

Notat: Efterafgrøder set fra himlen

31-05-2022

Philipp Trénel, ph.d., seniorkonsulent, faglig leder
phtr@teknologisk.dk, +45 72 20 33 92

Formål

- At udvikle en statistisk funktionel dataanalyse (FDA) model til bestemmelse af biomasse og kvælstofindhold (N) i efterafgrøder ud fra Sentinel 2 satellitdata.

Overordnede konklusioner

- Ved at anvende en funktionel oprensning og modellering af satellitdata vha. machine learning (ML) kan satellitdata alene forklare 52% hhv. 55% af variationen (R^2) i høstet efterafgrødebiomasse og N-optag.
- Dette svarer til en prædiktionsusikkerhed på 24.9% og 27.2% af den prædikterede værdi af hhv. biomasse og N-optag. Dvs. prædikterer modellen 25 kg N/ha er fejlmarginen fra 18.2 til 31.8 kgN/ha.
- Den mest betydende feature ved NDVI-vækstkurven er for begge responsvariabler arealet under vækstkurven.
- Dette indikerer, at det er vigtigt at modellere satellitdata som funktionelle data (=som vækstkurver) i stedet for at fokusere på en enkelt satellitobservation fra en mere eller mindre tilfældig dato. Satellitdata er støjfyldte (pga. skydække og atmosfærisk interferens) og forekommer på et irregulært tidsgrid. Det er derfor vigtig med en oprensings- og udglætningsprocedure før en eventuel modellering til prædiktions af f.eks. biomasse og N-optag.
- Der er undersøgt, om modellens prædiktionssevne kan forbedres ved at tage højde for en række yderligere informationer, såsom økologisk drift (M2), efterafgrødetype (M3), dyrkningsår (M4) og landsdel (M5).
- Prædiktionssevnen steg til $R^2 = 0.58$ og $R^2 = 0.61$ for hhv. biomasse og N-optag ved inklusion af efterafgrødetype (M3), resulterende i en mindsket prædiktionsusikkerhed (21.4% og 22.9%). Dvs. prædikterer modellen 25 kg N/ha er fejlmarginen nu fra 19.3 til 30.7 kgN/ha.
- Prædiktionssevnen blev kun marginalt forbedret ved yderligere inkludering af høstår og landsdel, mens økologisk dyrkningsform ja/nej bidrog med ingen forbedring og kan dermed betragtes til at være ubetydeligt.
- Det er også undersøgt, om modellernes prædiktionssevne kan forventes at variere fra efterafgrøde til efterafgrøde. Der er fundet, at modellens forudsigelser er bedst for Kløvergræs (usikkerhed $N_{\text{-optag}} = 16.6\%$) efterfulgt af Korsblomstrede (usikkerhed $N_{\text{-optag}} = 20.4\%$), men det understreges at sample størrelsen er for lille til en egentlig vurdering af dette.

Metoder

- Analyser og modeludvikling i indeværende notat er baseret på metoderne beskrevet i bilag1 ("Notat - Satellit-bestemt N i efterafgrøder") fra den 15.12.2021.

Data

- Data er vedhæftet som excelarket "Data_Satellitbestemt_N_i_efterafgroeder_2019_til_2021.xlsx", se bilag 2. Data er sammenstillet ud fra data beskrevet i "Notat - Satellit-bestemt N i efterafgrøder" og data fra projektet Efterafgrøder set fra himlen, NFTS forsøgsserierne 020411919, 020412020 og 020412121.

