1

Figur 7.8.B Klynge forsøk2

Klynge1 består av 6643 piksler og har en AGB på 6,595. Resultatene for funksjonsuttrykkene ble slik:

Funksjonsuttrykk	Estimert AGB	Avvik	Avvik / antall trær
Areal	6,924	0,329	0,033
Vegetasjonsindekser	9,974	3,051	0,305
Bånd	7,197	0,602	0,060

Tabell 7.8.C Estimat av treklynge1

Det beste estimater for klyngen kommer fra arealfunksjonen i dette tilfellet. Den bommer med 0,329 tonn. Dette tallet høres kanskje stort ut, men dette er estimatet for 10 trær. Det tilsvarer en feil på 33kg per tre og er langt bedre enn det som forventes av modellen. Funksjonsuttrykket for bånd overestimerer med 0,6 tonn og er også bedre enn forventet. Vegetasjonsindekser får den største feilen med 3 tonn. Dette er en stor feil. Modellen forventer ikke trær som er over 2,5 tonn og dette ser ut til å påvirke den mest. Eventuelt har dette området større verdier i LI_ratio eller GRVI enn det som er normalt, og lavere SR1 og Vogelman1.

Klynge nummer 2 består av 3579 piksler og har en AGB på 2,888 og fikk dette resultatet:

Funksjonsuttrykk	Estimert AGB	Avvik	Avvik / antall trær
Areal	3,676	0,788	0,079
Vegetasjonsindekser	3,110	0,566	0,057
Bånd	1,965	0,922	0,092

Tabell 7.8.D Estimat treklynge2

For denne klyngen blir alle resultatene bedre enn de forrige. Det er nok fordi det er mindre overlapp, og nærmere modellens anbefalte rekkevidde. 2,888 AGB er nær grensen på 2,5 AGB. Det beste estimatet her kommer fra vegetasjonsindeksene som gir et svært nøyaktig resultat på 57 kg feil i estimeringen. Det største kommer fra båndene som gir en feil på 92 kg. Alle feilene her er svært små og bedre enn hva jeg hadde forventet.

På grunn av lite tid fikk jeg bare gjennomgått to klynger, men resultatet for klyngene var bedre enn forventet. Jeg var redd for at hele modellen skulle gi store feil og nærmest tilfeldige verdier i klynger og «worst case» senarioer. Fra testene her ser det ikke slik ut. Det kan virke som om vegetasjonsindeksen er svært nøyaktig for trær og treklynger ved lav biomasse, og at feilen i estimatet øker veldig fort når klyngens størrelse øker.

8 Analyse og Resultater

8.1 Sammenligning av modeller

8.1.1 Sammenligning av resultater fra maskinlæring

Det første jeg ønsker å sammenligne er maskinlæringsmodellene. Hvilke metoder er best for å estimere AGB? Ved å sammenligne R^2 fra lineær regresjon i Orange kan vi finne ut hvilke modeller som er best. Modellen med høyest mulig R^2 er den beste dersom den ikke overfitter. I orange har det vært forsøkt å hindre overfit for alle resultater. Trolig har nok en del modeller og forsøk fått en del underfit der modellen ikke blir tillatt å justere seg nok til å passe dataene. Sluttresultatene burde verken ha stor grad av under eller overfit.

Metode	MSE	R^2	RMSE
Areal	0,080	0,646	0,283
Areal + høyde	0,073	0,679	0,269
Vegetasjonsindekser	0,088	0,612	0,296
LI_ratio, SR1, GRVI, SAVI, TVI, Vogelman1	0,070	0,692	0,264
LI_ratio	0,078	0,656	0,279
Bånd	0,082	0,638	0,286
NDVI funksjon	0,079	0,652	0,281

Tabell 8.1.1 MSE, RMSE og R^2 fra lineær regresjon i Orange

Fra tabellen ser vi at den som oppnår høyest grad av fitting er den som bruker summerte vegetasjonsindekser. Det er ikke den som bruker alle som får best resultat, men den som bruker noen utvalgte. Det er verdt å legge merke til at alle indeksene som var viktige for biomassen var indekser som ikke er normaliserte. Det kan hende at noen av de normaliserte indeksene er svært betydelige, men det vises ikke fordi bildene som er brukt er normalisert. R^2 på 0,692 er en del bedre enn resten av modellene og dette er min anbefalte modell å bruke. Den krever 2 bånd i SWIR og 6 bånd i VNIR. Likevel er det verdt å legge merke til areal uten

noen spektralinformasjon har en R^2 på 0,646. Den gir en MSE på 0,080. Det vil si at modellen har en gjennomsnittfeil i estimert biomasse på 80 kg. Gjennomsnittverdien for biomasse over bakken er ca. 480 kg for trær brukt analysen. Det betyr at modellen vil feilestimere biomassen med ca. 16%. Om dette er et godt nok resultat spørs på hvor nøyaktig data en trenger. Resultatet vil være godt nok for en storskala analyse som ikke krever spesielt nøyaktige resultat. Modellen vil derimot ikke være spesielt god dersom en ser på et lite område med bare noen få trær. Da ønsker en som oftest bedre resultat. Den beste modellen gir en MSE på 0,070. Det er tydelig forskjell mellom dem, men i prosent er feilestimeringen av biomassen er omtrent lik.

Resultatet fra Orange indikerer til at det ikke er mye hjelp i å bruke hele det hyperspektrale bildet for å beregne biomasse. For å få et godt estimat var det nødvendig å øke alfa for en lasso regression. Da fjerner den veldig mange bånd og lar bare de som er svært signifikante være med i regresjonsanalysen. Dette betyr at en kunne like godt forsøkt å finne de viktige båndene istedenfor å bruke alle. Dette vil spare tid og gjøre det lettere å skaffe sensorer som er gode for biomasseestimering. PLS analysen i Origin finner ut hvilke bånd som ikke er relevante i 8.1.2.

Dersom en ønsker å bruke et enkelt bånd er NDVI eller LI_ratio det beste ut ifra resultatene fra maskinlæringen. LI_ratio er noe bedre enn NDVI-funksjonen. NDVI har fordelen av å være mulig å lage med omtrent alle sensorer siden det eneste som trengs er bånd i NIR og rødt lys. LI_ratio derimot trenger båndene 1754nm og 1680 nm som er noe sjeldnere. Nettsiden Index Database har notert 10 ulike sensorer som kan beregne NDLI.

I artikkelen til Bernasconi brukes en single tree segmentation for å estimere biomasse. Dette er den samme metoden som areal-biomasse beregningen fra maskinlæringen beregner. Begge bruker lineær regresjon og ingen spektralinformasjon for å estimere biomasse. I denne testen ender resultatet opp med en R^2 0,7. I resultatet fra Orange ble R^2 0,646. Dette er noe lavere. Med bedre romlig oppløsning var det forventet å få noe bedre resultat enn det som Bernasconi oppnår. Grunnen til at det er lavere kan være fordi trær i urbane områder er kuttet og

vedlikeholdt. I Bernasconis tilfelle var det trær som vokste fritt og arealet av alle kroner passet nok bedre med stammen. I urbane områder kan trær ha store stammer uten å nødvendigvis ha en stor krone fordi den kan være klippet ned flere ganger. Områder som Gamle Oslo bygårdene og Galgeberg hadde tydelig vedlikeholdte trær som var klippet flere ganger. Unaturlige trekroner gjør at det er vanskeligere å estimere biomassen basert på areal.

8.1.2 Sammenligning av resultater fra Origin

Her sammenligner jeg de 6 ulike metodene for regresjonsanalyse som jeg gjorde i kapittel 7.7. Jeg vil starte med å se på PLS analysen fra 7.7.5. Denne hadde svært merkelige tall som ga lite mening. Med tvilsomme verdier for NDNI, NI_ratio og Bleaf_ratio har jeg lite tro på at denne modellen er god. Grunnen til at denne fungerer er trolig bare at den tilfeldigvis passer bra for feltdataen. Dette er nok overfit og modellen er trolig svært dårlig i andre tilfeller. De 5 andre modellene er mer interessante.

Metode	Adjusted R^2	RSS	RMSE
Areal	0,659	7,086	0,284
Areal + høyde	0,689	6,329	0,270
NDVI-funksjon	0,660	7,001	2,646
Ikke-normaliserte vegetasjonsindekser	0,826	5,880	0,261
PLS bånd	0,876	4,099	0,221

Tabell 8.1.2.A Adjusted R^2, RSS og RMSE fra lineær regresjon i Origin

De statistiske resultatene viser at spektralinformasjonen fra bånd gir best resultat. Denne gir helt tydelig best R^2, RSS og RMSE, men likevel er jeg ikke helt sikker på om modellen er god. Som nevnt i 7.7.6 er det rart at bånd som er nær hverandre har helt ulike koeffisienter. Det er også spesielt at bånd rundt 1130nm og 1155nm skal være særs viktig for

biomassestimeringen. Det kan være at dette er en god modell, men jeg har ikke nok kunnskap om absorberingen i dette området.

