STØTTET AF
Promilleafgiftsfonden for landbrug

Yield prediction in winter wheat using machine learning; improving implemented farm management tool.

M. K. Langgaard[1], L. R. Malskær[2], and J. Hein[2]
[1] *Department of Crop and Environment, SEGES Innovation P/S, Denmark*
[2] *Department of Digital solutions, SEGES Innovation P/S, Denmark*
Agro Food Park 15, 8200 Aarhus N, Denmark
mlje@seges.dk

Abstract

Accurate estimation of expected yield before harvest is important to regulate nitrogen application according to crop demand. Therefore, a forecast model predicting winter wheat yield before harvest was developed and implemented in 2020. In this study more data and new features were added to improve the accuracy and robustness of the model. The new model was based on yield data from combine harvesters, Sentinel-2 L1C data, data on terrain height, weather, soil type, crop rotation and wheat variety giving a total of 293,829 observations from 2016-2021.The main results showed that machine learning (ML) models were able to predict winter wheat yields at field level with a mean absolute error (MAE) of 0.65 and 0.55 t ha^{-1} on May 4th (before third nitrogen application) and July 27th (before harvest) when cross-validating the models with years.

Keywords: Forecast models, nitrogen regulation, geodata, Sentinel-2 L1C

Introduction

Since 2020, Danish farmers have had access to a yield forecast model in the web-based management tool, CropManager ([a]SEGES Innovation P/S, 2022). The model in production uses Sentinel-2 L1C data to predict winter wheat yield four times during the growth season from April to August.
Accurate quantification of expected yield before harvest is important to estimate the absolute nitrogen requirement of the crop. Winter wheat is the most important crop in Denmark, covering an area of 19% of arable land with an average yield of approximately 7.7 t ha^{-1} (Statistics Denmark, 2022 and [b]SEGES Innovation P/S, 2022)
If the farmer/advisor estimates winter wheat yield incorrectly by ±1.0 t ha^{-1} the nitrogen application will be inaccurate by ± 15 kg N ha^{-1} under Danish conditions ([a]The Danish Agricultural Agency, 2022). Consequently, accurate yield estimation is essential for regulating nitrogen application to crop demand to optimize the financial return of crop production (economic optimum) and to minimize discharge to water bodies and the environment.
The current forecast model in CropManager is based on data from 2016 and 2017 sampled in 2018. The data consist of yield data from combine harvesters (1,125 ha), field polygons (106 winter wheat fields) and Sentinel-2 L1C data (13 spectral bands and three indices). The yield potential varies between seasons and depends on soil types, geological origin, wheat variety, cultivation history etc. Consequently, it is important that the forecast model represents the spatial and temporal variation of winter wheats fields in Denmark.

Thus, the objectives of this study were to: 1) increase the accuracy and robustness of the yield prediction model in winter wheat by adding more data and new features to the model and 2) implement the new model in the web-based management tool CropManager used by Danish Farmers.

Materials and Methods

Multiple data sources were combined together to predict winter wheat yield on a calculated grid of 10 x 10 m and on field level. Concerning georefernce, all data were transformed to the WGS84/UTM32N (EPSG:32632) system and afterwards divided into a 10 x 10 m grid resulting in a dataset with 293,829 data points/pixels and a total of 791 features and one target.

Data layers and processing/feature engineering

Yield data: Yield data was collected for the period 2016-2021 from 19 Danish farms using the farmers access to datahubs where yield data from combine harvesters are stored. Data was retrieved at two sampling times; in 2018 and again in 2022. In 2018 yield data from 2016 and 2017 were sampled. In 2022 yield data from 2016-2021 were sampled. The data was cleaned resulting in 287 fields or 2,938 ha of winter wheat data in total from 2016-2021.

The cleaning process consisted of sorting yield measurements by timestamp, removal of measurements (outliers) outside an 1.0-25.0 t ha^{-1} interval, removal of statistical outliers using distance-to-yield ratio (-log(distance/yield)), and finally calculation of a moving average (MA) of the last 10 distance-to-yield ratios and removal of measurements above the MA + 2.5 standard deviations (SD). The last step was repeated four times. Subsequently, the cleaned yield point data were interpolated using the inverse distance algorithm (python library GDAL) and normalized.

