STØTTET AF
Promilleafgiftsfonden for landbrug

Udbytteprognosemodel for vinterhvede

December 29, 2022

Dette dokument beskriver udviklingen af en udbytteprognosemodel for vinterhvede foretaget i 2022.

Kontaktinformation: SEGES-datascience@seges.dk

1. Metode

1.1 Datagrundlag og features

Der er anvendt både offentlige og fortrolige datakilder til at udvikle udbytteprognosemodellen. Modellen er baseret på:

- Udbyttedata
- Satellitdata
- Terrænhøjdedata
- Vejrdata
- Markdata

Alt data er blevet konverteret til WGS84/UTM32N (EPSG:32632) og inddelt i grids på 10x10 meter.

1.1.1 Udbyttedata

Til modellen er anvendt positionsbestemte udbyttedata fra mejetærskere. Der er udbyttedata fra i alt 495 marker fra 2016-2021. Udbyttedataene er klargjort ved brug af følgende metode:

Kvalitetssikring:

- Fjern marker med lav datakvalitet (manuel gennemgang af marker).
- Fjern data udenfor markpolygon.
- Standardiser data
- Tjek afgrødetype på markerne (skal være 'vinterhvede' eller 'vinterhvede, brødhvede')

Oprensning:

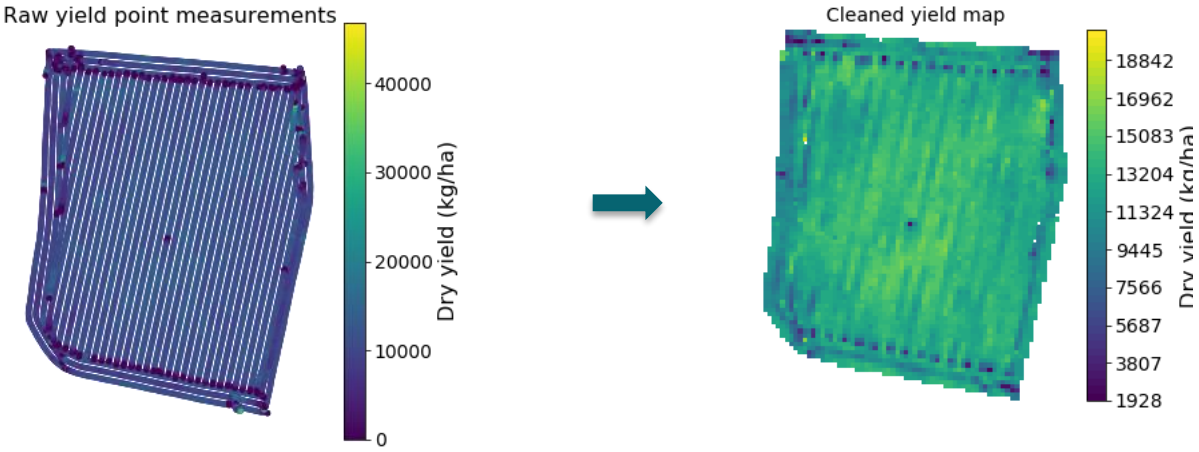
- Sorter udbyttemålinger ud fra tidsstempel.
- Fjern små udbyttemålinger (alle under 5-percentilen), samt nulværdier og værdier større end 250 hkg/ha.
- Fjern outliers:
  Mejetærskersensorerne kan give urealistiske udbyttemålinger, når den tilbagelagte distance er minimal. For at identificere disse outliers bruges distance-to-yield ratioen, -log(distance/yield). Der beregnes et glidende gennemsnit (moving average) over de seneste 10 distance-to-yield ratioer. Alle udbyttemålinger, som falder uden for 2,5*standardafvigelsen af det glidende distance-to-yield gennemsnit, fjernes. Denne procedure gentages 4 gange i alt med udbyttemålingerne, som ikke blev sorteret fra i den foregående iteration. Dette gøres, fordi variansen bliver mindre hver gang, der fjernes outliers, og metoden fjerner således færre og færre outliers med højere præcision for hver gentagelse.
- Interpoler udbyttedataene til 10x10 meter grids ved brug af "Inverse distance to a power"-algoritme.

Kvalitetssikring efter oprensning:

- Manuel gennemgang af marker efter oprensning.

Efter oprensning er der 300 marker tilbage, som opfylder kvalitetskravet. Det resulterer i 302.612 pixels (datapunkter). De interpolerede udbyttedata fungerer som target i modellen, som vi gerne vil forudsige.

I figur 1 ses oprensningen og interpoleringen af en mark.