Ikke-normaliserte vegetasjonsindekser gir også gode statistiske resultater. Ved å bruke 4 vegetasjonsindekser (7 ulike bånd) blir resultatet bra og verdiene er logiske. Det gir mening at koeffisienten til LI_ratio er positiv og det er forventet at Vogelman1 skal ha en negativ koeffisient. Dersom en skal gjøre en god estimering av biomasse i urbane strøk anbefaler jeg å bruke denne modellen:

$$AGB = \frac{1{,}74 \times GRVI + 4{,}68 \times LI_{ratio} - 1{,}88 \times SR_1 - 6{,}70 \times Vogelman_1}{1000}$$

Modellene for Areal er den svakeste. Denne har fordelen av å kunne brukes for pankromatiske bilder. Det trengs ingen bånd med spesifikke bølgelengde for å bruke denne modellen. Dersom en har en sensor med de vanligste båndene kan en vurdere å bruke for eksempel NDVI-funksjonen istedenfor. NDVI funksjonen gir bedre resultater enn arealfunksjonen, men forskjellen er svært liten. Dersom valget står mellom pankromatiske bilder med høy romlig oppløsning, eller NDVI med noe lavere romlig oppløsning, bør en velge de pankromatiske bildene. Romlig oppløsning er mye viktigere for biomasseestimering enn informasjonen en får fra NDVI. Når metoden skal brukes kan en bytte pikselstørrelse med areal siden pikslene må være 0,3 meter for at modellen fungerer. Dette gjelder alle modellene.

$$AGB = -0{,}11805 + 0{,}00106 \times Pixels$$

$$AGB = -0{,}11805 + 0{,}00106 \times \frac{Areal}{0{,}09}$$

Formel 8.1.2.B Areal formelen der antall 0,3 meters piksler byttes ut med areal i kvadratmeter

Når en legger til høyden blir modellene bedre. Det som gjør det vanskelig med høyder er at det ikke er like lett å skaffe. For å skaffe høydedata må en enten ha laser eller bruke stereobilder.

Modellene kan sammenlignes med resultatene fra Bernasconi. I forsøket hans fikk han R^2 = 0,7 og 0,9. Mine modeller med bare areal får dårligere resultater. Dette kommer nok av variasjonen i biomassen for trær i byen. Flere ulike treslag gjør at arealet ikke blir et like godt mål for biomasse alene i min oppgave. Trekronene kan også være klippet som skaper mindre korrelasjon mellom trekrone AGB. Modellene som bruker spektralinformasjon i tillegg får R^2 over 0,8 og er bedre enn hans single tree metode. Ingen av mine modeller klarer å oppnå en verdi på 0,9 slik som Bernasconis arealbaserte metode klarer. Dette er en helt annerledes metode og i tillegg er det helt ulike datasett. Det er lite hensiktsmessig å sammenligne med dette resultatet når alt er ulikt.

8.2 Hvordan klarer vegetasjonsindeksene å forbedre estimatene?

Den lineære regresjonen av bånd klarer å forbedre resultatene betraktelig fra det resultatet som en får fra areal estimatene, men hvordan klarer den egentlig det? Det er fordi vegetasjonsindekser har ulike verdier i piksler basert på trærs alder, helse og treslag. For eksempel klarer modellene tydelig å vise at store mengder lignin i pikslene gir stor biomasse og høy verdi i red edge betyr at treet har lav biomasse. Dette er informasjon som laser ikke har tilgjengelig og det kan erstatte informasjonen vi får fra høyden. Siden trær med svært bratt red edge vanligvis er yngre trær i vekstfasen kan denne indirekte fortelle at stammens biomasse er liten. Og høye verdier i LI_ratio betyr at kronen må inneholder mye lignin, som er en viktig faktor for biomasse. Samlet sett klarer vegetasjonsindeksene å gi svært mye informasjon om treets oppbygning og regresjonsmodellene klarer da å tolke hvordan treets biomasse antageligvis er. Dette gjør at høydedata og stamme ikke er like viktig å vite. Det vil alltid være noen trær som er utenfor normalen. For eksempel syke trær, døde trær og i noen tilfeller helt vanlige trær vil ha ulik spektralsignatur og gjør at modellen bommer stort. I en storskala gjør det lite om noen trær bommer så lengde flertallet av de vanlige trærne får solide estimat.

8.3 SWIR mot VNIR

Maskinlæringsresultatene viser at både SWIR og VNIR kan brukes for å estimere biomasse. Den beste modellen fra maskinlæringen bruker 8 bånd. 6 bånd er VNIR, og 2 er SWIR. Den mest signifikante vegetasjonsindeksen i modellen er LI_ratio. Indeksene som er relevante i VNIR er hovedsakelig indekser som baserer seg på red edge området. Modellen viser at både VNIR og SWIR er nyttig for biomasseestimering og ut ifra maskinlæringsresultatene er det ikke klart at det ene skal være noe bedre enn det andre. Det som er klart er at dersom en vegetasjonsindeks skal brukes, så er SWIR bedre enn VNIR. VNIR har langt flere bånd som korrelerer med biomasse enn SWIR, men SWIR har noe få som utmerker seg som ekstra gode.

I regresjonsanalysene fra Origin er både SWIR og VNIR relevant for biomasseestimering. I SWIR er det likt som i maskinlæring. LI_ratio er en viktig og god indikator for biomasse. Resten av SWIR vegetasjonsindeksene hadde lite eller ingen sammenheng med AGB i analysene. I 7.7.6 viste det seg at bølgelengdene 1130nm, 1152nm og 1157nm kunne brukes for å estimere biomasse. Det forventes å finne ligninabsorbsjon rundt 1120nm og kanskje har 1130nm noe med dette å gjøre. For 1152nm og 1157nm vet jeg faktisk ikke hva årsaken kan være. Kanskje er det overfit og tilfeldigheter, eller kanskje er det noe her som være nyttig for å estimere biomasse? Dette er kanskje noe som er verdt å se videre på.

8.4 Sammenheng mellom kroneutbredelse og biomasse

Det er ingen tvil om at kroneutbredelse er den viktigste faktoren for biomasseestimering i 2-dimensjonale bilder er arealet av trekronen. Alle modeller kommer frem til at arealet er det som forklarer best hva biomassen til et tre er. Spesielt godt vises dette i PLS analysen og tabellen 7.7.3.A. VIP verdien til antall piksler er på 8,6 og båndene ligger på verdier mellom 0 og 2,1. Det er tydelig at arealet er viktigere enn alt av spektralinformasjon samlet. Det betyr ikke at spektralinformasjon er unyttig, men den kan ikke gi gode resultat uten arealet av trekroner. For å forbedre arealberegningene slik en får enda bedre biomasseestimeringer, må

en ha bedre romlig oppløsning. Mindre piksler gjør at en får mer nøyaktig geometri for trær. Dette gir mer korrekte areal, som igjen gir bedre estimat for biomasse.

I urbane områder er det mye mer usikkert hvor stor korrelasjonen mellom trekrone og biomasse passer. I skog har alle trær med stor stamme stor trekrone, men i byen er det mye klipping og vedlikehold av trær. Dette gjør at korrelasjonen mellom trekrone-stamme og trekrone-biomasse er lavere. Dette er et problem som alle flybårne sensorer får siden det er omtrent umulig å vite om en trekrone er klippet eller ikke. Mitt beste forslag for å unngå dette problemet er å bruke satellittbilder. Da kan en oppdage om arealet av en trekrone har blitt redusert. Ved å sjekke eldre bilder kan en finne ut hvor stor trekronen var på sitt største og bruke dette arealet for å estimere biomasse.

8.5 Sammenheng mellom nitrogen, lignin og biomasse

Jeg hadde forventet at høye nitrogenverdier skulle ha en sammenheng med lav biomasse. Detter fordi nitrogenmengden i trær er høy i vekstfasen når stammen er liten. Ved regresjonsanalysen av vegetasjonsindekser (7.7.4) var det en sammenheng mellom NI_ratio og lav biomasse, men denne var ikke stor. Den hadde også en høy P-verdi. Det var ikke noen signifikante resultater som sa at NDNI eller NI_ratio hadde sammenheng med biomasse. I PLS analysen og regresjonen i 7.7.5 fikk jeg signifikante resultater med indeksene NDNI og NI_ratio, men indeksene var motsatte. Jeg tror ikke at 7.7.5 er en god modell og kan derfor ikke konkludere med at biomasse og nitrogen har noen signifikant sammenheng.