Table 1 shows the distribution of data between years. 2018 was a year with severe drought in Denmark which are reflected in the lower average yield this year compared to other years. The yield distribution for 2016 and 2017 respectively shows significant differences in the distributions of the data collected in 2018 compared to data collected in 2022 (analysis not shown).

Table 1. Yield data from combine harvesters from 2016 to 2021 showing the average yield level (t ha^{-1}) with the standard deviation (SD) in brackets, the amount of data (hectares and pixel) and the distribution of data between years, fields and farmers.

Year	**Yield data**				
	Number of Fields	**Number of farmers**	**Hectare**	**Pixels[1]**	**Avg. Yield,** t ha^{-1}
2016[2]	33	7	289	28,898	10.4 (1.8)
2017[2]	95	15	856	22,491	9.8 (2.2)
2018	35	6	356	35,580	6.8 (1.5)
2019	26	5	221	22,062	7.2 (1.3)
2020	29	4	233	23,322	7.3 (1.6)
2021	69	5	984	98,356	7.9 (1.3)
Sum:	**287**		**2,938**	**293,829**	

1) Pixels of 10 x 10 m.

2) Part of the yield data from 2016 and 2017 was collected in 2018 (69 % in 2016 and 74 % in 2017, respectively). The rest of the data was collected in 2022.

Satellite data: Sentinel 2 L1C data from March 9th to July 27th were downloaded for each year from 2016-2021 using the Sentinel-Hub services (Sinergise Laboratory for geographical information systems, 2022 and [a]The European Space Agency, 2022). For each field images with clouds were removed using S2_cloudless algorithm with a cloud threshold of 70% (Sinergise Laboratory for geographical information systems, 2022). Sentinel 2 L1C data consisted of the spectral bands B01, B02, B03, B04, B05, B06, B07, B08, B8A, B09, B10, B11, B12 and the vegetation indices Normalized Difference Vegetation index (NDVI), Normalized Difference Red Edge Index (NDRE), and Modified Soil Adjusted Vegetation index (MSAVI2) were calculated from the bands ([b]The European Space Agency, 2022).

$$NDVI = \frac{(B08 - B04)}{(B05 + B04)} \quad (1)$$

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

$$MSAVI2 = \frac{\left(2\,x\,B08 + 1 - \sqrt{(2\,x\,B08 + 1)^2 - 8\,x\,(B08 - B04)}\right)}{2} \quad (3)$$

The data were linearly interpolated in the time dimension using inverse distance interpolation and then resampled to 14 days' interval for each 10 x 10 m pixel. It resulted in 14 temporal features in the growth season from March 9th to July 27th. Furthermore, for each of the temporal features, the relative change since March 9th was calculated resulting in a total of 336 features from Sentinel 2 L1C.

Terrain Elevation: The Danish Terrain Elevation model (DEM) describes the height of the terrain above sea level in meters and has a resolution of 0.4 x 0,4 m (The Danish Agency for Data Supply and Efficiency, 2022). From the DEM data, the following 5 features were calculated for each 10 x 10 m pixel;

- Height of the terrain above sea level.
- Relative height of the terrain compared to the lowest point on the field.
- Slope percentage: $100\% \cdot \frac{\delta f}{\delta p}$ where *f* is the DEM field and *p* is the direction of the gradient vector, i.e. $\left|\left|\frac{\delta f}{\delta p}\right|\right|$ is the magnitude of the gradient vector.
- Slope angle: $\arctan\left(\left|\left|\frac{\delta f}{\delta p}\right|\right|\right)$ which results in a degree in [0;90].
- Aspect (clockwise orientation of the gradient relative to North): $\arctan\left(\frac{\frac{\delta f}{\delta E}}{\frac{\delta f}{\delta N}}\right)$ which results in a degree in [0:360) with 0 being North. E, N refer to easting and northing.

The gradients were approximated numerically using Scharr kernel (Huisman and de By, 2009). The easting and northing coordinates in WGS84/UTM32N (EPSG:32632) were used as two additional features.

Weather data: Climate variables consisted of air temperature, soil temperature, precipitation and global radiation and were available from the Danish Meteorological Institute's (DMI) Open Data API (Danish Meteorological Institute, 2022). For each field, data from the closest meteorological station was used. Consequently, the same measurements were used for each pixel in a given field. The weather data was aggregated at intervals of 14 days. The mean, standard deviation, minimum, and maximum were calculated for all climate variables resulting in 440 features.

