

Prognose for tørstofindholdet i majs helsæd	Ansvarlig	HJN
	Oprettet	14-12-2020
Projekt: [4584, Projektnavn]	Side	1 af 9

Formål

Formålet med dette notat er at få beskrevet og dokumenteret datagrundlag, modeller og drift af den eksisterende majsprognose for tørstofindholdet i majs helsæd i CropManager.

Majsprognosen er udviklet af Teknologisk Institut (tidligere AgroTech) i samarbejde med SEGES i årene 2017 til 2019. Dette notat er opstartet af Teknologisk institut men færdigskrevet af SEGES i tæt samarbejde med Teknologisk institut.

Majsprognosen

Majsprognosen er en forecast model der prædikerer en majssorts tidsmæssige udvikling af tørstofprocenten (TS%) i en given aktuel mark. Majsprognosen inddrager data fra en aktuel mark (geografisk position, sådato, sort, JB-nr.) og klimadata (historiske klimadata, 1-uges vejrudsigt og normal-klimadata) for pågældende mark. Modellen er bygget på baggrund af historiske klimadata samt data fra landsforsøgene, sortsafprøvninger i DLBR-regi eller i forskningsmæssige sammenhæng (data fra Aarhus Universitet), samt data fra bedrifter indsamlet af SEGES, kvæg.

Forecast modellerne er modul-baserede i tre moduler (Præprocessering, Prædiktion og Postprocesering) så de til enhver tid kan opdateres samt forbedres.

Da majssorterne er meget forskellige i tidlighed, har sorten en afgørende betydning for tørstofindholdet, og derfor er det også nødvendigt at angive sorten. Prognosen håndterer majssorter, som har deltaget i Landsforsøgene med majssorter til helsæd de to seneste år. Hvis marken er tørkepræget, påvirket af frost, unormal eller uens udviklet eller præget af ukrudt skal man anvende prognosen med forsigtighed.

Datagrundlag og modelbeskrivelse

Teknisk specifikation:

R-version og R pakker brugt til træning og udvikling af model

```

R version 3.3.3 (2017-03-06)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)

locale:
[1] LC_COLLATE=Danish_Denmark.1252 LC_CTYPE=Danish_Denmark.1252 LC_MONETARY=Danish_Denmark.1252
[4] LC_NUMERIC=C LC_TIME=Danish_Denmark.1252

attached base packages:
[1] stats graphics grDevices utils datasets methods base

other attached packages:
[1] scam_1.2-1 randomForestSRC_2.4.2 mgcv_1.8-17 nlme_3.1-131
[5] zoo_1.7-14 doBy_4.5-15 jsonlite_1.3 Rcurl_1.95-4.8
[9] bitops_1.0-6 dplyr_0.5.0 RevUtilsMath_10.0.0

loaded via a namespace (and not attached):
[1] Rcpp_0.12.9 lattice_0.20-34 assertthat_0.1 MASS_7.3-45 grid_3.3.3 R6_2.2.0
[7] DBI_0.6 magrittr_1.5 Matrix_1.2-8 splines_3.3.3 RevUtils_10.0.3 tools_3.3.3
[13] parallel_3.3.3 tibble_1.2
  
```

Datasæt til træning af modellerne

- Modellerne er bygget på baggrund af følgende data:
 - o Forsøgs el. sortsdata (1992 - 2016), i alt 19.389 observationer fordelt over 10 forskellige datakilder.
Se Tabel 1 for at se hvilke datakilder TS%-data stammer fra. De stammer fra 346 forskellige lokaliteter (unikke GPS-koordinater).
Alle datakilder er leveret af Ib Sillebak Kristensen, AU undtagen NFTS-data som er trukket

fra Nordic Field Trial System og data fra kvæg, leveret fra SEGES kvæg. Tabel 2 viser et udsnit af data.

- Klimadata (1985-2016, 40 km grid).
Her skal det nævnes at modellen er trænet på et 40 km grid, men i praksis kører den på et 10 km grid.
- Modellerne foreligger som R-filer.
 - De aktuelle modeller er 3 steps generaliserede additive (GAM) modeller samt en randomforest model.

Tabel 1: Antal observationer indenfor hver af de 10 "serier".