For lignin var det klarere resultater. Trær med høye verdier i lignin har som oftest høy biomasse. Dette er tydelig i 7.7.4 og 7.7.5. Spesielt viser regresjonen i 7.7.4 at LI_ratio verdiene bestemmer biomassen for modellen. LI_ratio gir også gode resultater i maskinlæringen og er alene en god indikator for å estimere AGB.

8.6 Hvor mange bånd/vegetasjons indekser er nødvendig?

Fra mine regresjoner er svaret at svært få bånd trengs for å estimere AGB. Å bruke 369 bånd i maskinlæringsalgoritmene ga ingen tegn til forbedring i estimatene. De beste estimeringene bruker bare 6-7 bånd og opp til fire vegetasjonsindekser. For en enkel estimering holder det å vite arealet til trekroner slik som (Bernasconi et al., 2017) gjør. Mine resultater viser også at areal gir estimater som er nærme det en får når det brukes spektralinformasjon. I maskinlæringen fra Orange viste det seg at spesielt LI_ratio var god for å estimere AGB. En vegetasjonsindeks gir gode estimat for biomasse og 4 vegetasjonsindekser gir svært gode estimat.

8.7 Kan biomassen estimeres ved hjelp av satellitt eller omløpsfoto?

Jeg tror både satellittbilder eller omløpsfoto kan være svært gode for estimeringer av biomasse. Fra regresjonsanalysene er det ingen tvil om at romlig oppløsning er den viktigste faktoren og dette er bra for både satellitt og omløpsfoto. Omløpsfoto gir svært god romlig oppløsning og har langt mye lavere kostnad enn hyperspektrale foto. Satelittbilder må ha god romlig oppløsning for å fungere for denne typen biomasseestimering. Oppløsningen må være god nok til å kunne identifisere enkeltstående trær. Satellitter som Quickbird og Worldview gir god oppløsning som klarer å gi arealer på trekroner. Derimot satellitter som Sentinel-2 vil ikke ha noen mulighet fordi oppløsningen er alt for lav. Utfordringen for satellitter er å klare å gjøre gode nok segmenteringer. Dette er langt mye enklere når en har flere bånd.

For de aller beste estimatene bør en nok vurdere å bruke noe annet enn satellitt. Da er det nok hyperspektrale eller multispektrale flyfoto som er best så lenge de har god romlig oppløsning. Den optimale sensoren ville vært en multispektral sensor med svært god romlig oppløsning som tar opp synlig lys, red edge, 1510nm og 1754 nm. Da ville sensoren hatt alt som trengs for å kjøre funksjonsuttrykket fra 7.7.4 og LI_ratio regresjonen som ble gjort i Orange.

8.8 Hvor gode er estimeringsmodellene og hvilke utfordringer har modellene?

Det er vanskelig å si hvor gode estimeringene er. For trær uten overlapp og med en trekrone som ikke klippet ned mye blir resultatene svært gode. Problemet er at i reelle tilfeller er det sjeldent slik. De fleste områder vil ha overlappende trær, klippede trekroner og trær som er for store for modellen. Dette er de tre utfordringene denne modellen har og det er ikke lett å vite hvordan problemene skal løses. Dette er problemer som også flybåren laser vil støte på når den bruker single tree estimeringer.

For klippede trekroner kan kanskje satellitt med temporal oppløsning kunne rette problemet med å ta i bruk arealet når kronen var størst. Overlapp et er problems som jeg tror er vanskelig å løse. Den beste muligheten er kanskje å prøve å bruke noen form for 3-dimaensjonale data i tillegg for å oppdage om det er trær som overlapper eller står under hverandre. 3-dimensjonal data kan kanskje også sjekke om stammer er splittet. Problemet med at modellen ikke passer for større trær er enkel å løse. Det som kreves er trær i alle størrelser og langt flere trær. Et feltarbeid på flere hundre trær i ville gjort modellen bedre egent for å estimere trær i alle størrelser. At datasettet bare inneholder 91 trær er en stor svakhet for modellene. Dette er et lite utvalg og gjorde at modellen ikke er egnet å estimere biomasser for trær som er over 2,5 tonn tørrvekt over bakken.

8.9 Hvordan bruke biomassefunksjonen

Når en skal gjøre en estimering av AGB trenger en ikke gjøre alt som står i kapittel 5.5. Istedenfor kan en bruke funksjonsuttrykkene som er laget. Dersom en bare har areal kan en bruke arealfunksjonen (Formel 7.7.1.C) og har en hyperspektrale bilder eller multispektrale bilder kan en bruke formel 7.4.4.E for vegetasjonsindekser. Ved å bruke funksjonsuttrykkene kan en hoppe over steget med regresjon og maskinlæring. For å kjøre funksjonsuttrykket trengs det fortsatt normalisert data. Funksjonsuttrykkene laget i disse oppgavene er ikke atmosfærekorrigert og det er ikke nødvendig.

Når datasettet er normalisert kan en segmentere slik at en får et datasett med et objekt for hvert tre. For objektene må antall piksler lagres for arealfunksjonen. Segmenteringen og lagring av egenskaper anbefales å gjøre i et program som eCongintion. Nå kan en regne ut biomassen for hvert tre med formel Formel 7.7.1.C. Antallet piksler i objektet er eneste ukjente og resultatet fra formelen blir AGB for hvert tre.

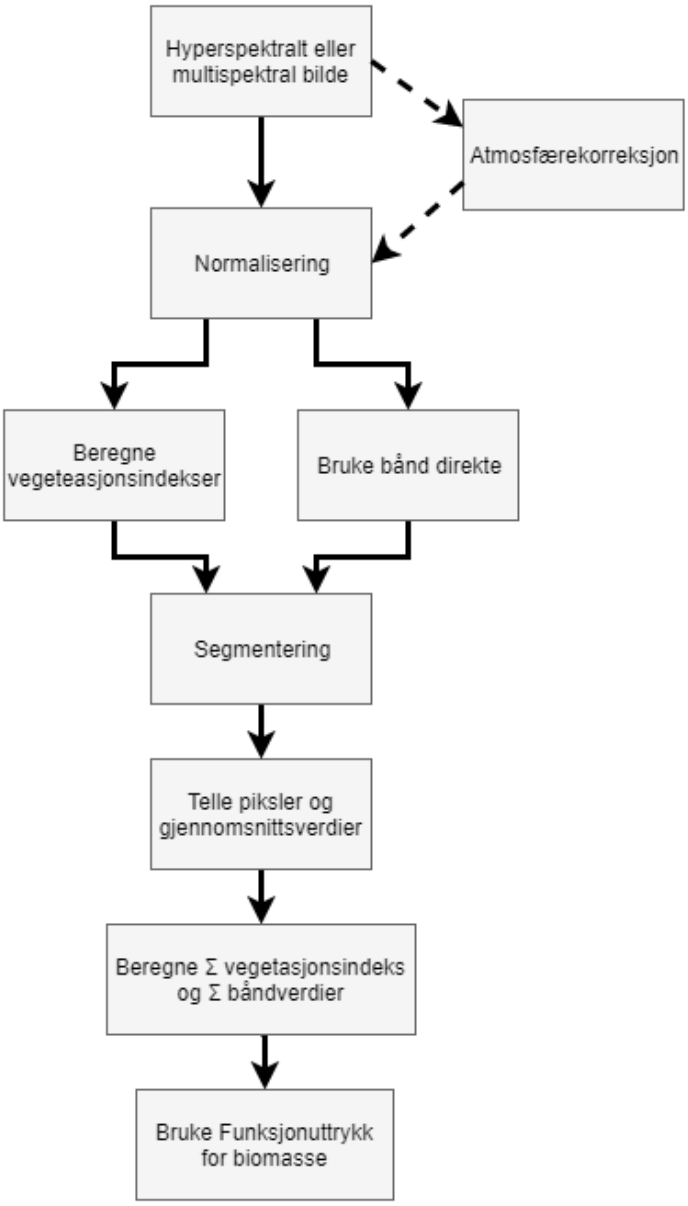
De andre funksjonsuttrykkene trenger også å lagre gjennomsnittsverdier for båndene. Verdiene brukes videre for å beregne vegetasjonsindeksene. Vi får da gjennomsnittsverdien til vegetasjonsindeksene for hvert tre. Gjennomsnittsverdiene multipliseres med antall piksler. Verdiene en har da er summerte vegetasjonsindekser som kan brukes i formelen 7.4.4.E. Resultatet fra formelen vil da bli AGB oppgitt i antall tonn tørrvekt for hvert tre.

