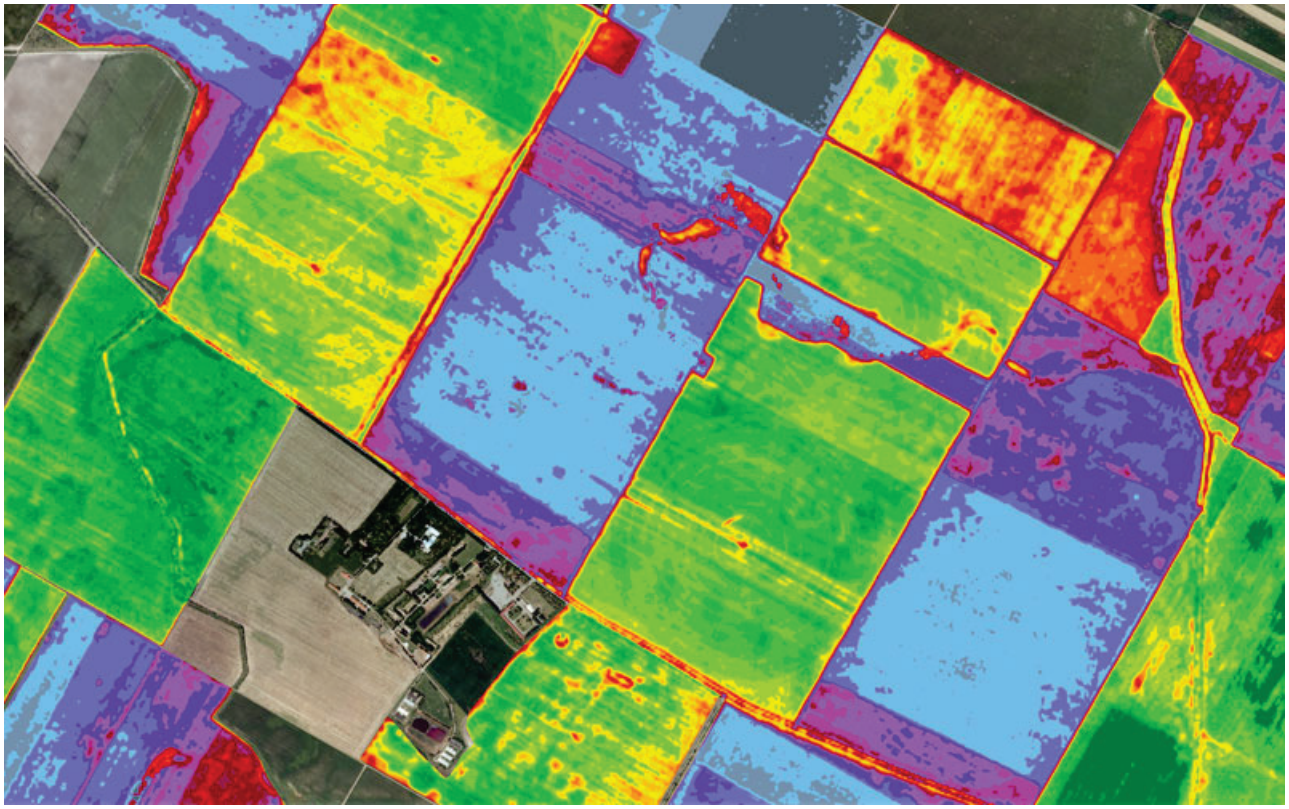
NYE MODELLER BASERET PÅ **SATTELITBASEREDE BIOMASSEMÅLINGER**

SEGES

Nye modeller baseret på satellitbaserede biomassemålinger

Remote Sensing-baserede dataprodukter har åbnet for nye muligheder inden for overvågning og forudsigelse af planteproduktion. De seneste år har budt på en eksplosion af datakilder fra både satellitter og droner, og fælles for dem er, at de hver især tillader et indblik i afgrødens udvikling både rumligt og i tid. Remote Sensing-litteraturen har budt på et hav af afledte dataprodukter til formålet, de såkaldte vegetationsindekser[1], der hver især forsøger at måle på afgrødens fysiske, kemiske eller biologiske karaktertræk. Et populært vegetationsindeks i den videnskabelige litteratur er det såkaldte Normalized Difference Vegetation Index (NDVI), der er stærkt korreleret med afgrødens fotosyntetiske aktivitet (fig. 1)



Figur 1: False color billede af NDVI målt på adskillige marker

NDVI-produktet har givet anledning til en række teknologier inden for præcisionsjordbrug, der alle har det til fælles at de (potentielt) inddeler marken i managementzoner – eksempelvis er CropSat-programmet[2] et forsøg på at lave positionsbestemt tildeling af kvælstof.