Datahåndtering

- Step1: Satellitdata blev rensed jf. procedure beskrevet i bilag 1, nemlig vha. piece-wise linear models (PLM). Dette fjernede formodede NDVI-outlier i satellitdata pga. skydække eller andre forhold, der forstyrrer NDVI målingen.
- Step 2: Derefter blev de rensede satellitdata modelleret i en Functional Principal Component Analysis (fPCA) som beskrevet i bilag 1, hvorved der for hver observationsfelt i marken fremkommer en glat NDVI-vækstkurve over et fast dato-grid.
- Step 3: Der blev beregnet såkaldte *features* ud fra de estimerede NDVI-vækstkurver, såsom arealet under vækstkurven (*AUC*), den maksimalt opnåede NDVI, dato for den maksimale NDVI og hældningen for en lineær vækst fra 1.9. til den maksimale NDVI (*growth*)
- Biomasse og N-optag blev log-transformeret inden modellering.
- Ud af de oprindelige 320 observationsfelter/marker passerede 233 (73%) oprensings- og udglætningsproceduren og leverede valide NDVI-vækstkurve features til modellering.
- Disse fordelte sig som følger:
 - o Efterafgrødetype: Klgr = 36, Korsbl = 150, Korn = 34 og H.urt = 13
 - o Dyrkningsår: 2019 = 90, 2020 = 108, 2021 = 35
 - o Landsdel: ØJ = 89, NJ = 16, Thy = 10, VJ = 17, MJ = 17, Fyn = 49, Sj. = 25, B. = 10.
 - o Økologisk dyrkning: ja = 60, nej = 173.

Model

- Den log-transformerede biomasse og N-optag blev modelleret som funktion af de beregnede NDVI-vækstkurve features i en random forest maschine learning model med standard default indstillinger (model M1).
- Modellen blev derefter skridtvist udvidet med nye variabler i følgende rækkefølge: Økologisk dyrkning (M2), afgrødetype (M3, se bilag 1), dyrknings-år (M4) og landsdel (M5). Sidstnævnte to er indikatorer for de klimatiske betingelser, som efterafgrøden er dyrket under.
- Alle analyser er udført i R, specielt med pakkerne `segmented` (PLM), `fdapace` (fPCA) og `randomForestSRC`.