Figur 1. Oprensning og interpolering af rå udbyttedata. Til venstre ses de rå udbyttedata fra en mejetærsker (punktdata) og til højre udbyttedataene oprenset og interpoleret.

1.1.2 Satellitdata

I analysen er der anvendt satellitdata fra Sentinel 2 (L1C), som er downloaded via en service fra Sentinel-Hub. Sentinel 2 består af to satellitter (S2A og S2B), som leverer data fra 13 spektrale bånd, der alle er anvendt i udbytteprognosen (Se tabel 1). Der er hentet satellitbilleder fra 9. marts til 27. juli for hvert år. Satellitbilleder med skyer er blevet fjernet ved brug af S2_cloudless algoritmen.

	S2A		S2B		
Bånd nummer	**Centrale bølgelængde (nm)**	**båndbredde (nm)**	**Centrale bølgelængde (nm)**	**Båndbredde (nm)**	**Spatial resolution (m)**
B01	442.7	21	442.2	21	60
B02	492.4	66	492.1	66	10
B03	559.8	36	559	36	10
B04	664.6	31	664.9	31	10
B05	704.1	15	703.8	16	20
B06	740.5	15	739.1	15	20
B07	782.8	20	779.7	20	20
B08	832.8	106	832.9	106	10
B08a	864.7	21	864	22	20
B09	945.1	20	943.2	21	60
B10	1373.5	31	1376.9	30	60
B11	1613.7	91	1610.4	94	20
B12	2202.4	175	2185.7	185	20

Tabel 1. Spektrale bånd tilgængelig fra Sentinel 2 (S2A og S2B) samt båndbredde (nm) og opløselighed (m). Bånd markeret med orange indgår i beregningen af vegetationsindeksene NDVI, NDRE og MSAVI2.

Ud fra bånd B04, B05 og B08 er vegetationsindeksene NDVI, NDRE og MSAVI2 udregnet.

(1)
$$NDVI = \frac{(B08 - B04)}{(B08 + B04)}$$

(2)
$$NDRE = \frac{B08 - B05}{B05 + B05}$$

(3)
$$MSAVI2 = \frac{2 \cdot B08 + 1 \ - \sqrt{(2 \cdot B08 + 1)^2} \ - 8 \cdot (B08 - B04)}{2}$$

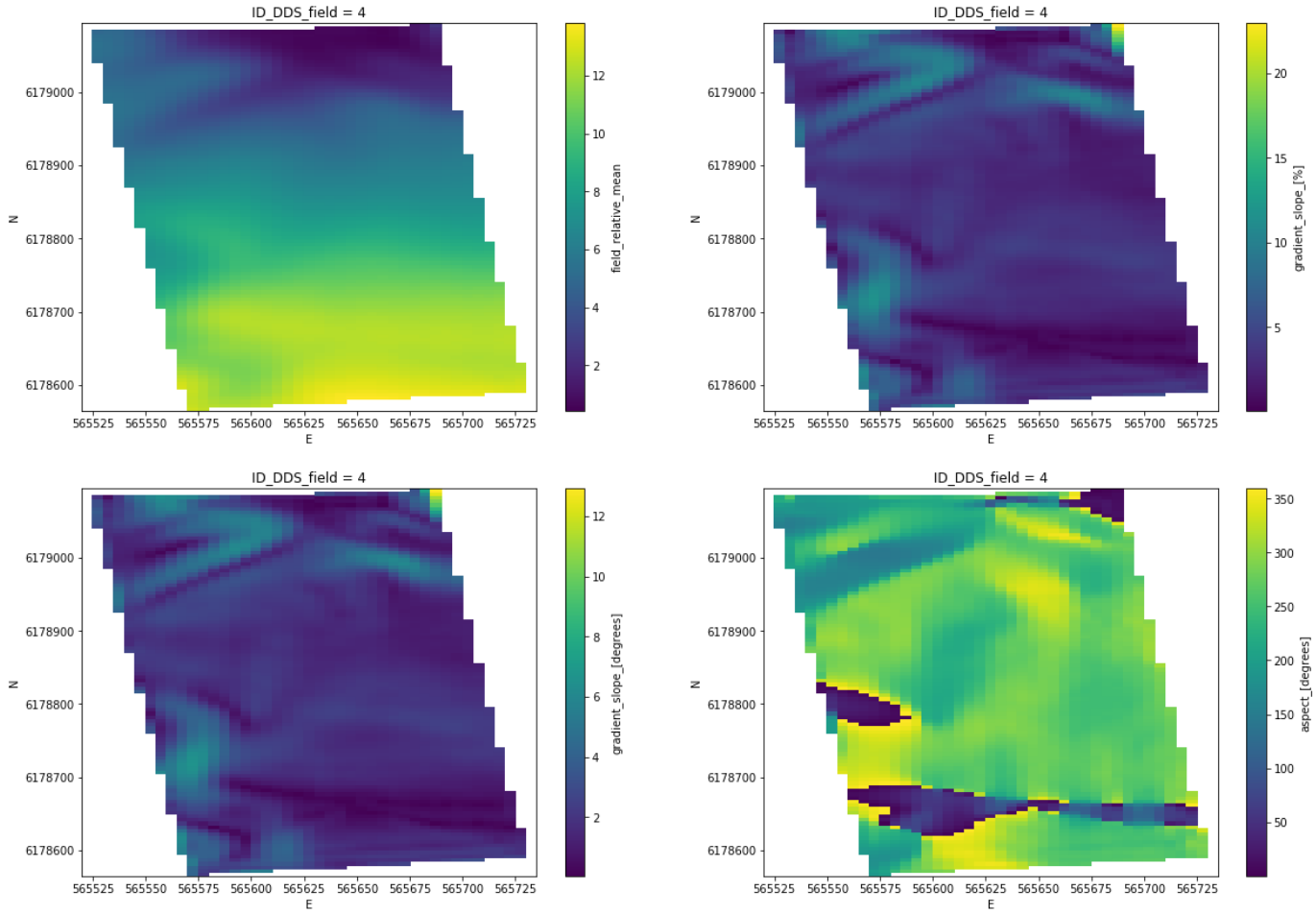
Satallitdataene er dernæst lineært interpoleret. For hver 10x10 meter pixel er værdien hver 14. dag i vækstsæsonen fra 9. marts til 27. juli beregnet for de 13 spektrale bånd, NDVI, NDRE og MSAVI2. Disse er anvendt som features i modellen.