Arealfunksjonen er beregnet på at en bruker antall 0,3 meter piksler en trekrone dekker og vegetasjonsindeksene er summert på antall 0,3 meter piksler som trekronen blir dekket av. Bruker en mindre piksler blir estimatet svært høyt, og bruker en større piksler blir estimatet svært lavt. For å unngå dette problemet må en bytte ut piksler med areal i en vanlig enhet, som for eksempel kvadratmeter. Dersom en skal bruke funksjonsuttrykkene og ikke har en GSD på 0,3 meter kan en fortsatt bruke funksjonene, men det data må datasettet enten resamples eller omregnes til piksler. Resamplingen gjøres ved at en setter pikselstørrelsen til å være 0,3 meter GSD i selve bildet og bruker nærmeste nabo samplingsmetode. Da får en 0,3 meter piksler i bildet og resten kan gjøres normlt. Svakheten med dette er at en svekker

datasettet med å endre romlig oppløsning. Det andre alternativet er å finne arealet i kvadratmeter, og deretter finne ut hvor mange piksler med GSD 0,3 meter som er passer for det arealet. Ved å gjøre dette taper en ingen informasjon.

$$Pixels = \frac{Areal}{0{,}3^2}$$

Formel 8.9.A Omgjøring av areal til piksler med 0,3 meters oppløsning



Figur 8.9.B Figuren viser fremgangsmåten for å estimere biomasse ved hjelp av funksjonsuttrykk

8.10 Bruke en treslagsklassifisering som segmentering

Istedenfor å kjøre en segmentering kan en bruke en treslagsklassifisering. En god klassifisering vil ha klassifisert treslag hver for seg uten å ha med spesielt mye gress. Bygninger vil heller ikke havne i treslagsklassen. Dersom en slår sammen alle treslag til en klasse og fjerner alle andre klasser sitter vi igjen med et godt utgangspunkt for biomasseestimering. For å sikre at veldig små trær og gress ikke havner i modellen kan vi også bruke morfologiske operasjoner. Ved å først kjøre en erode, og deretter en dilate med en kernel size på rundt 5 fjernes de minste trærne, gress og støy. Om dette er gjort i klassifiseringen er det ikke nødvendig. Det vi sitter igjen med er et datasett hvor alt som ikke er trær er filtrert ut uten å segmentere eller tegne noe manuelt enda.

For det nye datasettet vil det være enkelt med en segmenteringsalgoritme å detektere trekronene. Eventuelt kan en også tegne inn manuelt som også er enklere siden alt som ikke er trær er filtrert bort.

Biomasseestimeringsmetoden i kapittel 8.8 forkortes ned mye når klassifiseringen allerede er gjort. Dette er også en god måte å kombinere en klassifisering og biomasseestimering med det samme datasettet. Et hyperspektralt datasett kan da kjøre både klassifiseringsmetoden og biomasseestimeringen uten å trenge laserdata. Alternativt kan en bruke multispektrale bilder og gjøre det samme. Ved å bruke omløpsfoto eller flyfoto med bare synlig lys kan en også få gode resultater fra biomasseestimeringen siden kronearealet er det viktigste, men det spørs hvor god treslagsklassifiseringen blir da.

Om en bruker hyperspektral data og ønsker både treslagsklassifisering og biomassestimering i urbane områder anbefales det å se på Erik Røstads oppgave. Der kjøres en solid klassifisering av treslag ved hjelp av hyperspektrale flyfoto. Oppgaven bruker de samme fotoene som denne oppgaven og deler av de samme områdene. Ved å først kjøre en treslagsklassifisering og deretter bruke den som en maske for biomassen får en både treslag og biomasseestimer for de samme trærne.

8.11 Sammneligning av HySpex og Laserdata

Laserdataene er gjort av Kristoffer Ceballos og bruker de samme testområdene som jeg bruker for HySpex dataene. Laserdataene bruker noe færre trær. Laserdata bruker totalt 58 trær. Metoden som er brukt er også en single tree metode slik som HySpex datene. Modellene bruker feltdata kombinert med laserdata for å lage en lineær regresjonsmodell som estimerer høyden på treet og stammens diameter. Resultatene fra modellene brukes til å beregne AGB for trærne. Modellen får en R^2 på 0,711 og RMSE på 0,443. Modellene er gjort uten regularization og valideringssett. Siden mine origin regresjonsmodeller heller ikke har regularization og validering passer det å sammenligne resultatene.

Modellene mine som bruker bare areal får lavere R^2. Dette er ikke overraskende siden dette er en modell som bruker mindre informasjon. Derimot mine analyser som bruker spektralinformasjon har høyere R^2 og lavere RMSE. For eksempel 7.7.4 regresjonen med vegetasjonsindekser har en mye høyere R^2. Jeg tror grunnen til dette er fordi laserdata ikke er optimalt for urbane områder. Trær i urbane områder har større variasjon i korrelasjonen mellom kroneareal og stammediameter. Dette er fordi det er mange ulike treslag, dårligere forhold for trærne og fordi trærne kan ha vært klippet og behandlet. Når laser støter på slike problemer som at trekronen er klippet og ikke passer med stammediameter, har ikke laser noen måte å korrigere for dette. Den vil ende opp med å estimere for lav stammediameter på grunn av liten trekrone, og dette gir feil i estimater. Derimot hyperspektral data kan korrigere med å vite nitrogeninnhold, ligninninnhold, bratthet på red edge etc. Dette gjør at regresjonsmodellene kan korrigere biomassen selv om arealet er mindre enn vanlig. Dette gjør at HySpex takler bedre trær som har unormal trekrone eller unormal stammediameter. Siden modellene ikke tar hensyn til stammen har denne ingen påvirkning for modellen. Dersom en bruker satellittbilder med synlig lys eller multispektrale flyfoto mister vi mye av denne fordelen med at spektralinformasjonen kan korrigere biomasse. For multispektrale bilder kan noe informasjon beholdes fra bredbåndsindekser som for eksempel GRVI. Likevel tror jeg ikke dette er nok til å være bedre enn laser. Når det gjelder satellittbilder kan de trolig konkurrere på tilgjengelighet og pris. Satellittbilder kan gi tilfredsstillende resultater med å bare bruke arealet dersom en har god romlig oppløsning. Resultatet vil være dårligere enn

laser, men det vil være tilgjengelig å kjøpe når som helst og kanskje er det også en god del billigere? For store områder vil det være mye enklere å bruke satellittbilder enn laser.

Det er verdt å legge merke til at mengden trær er forskjellige for laser og HySpex. Dette gjør at HySpex får noe bedre modeller enn laser automatisk av å ha større datasett. Dette merkes spesielt med at RMSE blir veldig høy i laser. Det gir lite mening i å sammenligne RMSE for to datasett som er ulike.

For områder med helt ordinære trær tror jeg laser er noe bedre enn hyperpsektral data. Da vil høyde og stamme for laser bli korrekt, og resultatene blir gode. Der vil hyperspektral også treffe godt på arealet, men vil ikke ha noen informasjon om høyden. Måten HySpex klarer finne noe om biomassen er igjen med å se på brattheten i red edge og andre slike egenskaper. Dette blir ikke like nøyaktig som når høyden/stammen er korrekt for laser.

Lasermetoden som det sammenlignes med baserer seg på at en må ha feltdata i området for å fungere. Dette trenger ikke metoden fra hyperspektral. Hyperspektralmetoden trenger bare å lage et funksjonsuttrykk og etterpå kan den brukes i hvilket som helst område.

9 Refleksjon

9.1 Er HySpex verdt å skaffe for biomasseestimeringer?

For å gjøre biomasseestimeringer tror jeg ikke at HySpex er nødvendig. HySpex er dyrt og gir langt mye mer informasjon som trengs. En betaler for mye som ikke skal brukes for estimeringene. Det som er viktigst er å ha god romlig oppløsning, og da er vanlige flyfotograferinger billigere og kan gi bedre romlig oppløsning.

HySpex passer bra for å lage modeller. Med HySpex får en mulighet til å teste alle bånd og vegetasjonsindekser i tillegg til å ha en tilfredsstillende romlig oppløsning. HySpex er optimal for å bygge opp og teste modeller. HySpex gir oss muligheten til å finne den beste måten å estimere biomasse på ved hjelp av bilder. Uten HySpex ville det vært vanskelig å si hvilke bølgelengder som er best for estimering av AGB.