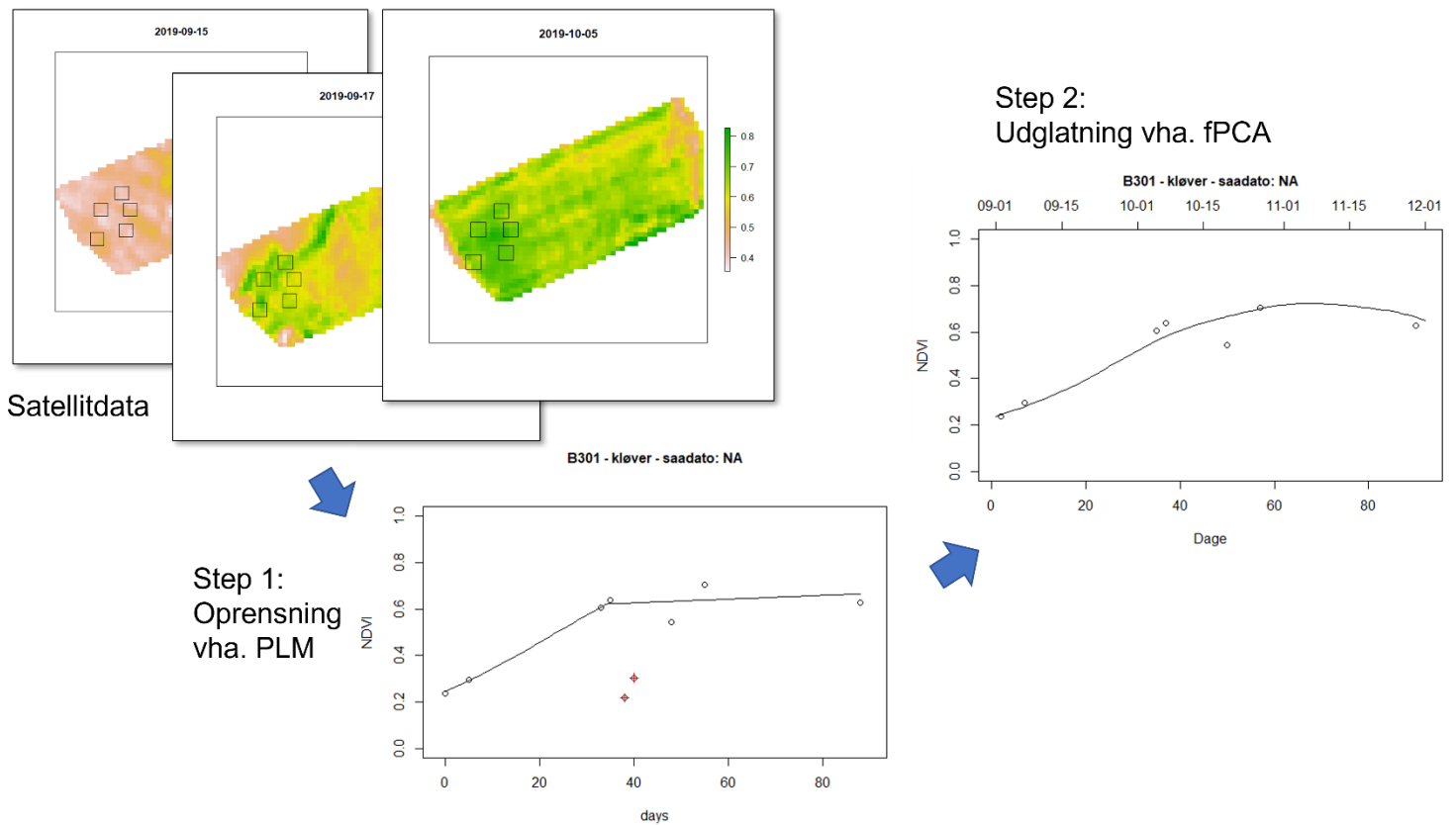
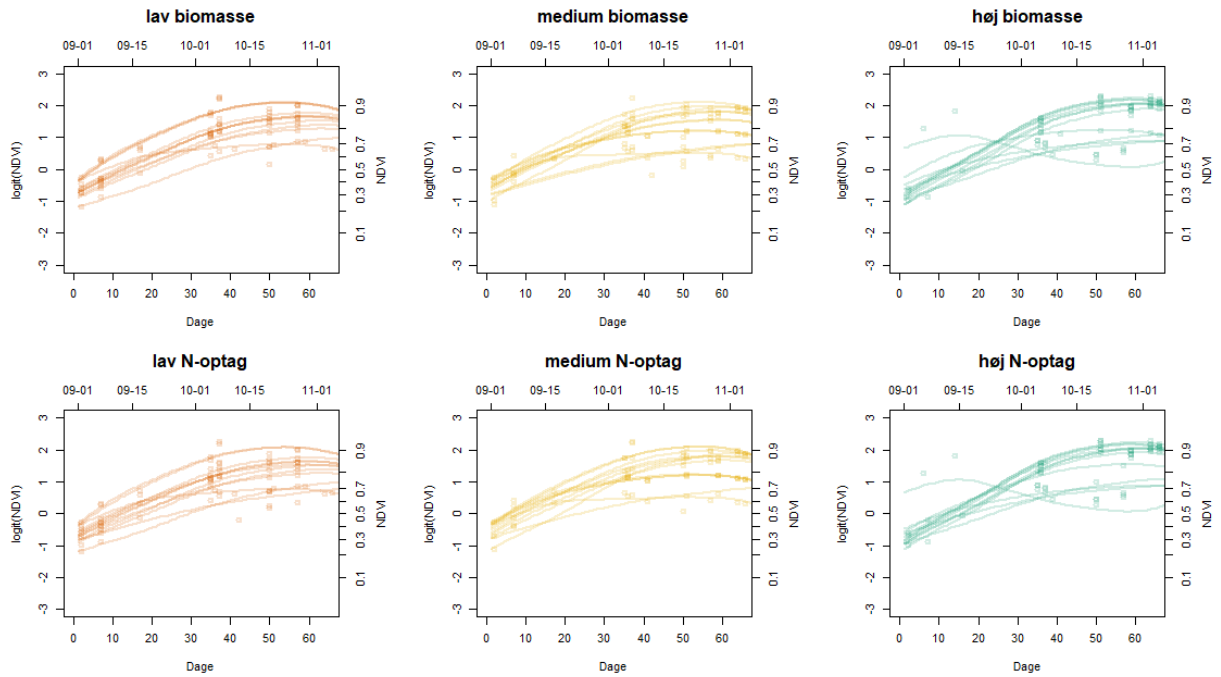


Fig. 1: Visualisering af den anvendte procedure.