1.1.3 Terrænhøjdedata

Der anvendes terrænhøjdedata fra den danske højdemodel (DEM), som beskriver terrænets højde i forhold til det gennemsnitlige havniveau (opløsning på 0,4 meter). For hver 10x10 meter pixel er der lavet følgende features:

- Højden ift. det gennemsnitlige havniveau.
- Den relative højde ift. det laveste punkt i marken.
- Hældningen, både i grader fra 0-90 og den procentvise hældning.
- Orienteringen af hældningen (0° = nord, 90 °= øst, 180° = syd og 270° = vest).

I figur 2 ses de anvendte features for en mark.



Figur 2. Afledte beregninger ud fra den danske højde model i en tilfældig mark. Øverst til venstre ses højden i marken relativt til det laveste punkt i marken, nederst til venstre ses hældningen i grader (0-90 grader), øverst til højre observeres hældningen i procent, og nederst til højre orienteringen af hældningen (0-360 grader, hvor 0 = nord, 90 = øst, 180 = syd og 270 =vest).

1.1.4 Vejrdata

Vejrdata som luft- og jordtemperatur, nedbør, fordampning og indstråling er hentet fra DMI fra den nærmeste vejrstation for hver mark. For alle vejrparametre er der for hver 14. dag beregnet: gennemsnit, standard afvigelse (SD), minimum, maksimum. Disse anvendes som features i modellen.

1.1.5 Markdata

Til sidst er der anvendt følgende oplysninger fra hver mark som features:

- Afgrøde
- Afgrødesort
- Afgrøde 5 år tilbage
- Jordbundstype

1.2 Machine learning model

Machine learning (ML) algoritmen *Gradient Boosting* er anvendt til udbytteprognosen i vinterhvede.

1.3 Modelevaluering

For at vurdere modellens forudsigelsesnøjagtighed er der anvendt en række metrikker:

- Mean absolute error (MAE)
- Root mean squared error (RMSE)
- R^2
- Standard deviation of absolute error (STD of AE)

Efter hvert eksperiment er de fire metrikker beregnet på både træningssættet og testsættet. Det er dog metrikkerne på testsættet, som vi er interesseret i, da disse siger noget om modellens forudsigelsesnøjagtighed out-of-sample.

2. Resultater

For at finde den bedste udbytteprognosemodel er der foretaget en række eksperimenter. Resultaterne af eksperimenterne er opsummeret i tabel 2. Tabellen viser, hvilket eksperiment det er, hvor tidligt i vækstsæsonen udbyttet forudsiges (model), hvilke features som er med, og hvordan datasættet er inddelt i trænings- og testsæt. De fire metrikker er inddelt i punktmetrikker (pixel) og markmetrikker (hvor forudsigelserne er opsummeret på markniveau), og i trænings- og testsæt. Target og modellens forudsigelser (og derfor også MAE) er i hektokilogram pr. hektar.