Dersom en skal gjøre klassifiseringer, sykdomsanalyser og biomasseestimeringer ville jeg anbefalt å skaffe HySpex. Biomasseestimeringer passer bra å gjøre sammen med analyser som dette og HySpex hadde vært et godt datasett som kan gjøre alle analysene på en gang. Å skaffe HySpex for biomassestimeringer alene er kostbart, men om datasettet kan brukes til flere analyser er det trolig hensiktsmessig.

9.2 Utfordringene med feltdata

For å få bedre modeller trengs større feltdata. 91 trær er for lite. Modellene trenger flere trær for å gi gode resultater som ikke lider av overfit. Det bør ikke bare være mange trær, men de bør også være godt fordelt i både treslag og størrelse. Det bør være trær i alle størrelser som en urban by inneholder. Det bør også være slik at hvert treslag har både store trær og små trær. Dersom alle er store vil deres spektralinformasjon bli definisjonen på hva som

bestemmer om et tre er stort. Tre nummer 128 måtte fjernes for å forhindre dette i mine analyser.

Feltdata bør også ha mer nøyaktige høyder. Ved å se på tallene fra laserdataen til Ceballos, kan en se at nøyaktigheten er noe varierende i høyden. Feilen i høydemålingene gjør at selve fasitdataen blir noe unøyaktig. Det er vanskelig å si nøyaktig hvor mye dette påvirker kvaliteten på modellene, men det gjør nok modellene noe svakere. Jeg tenker at metoden og analysene som er blitt gjort gjennom oppgaven fungerer bra, men at datasettet med fasittrær er fot lite i tillegg til å være litt upresist.

9.3 Videre arbeid

Det er en del en kan jobbe videre med i oppgaven. Jeg synes det er spesielt to ting som er verdt å jobbe videre med. Den første er å forbedre feltdata og rekonstruere modellene på nytt. En bør skaffe flere trær og en samling av små og store trær, enkle trær og klynger. Det bør også være flere trær av hvert treslag som representeres i oppgaven. Trærne må ha ulik størrelse og det må ikke være slik at et treslag har alle de store trærne. Ved å bruke større datasett kan en få bedre funksjonsuttrykk og mer valideringsdata. Større trær gjør også at en kan garantere at modellen virker for områder med store trær. Modellene i denne oppgaven er opprinnelig laget for trær som har under 2,5 tonn biomasse og stammediameter under 0,81 meter. Når en lager et nytt datasett kan en også få bedre høyder. Høydemålingene har vært noe svake i denne oppgaven. For videre arbeid bør en skaffe et instrument eller bruke laser for å kontrollere at alle høyder er korrekt. Laser er til en viss grad brukt for å sjekke høyder i oppgaven min, men ikke nøyaktig nok. Når en har større og bedre datasett kan en validere regresjonsanalysene når en lager funksjonsuttrykk. Med å kjøres kryssvalidering eller random sampling får man et funksjonsuttrykk en kan stole mer på. I arbeidet mitt var feltarbeidet lite og det var knapt med trær igjen som kunne brukes for regresjonsanalysen, og dette førte til at funksjonsuttrykkene ble laget uten noen form for validering. Dette gjør at funksjonsuttrykkene kan ha overfit eller være dårlige uten at det vises fra de statistiske resultatene. Et forsøk på manuell validering prøves i 7.8, men dette er ikke i nærheten like sikkert som en vanlig kryssvalidering eller random sampling.

Det andre som jeg tror er verdt å jobbe videre med er automatisering. Omtrent alt i oppgaven kan automatiseres. Dette fikk jeg aldri gjort fordi planlegging, testing og forbedringer av modeller tok svært lang tid. Det optimale ville vært å lage et Python skript som tar inn en segmentert fil og gir brukeren muligheten til å velge hvilken type modell som skal brukes, og da blir svaret beregnet ut og gitt som en ny fil til brukeren. Dersom en vil gjøre den enda enklere å bruke kan teknisk sett også Python programmet gjøre segmenteringen, og da vil brukeren bare trenge å levere inn en bildefil.

En enkel sak som også burde gjøres er å bytte mengden piksler i trekronen med areal i kvadratmeter for trekronen. Dette gjør at funksjonsuttrykkene kan brukes for sensorer med hvilken som helst GSD. Dette er en enkel prosess. Alt en trenger å gjøre er å ha antall piksler og GSD, og lage en ny egenskap som er areal m^2 i CSV filen før en starter med lineær regresjon og oppbygging av funksjonsuttrykk. Resultatet vil bli det samme som om en bruker piksler. For min testing hadde det ingenting å si om jeg brukte piksler eller areal i kvadratmeter, men for videre bruk er det viktig at areal regnes ut som en SI-enhet. Siden det er såpass lite arbeid å gjøre om til areal var det lite gunstig av meg å ikke gjøre dette, og jeg hadde ikke tid til å kjøre analysen på nytt med areal i kvadratmeter.

$$Areal = GSD^2 \times pixels$$

Formel 9.3 Gjøre piksler om til areal

Hadde jeg hatt mer tid ville jeg valgt å bruke litt tid på å forske videre på bølgelengdene 1130 nm, 1152 nm og 1157 nm. Jeg rakk aldri å se på hvorfor de hadde en sammenheng med biomasse i kapittel 7.7.6 analysen min. Her kan det kanskje være noe interessant.

9.4 Kan modellene konkurrere med laser?

Ut ifra resultatene ser det ut som om 2-dimensjonal bildedata kan konkurrere med laserdata. Laser gir svakere verdier i urbane områder enn i skog. Dette er fordi det er større variasjon i treslag, stammer og trekroner. Trekronen er ofte slik at den ikke samsvarer bra med stammens størrelse. Trekronene er ofte klippet og mindre enn de ville vært naturlig. Laserdata har ingen måte å kompensere for dette siden den bare bruker høydedata og romlig informasjon. Ved å bruke den spektrale oppløsningen kan hyperspektrale data få bedre estimater enn laser. Det er fordi mengden lignin i pikslene og brattheten på red edge kan fortelle oss noe om biomassen som er nevnt i 8.2. I urbane områder tror jeg at satellitt, multispektral data og hyperspektral data kan konkurrere med laser dersom en har de rette modellene og ikke ekstreme mengder overlapp for trær.

10 Konklusjon

Ved å bruke maskinlæring og regresjonsmodeller er det mulig å bruke hyperspektral data til å estimere biomasse i urbane områder. Det er laget nøyaktige estimatmetoder basert på maskinlæring som estimerer biomasse for segmenterte trær. Metodene som er laget baserer seg på ulike egenskaper til trær. Den enkleste modellen krever bare arealet til trær og kan brukes med satellittbilder, omløpsfoto, multispektrale bilder og hyperspektrale bilder. Modellen har en R^2 på 0,656.

For HySpex: $AGB = -0{,}11805 + 0{,}00106 \times Pixels$

For hvilken som helst sensor: $AGB = -0{,}11805 + 0{,}00106 \times \frac{Areal}{0{,}09}$

Det er laget modeller som tar i bruk spektralinformasjon for å estimere biomasse mer nøyaktig. Slike modeller tar i bruk ratio vegetasjonsindekser som en kan lage fra hyperspektrale bilder. Vegetasjonsindeksene som brukes er summert med antall piksler slik at de inneholder både informasjon om arealet til trær i tillegg til spektralinformasjon. Denne modellen har en R^2 på 0,826.

For HySpex: $AGB = \frac{1{,}74 \times GRVI + 4{,}68 \times LI_{ratio} - 1{,}88 \times SR_1 - 6{,}70 \times Vogelman_1}{1000}$

Ved å analysere videre på resultater fra maskinlæring og regresjonene var det tydelig at noen areal, egenskaper, bølgelengder og vegetasjonsindekser har en korrelasjon med biomasse. Arealet er den viktigste faktoren for å estimere biomasse. Fra PLS analysen kan en se at arealet inneholder mest nyttig informasjon for å estimere biomasse. Dette gjør at den viktigste faktoren for estimering er god romlig oppløsning. Andre faktorer som korrelerte med biomasse var vegetasjonsindeksene GRVI, LI_ratio, SR1 og Vogelman1. Spesielt LI_ratio er en viktig faktor for å bestemme biomasse. Dette var tydelig i både analysene fra maskinlæring og lineære regresjoner i Origin. Høye verdier i LI_ratio korrelerte med høy biomasse. Dette gjør betyr at ligninninnhold i trær kan brukes for å definere biomasse. Vogelman1 og red edge bratthet hadde også en korrelasjon med biomasse, men denne var negativ. Høye verdier i Vogelman1 og bratt red edge kurve ga lavere biomasseverdier. Dette kommer trolig av at trær i vekstfasen er friske og har en bratt red edge, men stammen er liten og gir liten biomasse.