Kløvergræs



Korsblomstret

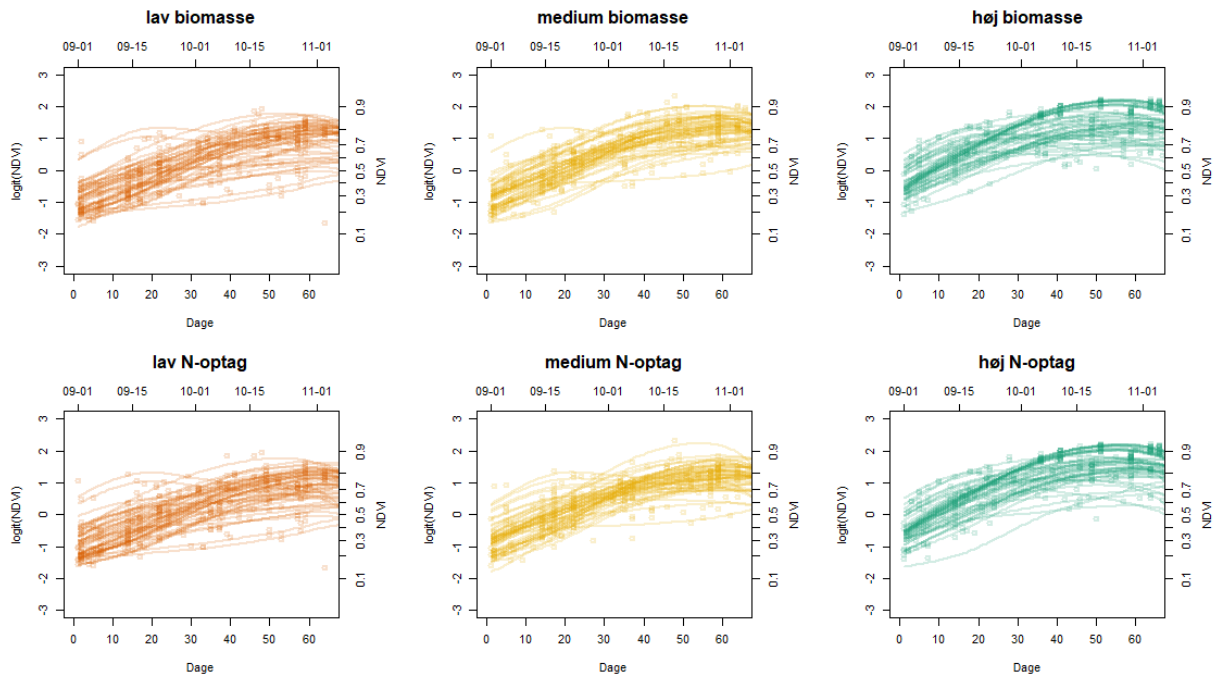


Fig. 2: NDVI-vækstkurver for efterafgrødetyperne kløvergræs og korsblomstrede grupperet efter den laveste, midterste og højeste tredjedel af observationerne i forhold til hhv. biomasse og N-optag. Det ses, at det er svært at se med det blotte øje, at NDVI-vækstkurverne varierer, om end muligvis lidt højere NDVI maksima for de højt-ydende korsblomstrede efterafgrøder.

Resultater

- Resultater fremgår af Tabel 1 og Fig. 3 her i notatet.

Tabel 1: Resultater fra modellerne M1 (kun information fra satellitdata NDVI) til M5. Usikkerheden er en (out-of-bag test-set) prædiktionsfejl angivet i procent af den prædikterede værdi. Variable importance er udtrykt som en partiel forklaringsgrad (η^2). Summen af η^2 giver R^2 .

