STØTTET AF
**Promille**afgiftsfonden for landbrug

# YMP re-training on 2016-2021 data

Yield Map Prognosis

Exported on 12/15/2021

# Table of Contents

**Table of Content**

# 1 Cleaning of yield maps (YMCA)

- It was complicated to use the code from 2019, as it relied on a different data structure and code-environment.
- There is no matching polygon/yield map sets in the timespan 2018-2020 (both included) resulting in no useful data for this period, for 2021 only 7 unique yield maps was found.
- Use of field polygons 'Dansk Mark Database' (DMDB) instead of 'Internet Mark Kort' (IMK) resulted in a vastly different amount of matching polygon/yield map sets, even when adding the 7 new yield maps (100 unique sets in 2019, and 77 now)
- The location of the fields in the old and new set of yield maps also differs, as the new yield maps also covers the area of Randers and Ringkøbing, as seen the these plots:

  Old yield maps: 
  
  - New yield maps: 
  
- **Conclusion**: For this pipeline to be automated, and to ensure an easy and reliable way to retrain and expand the data, this solution needs a lot of work. We strongly recommend a full restructure on a more automated pipeline.
  - Ensure all needed field polygons exists in DMDB - else add the field polygons from IMK to DMDB, such that we can still rely on our SEGES database, instead of an external data source.
  - Heighten the consistency of the yield data, a way to achieve this is by using the ISOXML files from CropManager instead of the current shape files (manually received from Danish farmers).
    - The developer time spend on normalizing the inconsistencies in the data, makes the current setup unfeasible for reuse in coming years.
  - Our opinion is that YMCA should become a data product, owned and maintained by a dedicated data-team, thus avoiding the technical debt which prevented improvements in the current rerun.

# 2 Fetch satellite images (YMP)

- It was (again) complicated to use the code from 2019, as it relied on a different data structure and code-environment.
- Instead of downloading the satellite images for each rerun, we should rely on all ready downloaded images, stored in an easily accessible way - e.g. the 'sat-lake'-solution, which pre-cashes all sentinel 2 images in .zarr format on Azure Data Lake.

# 3 Retraining of YMP

- It was complicated to use the code from 2019, as it relied on a different data structure and code-environment.
- Model score results:

| | OLD - trained on 2016-2017 yield maps | | | | NEW - trained on 2016-2021 yield maps | | | | Difference | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Field overall scores** | April 10[1] | May 10[2] | June 5[3] | August 1[4] | April 10[5] | May 10[6] | June 5[7] | August 1[8] | **April 10** | **May 10** | **June 5** | **August 1** |
| **train_field_mean_mae_kg/ha** | 327.9 | 333.9 | 290.8 | 283.1 | 354.9 | 362.7 | 331.2 | 223.7 | 27,00 | 28,80 | 40,40 | -59,40 |
| **train_field_mean_std_kg/ha** | 451.9 | 454.3 | 419.7 | 381.4 | 472.4 | 472.6 | 458.4 | 301 | 20,50 | 18,30 | 38,70 | -80,40 |
| **train_field_mean_samples** | 79 | 79 | 79 | 79 | 61 | 61 | 61 | 61 | -18 | -18 | -18 | -18 |
| **val_field_mean_mae_kg/ha** | 835.7 | 926.3 | 962.8 | 717.5 | 648.2 | 747 | 608.6 | 674.7 | -187,50 | -179,30 | -354,20 | -42,80 |
| **val_field_mean_std_kg/ha** | 1127 | 1220.7 | 1240.1 | 863.7 | 879 | 969.5 | 799.9 | 859.4 | -248,00 | -251,20 | -440,20 | -4,30 |
| **val_field_mean_samples** | 22 | 22 | 22 | 22 | 15 | 15 | 15 | 15 | -7 | -7 | -7 | -7 |
| **Position overall scores** | | | | | | | | | | | | |

1 http://localhost:5000/#/experiments/4/runs/36759b7dea8045a1ab8ab1ed2b6bc427
2 http://localhost:5000/#/experiments/4/runs/8e24b64ef9194a10aeaa776f1d5cc6ae
3 http://localhost:5000/#/experiments/4/runs/a16a72324a5849819ae37f4e5e7e04ee
4 http://localhost:5000/#/experiments/4/runs/8616813e09c64e44a6e8e3ea3a566f56
5 http://localhost:5000/#/experiments/4/runs/642826c95b66489db8199587bd920d94
6 http://localhost:5000/#/experiments/4/runs/b840288d70a34d8c8e8299dc40d0c130
7 http://localhost:5000/#/experiments/4/runs/cf7a340a973e4accb98a3f1951f6aa7b
8 http://localhost:5000/#/experiments/4/runs/9673b10a2eaf43a9b17a7b576a57039d