NFTS	DJF	HeKo	HoHe	KoId	Kvæg	Ma_N	Sort	Tyst	UK
10229	211	15	184	6	681	102	6129	394	1438

Tabel 2: Udsnit af forsøgs/sortsdata

id	NFTS	serie	locality	PLANNR	LBNR	Grid40km	EASTING	NORTHING	UTMZONE	HOSTAAR	saadato
1	1	Nej	Sort	5e+05.6300000	030229292	001	7	5e+05	6300000	32	1992 1992-05-12
2	2	Nej	Sort	5e+05.6300000	030229292	001	7	5e+05	6300000	32	1992 1992-05-12
3	3	Nej	Sort	5e+05.6300000	030229292	001	7	5e+05	6300000	32	1992 1992-05-12
4	4	Nej	Sort	5e+05.6300000	030229292	001	7	5e+05	6300000	32	1992 1992-05-12
5	5	Nej	Sort	5e+05.6300000	030229292	001	7	5e+05	6300000	32	1992 1992-05-12
6	6	Nej	Sort	5e+05.6300000	030229292	001	7	5e+05	6300000	32	1992 1992-05-12
hostdato	host	hostDayNr	saa	GrowingDays	Forfbet	JB_nr	LSNBET	TS_helsæd	TS_kolbe		
1	1992-10-07	0000-10-07	49	0000-05-12	148	<NA>	1	Fronica	30.24819	NA	
2	1992-10-07	0000-10-07	49	0000-05-12	148	<NA>	1	Hiro	36.44400	NA	
3	1992-10-07	0000-10-07	49	0000-05-12	148	<NA>	1	Erlevo	35.24400	NA	
4	1992-10-07	0000-10-07	49	0000-05-12	148	<NA>	1	Challenger	35.94400	NA	
5	1992-10-07	0000-10-07	49	0000-05-12	148	<NA>	1	Aviso	36.54400	NA	
6	1992-10-07	0000-10-07	49	0000-05-12	148	<NA>	1	Solidor	33.42063	NA	
VAND_kerne	saa2	host2	KTGr	JB	HOSTAAR_fac	y	tidlighed_ts				
1	NA	11	6	Ukendt	JB1,JB3	1992	30.24819	3.8956492			
2	NA	11	6	Ukendt	JB1,JB3	1992	36.44400	0.9686683			
3	NA	11	6	Ukendt	JB1,JB3	1992	35.24400	1.2330113			
4	NA	11	6	Ukendt	JB1,JB3	1992	35.94400	0.3272476			
5	NA	11	6	Ukendt	JB1,JB3	1992	36.54400	-0.8688007			
6	NA	11	6	Ukendt	JB1,JB3	1992	33.42063	0.8212909			

Tabel 3: Udsnit af klimadata

Grid40Km	Year	DayNo	Date	Airtemp	Mintemp	Maxtemp	Glorad	Prec
1	1	1985	1 1985-01-01	-1.5	-2.4	-0.3	0.5	5.5
2	1	1985	2 1985-01-02	-5.0	-8.9	-1.6	1.1	0.5
3	1	1985	3 1985-01-03	-6.4	-8.4	-4.7	1.6	0.1
4	1	1985	4 1985-01-04	-8.0	-8.4	-7.5	0.9	0.1
5	1	1985	5 1985-01-05	-8.7	-9.5	-7.9	1.6	0.2
6	1	1985	6 1985-01-06	-10.5	-12.3	-9.0	1.3	2.5

Datamodel til forecast

Når landmanden laver en forecast forespørgsel er der brug for følgende input.

- Data input:
 - **Klimadata** (fra ClimateData tabellen se Tabel 5)
 - Klimadata fra aktuell lokalitet bestående af historisk observerede klimadata i indeværende år op til forespørgselsdatoen.
 - En uges prognose vejrdato tilgængeligt på forespørgselsdatoen.
 - Normklimadata (30 års normen) for resten af året indtil den 1.november.
 - OBS. Der hentes data ud fra DMI i proceduren som gemmes ned med JobId, disse data bliver slettet igen efter prognosen så vil ikke være at genfinde i tabellen.
 - **Sortsnavn**, f.eks. "Atrium"
 - **Saadato**
 - **Forespørgselsdato**
 - **JB-nr.** i kategorierne "JB1, JB3" eller "JB2, JB4+".
 - **UTM-koordinater.** OBS: SEGES oversætter kommune til kommunens centroid-koordinat.
 - **TS_DK** i form af en JSON tabel. TS_DK dækker over tørstofprøver indhentet af SEGES i løbet af sæsonen og om muligt jævnt fordelt over Danmark for at kunne levere en mulighed for at kalibrere prædiktionen ind til indeværende år. Selvom modellen søger at indfang årsspecifikke tendenser i tørstof-udvikling via inddragelse af de historiske klimavariabler fra indeværende år, så forventes det, at inddragelsen af kalibreringsprøver i