Model		Biomasse			N-optag		
		Usikkerhed (%)	R^2	Variable Importance (η^2)	Usikkerhed (%)	R^2	Variable Importance (η^2)
M1	NDVI-vækstkurve features	24.9	0.52	NDVI_auc (0.25), NDVI_0 (0.11), NDVI_maxday (0.06), NDVI_growth_pr_week (0.05), NDVI_max (0.04)	27.2	0.55	NDVI_auc (0.28), NDVI_maxday (0.07), NDVI_growth_pr_week (0.07), NDVI_0 (0.07), NDVI_max (0.06)
M2	+øko	24.5	0.53	NDVI_auc (0.21), NDVI_0 (0.13), NDVI_maxday (0.06), NDVI_max (0.05), NDVI_growth_pr_week (0.05), Eco (0.03)	26.8	0.55	NDVI_auc (0.24), NDVI_0 (0.08), NDVI_max (0.07), NDVI_maxday (0.07), NDVI_growth_pr_week (0.06), Eco (0.03)
M3	+afgr.	21.4	0.58	NDVI_auc (0.17), Afgrodekate_phtr (0.14), NDVI_0 (0.11), NDVI_maxday (0.07), NDVI_growth_pr_week (0.05), NDVI_max (0.04)	22.9	0.61	NDVI_auc (0.21), Afgrodekate_phtr (0.13), NDVI_maxday (0.08), NDVI_0 (0.07), NDVI_max (0.06), NDVI_growth_pr_week (0.06)
M4	+afgr. +år	20.3	0.60	NDVI_auc (0.16), Afgrodekate_phtr (0.13), NDVI_0 (0.1), NDVI_maxday (0.07), Aar (0.07), NDVI_growth_pr_week (0.04), NDVI_max (0.04)	22.5	0.62	NDVI_auc (0.19), Afgrodekate_phtr (0.12), Aar (0.08), NDVI_maxday (0.07), NDVI_max (0.06), NDVI_0 (0.05), NDVI_growth_pr_week (0.05)
M5	+afgr. +år +landsdel	19.8	0.61	NDVI_auc (0.14), Afgrodekate_phtr (0.12), NDVI_0 (0.1), Aar (0.07), NDVI_maxday (0.07), NDVI_max (0.04), Landsdel (0.04), NDVI_growth_pr_week (0.03)	21.9	0.63	NDVI_auc (0.17), Afgrodekate_phtr (0.12), Aar (0.08), NDVI_maxday (0.07), NDVI_max (0.06), NDVI_0 (0.06), NDVI_growth_pr_week (0.05), Landsdel (0.03)

Tabel 2: Resultater fra model M1 kørt for hver afgrøde særskilt

Model M1	Biomasse		N-optag	
	Usikkerhed (%)	R^2	Usikkerhed (%)	R^2
Kløvergræs	15.5	0.29	16.7	0.26
Korsblomstrede	18.6	0.46	20.4	0.56
Korn	39.6	0.39	35.9	0.38
Honningurt	46.2	0.00	35.7	0.00