Exp	Model	Features	Split	MAE, hkg/h				RMSE				R^2				Std. of AE			
				Point		Field		Point		Field		Point		Field		Point		Field	
				Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
1	July 27	All	Random split, 32870 obs. in test set	4,06	2,15	0,59	0,034	6,18	2,96	0,876	0,062	0,921	0,982	0,998	1	4,66	2,04	0,647	0,052
2	April 6	All	Fields, 40 fields in test set	9,65	8,56	6,72	3,43	12,6	11,44	9,34	4,87	0,56	0,74	0,74	0,95	8,09	7,59	6,5	3,46
	May 4			9,42	8,69	6,23	3,82	12,17	11,58	8,45	5,2	0,59	0,73	0,79	0,95	7,71	7,66	5,78	3,53
	June 1			9,31	8,69	5,86	3,88	12,15	11,6	7,87	5,14	0,59	0,73	0,82	0,95	7,8	7,68	5,25	3,37
	July 27			9,21	8,57	5,64	3,81	11,93	11,49	7,54	5,31	0,61	0,74	0,83	0,94	7,58	7,65	5,01	3,7
3	July 27	Feature elimination	Fields, 40 fields in test set	8,53	8,53	4,86	3,39	11,22	11,44	7,08	4,94	0,654	0,741	0,852	0,95	7,28	7,62	5,14	3,59
		Exclude DMI		9,63	8,98	7,41	4,86	12,65	11,98	9,98	6,63	0,561	0,716	0,705	0,911	8,2	7,93	6,69	4,5
		Exclude DTM		9,2	8,74	5,74	4,2	11,94	11,68	7,66	5,65	0,61	0,73	0,826	0,935	7,61	7,75	5,07	3,78
		Exclude S2		9,38	8,88	5,9	4,07	12,17	11,81	8,24	5,63	0,594	0,724	0,8	0,936	7,75	7,78	5,75	3,88
4	July 27	All	Fields, 37 fields in test (removed year 2016 in both test and train)	9,38	8,76	5,88	4,4	12,17	11,67	8,04	5,95	0,588	0,73	0,8	0,931	7,77	7,72	5,48	4
			Fields, 37 fields in test (removed year 2016 in only test)	9,17	8,57	5,48	3,81	11,88	11,49	7,45	5,31	0,608	0,738	0,828	0,942	7,56	7,65	5,05	3,7
5	July 27	Aggregate by field - All	Fields, 40 fields in test set			5,39	0,325			7,12	0,382			0,85	0,999			4,66	0,2
		Aggregate by field - Feature elimination				4,06	0,213			5,39	0,253			0,914	0,999			3,55	0,135
6	July 27	Aggregate by field - Feature elimination	Years, 2021 in test set			4,12	0,449			5,38	0,54			0,826	0,999			3,46	0,3
			Years, 2020 in test set			3,95	0,546			4,82	0,671			0,902	0,999			2,77	0,391
			Years, 2019 in test set			9,95	0,475			14,76	0,599			0,812	0,999			11,01	0,635
			Years, 2018 in test set			7,13	0,656			9,47	0,875			0,6	0,998			6,24	0,578
			Years, 2017 in test set			2,36	0,723			3,49	0,875			0,835	0,998			2,57	0,492
			Years, 2016 in test set			3,59	0,577			4,64	0,711			0,893	0,999			2,94	0,416

Tabel 2 opsummerer resultaterne fra eksperimenterne. Tabellen viser, hvilket eksperiment det er, hvor tidligt i vækstsæsonen udbyttet forudsiges (model), hvilke features som er med, og hvordan datasættet er inddelt i trænings- og testsæt. De fire metrikker er inddelt i punktmetrikker og markmetrikker (hvor forudsigelserne er opsummeret på markniveau), og i trænings- og testsæt.

Eksperimenterne er grupperet i 6 overordnede eksperimenter:

Eksperiment 1: I eksperiment 1 er der anvendt alle features frem til d. 27. juli. Der er anvendt et random split på tværs af alle pixels i datasættet med 32.870 pixels i testsættet. Der opnås en rigtig god nøjagtighed på testsættet med MAE på 4,06 hkg/h på punktniveau og 0,59 hkg/h på markniveau. Pga. det random split optræder der dog pixels fra samme marker i både trænings- og testsættet. Dette resulterer i en kunstig god performance, og det afspejler ikke virkelighedens brug af modellen.

Eksperiment 2: I eksperiment 2 løses dette ved at splitte på hele marker med 40 marker i testsættet. Igen anvendes alle features, men der er fire undereksperimenter. Disse omhandler, hvor tidligt i vækstsæsonen udbyttet forudsiges, dvs. til hvor langt frem der er anvendt data i modellen. De fire datoer er:
6. april, 4. maj, 1. juni, 27. juli.
Som det ses, er MAE noget højere end i eksperiment 1 på både punktniveau og markniveau. Dette er at forvente, når vi splitter på markniveau, da dette afspejler virkelighedens brug mere. Det ses dog også, at MAE bliver lavere, når der anvendes mere data længere henne i vækstsæsonen.

Eksperiment 3: I eksperiment 3 splittes der igen på hele marker med 40 marker i testsættet. Der anvendes data frem til d. 27. juli, men til gengæld fjernes features fra modellen for at se hvordan performance ændres. Ved 'feature elimination' trænes modellen først, hvorefter de vigtigste features findes. De mindst vigtige features fjernes, hvorefter modellen trænes igen, og de mindst vigtige features endnu engang fjernes. Sådan fortsætter det i 20 skridt. Det ses i tabellen, at MAE falder signifikant ift. eksperiment 2, hvor alle features er med (det testes på de samme marker). Det kan derfor betale sig at bruge et mindre antal betydningsfulde features sammenlignet med at bruge alle features.
Dernæst forsøges det at fjerne hele grupper af features, vejrdata features, satellitdata features, og terrænhøjdedata features. At dømme ud fra MAE er vejrdata features og satellitdata features vigtige, da MAE stiger meget, hvis disse udelades hver især. Terrænhøjdedata features kan dog umiddelbart udelades uden nogen indvirkning på MAE.

Eksperiment 4: I eksperiment 4 splittes der igen på hele marker, og der anvendes data frem til d. 27. juli. I dette eksperiment undersøges, hvad der sker med performance, når vi udelader marker fra 2016. Satellitdata features fra 2016 kan være af tvivlsom kvalitet, da kun den ene satellit S2A af de to (S2A og S2B) var lanceret på daværende tidspunkt. I testsættet er der dermed 37 marker (minus 3 fra 2016) og så trænes modellen med og uden marker fra 2016 for at sammenligne performance. Det ses, at performance på testsættet er bedre, når modellen får lov til også at træne på marker fra 2016 til trods for, at der ingen marker fra 2016 er i testsættet. Det kan tyde på, at modellen lærer nogle generelle sammenhænge fra marker fra 2016, som er mere universelle på tværs af år. Det giver derfor umiddelbart mening at beholde marker fra 2016.

Eksperiment 5: Da vi i sidste ende er interesseret i performance på markniveau, undersøges der i eksperiment 5, hvordan performance ændrer sig, når data først aggregeres på markniveau og modellen dermed trænes på markniveau. Igen splittes der på hele marker med 40 marker i testsættet (de samme som tidligere) og der anvendes data frem til d. 27. juli. Der eksperimenteres så med at bruge alle features og anvende feature elimination, som beskrevet i eksperiment 3. Det ses, at performance er bedre ved at træne modellen på markniveau sammenlignet med at træne modellen på punktniveau og bagefter aggregere forudsigelserne på markniveau. Igen er performance signifikant bedre, når der anvendes feature elimination.

Eksperiment 6: Når der splittes på hele marker, så optræder der marker fra samme år i trænings- og testsættet. Dette kan resultere i en kunstig god performance, da modellen derved kan finde frem til det 'generelle' udbytteniveau for et givet år. 'År' optræder ikke som en feature i modellen, men andre features (som vejrdata features) kan komme til at agere proxyvariabler for året. Dette betyder, at testsættet ikke afspejler virkelighedens brug af modellen. For at efterligne virkelighedens brug af modellen fjernes der i eksperiment 6 hele år fra træningssættet, som så fungerer som testsættet. På denne måde kan modellen testes præcist, som den vil blive anvendt i virkeligheden. Grundet resultaterne i eksperiment 5 trænes modellen på markniveau, og der anvendes feature elimination og data frem til 27. juli. Der testes så med hvert år som testsæt, hvor det fjernes fra træningssættet. Det ses, at performance varierer meget på tværs af år. Performance er dårlig for 2018 og 2019, hvorimod performance for 2016, 2017, 2020 og 2021 er god med MAE helt ned til 2,36 hkg/h.