Soil texture: Geodata on soil type describes soil texture in 0-20 cm depth and covers all cultivated land in Denmark. The map is based on approx. 36,000 soil samples distributed throughout the country ([b]The Danish Agricultural Agency, 2022). Soils are divided into eight soil type classes and, for each field, the dominant soil type was used in the model resulting in one feature.

Registration data: Most Danish farmers or their advisors register management practices in the Danish Field database covering 88 % of cultivated land in Denmark ([c]SEGES Innovation P/S, 2022). Information on winter wheat variety was available for 163 of the fields used in the study. Twelve different varieties were registered (Torp, Benchmark, Graham, Informer, Sheriff, Kalmar, KWS Lili, Pistoria, Chevignon, Kvium, KWS Scimitar, Ohia) along with a variety mix resulting in one feature.

Crop type and crop rotation: The Danish Agency for Agriculture displays public geodata on field crop type. Information on crop rotation the last five years was extracted giving an additional six features ([c] The Danish Agricultural Agency, 2022).

Models

Machine learning model: For this study, a *Gradient Boosting* ML algorithm, Catboost (Prokhorenkova *et al.*, 2019), was used. The algorithm uses binary decision trees as base predictors, and it works by combining several decision trees into one model by growing each tree sequentially on the previous tree's residuals.

Prediction dates: If the farmer should be able to regulate the input of nitrogen fertilizer to winter wheat fields using a yield forecast model, multiple predictions dates are of interest. Thus, multiple models were trained providing predictions for harvest dry yield on April 6th (before second nitrogen application), May 4th (before third nitrogen application), June 1st and July 27th (before harvest).

Validation

For validating the model, several validation methods were used. In the first experiment, the data were randomly split such that 85% of the fields were in the training set and 15% were in the validation set.

In experiment 2-5, cross validation was used with entire harvest years as folds. Thus, the data were split based on harvest year resulting in 6 folds of data (2016-2021). For each year, a model was trained on the remaining 5 years of data and evaluated on the hold-out

year not in the training data. The predictions for each hold-out year were saved and later combined, and the metrics was calculated on this combined set.
Due to the large number of features in the model, feature elimination was performed in experiment 2-5 to avoid overfitting. This was performed by iteratively removing the least important features (based on the training set) in 20 steps until only 10-20 features were left in the different models.
To reduce pixel level noise, in experiment 2 and 3, the data was aggregated on field level before training.
Yield data from 2016 and 2017 shows significant differences in the distributions of the data collected in 2018 compared to data collected in 2022. Consequently, experiment 2 and 4 were tested on all data while experiment 3 and 5 were tested on data collected in 2022 only.
The performance of the models was evaluated using mean absolute error (MAE), root mean square error (RMSE), R^2, and standard deviation of absolute error (SD of AE).

Results

Several iterative experiments were conducted. The results are summarized in table 2, where the MAE, RMSE, R^2 and SD of AE on the validation data in t ha^{-1} can be seen for the 5 experiments. The prediction performance is shown on pixel level and summed up to field level. In experiment 1, a limited amount of hyperparameter tuning was performed to avoid a large optimistic bias in the validation set.

Experiment 1. In the first experiment data from a field was either in the training or the validation set. Models were trained on pixel level using all features for all four prediction dates. This resulted in a MAE of 0.67, 0.62, 0.59 and 0.56 t ha^{-1} on April 6th, May 4th, June 1st and July 27th on field level with a R^2 of 0,74-0,83. Model predictions on pixel level resulted in a MAE of 0.97, 0.94, 0.93 and 0.92 t ha^{-1} on April 6th, May 4th, June 1st and July 27th with a R^2 of 0,56-0,61. In general, the MAE decreases with the prediction date. The MAE also decreases when summed up to field level compared to pixel level.
Experiment 2. In experiment 2 cross-validation with years as folds was used as the validation method. Data was aggregated to field level before training and feature elimination was performed removing the least important features in the model. This resulted in MAEs of 0.90 and 0.88 t ha^{-1} for prediction dates May 4th and July 27th, respectively. The R^2 decreased to 0.68-0.69 compared to the models in experiments 1.
Experiment 3. Since yield data (from 2016 and 2017) sampled in 2018 was distinctly different from yield data sampled in 2022 for unknown reasons, a model was trained using only data collected in 2022. The setup was otherwise identical to experiment 2 but with less yield data from 2016 and 2017. This resulted in MAEs of 0.65 and 0.55 t ha^{-1} for prediction dates May 4th and July 27th, respectively. The R^2 increased to 0.72-0.80 compared to experiment 2. The performance of the models improved substantially when data sampled in 2018 was not used in the models.
Experiment 4: The setup for experiment 4 was identical to experiment 2, except that the models were trained on pixel level instead of fields. When summed up to field level the MAE increased to 1.02 and 0.94 t ha^{-1} for prediction dates May 4th and July 27th with a R^2 of 0.65-0.68.
Experiment 5: The setup for experiment 5 was identical to experiment 3, except that the models were trained on pixel level. This resulted in MAEs of 0.71 and 0.68 t ha^{-1} with a