- løbet af vækstsæsonen forbedrer prædiktionssevnen af modellen. JSON-tabellen skal indeholde felterne: "saadato", "hostdato", "UTME", "UTMN", "UTMzone", "KIGr" (kløvergræs med mulighederne "Ja" el. Nej"), "JB" (se JB-nr), "sort" og "TS" (TS%).
- **TS_inField** i form af en JSON tabel. TS_inField dækker over tørstofprøver indhentet af landmanden i aktuel mark for at kunne levere en mulighed for at kalibrere prædiktionen ind til den aktuelle mark. JSON-tabellen skal indeholde felterne: "dateDay", "TS" (TS%).
- Yderlig data input:
- **modeltabel.rda**. Modeltabellen indeholder den aktuelle model til brug for forecast, R data objekter der skal loades, klimadata variable der skal anvendes eller beregnes, navnet på den variabel, der beskriver sortens afvigelse, og endeligt navnene på de funktioner der præprocesserer, prædikterer og postprocesserer. Se Tabel 4 for et eksempel. Alle funktioner listet i modeltabellen skal være tilstede i R-filen CommonFunctions.r. Alle R-objekter, der skal loades skal findes i en undermappe "data" i det workspace hvor der afvikles.
 - **majstabel.rda**. Majstabellen indeholder information over de enkelt sorters "tidlighed" i form af en estimeret afvigelse fra gennemsnitstørstofprocent. Denne er estimeret på baggrund af modelprædikterede LSmeans tørstofprocenter for alle sorter over alle forsøgsår. Estimeringen af LSmeans stammer fra den årligt opdaterede modelkørsel i Sortsvalg majs værktøjet, se her: <https://sortsvalgmajs.dlbr.dk/da-dk>. OBS: sorter der ikke er inkluderet i denne tabel kan der stadigvæk laves en prædiktions for. Disse vil så basere sig på en relativ tidlighed i TS% afvigelse fra gennemsnittet på 0, dvs. en gennemsnitssort. Se Tabel 6 for et eksempel.
- I modsætning til tidligere majsprognosemodeller, der byggede på konceptet majs dage som en skaleret afvigelse i TS% fra en referencesort, så bygger denne model på antagelsen, at den gennemsnitlige tørstofprocent på tværs af år og sorter vil være stabil og ikke ændre sig betydeligt, f.eks. en target-TS% på 32% ved høst. Hvis nu den optimale ensilage tørstof procent ændrer sig fordi der bruges andre teknikker, så vil denne antagelse ikke holde mere.
- Majstabellen bliver opdateret hvert år hos TI på baggrund af landsforsøgene via sortsvalg majs. Her er det dog kun for helsæd, men der bliver også beregnet tilsvarende for kolbe- /kernemajs.
- Modellen der her bruges for at beregne tidlighed er en mixed effects model hvor der efter korrektion for årseffekt og lokalitet beregnes en LSmeans for tørstofprocenten for hver sort. Her fratrækkes tørstofprocenten for gennemsnitssorten for at se hvor meget tidligere den givne sort er i tørstofprocent.