Satellitbilleder fra Sentinel-2 satellitten leverer i skrivende stund billeder ca. hver uge, og således er der også potentiale for at følge afgrødens udvikling over tid ved at kigge på den løbende udvikling af NDVI.

[1] Johnson et. al. 2014
[2] www.cropsat.dk

NDVI i sig selv er blevet grundigt undersøgt i den akademiske litteratur til forudsigelse af høstudbytter[3],[4],[5] men uden at være særdeles overbevisende; evnen til at modellere variation i udbytterne er typisk ganske lav, med en R^2–værdi i omegnen af 0.2-0.4, og meget få studier forsøger at modellere udbytter fra enkelte marker men i højere grad for hele regioner.

Den nærværende undersøgelse tager udgangspunkt i at NDVI som forklarende variabel kun er en lille del af historien, og at bedre modeller sandsynligvis skal findes i at inkorporere andre, muligvis også svage former for evidens. Således undersøger vi effekten af at kombinere klimavariable, NDVI og disses udvikling over tid. Responsvariablet i undersøgelsen er udbyttemålinger der er kvalitetskontrolleret med en brovægt.

Der er tre primære problemstillinger:

1) Estimering af NDVI-vækstkurven for den individuelle mark over vækstsæsonen. Satellitbilleder er ubrugelige hvis den pågældende mark tilfældigvis er overskyet når billedet tages, og således er individuelle marker i datasættet fotograferet alt fra 3 til 12 gange over sæsonen. Det er et såkaldt *missing data*-problem, idet NDVI-værdien i princippet kan måles hver dag, men af tilfælde årsager (skydække) ikke bliver det. Dette er et åbent problem, og der er i skrivende stund ingen akademiske studier der takler det på en fyldestgørende måde. Denne undersøgelse skal således betragtes som et slags første udkast, idet metoden til at estimere vækstkurven nødvendigvis spiller en stor rolle for den endelige model.

2) Indsamling og validering af klimavariable. For den enkelte mark, der kan være fordelt over hele Danmark, skal tilknyttes klimavariable fra vejrstationer der i så høj grad som muligt beskriver de lokale klimaforhold.

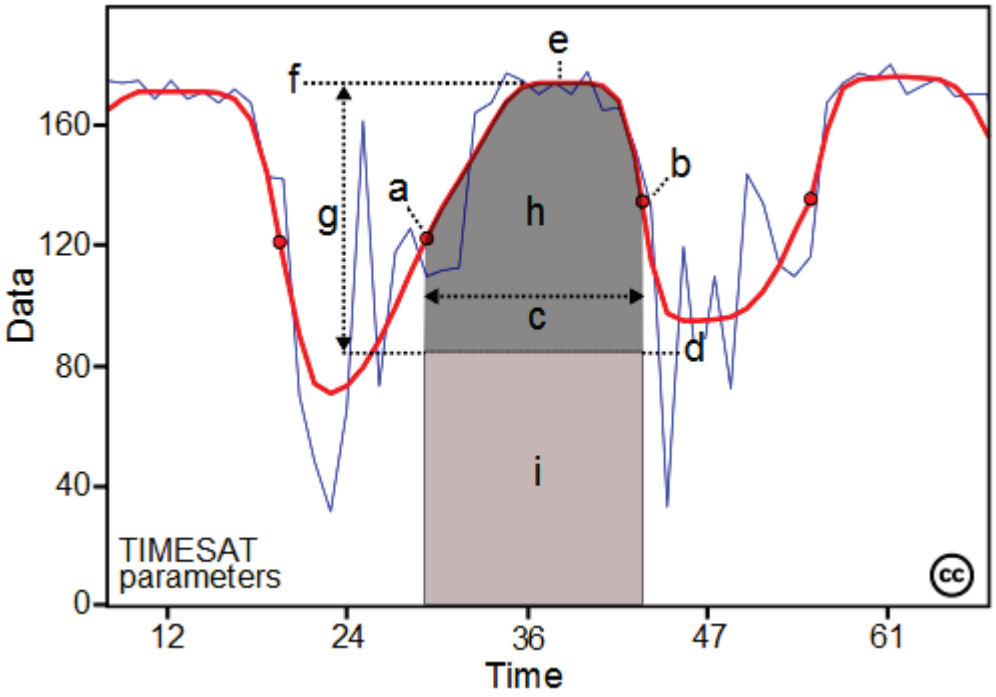
3) Optimering af en tilpas nøjagtig regressionsmodel, og undersøgelse af om den temporale dimension spiller en rolle for modellens nøjagtighed. Klassiske metoder indenfor timeseries regression antager at responsvariablet også udvikler sig over tid – hvilket ikke er tilfældet her. Vi er således statistisk i en gråzone mellem timeseries analyse og klassisk regression, og det primære problem er at finde på en fyldestgørende repræsentation af data.

Estimering af NDVI-vækstkurven

Størstedelen af den akademiske litteratur på området beskæftiger sig med data på regionsbasis der typisk også er målt over en årrække[6]. Problemet er at estimere en kurve givet få, pletvise målinger. Kurven kan forstås som en funktion $f(t) = NDVI_t$, hvor t angiver et tidspunkt målt i en eller anden tidsenhed (fig. 2). For at bruge NDVI-målinger fra den enkelte mark til udbytteforudsigelse, så er det vigtigt at denne kurve reflekterer den sande kurve i så høj grad som muligt – men givet de sparsomme målinger fra Sentinel-2 satellitten, så er det klart at denne proces aldrig bliver eksakt. Der findes et hav af mulige algoritmer til

[3] Sudhanshu et. al. 2010
[4] Jurecka et. al. 2016
[5] Becker-Reshef et. al. 2010
[6] Moreno et. al. 2014

denne slags *curve fitting*[7], så en naturlig fremgangsmåde er at opstille to datasæt – et til estimering og et andet til validering – og vælge den metode der klarer sig bedst, under et passende mål for nøjagtighed, på valideringssættet.



Figur 2: Estimering af en kontinuer kurve (rød) givet lejlighedsvise, støjbehæftede målinger (blå). Billede fra dokumentationen til TIMESAT-algoritmen.

Musial et. al. 2011 vurderede adskillige nøjagtighedsmål og algoritmer til bl.a. Remote Sensing-formål, og fandt at Mean Absolute Error er mest hensigtsmæssig:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |forudsigelse_i - aktuel_i|$$

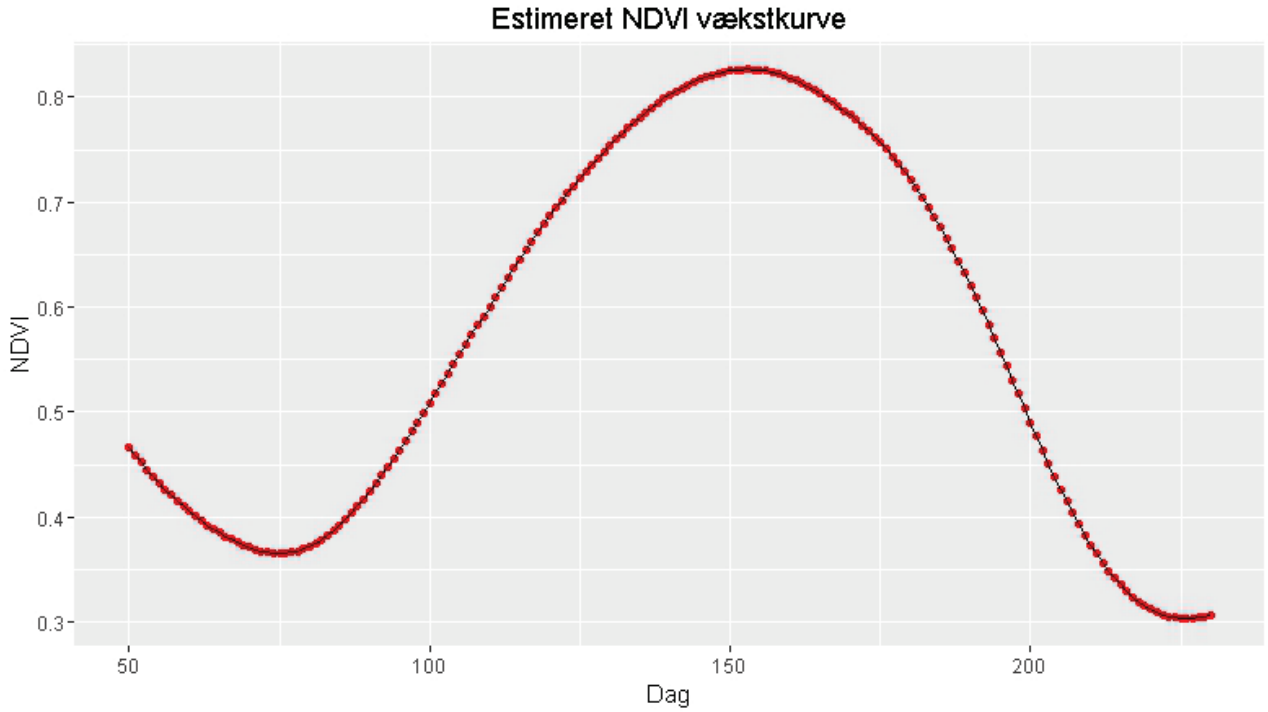
Den absolutte værdi er et udtryk for, at forudsigelser under og over det observerede tal straffes lige hårdt.

Den mest nøjagtige metode på valideringssættet var "smoothing spline"-metoden, implementeret i sproget R. Denne var væsentligt mere nøjagtig end fx Savitzky-Golay Filtering, der danner basis for den anerkendte TIMESAT-algoritme.

For at gøre kurverne så repræsentative som muligt, så er kurven for hver mark i datasættet estimeret ud fra punkterne for marker inden for en radius af 50 km med samme afgrødetype. Dette skyldes to antagelser: 1) afgrøder har unikke NDVI-vækstkurver, dvs. man kan tydeligt se forskel på fx vinterraps og vinterhvede, og 2) marker tæt på hinanden har sandsynligvis korrelerede vækstkurver grundet ensartede klimaforhold.

[7] Jönsson et. al. 2004

Det er dog stadig en fejlkilde at NDVI-dataen er så fejlbehæftet. Marker med få konkrete observationer har en kurve der stort set kun er gennemsnittet af de omkringliggende marker (fig. 3), og er derfor ikke særligt godt repræsenteret. Det er en uundgåelighed ved satellitdata.

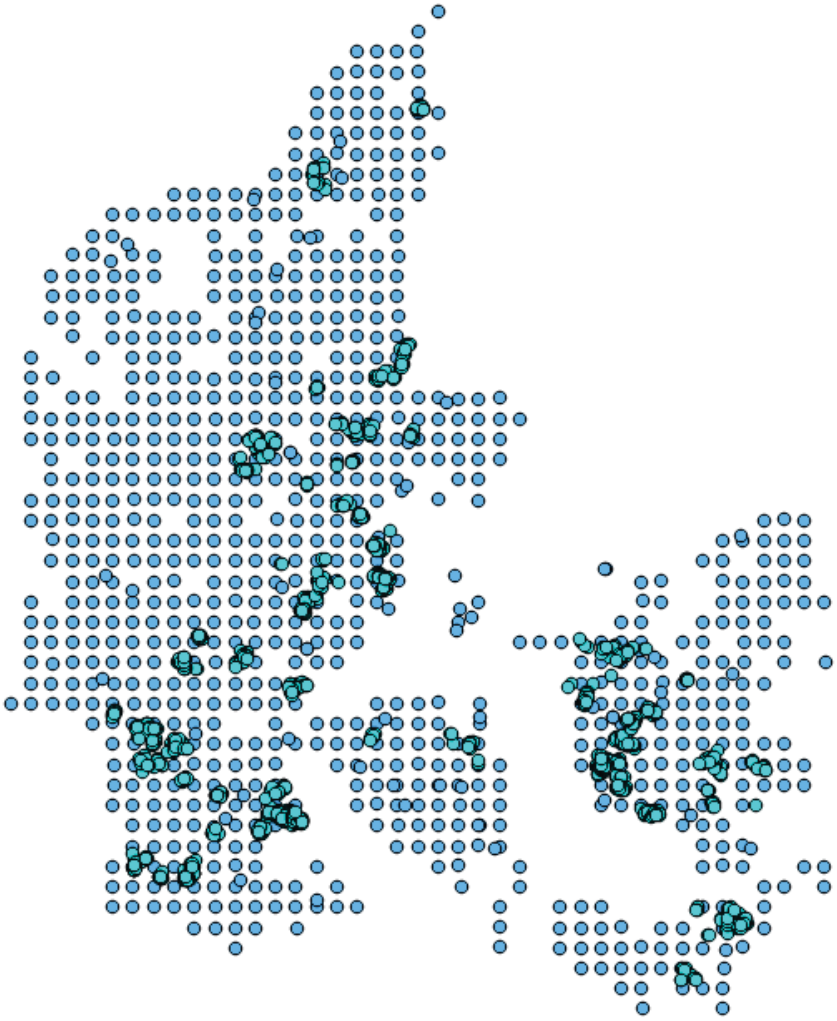


Figur 3: Vækstkurve for en vinterhvedemark

Indsamling og validering af klimadata.

5 klimavariable blev stillet til rådighed igennem SEGES' egne databaser fra KvadratNet-projektet: Hver dag i vækstsæsonen er der registreret værdier for temperatur, jordtemperatur, græstemperatur, nedbør, global stråling. Disse målinger er registrerede af DMI på forskellige vejrstationer, og statistiske estimater bliver lavet på et grid af punkter over hele Danmark. For hver mark i datasættet tilknyttedes data fra det nærmeste gridpunkt (fig. 4).

Den spatiale databehandling blev udført i programmet QGIS. Det var en udfordring at få klimavariablene til at stemme overens, idet der både var mangelfuld data i nogle tilfælde (NaN-værdier), og fordi KvadratNet-databasen refererer til gridpunkterne med nogle arbitrære nøgler og ikke den geografiske beliggenhed.



Figur 4: Registrerede udbytter (lyseblå) knyttes til nærmeste gridpunkt (mørkeblå)

Repræsentation og regression

I sin umiddelbare form er regressionsproblemet et, hvor man forsøger at forudsige en skalarværdi udfra en mængde af tidsserier. Dette præsenterer dog det problem at modellen skal vide hvor langt tilbage i tiden den skal kigge når den laver en forudsigelse, og denne grænse er dybest set arbitrær – måske er det hele sæsonen der er afgørende, måske er det en eller anden delmængde. En provisorisk fremgangsmåde er at omgå problemet ved at inddele tidsserierne i perioder, og så træne en model for hver periode. Problemet er således reduceret til klassisk regression, som kan løses af en mængde af metoder fra statistik og Machine Learning. Valget af model er som sådan underordnet, men for fortolkningens skyld faldt valget på en algoritme kaldet ”Conditional Inference Trees”.

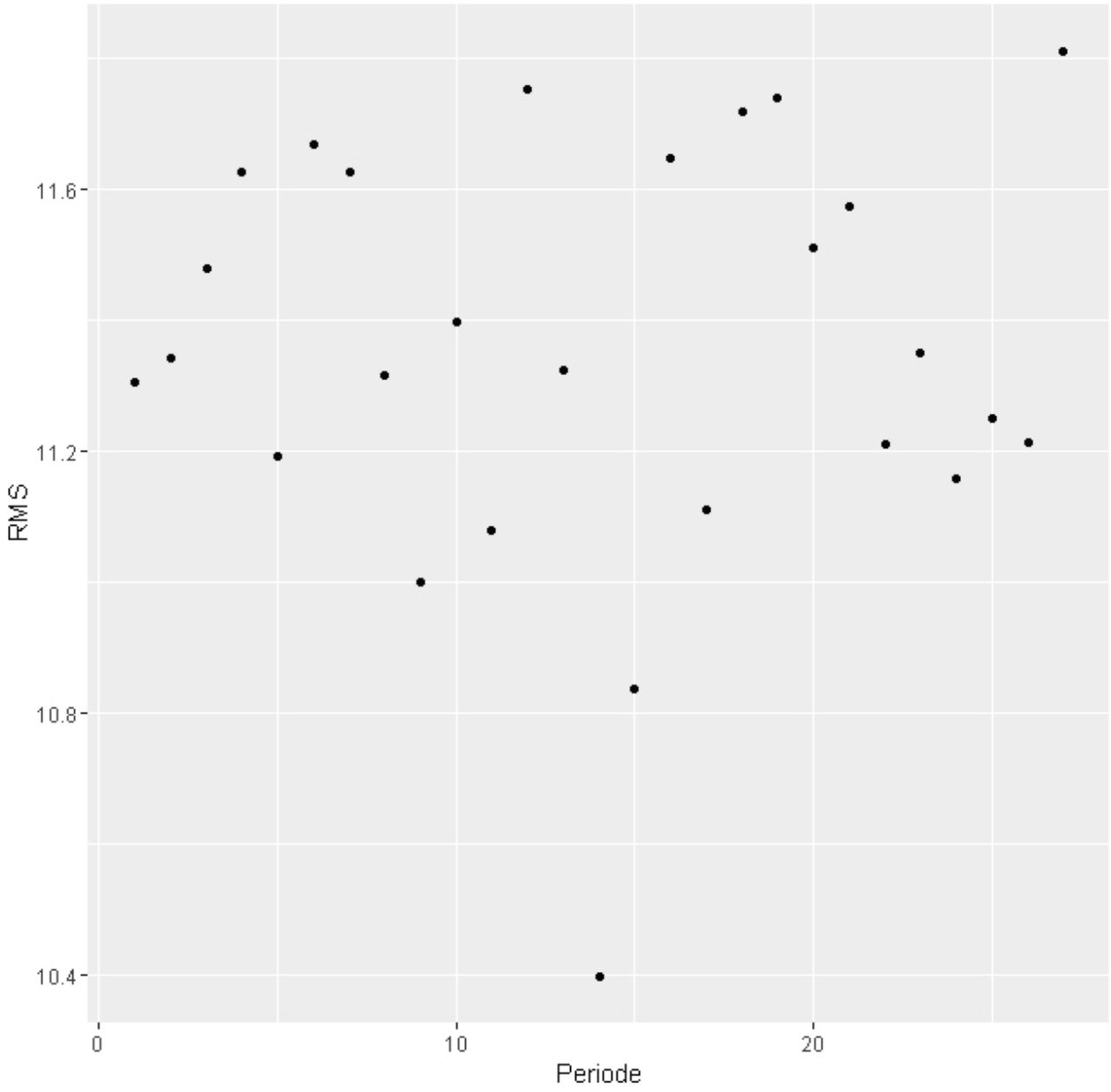
Hver periode havde en længde på 9 dage. For klimavariablene beregnedes gennemsnittene for perioden, og for NDVI beregnedes tilvæksten i perioden, arealet under kurven, samt den gennemsnitlige værdi. Alt sammenfatning af data, samt kørsel af modeller m.m. blev foretaget i programmeringssproget R (fig. 5).

Træning af modellerne foregår i det traditionelle supervised learning-framework: datasættet blandes tilfældigt og deles derefter tilfældigt op i et træningssæt bestående af 70% af eksemplerne, og et valideringssæt bestående af resterende 30%.

Modellernes nøjagtighed varierer noget over perioderne (fig. 6), men der er et interval fra slutningen af april til midten af maj hvor unøjagtigheden er lavest – i gennemsnit en unøjagtighed på 10,4 hektokilo pr. hektar.

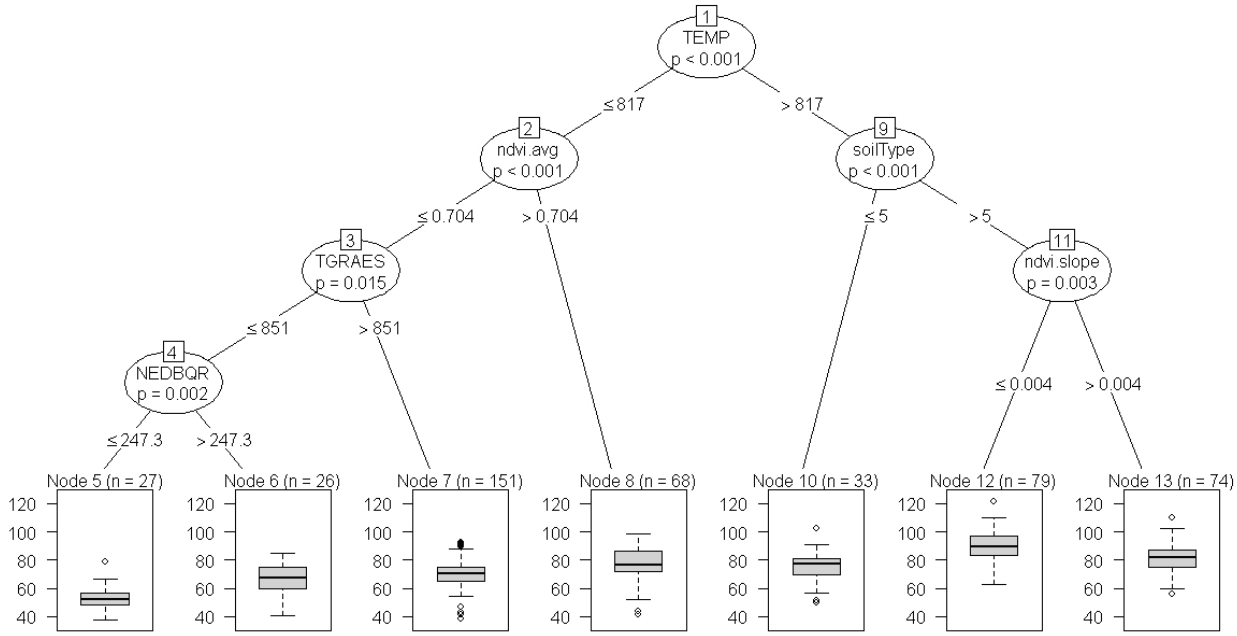
```
287
288 yieldIntervals <- function(intervalSize = 9) {
289   # ======================================
290   # Construct (to, from) pairs to cover the
291   # entire season (on which is recorded data)
292   #
293   # Accepts:
294   #   int intervalSize: number of days
295   #
296   # Returns:
297   #   list of size = intervalSize,
298   #   where list[i] contains (to, from) of
299   #   interval i
300   # ======================================
301   numIntervals <- floor(243 / intervalSize) # 243 is the maximum date of climate data
302
303   outlist <- vector(mode="list", length=intervalSize)
304
305   ctr <- 1
306   from <- 0
307   to <- intervalSize
308   while(ctr <= numIntervals){
309     outlist[[ctr]] <- list(from=convertDateInt(from), to=convertDateInt(to))
310     # Increment the interval
311     from <- from + intervalSize
312     to <- to + intervalSize
313     # increment counter
314     ctr <- ctr + 1
315   }
316   outlist
317 }
318 #testlist <- yieldIntervals() # works
319 #convertDateInt(testlist[[27]]$from);convertDateInt(testlist[[27]]$to) # returns as expected
320
321 # ======================================
322 # Prelim test on all data with ctree
323 # ======================================
324
325 dates <- yieldIntervals()
326 testdf <- makeDF(vinterhv.df, VHV.spline, dates[[1]]$from, dates[[1]]$to)
327
328 for(i in 1:length(dates)) {
329   this.df <- makeDF(vinterhv.df, VHV.spline, dates[[i]]$from, dates[[i]]$to)
330   this.tree <- ctree(yieldQuant ~ ., data=this.df)
331   rms <- sqrt(sum((this.df$yieldQuant - predict(this.tree, newdata = this.df))^2)/nrow(this.df))
332   dates[[i]]$thisdf <- this.df
333   dates[[i]]$rms <- rms
334   dates[[i]]$tree <- this.tree
```

Figur 5: Databehandling i R



Figur 6: Root Mean Square Error pr. periode

Conditional Inference Trees (fig. 7) har en glimrende grafisk fortolkning: modellen består af en serie af ja-nej-spørgsmål, som man kan stille den enkelte mark. Toppen (roden) af træet er det variabel der gør den største forskel i udbyttet, hvorefter man kan følge stierne ned af træet indtil man kommer til forudsigelsen. Modellen stemmer nogenlunde overens med ens forventninger: de marker med de laveste udbytter var også de koldeste samt dem med mindst nedbør.



Figur 7: et Conditional Inference Tree for perioden i slutningen af april for Vinterhvede

Konklusion

Der er stadig meget der skal undersøges, og der er meget plads til forbedring: fremgangsmåden i denne undersøgelse har kun indirekte påvist udbyttets afhængighed af markens klima- og biomasseforhold over tid. Conditional Inference Trees er heller ikke særligt nøjagtige modeller – prisen for deres fortolkelighed – så præcisionen kunne tænkes at blive bedre med ikke-lineære metoder samt ensembling af flere forskellige modeller. Men ikke desto mindre optræder klimavariable i samtlige perioder, hvilket styrker hypotesen om at NDVI alene ikke er nok til udbytteforudsigelse.

Kilder

”CROP YIELD ESTIMATION IN THE FIELD LEVEL USING VEGETATION INDICES”. FRANTISEK JURECKA, PETR HLAVINKA, VOJTECH LUKAS, MIROSLAV TRNKA, ZDENEK ZALUD. MendelNet 2016. http://mendelnet.cz/artkey/mnt-201601-0014_CROP_YIELD_ESTIMATION_IN_THE_FIELD_LEVEL_USING_VEGETATION_INDICES.php?back=/magno/mnt/2016/mn1.php?secid=9

“Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques”. Sudhanshu Sekhar Panda, Daniel P. Ames, and Suranjan Panigrahi. Remote Sensing 2010, 2, 673-696; doi:10.3390/rs2030673

Becker-Reshef, I., Vermote, E., Lindeman, M., Justice, C.O. 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sensing of Environment, 114(6): 1312–1323.

Johnson, D.M. 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sensing of Environment, 141: 116–128.

“Noise Reduction and Gap Filling of fAPAR Time Series Using an Adapted Local Regression Filter”. Álvaro Moreno *, Francisco Javier García-Haro, Beatriz Martínez and María Amparo Gilabert. Remote Sensing 2014.

“TIMESAT—a program for analyzing time-series of satellite sensor data”. Per Jönsson, Lars Eklundh. Computers & Geosciences, Volume 30, Issue 8, October 2004, Pages 833-845, ISSN 0098-3004, http://dx.doi.org/10.1016/j.cageo.2004.05.006. (http://www.sciencedirect.com/science/article/pii/S0098300404000974)

“Technical note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series”. J.P. Musial, M.M. Verstraete, N. Gobron. Atmospheric Chemistry and Physics 2011.

NYE MODELLER BASERET PÅ SATTELITBASEREDE BIOMASSE-MÅLINGER
er udgivet af

SEGES
Landbrug & Fødevarer F.m.b.A.
Agro Food Park 15, Skejby
DK 8200 Aarhus N

Kontakt
Nicolai Cryer, nicr@seges.dk
D +45 8740 5271

December 2016

Redaktion
Nicolai Cryer, SEGES

Denne publikation er finansieret af

STØTTET AF
promilleafgiftsfonden
for landbrug

SEGES skaber løsninger til fremtidens landbrugs- og fødevareerhverv. Vi udvikler forretningsmuligheder i tæt samarbejde med vores kunder, forsknings-institutioner og virksomheder over hele verden. SEGES er en del af Landbrug & Fødevarer F.m.b.A.