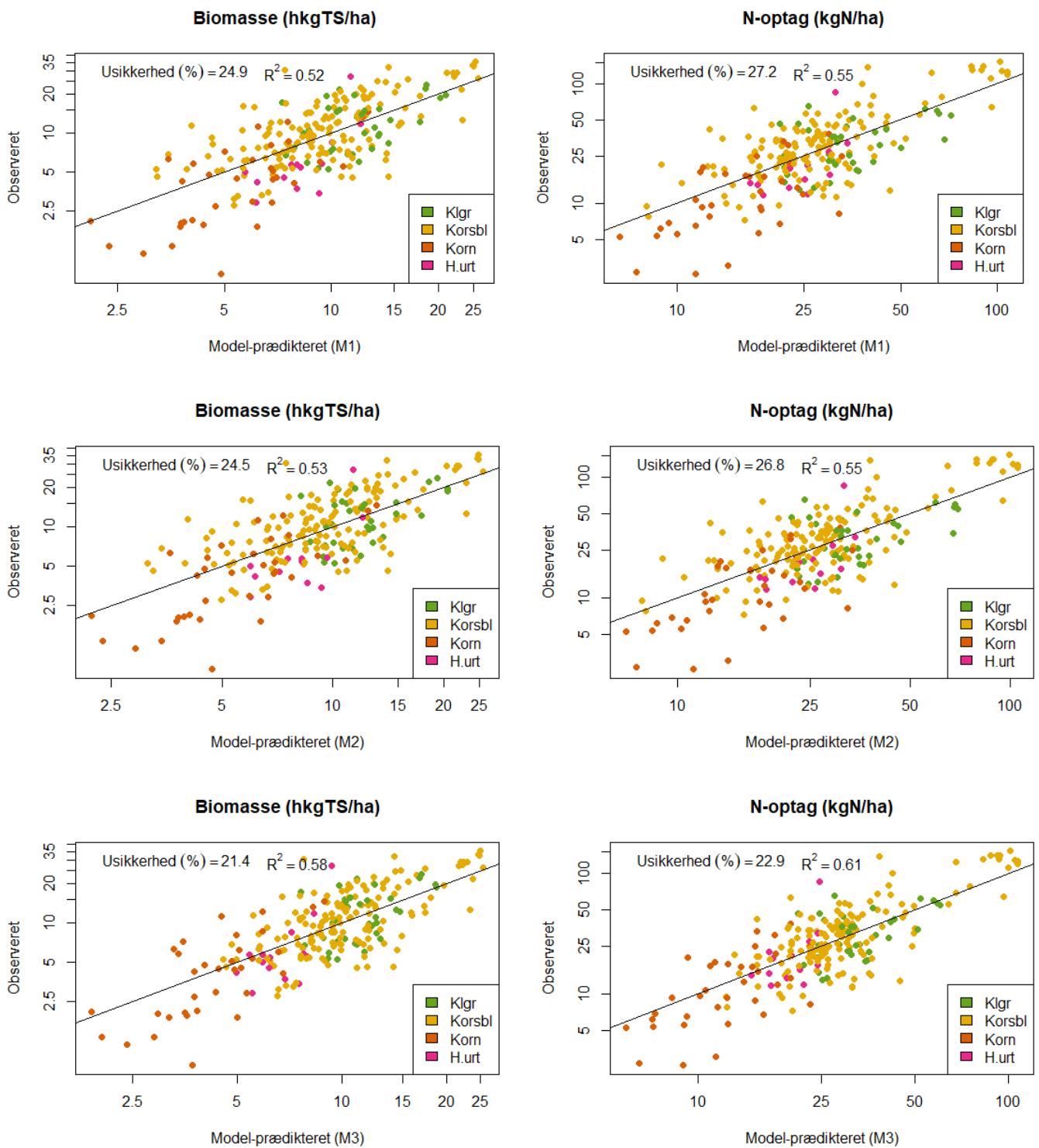


Fig. 3a: Observeret over prædikeret for modellerne M1 til M3, se tabel 1.