Tabel 4: modeltabel.rda

--

```

__id  afgrbet  model_name active  update_dato  reference  rmsep
1 1 1 Majshelsaed Ibs model 2013 0 2013-12-07 00:00:00.0000000 /dokumenter/Majsprgnose_Rapport_2013_ISK.docx 2,2
2 2 2 Majshelsaed Model v.1 1 2018-06-08 00:00:00.0000000 <NA> <NA>
3 3 3 Kolbemaajs Model v.1 0 2018-06-08 00:00:00.0000000 <NA> <NA>
4 4 4 Kernemaajs Model v.1 0 2018-06-08 00:00:00.0000000 <NA> <NA>
5 5 5 Kolbemaajs Model v.2 1 2019-07-05 00:00:00.0000000 <NA> <NA>
6 6 6 Kernemaajs Model v.2 1 2019-07-05 00:00:00.0000000 <NA> <NA>
7 7 7 Majshelsaed Model v.2 0 2019-09-23 00:00:00.0000000 <NA> <NA>

R_Objects

1

2
<NA>

3
helsaed_m_step1 <- readRDS('./data/helsaed_m_step1.RDS'); helsaed_m_step2 <- readRDS('./data/helsaed_m_step2.R
DS'); klima_pca_model <- readRDS('./data/klima_pca_model.RDS'); m_kalibrering <- readRDS('./data/m_kalibrering.RDS')
4
kolbe_m_step1 <- readRDS('./data/kolbe_m_step1.RDS'); kolbe_m_step2 <- readR
DS('./data/kolbe_m_step2.RDS'); helsaed_m_step1 <- readRDS('./data/helsaed_m_step1.RDS'); helsaed_m_step2 <- readRDS('./data/helsaed_m_step2.R
DS'); klima_pca_model <- readRDS('./data/klima_pca_model.RDS'); m_kalibrering <- readRDS('./data/m_kalibrering.RDS')
5
predict_VAND_kerne <- readRDS('./data/predict_VAND_kerne.RDS'); kolbe_m_step1 <- readRDS('./data/kolbe_m_step1.RDS'); kolbe_m_step2 <- readR
DS('./data/kolbe_m_step2.RDS'); helsaed_m_step1 <- readRDS('./data/helsaed_m_step1.RDS'); helsaed_m_step2 <- readRDS('./data/helsaed_m_step2.R
DS'); klima_pca_model <- readRDS('./data/klima_pca_model.RDS'); m_kalibrering <- readRDS('./data/m_kalibrering.RDS')
6
kolbe_m_step1 <- readRDS('./data/kolbe_m_step1_v2.RDS'); kolbe_m_step2 <- readRDS('./data/kolbe_m_step2_v2.
RDS'); klima_pca_model <- readRDS('./data/klima_pca_model.RDS'); m_kalibrering <- readRDS('./data/m_kalibrering.RDS')
7
predict_VAND_kerne <- readRDS
('./data/predict_VAND_kerne.RDS'); kolbe_m_step1 <- readRDS('./data/kolbe_m_step1_v2.RDS'); kolbe_m_step2 <- readRDS('./data/kolbe_m_step2_v2.
RDS'); klima_pca_model <- readRDS('./data/klima_pca_model.RDS'); m_kalibrering <- readRDS('./data/m_kalibrering.RDS')
8
helsaed_m_step1 <- readRDS('./data/helsaed_m_step1.RDS'); helsaed_m_step2 <- readRDS('./data/helsaed_m_step2.R
DS'); klima_pca_model <- readRDS('./data/klima_pca_model.RDS'); m_kalibrering <- readRDS('./data/m_kalibrering.RDS')

climate_variables variety_variable preprocessFUN predictFUN
1 Airtemp, Mintemp, Maxtemp, Glorad, Prec, MVE, T8, TO_8, T8_12, T12, coolITS Majsdage convert2Ibsklima predict_TS_Ib
2 Airtemp, Mintemp, Maxtemp, Glorad, Prec, MVE, T8, TO_8, T8_12, T12, coolITS tidligned_ts prepareclimate_model_v1 predictCTS_model_v1
3 Airtemp, Mintemp, Maxtemp, Glorad, Prec, MVE, T8, TO_8, T8_12, T12, coolITS tidligned_ts prepareclimate_model_v1 predictCTS_model_v1
4 Airtemp, Mintemp, Maxtemp, Glorad, Prec, MVE, T8, TO_8, T8_12, T12, coolITS tidligned_ts prepareclimate_model_v1 predictCTS_model_v1
5 Airtemp, Mintemp, Maxtemp, Glorad, Prec, MVE, T8, TO_8, T8_12, T12, coolITS tidligned_ts prepareclimate_model_v1 predictCTS_model_v2
6 Airtemp, Mintemp, Maxtemp, Glorad, Prec, MVE, T8, TO_8, T8_12, T12, coolITS tidligned_ts prepareclimate_model_v1 predictCTS_model_v2
7 Airtemp, Mintemp, Maxtemp, Glorad, Prec, MVE, T8, TO_8, T8_12, T12, coolITS tidligned_ts prepareclimate_model_v1 predictCTS_model_v1

postprocessFUN
1 identity
2 postProcess_TS_model_v1
3 postProcess_TS_model_v1
4 postProcess_TS_model_v1
5 postProcess_TS_model_v1
6 postProcess_TS_model_v1
7 postProcess_TS_model_v2