R^2 of 0.66-0.68. Once again, the performance of the models improved substantially when data sampled in 2018 was not used in the models. However, the performance of the models in this experiment decreased compared to experiment 3 where data were aggregated to field level before training.

Discussion

In experiment 1 models were trained providing prediction on four dates from April 6th to July 27th. Even though the prediction accuracy improves the closer to harvest the prediction is done, it is striking that the model on April 6th, with less than a month of satellite and weather data, are able to give a prediction with a MAE below 1.0 t ha^{-1}. The growth condition the rest of the season (precipitation, temperature, and radiation) are normally considered to be crucial to the yield level achieved. However, in this experiment the MAE only decrease with 0.11 t ha^{-1} from a prediction on April 6th to a prediction on July 27th.
In experiment 1 fields from the same year can occur in both the training and validation set, which can give an optimistic bias. Although no 'year' feature is given to the model, other features (such as weather data features) can act as proxy variables for the year by which the model can learn the mean yield level for each year.
The main focus of this study was to develop models that represents the spatial and temporal variation of winter wheats fields in Denmark. Consequently, the ultimate test is the ability of a model to predict winter wheat yield in a year not used for training the model, which was done in experiment 2-5. This validation method produces the least biased estimator of the out-of-sample performance. Petersen et al., 2023 used same validation method in their study in spring barley which resulted in a MAE of 0.38 t ha^{-1} approximately one month before harvest. The study illustrates the potential of yield prediction in cereals. However, Petersen et al., 2023 had nine years of yield data, RVI measurements and weather data from the same field compared to this study with only six years of data and data from all over Denmark.
It has become common practice to collect data from machinery in hubs. In this study yield data was retrieved and cleaned in 2018 and again in 2022, which means that yield data from 2016 and 2017 were a mix from the two sampling times. For unknown reasons the distribution of data differs between the two samplings. This could be due to small variations in the methods used to extract, translate, and clean the yield data even though the processes have been standardized. Additionally, yield data from combine harvesters vary in quality because the practice of calibrating the machinery varies between farmers and harvest year. The farmers were asked for calibrated data only, but some farmers might use more time on calibration than others. The focus on calibrated data might be more pronounced in the data collected in 2022 compared to data sampled in 2018 which could explain the difference between the two datasets. However, because it is unknown what caused the data to differ in distribution, the yield prediction models from experiment 3 and 5 are incorporated in the web-based management tool CropManager in 2023, where Danish farmers can access the models. The models from experiment 3 are used to give a forecast on field level, while models from experiment 5 are used to predict the variation within the field.

Table 2. Results from experiment 1-5 showing the prediction date, features included, total number of observations, split between training and validation data and the prediction performance MAE, RMSE, R^2 and SD of AE.

Experiment	Prediction date	Features	Observations	Split of data	MAE, t ha^{-1}		RMSE		R^2		SD of AE	
					Validation		Validation		Validation		Validation	
					Field	Pixel	Field	Pixel	Field	Pixel	Field	Pixel
1	April 6th	All	293,829	Field level (40 fields in validation set)	**0.67**	0.97	**0.93**	1.26	**0.74**	0.56	**0.65**	0.81
	May 4th				**0.62**	0.94	**0.85**	1.22	**0.79**	0.59	**0.58**	0.77
	June 1st				**0.59**	0.93	**0.79**	1.22	**0.82**	0.59	**0.53**	0.78
	July 27th				**0.56**	0.92	**0.75**	1.19	**0.83**	0.61	**0.50**	0.76
2	May 4th	Aggregate by field + Feature elimination	287	Cross-validation with years as folds	**0.90**		**1.18**		**0.69**		**0.76**	
	July 27th				**0.88**		**1.19**		**0.68**		**0.80**	
3	May 4th	Aggregate by field + Feature elimination	195	Cross-validation with years as folds (only data collected in 2022)	**0.65**		**0.86**		**0.72**		**0.56**	
	July 27th				**0.55**		**0.73**		**0.80**		**0.47**	
4	May 4th	Feature elimination	293,829	Cross-validation with years as folds	**1.02**	1.40	**1.30**	1.86	**0.65**	0.45	**0.73**	1.17
	July 27th				**0.94**	1.36	**1.21**	1.81	**0.68**	0.47	**0.76**	1.18
5	May 4th	Feature elimination	220,644	Cross-validation with years as folds (only data collected in 2022)	**0.71**	1.08	**0.94**	1.42	**0.66**	0.41	**0.62**	0.92
	July 27th				**0.68**	1.07	**0.91**	1.41	**0.68**	0.42	**0.61**	0.92

Conclusion

The results showed that ML models were able to predict winter wheat yield with a MAE of 0.65 and 0.55 t ha^{-1} on May 4th and July 27th respectively when cross-validating with years (only using data collected in 2022). The prediction accuracy on May 4th is acceptable to regulate nitrogen application to crop demand in third application. In 2023 the yield predictions models will be incorporated into CropManager.

References

Danish Meteorological Institute (DMI), 2022. Weather data: https://www.dmi.dk/frie-data/, last accessed December 2022.

[a]The European Space Agency (ESA), 2022. L1C Sentinel 2 product. https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/product-types/level-1c, last accessed December, 2022.

[b]The Eropean Space Agenscy (ESA), 2022. Spectral Resolution. https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/resolutions/spectral.

Huisman, O and de By, R. A. 2009. Principles of Geographic Information Systems.https://webapps.itc.utwente.nl/librarywww/papers_2009/general/PrinciplesGIS.pdf, pp 411-414, last accessed December 2022.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). CatBoost: unbiased boosting with categorical features. Non-peer reviewed preprint at ARXIV.1706.09516

Petersen, C. T., Langgaard, M. K., & Petersen, S. D. (2023). Yield prediction in spring barley from spectral reflectance and weather data using machine learning. Soil Use and Management, 00, 1–13. https://doi.org/10.1111/sum.12902

[a]SEGES Innovation P/S. 2022. CropManager, The Danish program for farmers. https://cropmanager.dk/, last accessed December 2022.

Landsforsøgene 2022. Forsøg og undersøgelser i Dansk Landbrugsrådgivning. Page 13. ISBN 978-87-93051-11-9.https://www.landbrugsinfo.dk/-/media/landbrugsinfo/public/a/5/c/planter_landsforsogene_2022.pdf

[c]SEGES Innovation P/S. 2022. Mark Online (The Danish Field Database) https://www.seges.dk/da-dk/software/plante/mark-online

Sinergise Laboratory for geographical information systems, Ltd, 2022. Sentinel 2 data using the Sentinel-Hub services: https://sentinel-hub.com/.

Statistics Denmark (Danmarks Statistik), 2022. Historisk høje udbytter i høsten. Nr. 396. https://www.dst.dk/Site/Dst/Udgivelser/nyt/GetPdf.aspx?cid=44589

The Danish Agency for Data Supply and Efficiency. 2022. Danmarks Højdemodel - Terræn. https://dataforsyningen.dk/data/930

[a]The Danish Agricultural Agency. 2022. Vejledning om gødsknings- og harmoniregler Planperioden 1. august 2022 til 31. juli 2023. https://lbst.dk/fileadmin/user_upload/NaturErhverv/Filer/Landbrug/Goedningsregnskab/Vejledning_om_goedskning_og_harmoniregler_2022_2023.pdf

[b]The Danish Agricultural Agency. 2022. Vi forbedrer jordbundstypekortet. https://lbst.dk/nyheder/nyhed/nyhed/vi-forbedrer-jordbundstypekortet/

[c]The Danish Agricultural Agency. 2022. Markkort og markblokke. https://lbst.dk/landbrug/kort-og-markblokke/markkort-og-markblokke