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **train_position_mae_kg/ha** | 1410.6 | 1359.8 | 1295.5 | 1223.6 | 1410.5 | 1349.9 | 1283.6 | 1199.2 | | -0,10 | -9,90 | -11,90 | -24,40 |
| **train_position_samples** | 383594 | 383594 | 383594 | 383594 | 295733 | 295733 | 295733 | 295733 | | -87861 | -87861 | -87861 | -87861 |
| **val_position_mae_kg/ha** | 1729.6 | 1714.4 | 1707.5 | 1473.7 | 1462.2 | 1552.2 | 1327.5 | 1271.9 | | -267,40 | -162,20 | -380,00 | -201,80 |
| **val_position_samples** | 81880 | 81880 | 81880 | 81880 | 77411 | 77411 | 77411 | 77411 | | -4469 | -4469 | -4469 | -4469 |

  - We do not see significance changes in any (i.e. field-level or position-level) of the training performance scores. The small changes are likely to be a coincidence, given the changes in the dataset.
  - However, both field-level and position-level validation Mean Absolute Error (MAE) scores are improved significantly (~200kg/ha) for all prediction dates.
  - But due to the smaller sample size we decided to perform a ad hoc cross-validation to conclude if other train/val splits results in the same model improvements.
- 5-fold cross-validations results of val_position_mae_kg/ha (see notebook YMP-678_crossvalidation_of_2016-2021_yield_data.ipynb[9]):

| | **Support (# pixels)** | **04-10_MAE** | **04-10_STD** | **05-10_MAE** | **05-10_STD** | **06-05_MAE** | **06-05_STD** | **08-01_MAE** | **08-01_STD** |
|---|---|---|---|---|---|---|---|---|---|
| **fold_num_1** | 108092 | 1715,81 | 2169.07 | 1745.27 | 2298.67 | 1617.63 | 2143.24 | 1369.40 | 1907.88 |
| **fold_num_2** | 178274 | 1558.85 | 2063.96 | 1493.90 | 2028.16 | 1578.41 | 2035.51 | 1505.94 | 1991.16 |
| **fold_num_3** | 155436 | 1726.96 | 2337.27 | 1722.46 | 2322.18 | 1727.66 | 2307.62 | 1461.70 | 1989.68 |
| **fold_num_4** | 190402 | 2901.53 | 3368.22 | 2791.30 | 3317.64 | 2762.46 | 3348.74 | 2449.48 | 3166.34 |
| **fold_num_5** | 114084 | 1742.17 | 2303.13 | 1708.38 | 2178.57 | 1594.68 | 2100.75 | 1458.95 | 1902.38 |

9 https://bitbucket.seges.dk/projects/DDS/repos/yield_map_prognosis/browse/data_analyses/YMP-678_crossvalidation_of_2016-2021_yield_data/YMP-678_crossvalidation_of_2016-2021_yield_data.ipynb?at=refs%2Fheads%2FYMP-678-cross-validation-notebook-on-new-ymp-models-with-2021-data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Mean over folds** | **---** | **1929.06** | **2448.33** | **1892.26** | **2429.04** | **1856.17** | **2387.17** | **1649.09** | **2191.49** |
| **Scores over total pixels** | **---** | **1987.18** | **2678.49** | **1941.71** | **2647.04** | **1919.75** | **2623.27** | **1710.49** | **2397.65** |
| **Scores over yield maps** | **---** | **1023.69** | **1555.24** | **1138.50** | **1607.36** | **1177.33** | **1705.19** | **902.73** | **1359.97** |

    - We see the MAE for each fold is higher than for the previous trained model, thus the mean over the folds are also significant higher (300-500 kg/ha higher). Thus the specific train/val split used previously benefitted the model, but its performance is not generalizable.
    - The standard deviation (STD) of each prediction error, shows that 68% of all validation samples has less than +/-2300kg/ha in prediction error.
    - However, our cross validation still shows that more data (closer to harvest) improves the model performance.
- **Conclusion**:
    - As a larger, and more consistent, pool of data greatly benefits the model, we think this should be the main focus going forward.
        - A way to achieve this by ensuring all data fetching to the YMP pipeline comes from data products used across SEGES, i.e.  yield data from CropManager.

# 4 Overall conclusion

- Our rerun of the YMCA pipeline resulted in fewer yield maps, i.e. 100 unique fields in 2019, and 77 now. However, the new fields also covers the area of Randers and Ringkøbing, i.e. more divergent soil/climate types.
- Our retrained YMP models on the new 77 yield maps, show models improvements of ~200kg/ha. However, following cross validations, the results were worse than our train/val model. This shows the dataset is not generalizable, thus resulting in better model-performance than expected in real life application.
- The smaller sample size is troubling, but due to the slightly improved train/val model results, we would recommend deploying the new models to production.

# 5 Future work

1. Implement automated pipeline for YMCA based on ISOXML yield maps.
2. Implement YMP training pipeline to fetch all geo-data from SEGES data cache store (Spatio-Temporal Data Store Definition[10] like Sat-Lake)
3. Move YMP training/deployment pipeline to Azure Machine learning.
4. Implement cross-validation in to the YMP training pipeline.

10 https://confluence.seges.dk/display/YMCA/Spatio-Temporal+Data+Store+Definition