Modellene klarte ikke påvise om nitrogeninnhold kunne si noe om biomassen til trær. Indekser som TVI, NDVI og SAVI hadde lite betydning for biomassen.

Ifølge PLS analysen er Bølgelengdene 732 nm ,736 nm og 761 nm de mest signifikante båndene for estimering av biomasse i VNIR. Det er mulig at bølgelengdene 1130 nm, 1152 nm og 1157 nm er signifikante for estimering av biomasse i SWIR. Det trengs mer forskning på dette for å finne ut om det er tilfeldig at de båndene ble sett på som viktige i PLS analysen, eller om det er noen klare egenskaper for trær i bølgelengdene som kan være med på å definere biomassen.

Fra maskinlæringen viser det seg at å bruke alle bånd i det hyperspektrale bildet ikke er nødvendig. Det ser ut som om 7 bånd er nok for å estimere AGB i urbane områder med høy nøyaktighet.

Sammenligning med laserdata viser at hyperspektrale modellene har bedre statistiske resultater enn laser. Laser fikk en R^2 på 0,711 som er lavere enn de beste modellene for hyperspektral data. Modellene som skal kunne brukes for multispektral data og satellitter har lavere R^2 enn laser, men kan konkurrere på tilgjengelighet og pris.

Litteraturliste

Aarshay, J. (2016). *A Complete Tutorial on Ridge and Lasso Regression in Python*: Analytics Vidhya. Tilgjengelig fra: https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/#four (lest 10.05.2018).

Aarsten, D. (2018). *Rapport for Hyperspektral tillegsleveranse*. Oslo.

Bannari, A., Morin, D., Bonn, F. & Huete, A. R. (1995). A review of vegetation indices. *Remote Sensing Reviews*, 13 (1-2): 95-120. doi: 10.1080/02757259509532298.

Baret, F., Guyot, G. & Major, D. J. (1989). Crop biomass evaluation using radiometric measurements. *Photogrammetria*, 43 (5): 241-256. doi: https://doi.org/10.1016/0031-8663(89)90001-X.

Bernasconi, L., Chirici, G. & Marchetti, M. (2017). *Biomass Estimation of Xerophytic Forests Using Visible Aerial Imagery: Contrasting Single-Tree and Area-Based Approaches*, b. 9.

Burger, W. & Burge, M. J. (2016). *Digital Image Processing: An Algorithmic Introduction Using Java*: Springer Publishing Company, Incorporated.

Cassotti, M. & Grisoni, F. *Variable selection methods: an introduction*. Milano: Milano Chemometrics and QSAR Research Group - Dept. of Environmental Sciences, University of Milano-Bicocca. Tilgjengelig fra: http://math.arizona.edu/~hzhang/waeso/vsTutorial.pdf (lest 12.05).

Cerdeira, C. (2018, 19.10). *FKB og Laser - datasettene*. Satellitt, klima og miljø, Heggedal, Norge.

Cherkassky, V. & Ma, Y. (2002). *Selection of Meta-parameters for Support Vector Regression*, Berlin, Heidelberg: Springer Berlin Heidelberg.

Darwish, A., Leukert, K. & Reinhardt, W. (2003, 21-25 July 2003). *Image segmentation for the purpose of object-based classification*. IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477).

Driss Haboudane, J. R. M., Elizabeth Pattey, Pablo J Zarco-Tejada, Ian B Strachan. (2004). Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment*, 90 (3): 341-342. doi: https://doi.org/10.1016/j.rse.2003.12.013.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. (1996). *Support vector regression machines*. Proceedings of the 9th International Conference on Neural Information Processing Systems, Denver, Colorado, s. 155-161. 2999003: MIT Press.

Dvergsdal, H. (2017a). *Nevralt nettverk*: Store norske leksikon. Tilgjengelig fra: https://snl.no/nevralt_nettverk (lest 12.05).

Dvergsdal, H. (2017b). *Python: programmeringsspråk*: Store norske leksikon. Tilgjengelig fra: https://snl.no/Python_-_programmeringsspr%C3%A5k (lest 14.05).

ENVI Help. (2007). *The ENVI Header Format*: Harris Geospatial solutions,.

ESRI. (1998). *ESRI Shapefile Technical Description*.

Field Studies Council. *Carbon Cycle: Data analysis*. Tilgjengelig fra: https://www.geography-fieldwork.org/a-level/water-carbon/carbon-cycle/data-analysis/ (lest 19.03).

Forestry Commission. (2014). *Biomass in live woodland*

trees in Britain. Edinburgh: Forestry Commission,.

Forestry Commission. (2018). *What is biomass?*: Forestry Commission. Tilgjengelig fra: https://www.forestry.gov.uk/fr/beeh-9uhlqv (lest 26.01).

GISGeography. (2018). *Multispectral vs Hyperspectral Imagery Explained*. Tilgjengelig fra: https://gisgeography.com/multispectral-vs-hyperspectral-imagery-explained/ (lest 11.05).
Harris Geospatial solutions. (2014). *How to figure out Principal Component Analysis band weightings*: Harris Spatial, (lest 17.01).
Harris Geospatial solutions. (2018a). *Clump Classes*. Tilgjengelig fra: http://www.harrisgeospatial.com/docs/ClumpingClasses.html (lest 02.04).
Harris Geospatial solutions. (2018b). *ENVI*: Harris Spatial solutions,. Tilgjengelig fra: http://www.harrisgeospatial.com/SoftwareTechnology/ENVI.aspx (lest 14.05).
Harris Geospatial solutions. (2018c). *ENVI Header Files*. Tilgjengelig fra: http://www.harrisgeospatial.com/docs/enviheaderfiles.html (lest 01.05).
Harris Geospatial solutions. (2018d). *ENVI Image Files*. Tilgjengelig fra: http://www.harrisgeospatial.com/docs/enviimagefiles.html (lest 01.05).
Harris Geospatial solutions. (2018e). *Region of Interest (ROI) Tool*. Tilgjengelig fra: http://www.harrisgeospatial.com/docs/RegionOfInterestTool.html (lest 30.01).
Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, 25 (3): 295-309. doi: https://doi.org/10.1016/0034-4257(88)90106-X.
Index DataBase. (2018). *Application: Vegetation Biomass*: Index DataBase, (lest 01.02).
Jenkins, T. A. R., Mackie, E. D., Matthews, R. W., Miller, G., Randle, T. J. & White, M. E. (2011). *FC Woodland Carbon Code: Carbon Assessment Protocol*: Forestry Commission, .
Jensen, R., J. Hardin, P. & J. Hardin, A. (2012). *Estimating Urban Leaf Area Index (LAI) of Individual Trees with Hyperspectral Data*, b. 78.
Jonassen, V. & Aarsten, D. (2017, 19.10). *Hyperspektrale bildedata og multispektrale lasertdata*. Satelitt, klima og miljø, Heggedal.
Jonassen, V. (2018). *Møte og diskusjon om bruk av atmosfærekorrigerte bilder, normalisering og PCA for HySpex bilder*.
Liu, W., Gao, W., Gao, Z. & Wang, X. (2006). *Correlation analysis between the biomass of oasis ecosystem and the vegetation index at Fukang - art. no. 62982M*.
Marklund, L. G. (1987). *Biomassafunktioner för tall, gran och björk i Sverige = Biomass functions for pine, spruce and birch in Sweden*. Rapport (Sveriges lantbruksuniversitetet. Institutionen för skogstaxering), b. 45. Umeå.
MathWorks. (2018). *What Is Machine Learning?* Tilgjengelig fra: https://se.mathworks.com/discovery/machine-learning.html (lest 10.05).
Mendenhall, W. & Sincich, T. (1997). *A Second Course in Statistics: Regression Analysis*, b. 92.
Nelson, R., Krabill, W. & Tonelli, J. (1988). Estimating forest biomass and volume using airborne laser data. *Remote Sensing of Environment*, 24 (2): 247-267. doi: https://doi.org/10.1016/0034-4257(88)90028-4.
NEO. (2018). *HySpex Main Specifications*. Tilgjengelig fra: https://www.hyspex.no/products/all_specs.php (lest 21.01.2018).
Orange. (2018). *Orange*: Orange. Tilgjengelig fra: https://orange.biolab.si/ (lest 15.01).
Orange Documentation. (2015a). *Linear Regression*. Tilgjengelig fra: https://docs.orange.biolab.si/3/visual-programming/widgets/model/linearregression.html#linear-regression (lest 11.05).
Orange Documentation. (2015b). *Random Forest*. Tilgjengelig fra: https://docs.orange.biolab.si/3/visual-programming/widgets/model/randomforest.html (lest 02.10).