```

Tabel 5: Klimadata

id	jobId	createdDate	editedDate	creator	editor	airtemp	mintemp	maxtemp	glorad	prec		
1	653899	83822018	2019-05-06 10:52:30	2019-05-06 10:52:30	app_dmdb app_dmdb	14.3730	12.8	18.5	5.1170	2.0		
2	654324	83822017	2019-05-06 13:46:49	2019-05-06 13:46:49	app_dmdb app_dmdb	15.5120	11.5	20.8	13.2150	6.6		
3	671719	636645875	2019-05-10 09:44:36	2019-05-10 09:44:36	app_dmdb app_dmdb	14.0600	11.5	16.3	6.3210	9.6		
4	707584	82002018	2019-05-31 13:40:23	2019-05-31 13:40:23	app_dmdb app_dmdb	14.6101	12.9	18.5	4.4392	1.7		
5	708254	82202018	2019-05-31 13:54:25	2019-05-31 13:54:25	app_dmdb app_dmdb	20.4160	13.8	26.5	23.5340	0.0		
6	741284	69502018	2019-06-19 10:25:04	2019-06-19 10:25:04	app_dmdb app_dmdb	7.2931	5.7	8.6	1.2042	2.2		
7	742881	69502017	2019-06-19 13:44:48	2019-06-19 13:44:48	app_dmdb app_dmdb	16.3125	13.4	18.8	15.9486	2.1		
8	743278	2017	2019-06-19 13:45:53	2019-06-19 13:45:53	app_dmdb app_dmdb	-0.2710	0.0	1.1	1.1000	2.2		
9	743672	55002017	2019-06-19 13:48:04	2019-06-19 13:48:04	app_dmdb app_dmdb	15.1000	6.8	21.1	14.0326	0.9		
10	745877	2019	2019-06-20 10:21:28	2019-06-20 10:21:28	app_dmdb app_dmdb	15.9000	12.4	19.4	18.1000	2.4		
precNoCor	dateDay	dataSource	expectedDataSource									
1	1.8220	2018-09-20	0	0								
2	6.0000	2017-07-11	0	0								
3	8.6590	2018-09-14	0	0								
4	1.5013	2018-09-20	0	0								
5	0.0000	2018-07-11	0	0								
6	1.8000	2018-11-30	0	0								
7	1.9472	2017-07-11	0	0								
8	1.5790	2017-12-10	0	0								
9	0.7728	2017-09-05	0	0								
10	2.2000	2019-07-11	2	2								

Tabel 6: majstabel.

id	Isnbet_LF1	AFGRNR	LSNNR_LF1	Majsdage	Majsdage_se	N_kardex	N_hostaar	HOSTAAR_min	HOSTAAR_max	tidligned_ts	afgrbet	
1	124	Atrium	NA	25078	32.04	1.47	NA	NA	NA	NA	2.4521538	Majshelsaed
2	125	Chavoxx	NA	25621	33.36	1.47	NA	NA	NA	NA	1.1321538	Majshelsaed
3	126	Activate	NA	26133	39.78	1.47	NA	NA	NA	NA	-5.2878462	Majshelsaed
4	127	Ambition	NA	26437	36.34	1.47	NA	NA	NA	NA	-1.8478462	Majshelsaed
5	128	Emblem	NA	26528	33.96	1.47	NA	NA	NA	NA	0.5321538	Majshelsaed
6	129	Fieldstar	NA	26842	35.36	1.47	NA	NA	NA	NA	-0.8678462	Majshelsaed
7	130	Exxtens	NA	27174	35.53	1.48	NA	NA	NA	NA	-1.0378462	Majshelsaed
8	131	Glory	NA	27186	36.78	1.48	NA	NA	NA	NA	-2.2878462	Majshelsaed
9	132	Sunlite	NA	27188	36.07	1.47	NA	NA	NA	NA	-1.5778462	Majshelsaed
10	133	Asgaard	NA	27189	32.60	1.47	NA	NA	NA	NA	1.8921538	Majshelsaed

Data flowet i en forecast er følgende:

- 1) Præprocessing af input
- 2) Prædiktion

3) Postprocessering af forecast

Procedurene, der skal afvikles, er specificeret i modeltabellen i databasen i den aktive linje. Det vil sige her specificeres for hver afgrøde (majshelsæd, kolbemajs eller kernemajs) hvilke klimavariabler der skal hentes, hvilke funktioner der skal køres samt hvilke R objekter der skal loades og bruges i de forskellige skridt i forecasten. Disse R objekter er RDS filer der er dannet da modellen blev trænet. I det følgende er der beskrevet de funktioner, R objekter og variable der pt. er aktive (i december 2020). Output fra en forecast kan ses i Fig 1.

1) **Præprocessering af input**

I dette skridt bearbejdes klimadata fra databasen så de kan indgå i prædiktionsmodellerne.

Der ses kun på prædiktioner af høstdatoer i høståret mellem datoerne 1. august og 1. november. Hvis der findes flere observationer på samme dag i klimatabellen i databasen da tages et gennemsnit af disse inden de bearbejdes yderligere.

Der afvikles funktionen *prepareClimate_model_v1*. I denne funktion loades R objektet *klima_pca_model*.