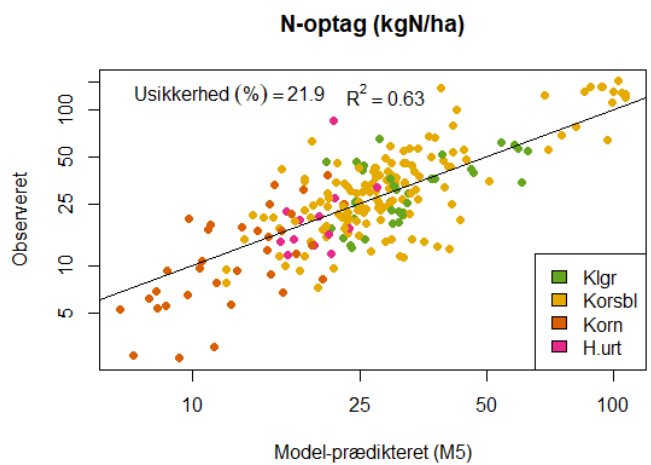
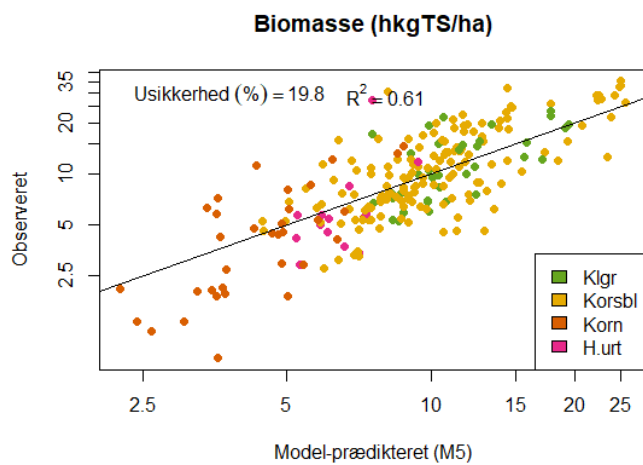
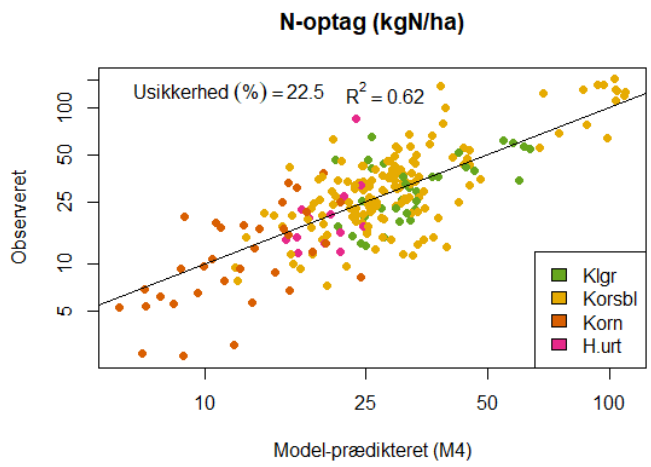
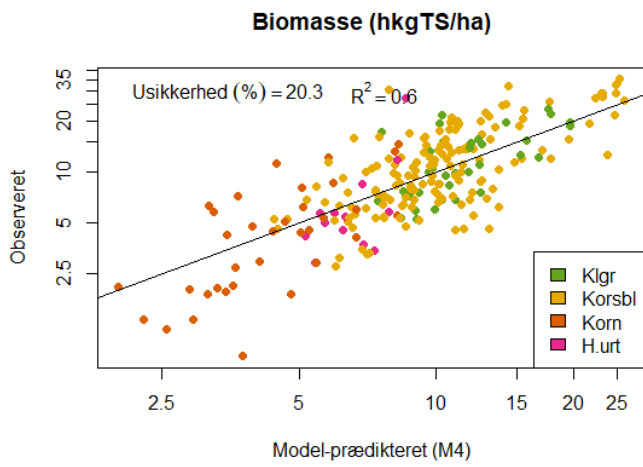


Fig. 3b: Observeret over prædikeret for modellerne M4 og M5, se tabel 1.

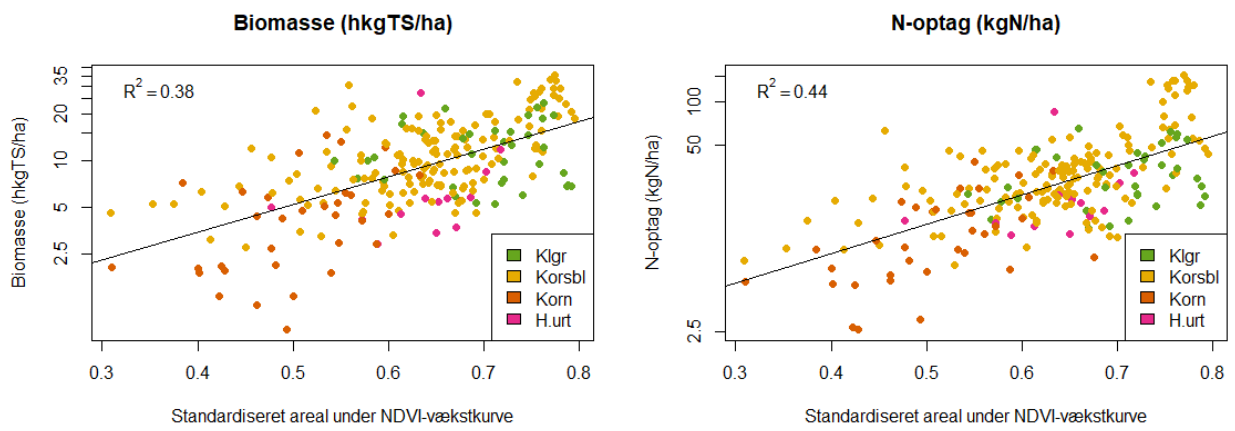


Fig. 4: Biomasse og N-optag som funktion af standardiseret areal under NDVI-vækstkurven.

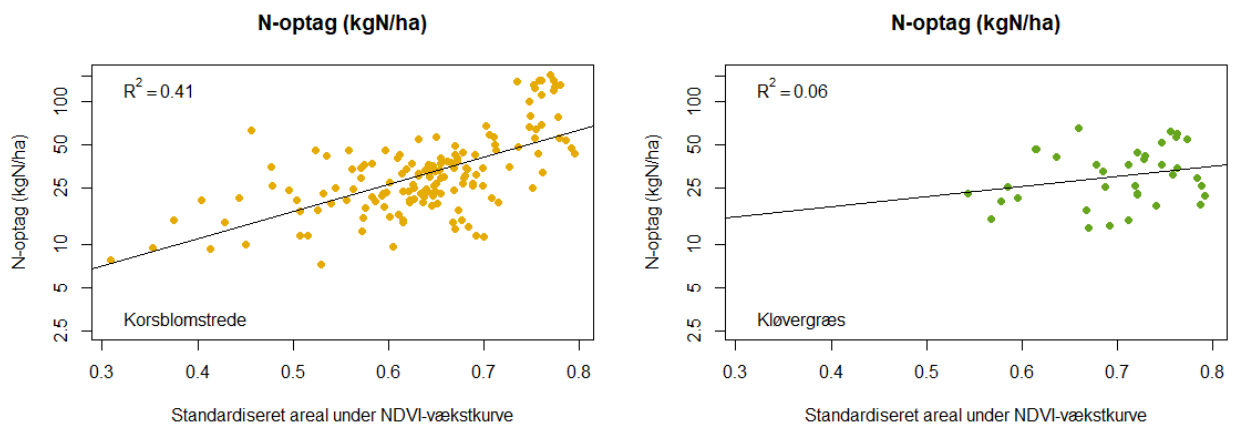


Fig. 5: som Fig 4., men for kun korsblomstrede og kløvergræs.

Referencer

Cody Carroll, Alvaro Gajardo, Yaqing Chen, Xiongtao Dai, Jianing Fan, Pantelis Z. Hadjipantelis, Kyunghee Han, Hao Ji, Hans-Georg Mueller and Jane-Ling Wang (2020). fdapace: Functional Data Analysis and Empirical Dynamics. R package version 0.5.5.
<https://CRAN.R-project.org/package=fdapace>

Jeff Goldsmith, Fabian Scheipl, Lei Huang, Julia Wrobel, Chongzhi Di, Jonathan Gellar, Jaroslaw Harezlak, Mathew W. McLean, Bruce Swihart, Luo Xiao, Ciprian Crainiceanu and Philip T. Reiss (2020). refund: Regression with Functional Data. R package version 0.1-23.
<https://CRAN.R-project.org/package=refund>

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Vito M. R. Muggeo (2017). Interval estimation for the breakpoint in segmented regression: a smoothed score-based approach. Australian & New Zealand Journal of Statistics, 59, 311-322.