Orange Documentation. (2015c). *SVM*: Orange,. Tilgjengelig fra: https://docs.orange.biolab.si/3/visual-programming/widgets/model/svm.html (lest 10.02).
Orange Documentation. (2015d). *Test & Score*. Tilgjengelig fra: https://docs.orange.biolab.si/3/visual-programming/widgets/evaluation/testandscore.html (lest 11.05).
Originlab Documentation. *Algorithm (Multiple Linear Regression)*. Tilgjengelig fra: https://www.originlab.com/doc/Origin-Help/Multi-Regression-Algorithm (lest 20.02).
Ouyang, Z. (2015). *Object-Based Classification & eCognition*. eCognition, O.-B. C. (red.).
Popescu, S. C. (2007). Estimating biomass of individual pine trees using airborne lidar. *Biomass and Bioenergy*, 31 (9): 646-655. doi: https://doi.org/10.1016/j.biombioe.2007.06.022.
Randle, T., Matthews, R. & Jenkins, T. (2011). *Technical Specification for the Biomass Equations Developed for the 2011 Forecast*: The Research Agency of the Forestry Commission,.
Richter, R. & Schläpfer, D. (2016). *Atmospheric / Topographic Correction fior Airborne Imagery*.
Rodarmel, C. & Shan, J. (2002). *Principal Component Analysis for Hyperspectral Image Classification*, b. 62.
Sayad, S. *Support Vector Machine - Regression (SVR)*. Tilgjengelig fra: http://www.saedsayad.com/support_vector_machine_reg.htm (lest 05.05).
Scikit-learn. (2017). *Neural network models (supervised)*. Tilgjengelig fra: http://scikit-learn.org/stable/modules/neural_networks_supervised.html (lest 13.05).
Serrano, L., Peñuelas, J. & Ustin, S. L. (2002). Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data: Decomposing biochemical from structural signals. *Remote Sensing of Environment*, 81 (2): 355-364. doi: https://doi.org/10.1016/S0034-4257(02)00011-1.
Shafranovich, Y. (2005). *Common Format and MIME Type for Comma-Separated Values (CSV) Files*: SolidMatrix Technologies, Inc. Tilgjengelig fra: https://tools.ietf.org/html/rfc4180 (lest 12.05).
Shapiro & Stockman. (2001). *Computer Vision*. 1 edition utg.: Pearson.
Shibayama, M., Salli, A., Häme, T., Iso-Iivari, L., Heino, S., Alanen, M., Morinaga, S., Inoue, Y. & Akiyama, T. (1999). Detecting Phenophases of Subarctic Shrub Canopies by Using Automated Reflectance Measurements. *Remote Sensing of Environment*, 67 (2): 160-180. doi: https://doi.org/10.1016/S0034-4257(98)00082-0.
Sripada, R. P., Heiniger, R. W., White, J. G. & Meijer, A. D. (2006). Aerial Color Infrared Photography for Determining Early In-Season Nitrogen Requirements in Corn This project was supported in part by Initiative for Future Agriculture and Food Systems Grant no. 00-52103-9644 from the USDA Cooperative State Research, Education, and Extension Service. *Agronomy Journal*, 98 (4): 968-977. doi: 10.2134/agronj2005.0200.
Thuy, M. (2012). *What are passive and active sensors?*: NASA. Tilgjengelig fra: https://www.nasa.gov/directorates/heo/scan/communications/outreach/funfacts/txt_passive_active.html (lest 01.05).
University of Pennsylvania. (2018). *Quantum GIS: What is GIS?* Tilgjengelig fra: https://guides.library.upenn.edu/c.php?g=475976&p=3255387 (lest 14.05).
Urban EEA vegetation survey sample summer 2017. (2017). *Skarpås, Olav.*: NINA (Norsk institutt for naturforskning),.
USGS. (2011). *Sensors - Hyperion*. Tilgjengelig fra: https://eo1.usgs.gov/sensors/hyperion (lest 01.05).

Vogelmann, J. E., Rock, B. N. & Moss, D. M. (1993). Red edge spectral measurements from sugar maple leaves. *International Journal of Remote Sensing*, 14 (8): 1563-1575. doi: 10.1080/01431169308953986.

Yu, B., Ostland, M., Gong, P. & Pu, R. (1999). Penalized discriminant analysis of in situ hyperspectral data for conifer species recognition. *IEEE Transactions on Geoscience and Remote Sensing*, 37 (5): 2569-2577. doi: 10.1109/36.789651.

Vedlegg A

Beregninger i rasterkalkulatoren til QGIS.

Radius stamme:

CASE

WHEN "stamme" IS NOT NULL OR "stamme" != 0 THEN "stamme"/(2*pi())

END

Diameter stamme:

"stamme_rad"*2

Volum av stammen:

Pi()*"Stamme_rad"^2*(hoyde/3)

Biomassen til stammen:

CASE

WHEN "treslag" = 'Blodbok' THEN "stamme_vol"*0.55

WHEN "treslag" = 'Bok' THEN "stamme_vol"*0.55

WHEN "treslag" = 'Bjork' THEN "stamme_vol"*0.53

WHEN "treslag" = 'Soyleagnbok' THEN "stamme_vol"*0.53

WHEN "treslag" = 'Hengebjork' THEN "stamme_vol"*0.53

WHEN "treslag" = 'Or' THEN "stamme_vol"*0.42

WHEN "treslag" = 'Pil' THEN "stamme_vol"*0.49

WHEN "treslag" = 'Lind' THEN "stamme_vol"*0.44

WHEN "treslag" = 'Lonn' THEN "stamme_vol"*0.49

WHEN "treslag" = 'Syrin' THEN "stamme_vol"*0.49

WHEN "treslag" = 'Bartre' THEN "stamme_vol"*0.39

WHEN "treslag" = 'Furu' THEN "stamme_vol"*0.42

WHEN "treslag" = 'Gran' THEN "stamme_vol"*0.33

WHEN "treslag" = 'Eik' THEN "stamme_vol"*0.56

```
 WHEN "treslag" = 'Osp' THEN "stamme_vol"*0.35
 WHEN "treslag" = 'Ask' THEN "stamme_vol"*0.53
 WHEN "treslag" = 'Kastanje' THEN "stamme_vol"*0.44
 WHEN "treslag" = 'Alm' THEN "stamme_vol"*0.43
 WHEN "treslag" = 'Tuja' THEN "stamme_vol"*0.33
 WHEN "treslag" = 'Soyletuja' THEN "stamme_vol"*0.33
 WHEN "treslag" = 'Graor' THEN "stamme_vol"*0.42
WHEN "treslag" = 'Hegg' THEN "stamme_vol"*0.50
WHEN "treslag" IS NULL THEN "stamme_vol"*0
 ELSE "stamme_vol" * 0.49
END
```

Konstant A:

```
CASE
 WHEN "treslag" = 'Blodbok' THEN 0.000017
 WHEN "treslag" = 'Bok' THEN 0.000017
 WHEN "treslag" = 'Bjork' THEN 0.000019
 WHEN "treslag" = 'Soyleagnbok' THEN 0.000019
 WHEN "treslag" = 'Hengebjork' THEN 0.000019
 WHEN "treslag" = 'Or' THEN 0.000017
 WHEN "treslag" = 'Pil' THEN 0.000017
 WHEN "treslag" = 'Lind' THEN 0.000017
 WHEN "treslag" = 'Lonn' THEN 0.000019
 WHEN "treslag" = 'Syrin' THEN 0.000017
 WHEN "treslag" = 'Bartre' THEN 0.000015
 WHEN "treslag" = 'Furu' THEN 0.000016
 WHEN "treslag" = 'Gran' THEN 0.000015
 WHEN "treslag" = 'Eik' THEN 0.000017
 WHEN "treslag" = 'Osp' THEN 0.000017
 WHEN "treslag" = 'Ask' THEN 0.000017
```

```
  WHEN "treslag" = 'Kastanje' THEN 0.000017
  WHEN "treslag" = 'Alm' THEN 0.000017
  WHEN "treslag" = 'Tuja' THEN 0.000015
  WHEN "treslag" = 'Soyletuja' THEN 0.000015
  WHEN "treslag" = 'Graor' THEN 0.000017
 WHEN "treslag" = 'Hegg' THEN 0.000017
 WHEN "treslag" IS NULL THEN 0
  ELSE 0.000017
END
```

