



Notat: Statistiske analyser til "Big Data i planteavl" – foreløbig version

30-01-2019

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

Dette notat er foreløbig og adresserer udelukkende formål C, se nedenfor.

Formål

- A) Mellem-mark-analyse:
Udvikling af en model til prædiktion af markernes absolute gns. udbytte på baggrund af funktionen af markernes gns. spektraldata, f.eks. NDVI og/eller NDRE, over tid igennem vækstsæsonen.
- B) Kontrast-analyse:
Analyse af sammenhængen mellem markernes absolute gns. udbytte og markernes gns. spektraldata (NDVI og/eller NDRE) som funktion af tid i gennem vækstsæsonen kontrasteret mellem marker med de (10%) laveste og (10%) højeste udbytter.
- C) Indenfor-mark-analyse:
Udvikling af en model til prædiktion af den relative udbytte indenfor markerne på baggrund af variationen af spektraldata, f.eks. NDVI og/eller NDRE, indenfor markerne over tid igennem vækstsæsonen, herunder prædiktion af områder/pixels/grid i marken, hvor der kan forventes et lavt udbytte (f.eks. under 5 tons pr. ha i vinterhvede og under 5 tons TS pr. ha i majs/helsæd).

Overordnede konklusioner

- Vedr. formål C): Analyse af sammenhængen mellem NDVI hhv. NDRE og udbytte indenfor marken viste markant store effektstørrelser (Cohen's effect size $d > 0.5$) i begge afgrøder.
- For vinterhvedens vedkommende ligger tidspunktet med NDVI-effektstørrelser > 0.5 fra primo maj til primo juli. NDRE-effektstørrelser > 0.5 har et marginalt større tidsvindue.
- For majsens vinterhvedens vedkommende ligger tidspunktet med NDVI-effektstørrelser > 0.5 fra ultimo juni til medio september og NDRE-effektstørrelser > 0.5 fra medio juni til medio september
- Analyserne underbygger, at NDVI og NDRE potentielt kan anvendes til at identificere områder i marken med et udbytte lavere end 5t/ha.

Metoder

Datasæt og datahåndtering

- Datasættet er beskrevet i detaljer i bilaget BDICG-115_dataset_documentation.html [her](#).
- Der er for hver mark (ID_DDS_field) foretaget følgende procedure:
 - 1) Markens maske er skrumpet med 20 m for at fjerne markskæleffekter (skygger fra træer m.m.).
 - 2) Alle punktobservationer (pixels) med en skymaske = 1 og en sky-sandsynlighed > 7.5% blev fjernet (sæt til NA).
 - 3) Alle datoer med andelen af NA-pixels i marken > 5% blev fjernet.
 - 4) Høstdatalaget blev udglattet med et moving average vindue a 3 x 3 pixels.
 - 5) Der er beregnet markens gennemsnitlige udbytte (Mean), gennemsnitlige 10%-trimmede udbytte (trimMean) og standard afvigelse (SD) af udbytte (udbyttevariation) ud fra det ikke-udglattede høstdatalag og Moran's I (homogenitet og grad af clustering af udbytter i marken i et mosaik, se Fig. 1) ud fra det udglattede høstdatalag.

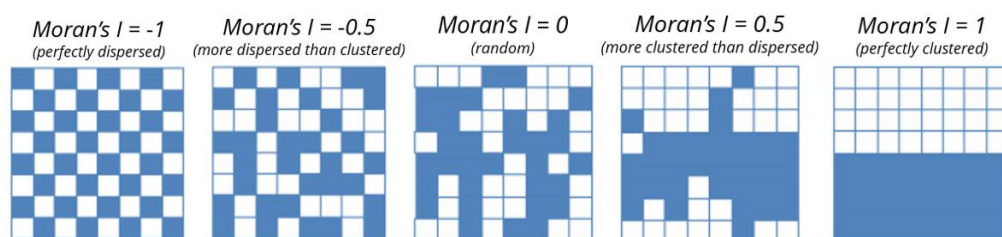


Fig. 1: Fortolkning af Moran's I .

- 6) Det samme er gjort for alle satellitdatalag-dato-kombinationer, hvis dato ligger før markens høstdato indenfor høståret (undtagen trimMean). Der er brugt følgende satellitdatalag: bånd B01 til B12, NDVI, NDRE, MSAVI2, og en NDRE/NDVI-ratio. Derudover er der beregnet følgende statistikker til kvantificering af indenfor-mark variationens sammenhæng med udbyttelaget: Pearson's korrelationskoefficient r_p (lineær korrelation), Spearman's korrelationskoefficient r_s (ikke parametrisk rank-baseret korrelation), begge beregnet over alle pixels indenfor marken, en median lokal Pearson's korrelationskoefficient $r_{p, local}$ beregnet ud fra et moving 11-naboersvindue, gennemsnit og SD af spektralværdier for udbytter under hhv. over 5000 kg/ha, og en effektstørrelse d for denne kontrast beregnet som $(\text{Mean}_{\geq 5000} - \text{Mean}_{< 5000})/\text{SD}$. Effektstørrelsen d er et udtryk for forskellen i et givet satellit-spektralbåndet mellem ikke-lave ($\geq 5\text{t/ha}$) og lave ($< 5\text{t/ha}$) udbytteområder (pixels) i den aktuelle mark relativ til markens variation i satellit-spektralbåndet. Effektstørrelsen d er også kaldt Cohen's effect size. En Cohen's effect size større end 0.5 betegnes til at være en stor effekt.
- Der er trukket jordbundsdata (JBnr) fra hjemmesiden Den Danske Jordklassificering, <http://dca.au.dk/forskning/den-danske-jordklassificering/> og flettet på hver mark.

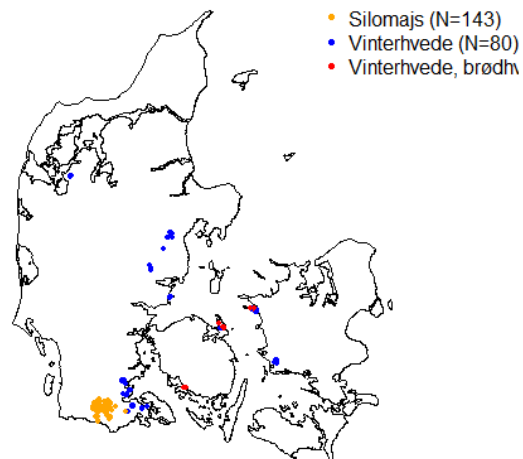


Fig. 2: Markernes placering i DK.

Modeller

- Vedr. formål C): En analyse af sammenhængen mellem NDVI hhv. NDRE og udbytte indenfor marken blev foretaget for hver afgrøde ved at beregne indenfor hver mark Pearson's og Spearman's korrelationer mellem NDVI hhv. NDRE ved hvert af måletidspunkter (satellitdage) og udbytte ved høst. Derudover blev der beregnet for hvert måletidspunkt og hver af de to spektralbånd forskellen i middelværdien mellem pixels i marken, hvis udbytte er over 5t/ha og under 5t/ha divideret med markens standardafvigelse af spektralbåndet. Denne størrelse betegnes Cohen's effect size d eller effektstørrelse og er et standardiseret udtryk for en forskel mellem grupper eller behandlinger. Er d større end 0.5 betegnes effekten eller forskellen til at være stor. De beregnede størrelser blev herefter modelleret i en Generaliseret Additiv model som en udglattet thin plate spline funktion af tiden, med udglatningsgraden bestemt vha. generaliseret krydsvalidering, hen over sæsonen med mark som en tilfældig effekt i modellen.

Resultater

Vedr. formål C)

- Se figurer 1 til 4.

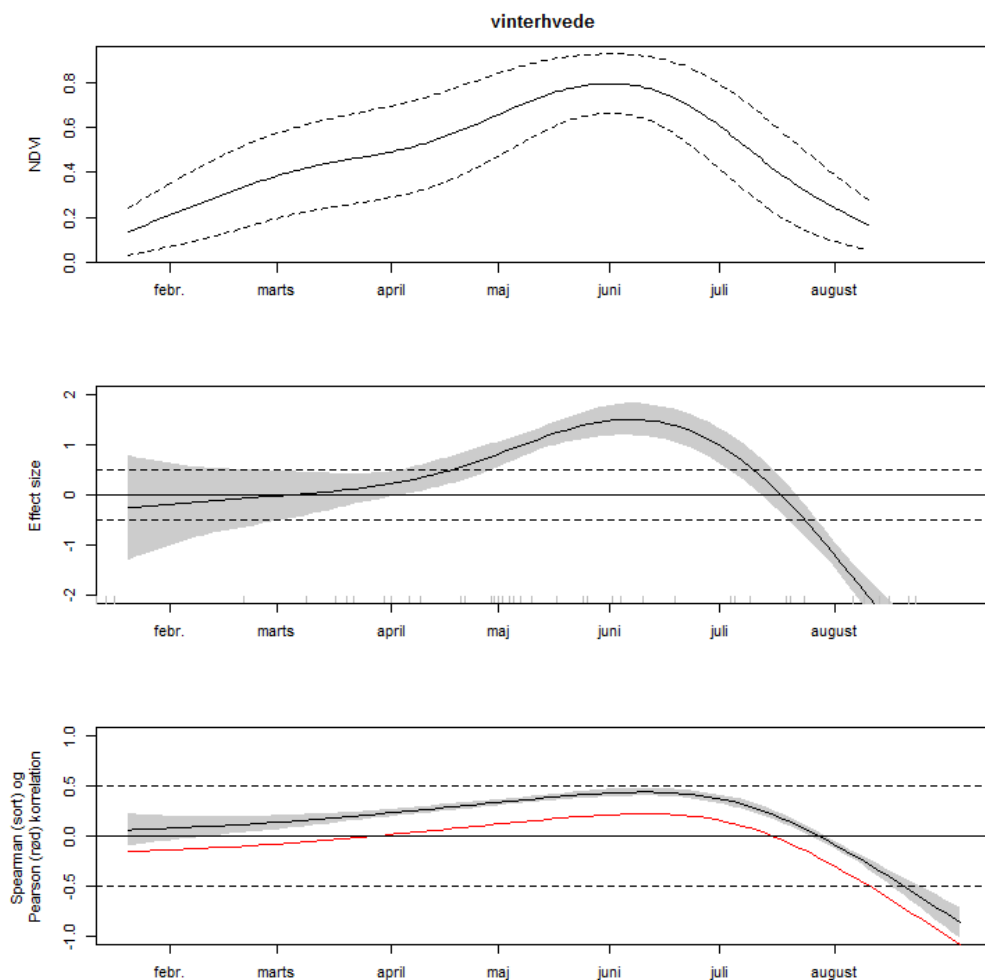


Fig. 1: **NDVI**'ens populationsudvikling i **vinterhvede** (GAM-model udglattet gennemsnitskurve) og 95% konfidensbåndet vises i øverste panel. Effektstørrelsen d beregnet som $\text{gns. NDVI}_{\text{udbytte} \geq 5\text{t/ha}} - \text{gns. NDVI}_{\text{udbytte} < 5\text{t/ha}} / \text{SD}(\text{NDVI})$ vises i midterste panel. En d på $|0.5|$ betegnes som værende en stor effekt. Spearman's og Pearson's korrelationskoefficienter for sammenhængen mellem NDVI og udbytte i vist i nederste panel.

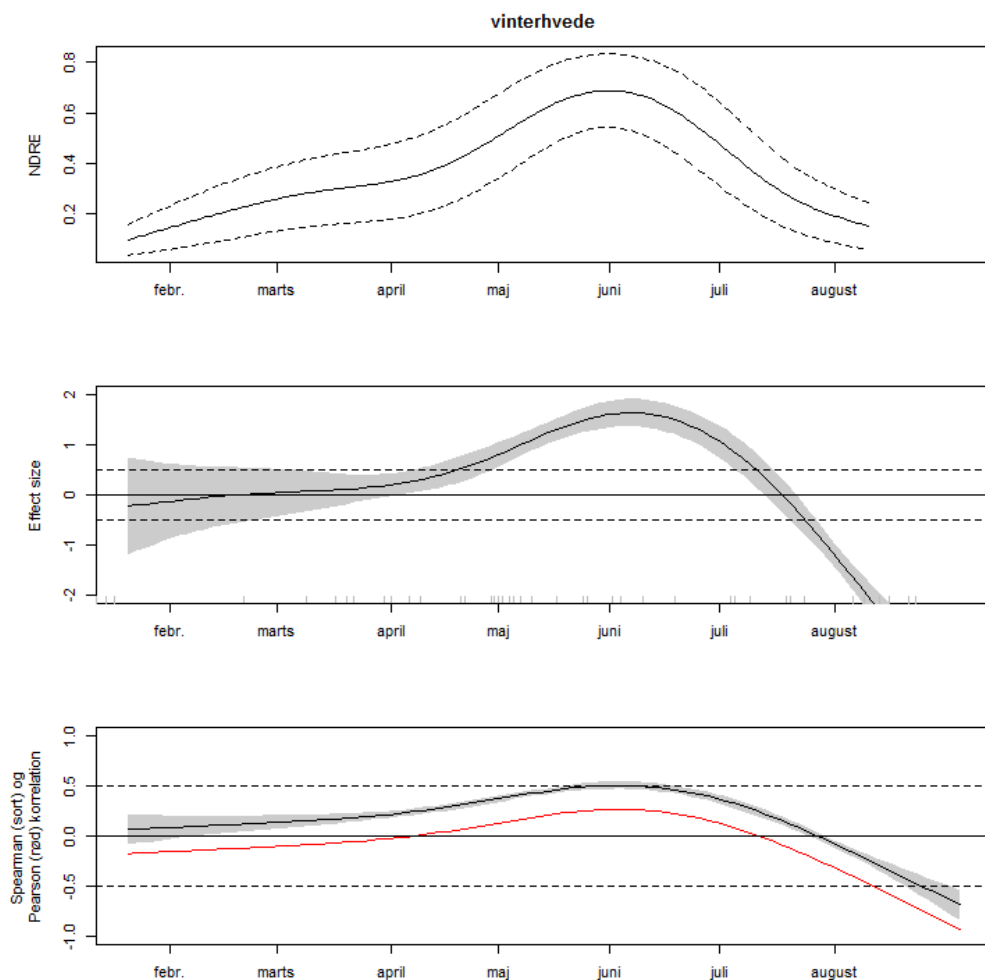


Fig. 2: **NDRE**'ens populationsudvikling i **vinterhvede** (GAM-model udglattet gennemsnitskurve) og 95% konfidensbåndet vises i øverste panel. Effektstørrelsen d beregnet som $\text{gns. NDRE}_{\text{udbytte} \geq 5\text{t/ha}} - \text{gns. NDRE}_{\text{udbytte} < 5\text{t/ha}} / \text{SD}(\text{NDRE})$ vises i midterste panel. En d på $|0.5|$ betegnes som værende en stor effekt. Spearman's og Pearson's korrelationskoefficienter for sammenhængen mellem NDRE og udbytte i vist i nederste panel.

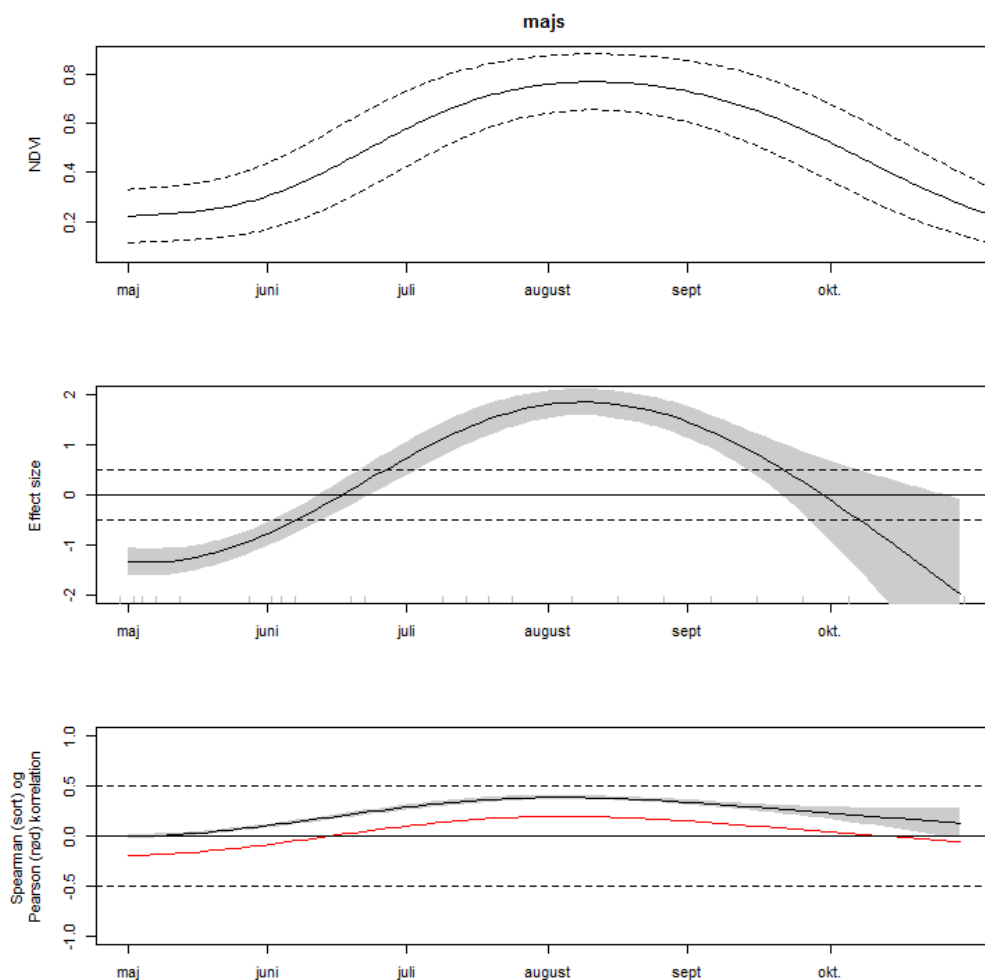


Fig. 3: **NDVI**'ens populationsudvikling i **majs** (GAM-model udglattet gennemsnitskurve) og 95% konfidensbåndet vises i øverste panel. Effektstørrelsen d beregnet som $\text{gns. NDVI}_{\text{udbytte} \geq 5\text{t/ha}} - \text{gns. NDVI}_{\text{udbytte} < 5\text{t/ha}} / \text{SD}(\text{NDVI})$ vises i midterste panel. En d på $|0.5|$ betegnes som værende en stor effekt. Spearman's og Pearson's korrelationskoefficienter for sammenhængen mellem NDVI og udbytte i vist i nederste panel.

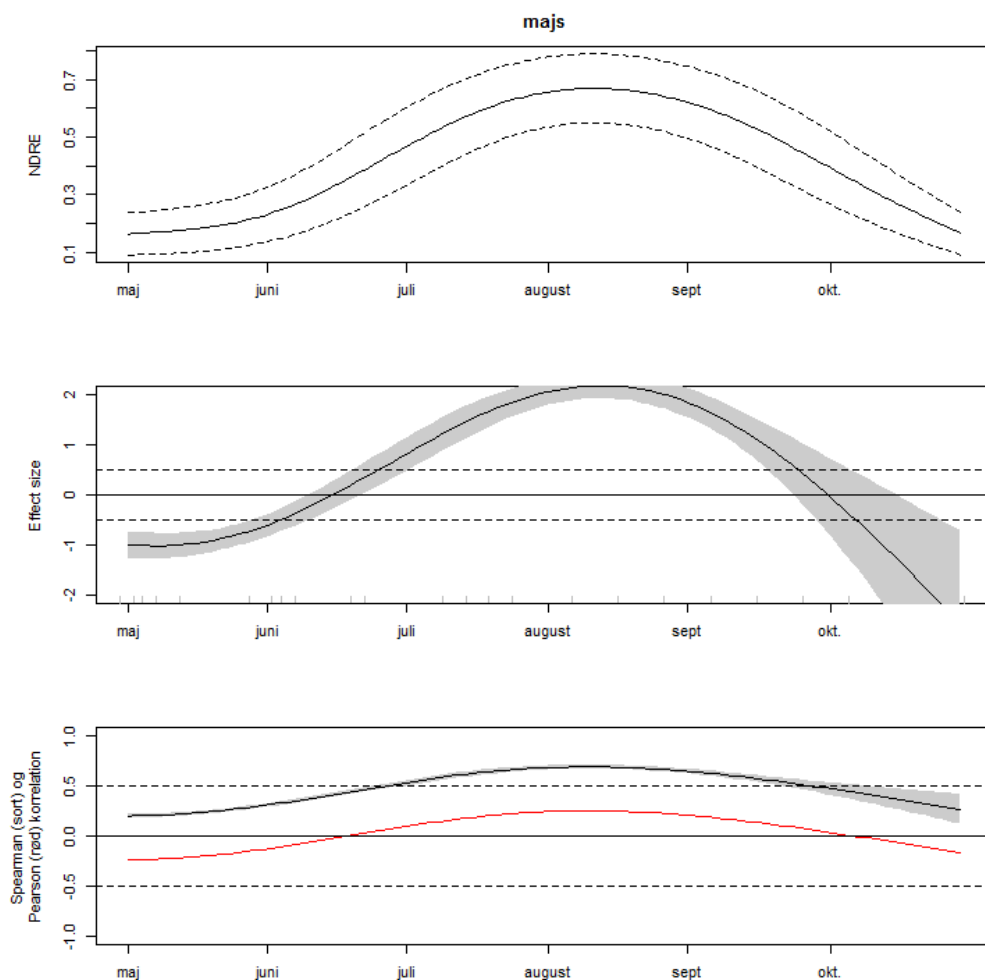


Fig. 4: **NDRE**'ens populationsudvikling i **majs** (GAM-model udglattet gennemsnitskurve) og 95% konfidensbåndet vises i øverste panel. Effektstørrelsen d beregnet som $\text{gns. NDRE}_{\text{udbytte} \geq 5\text{t/ha}} - \text{gns. NDRE}_{\text{udbytte} < 5\text{t/ha}} / \text{SD}(\text{NDRE})$ vises i midterste panel. En d på $|0.5|$ betegnes som værende en stor effekt. Spearman's og Pearson's korrelationskoefficienter for sammenhængen mellem NDRE og udbytte i vist i nederste panel.