Ud fra de eksisterende klimavariabler i databasetabellen: Airtemp, Glorad, Mintemp, Maxtemp, Prec og PrecNoCor beregnes der nu de følgende klimavariabler: *MVE*, *T8*, *T0_8*, *T8_12*, *T12* samt *coolT5*, der skal bruges i prædiktionsmodellerne for majshelsæd. Beregningerne er beskrevet nedenfor.

- $MVE = (y_{max} - y_{min})/2$, er Majs varme enheder hvor $y_{max} = 3.33 * (Mintemp - 10) - 0.084 * (Maxtemp - 10)^2$ og $y_{min} = 1.8 * (Mintemp - 4.4)$ se (https://www.landbrugsinfo.dk/public/0/d/5/plante_majsvarmeenhederproduktion_majsmark) for yderligere info.
- *T8*: Hvor meget luft temperaturen afviger/er højere end 8 grader. Det vil sige hvis temperaturen er over 8 da trækkes 8 fra ellers 0.
- *T0_8*: Den givne temperatur mellem 0 og 8, ellers 0.
- *T8_12*: Hvor meget luft temperaturen mellem 8 og 12 afviger/er højere end 8 grader. Det vil sige hvis temperaturen er mellem 8 og 12 da trækkes 8 fra ellers 0.
- *T12*: Hvor meget luft temperaturen afviger/er højere end 12 grader. Det vil sige hvis temperaturen er over 12 da trækkes 12 fra ellers 0.

Derudover beregnes også for hver dato en løbende uges-, måneds- og 2 måneders sum, samt standardafvigelse for disse variable.

Ud fra dette datasæt dannes der for hver mulig høstdato i vores høstperiode et nyt datasæt bestående af de ovenstående samt de løbende summer beregnet ud fra input sådatoen og de forskellige høstdage. For dette datasæt bestående af en høstdato samt 121 klimavariabler bliver der prædikeret 121 principalkomponenter ud fra træningsmodellen *klima_pca_model* en principal komponentanalyse (PCA).

klima_pca_model:

$$\text{princomp}(X, \text{covmat} = \text{cov.wt}(X, \text{wt} = w/\text{max}(w)))$$

Grundet den ulige fordeling af observationer på tværs af år blev hver observation vægtet med den inverse af observationsårets antal observationer. Dette forbedrer generaliserbarheden på tværs af år.

Disse 121 principalkomponenter samt høstdato, *Airtemp_weekSum_h*, *Glorad_weekSum_h*, *Prec_weekSum_h* og *MVE_weekSum_h* danner tilsammen det nye klimainput til prædiktionen. Se Tabel 7.

Tabel 7

--

indikatorvariablene for JB-nr. kategori, kløvergræs som forfrugt, en lineær effekt af såtidspunkt (saa2) og sorterens tidlighed (tidlighed_ts) beregnet som TS%-afvigelsen fra gennemsnits TS% på tværs af alle sorter. Derudover er der modelleret en tilfældig normalfordelt effekt af lokalitet og høstår. Grundet den ulige fordeling af observationer på tværs af år blev hver observation vægtet med den inverse af observationsårets antal observationer. Dette forbedrer generaliserbarheden på tværs af år. Den gennemsnitlige årsafvigelse på baggrund af TS_DK beregnes som den gennemsnitlige residual, dvs. forskel mellem prædikeret (*m_kalibrering*) og observeret (TS_DK), for kalibreringsprøverne.

Det nye datasæt består nu af de præprocesserede klimadata, en gennemsnitlig årsafvigelse, en tidlighed_ts for sorten, input koordinater, en forskel i dage til reference sådatoen (1. maj) og en forskel i dage til reference høstdatoen (1. oktober). Der prædikteres nu tørstofprocent i en 2 skridts model hvor tørstofprocenten prædikteres ud fra de fittede værdier i en GAM model *helsaed_m_step1* (1) samt residualerne fra en random forest model *helsaed_m_step2* (2).

(1) *helsaed_m_step1*:

$$y \sim s(\text{EASTING}, \text{NORTHING}, k = 30, \text{fx} = T) + \text{TS_year_deviation} + s(\text{host2}, k = k, \text{fx} = T) + \text{tidlighed_ts} + \text{saa2} + s(\text{Glorad_weekSum_h}, k = k, \text{fx} = T) + s(\text{sqrt}(\text{Prec_weekSum_h}), k = k - 1, \text{fx} = T) + \text{JB}$$

Denne GAM (Generalized additive model) model består af en glat funktion af koordinaterne easting og northing med en fixed udglatningsgrad, en lineær effekt af årsvariationen der indgår som en parallelforskydning, en glat funktion af host2 som er en numerisk værdi af høstdatoen, en lineær effekt af tidligheden, såtidspunkt samt jordbundstype og to klimavariabler Global radiation beregnet som ugesum ud fra høsttidspunkt og precipitation (nedbør).

Motivationen bag step 1 modellen er, at levere en model der kan levere prædiktioner for alle geografiske placeringer ved interpolering langs den estimerede glatte flerdimensionelle modelflade.

(2) *helsaed_m_step2*:

$$\text{res} \sim \text{tidlighed_ts} + \text{JB} + \text{fit_step1} + \text{EASTING} + \text{NORTHING} + \text{saa2} + \text{host2} + \text{TS_year_deviation} + \text{Comp_1} + \text{Comp_2} + \text{Comp_3} + \text{Comp_4} + \text{Comp_5} + \text{Comp_6} + \text{Comp_7} + \text{Comp_8} + \text{Comp_9}$$

Denne model er en random forest model. En random forest er en machine learning prædiktionsmodel, der prædikterer responsvariablen som den gennemsnitlige prædiktions på tværs af en stor samling (skov) af decision tree modellens prædiktioner, hvor hver decision tree model er trænet på baggrund af et subsample af observationer og et subsample af forklarende variabler for derefter gennem at opnå dekorrelerede modeller i skoven. En decision tree model er en model der består af en hierarkisk samling af if-else statements (tvegrenende træ). En random forest er dermed en ikke-parametrisk model, som kan approksimere en hvilken som helst sammenhæng mellem input og outputvariabler. Modellens formel skal forstås på den måde, at variablerne listet på højre side af ~ er brugt som forklarende variabler i random forest modellens decisions tree modeller.

Motivationen bag step 2 modellen er at indfange lokale effekter både i tid (klimaeffekter som PCA-scores) og rum (*EASTING*, *NORTHING*) ikke indfanget af den glatte step 1 model, dvs. at levere nuancer, der hvor step 1 modellen er upræcis.

Den samlede TS% prædiktions er summen af step1 og step 2 prædiktions.

3) **Postprocessing af forecast**

Funktionen *postProcess_TS_model_v1* afvikles. Tanken her er at der ønskes en udglatning af modellens prædiktioner da der er flere svingninger fra dag til dag end der var ønsket fra opgavestiller. Der blev vurderet at dette ville være forvirrende for landmanden at tolke på, og derfor blev der ønsket en udglatning af tørstofprædiktionerne.

Der er både en logit vægtning af observationer samt en SCAM model (Shape constrained additive model, som er en GAM model men hvor der er sat restriktioner på de splines der fittes) hvor der her for kernemajs er krævet en monotont faldende form og for de andre en monotont stigende form.

De prædikterede tørstofprocenter (*yhat*) bruges i en scam model

$$\text{yhat} \sim s(\text{host})$$

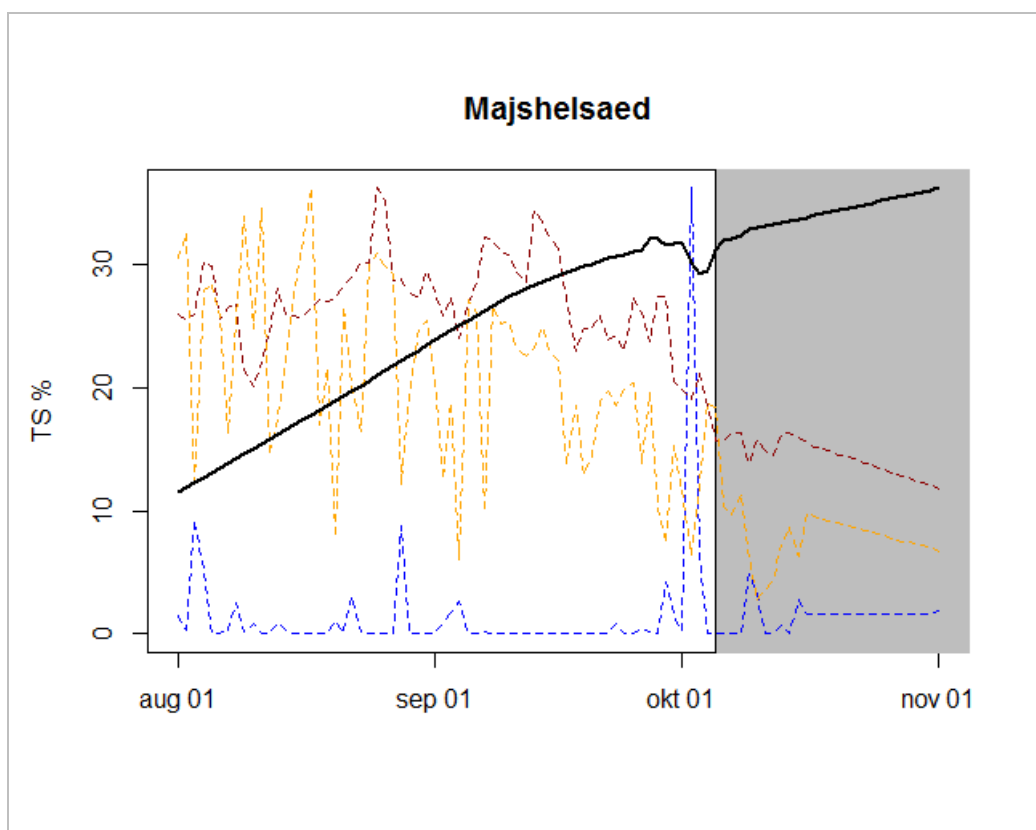
for at smoothe prædiktionerne. Høstdatoer tættere på midten af høstsæsonsperioden er vægtet højere end først eller sidst på sæsonen.

Ud over dette er disse brugt i en udglatningsfunktion. Se *postProcess_TS_model_v1* funktionen for nærmere uddybelse af udglatningsformlerne.

Dertil kommer at findes der helsædsmajs tørstofprøver indrapporteret af landmanden selv for aktuelle lokalitet, da beregnes der en gennemsnitlig afvigelse mellem den udglattede estimaterede tørstofprocent og den observerede tørstofprocent. Den prædikterede tørstofprocenten parallelforskydes derefter med denne korrektion. Det vil sige niveauet bliver korrigeret for lokationen såfremt at der er set et lavere eller højere niveau end forventet og dette styrker dermed prædiktionen.

Desuden afgrænses tørstofprocenten til at være mellem 0 og 100 hvis de prædikterede tørstofværdier skulle ligge udenfor dette interval.

Tilslidst samles outputtet så der for hver høstdato i intervallet består af sort, prædikteret tørstofprocent, sådato, JB-nummer, UTM koordinater samt klimadata.



Figur 1: Eksempel af forecast profilen for en majshelsædssort. (Sort = Atrium, sådato = 2016-05-01, forespørgselsdato = 2016-10-05, JB = "JB1, JB3", UTME = 606571, UTMN = 6131654, UTM zone = 32.)

Afvikling, drift og årlig opdatering

Den mere tekniske del som SEGES, digital står for vedrørende afvikling af R scripts i crop manager vil ikke blive beskrevet i dette dokument.

Hvert år den 1. december da genkøres sortsvalg majs hos TI. Her dannes der input til Sortsvalg majs programmet. Her beregnes også input til majstabelen for majshelsæd. I samme forbindelse afvikles der lignende for kolbe- og kernemajs. Resultater sendes til SEGES der opdaterer majstabelen i databasen til brug for majshøstprognosen.

Der har tidligere været indberettet kalibreringsprøver som SEGES har fået taget. Disse prøver er ikke blevet taget i 2019 og 2020. Det ville styrke prædiktionsmodellen hvis disse prøver blev taget og der burde i den forbindelse sikres en automatiseret proces, sådan at kalibreringsprøver bliver læst ind i tabellen TS_DK.

Udviklingsmuligheder

Der er trænet på 40 km grids hvor man i den kørende prædiktion har 10 km grids. Dette kan betyde at prædiktionerne rammer skævt.

Hvis der prædikteres for en kolbe/kerne sort hvor der ikke findes helsædsprøver, da parrer man op mod

en gennemsnitlig sort, her kunne der overvejes om man skulle vælge en kategori af sorter der minder mere om sorten.

Der er trænet på modeller hvor der findes kalibreringsprøver. Disse har været manglende i både 2019 og 2020. Der fjernes noget styrke fra modellen.

Der er ikke afrapporteret usikkerheder på prædiktionsestimaterne. Dette kunne man overveje, da der naturligvis er noget usikkerhed forbundet med estimaterne. Der skal dog gøres nogle overvejelser hvordan man præcis beregner disse usikkerheder for eksempelvis en blanding af en GAM model og en random-forestmodel.

Version to modellen taget i brug i 2019 burde måske ses på igen i forhold til at den ikke længere inddrager viden omkring lokalitet og klima fra helsædsprøverne.