Konstant B:

```
CASE
  WHEN "treslag" = 'Lerke' THEN 2.0291
  ELSE 2.4767
END
```

Konstant C:

```
CASE
  WHEN "treslag" = 'Blodbok' THEN -0.411551
  WHEN "treslag" = 'Bok' THEN -0.459519
  WHEN "treslag" = 'Bjork' THEN -0.459519
  WHEN "treslag" = 'Soyleagnbok' THEN -0.459519
  WHEN "treslag" = 'Hengebjork' THEN -0.411551
  WHEN "treslag" = 'Or' THEN -0.411551
  WHEN "treslag" = 'Pil' THEN -0.411551
  WHEN "treslag" = 'Lind' THEN -0.411551
  WHEN "treslag" = 'Lonn' THEN -0.459519
  WHEN "treslag" = 'Syrin' THEN -0.411551
  WHEN "treslag" = 'Bartre' THEN -0.353198
```

```
 WHEN "treslag" = 'Furu' THEN -0.394206
 WHEN "treslag" = 'Gran' THEN -0.353198
 WHEN "treslag" = 'Eik' THEN -0.411551
 WHEN "treslag" = 'Osp' THEN -0.411551
 WHEN "treslag" = 'Ask' THEN -0.411551
 WHEN "treslag" = 'Kastanje' THEN -0.411551
 WHEN "treslag" = 'Alm' THEN -0.411551
 WHEN "treslag" = 'Tuja' THEN -0.353198
 WHEN "treslag" = 'Soyletuja' THEN -0.353198
WHEN "treslag" = 'Graor' THEN -0.411551
WHEN "treslag" = 'Hegg' THEN -0.411551
WHEN "treslag" IS NULL THEN 0
 ELSE -0.411551
END
```

Konstant D:

```
CASE
 WHEN "treslag" = 'Blodbok' THEN 0.013670
 WHEN "treslag" = 'Bok' THEN 0.013670
 WHEN "treslag" = 'Bjork' THEN 0.015263
 WHEN "treslag" = 'Hengebjork' THEN 0.015263
 WHEN "treslag" = 'Soyleagnbok' THEN 0.015263
 WHEN "treslag" = 'Or' THEN 0.013670
 WHEN "treslag" = 'Pil' THEN 0.013670
 WHEN "treslag" = 'Lind' THEN 0.013670
 WHEN "treslag" = 'Lonn' THEN 0.015263
 WHEN "treslag" = 'Syrin' THEN 0.013670
 WHEN "treslag" = 'Bartre' THEN 0.011732
 WHEN "treslag" = 'Furu' THEN 0.013094
 WHEN "treslag" = 'Gran' THEN 0.011732
 WHEN "treslag" = 'Eik' THEN 0.013670
```

WHEN "treslag" = 'Osp' THEN 0.013670

WHEN "treslag" = 'Ask' THEN 0.013670

WHEN "treslag" = 'Kastanje' THEN 0.013670

WHEN "treslag" = 'Alm' THEN 0.013670

WHEN "treslag" = 'Tuja' THEN 0.011732

WHEN "treslag" = 'Soyletuja' THEN 0.011732

WHEN "treslag" = 'Graor' THEN 0.013670

WHEN "treslag" = 'Hegg' THEN 0.013670

WHEN "treslag" IS NULL THEN 0

ELSE 0.013670

END

Biomasse krone:

CASE

WHEN "stamme_dia" > 0.5 THEN "c" + ("d" * "stamme_dia" * 100)

ELSE "a" * ("stamme_dia"*100)^"b"

END

Konstant E:

WHEN "treslag" = 'Blodbok' THEN 0.000023

WHEN "treslag" = 'Bok' THEN 0.000023

WHEN "treslag" = 'Bjork' THEN 0.000023

WHEN "treslag" = 'Hengebjork' THEN 0.000023

WHEN "treslag" = 'Soyleagnbok' THEN 0.000023

WHEN "treslag" = 'Or' THEN 0.000023

WHEN "treslag" = 'Pil' THEN 0.000023

WHEN "treslag" = 'Lind' THEN 0.000023

WHEN "treslag" = 'Lonn' THEN 0.000023

WHEN "treslag" = 'Syrin' THEN 0.000023

WHEN "treslag" = 'Bartre' THEN 0.000012

```
  WHEN "treslag" = 'Furu' THEN 0.000015
  WHEN "treslag" = 'Gran' THEN 0.000012
  WHEN "treslag" = 'Eik' THEN 0.000023
  WHEN "treslag" = 'Osp' THEN 0.000023
  WHEN "treslag" = 'Ask' THEN 0.000023
  WHEN "treslag" = 'Kastanje' THEN 0.000023
  WHEN "treslag" = 'Alm' THEN 0.000023
  WHEN "treslag" = 'Tuja' THEN 0.000012
  WHEN "treslag" = 'Soyletuja' THEN 0.000012
  WHEN "treslag" = 'Graor' THEN 0.000023
  WHEN "treslag" = 'Hegg' THEN 0.000023
  WHEN "treslag" IS NULL THEN 0
  ELSE 0.000023
END
```

Konstant F:

```
CASE
  WHEN "treslag" = 'Blodbok' THEN -0.174882
  WHEN "treslag" = 'Bok' THEN -0.174882
  WHEN "treslag" = 'Bjork' THEN -0.174882
  WHEN "treslag" = 'Hengebjork' THEN -0.174882
  WHEN "treslag" = 'Or' THEN -0.174882
  WHEN "treslag" = 'Pil' THEN -0.174882
  WHEN "treslag" = 'Lind' THEN -0.174882
  WHEN "treslag" = 'Lonn' THEN -0.174882
  WHEN "treslag" = 'Syrin' THEN -0.174882
  WHEN "treslag" = 'Bartre' THEN -0.091547
  WHEN "treslag" = 'Furu' THEN -0.118673
  WHEN "treslag" = 'Gran' THEN -0.091547
  WHEN "treslag" = 'Eik' THEN -0.174882
```

WHEN "treslag" = 'Osp' THEN -0.174882

WHEN "treslag" = 'Ask' THEN -0.174882

WHEN "treslag" = 'Kastanje' THEN -0.174882

WHEN "treslag" = 'Alm' THEN -0.174882

WHEN "treslag" = 'Tuja' THEN -0.091547

WHEN "treslag" = 'Soyletuja' THEN -0.091547

WHEN "treslag" = 'Graor' THEN -0.174882

WHEN "treslag" = 'Hegg' THEN -0.174882

WHEN "treslag" IS NULL THEN 0

ELSE -0.174882

END

Konstant G:

CASE

WHEN "treslag" = 'Blodbok' THEN 0.009559

WHEN "treslag" = 'Bok' THEN 0.009559

WHEN "treslag" = 'Bjork' THEN 0.009559

WHEN "treslag" = 'Hengebjork' THEN 0.009559

WHEN "treslag" = 'Or' THEN 0.009559

WHEN "treslag" = 'Pil' THEN 0.009559

WHEN "treslag" = 'Lind' THEN 0.009559

WHEN "treslag" = 'Lonn' THEN 0.009559

WHEN "treslag" = 'Syrin' THEN 0.009559

WHEN "treslag" = 'Bartre' THEN 0.005004

WHEN "treslag" = 'Furu' THEN 0.006487

WHEN "treslag" = 'Gran' THEN 0.005004

WHEN "treslag" = 'Eik' THEN 0.009559

WHEN "treslag" = 'Osp' THEN 0.009559

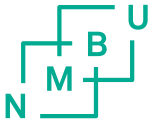
WHEN "treslag" = 'Ask' THEN 0.009559

WHEN "treslag" = 'Kastanje' THEN 0.009559

```
 WHEN "treslag" = 'Alm' THEN 0.009559
 WHEN "treslag" = 'Tuja' THEN 0.005004
 WHEN "treslag" = 'Soyletuja' THEN 0.005004
 WHEN "treslag" = 'Graor' THEN 0.009559
 WHEN "treslag" = 'Hegg' THEN 0.009559
 WHEN "treslag" IS NULL THEN 0
 ELSE 0.009559
END
```

Biomasse røtter:

```
CASE
        WHEN "stamme_dia" > 0.5 THEN "f" + ("g" * "stamme_dia"*100)
        ELSE "e" * ("stamme_dia"*100)^2.5
END
```



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway