EARLY CROP TYPE CLASSIFICATION WITH SATELLITE IMAGERY - AN EMPIRICAL ANALYSIS

Lukas Kondmann[1,2,3], **Sebastian Boeck**[3], **Rogerio Bonifacio**[3], **Xiao Xiang Zhu**[1,2*]
[1]Data Science in Earth Observation, Technical University of Munich (TUM)
[2]Earth Observation Center, German Aerospace Center (DLR)
[3]Research, Assessments and Monitoring Division, United Nations World Food Programme

ABSTRACT

Crop type mapping from satellite images is an essential input for food security monitoring systems. Many approaches focus on mapping crop types based on a full time series of a growing season. However, a variety of use cases require predictions already during the growing season which can be technically challenging. In this paper, we experiment with Sentinel-2 and Planet Fusion data to explore their potential for early season crop type classification at different points in the season. We use high-quality field collections from Germany and South Africa as reference data and find that daily revisit times can be advantageous but are no silver bullet for early season classification of crops.

1 INTRODUCTION

Remote sensing is at the heart of many agricultural monitoring systems to support food security across the globe. Particularly in low and middle-income countries, data on crop types and potential yields are often scarce (Tseng et al., 2021). Satellite images can help to fill these gaps in combination with artificial intelligence algorithms (Nakalembe et al., 2021). Significant methodological progress has been fueled by the increasing availability and temporal coverage of public satellite imagery in AI for Earth observation (EO) (Zhu et al., 2017). In combination with methodological progress in deep learning (LeCun et al., 2015), this has led to a variety of contributions in time-series monitoring for crop type classification, particularly with Sentinel-2 (S2) data (Pelletier et al., 2019; Rußwurm & Körner, 2017; Rußwurm & Körner, 2018; Sainte Fare Garnot et al., 2020). In time-series crop type mapping, the field boundaries of an agricultural field are taken as given and the field is observed throughout the season from space. This input is used to train an algorithm to classify the crop type in the respective field.

More recently, advances in satellite image technology and processing have further complemented this trend by providing data in up to daily intervals. One prominent example of these new generation EO products is Planet Fusion (PF) which generates cloud-free data in daily intervals across the globe. Kondmann et al. (2021) underline the potential of this data source for crop type mapping over the full season. Further, they hypothesize that these kinds of daily observations may be particularly useful in early crop type classification already during the season. This is because the temporal cadence allows an unprecedented density of observations already early in the year at higher spatial resolution. Early classification is a topic of particular interest since applications of crop type mapping can be time-critical. Methodological explorations to classify as early as possible (Mori et al., 2017; Rußwurm et al., 2019) exist but the task remains challenging.

In this paper, we therefore explore opportunities of S2 and PF data for early crop type classification with datasets from Germany and South Africa. We sample time-series imagery up to a variable day in the year for training and inference to simulate within-season classification. We rely on a Multi-Scale(MS) ResNet as temporal model for PF data and compare the results to a Random Forest (RF) baseline with S2. We find that the higher temporal density of observations with PF does not always automatically lead to better early-season performance compared with S2 with the models we tried. Early season performance is better with MSResNet based on PF than RF with S2 in the German

*Corresponding author: xiaoxiang.zhu@dlr.de.

dataset but the edge is largest at the end of the season rather than at the beginning. On the contrary, MSResNet struggles to keep up with the RF S2 model in the South African dataset generally but performance is similar very early in the season. Tailored methodological approaches may hence be necessary to exploit the full potential of next-generation data sources for early-season crop type mapping.

2 DATA

We explore early season crop type classification with two datasets. One from Germany where we use the DENETHOR dataset (Kondmann et al., 2021) and one from South Africa. Both datasets are structurally similar as they are designed for crop type mapping with the same input data although with different classes and geographies. Both were introduced in the recent AI4 Food Security challenge hosted by the European Space Agency (ESA).[1]

Crop Data. As targets, the dataset includes a collection of fields with a multi-class crop type label associated with it. The reference data in DENETHOR is collected as part of the Common Agricultural Policy of the European Union. The DENETHOR dataset (Kondmann et al., 2021) aggregates crop types into 9 classes: Wheat, Rye, Barley, Oats, Corn, Oil Seeds, Root Crops, Meadows, and Forage Crops. Frequencies are reported in the DENETHOR paper. The South African field data was collected as part of an agricultural census by the government and is provided by the Radiant Earth Foundation. In contrast to the German data, the field types are aggregated into 5 classes with respective training and test frequencies: Wheat (train: 437; test: 296), Barley (194;465), Lucerne/Medics(678;1114), Canola (231;281) and Small Grain Gazing(153;280). Therefore, there seems to be a significant distribution shift from training to test set with the SA data.

Satellite Imagery. As inputs, the dataset contains multispectral satellite images from two different sources: Sentinel-2 (S2) and Planet Fusion (PF). S2 data is publicly available as part of ESA's Copernicus Program and collects images across all continents at most every five days (Drusch et al., 2012). The spatial resolution per pixel is 10m and 13 spectral channels are collected. We use S2 data at preprocessing level L2A.[2] For S2, we download all available images but filter them at training time. We allow up to 40% cloud cover per scene unless this results in no images available. This can happen early in the season and we adjust the threshold to 80% in these cases. In contrast, the PF product provides daily data at 3m resolution per pixel. Clouds and shadows are already removed and missing gaps are filled from the next available point in time. Hence, such products are also referred to as analysis-ready because of the high level of preprocessing. The images have four channels: RGB and a near-infrared (NIR) channel. Therefore, the PF data comes in higher spatial and temporal but with lower spectral resolution as the number of bands is significantly lower compared to S2.

Table 1: Dataset Overview

	Fields Train	Fields Test	Year	Timeframe	# PF	# S2
Germany	2533	2028	2018	Jan - Dec	365	144
South Africa	1715	2436	2017	Apr - Nov	244	76

Time Series Characteristics. Table 1 gives an overview of the two areas of interest. Both datasets are geographically split into a train and test tile. The German dataset chooses a location north-east of Berlin in the state of Brandenburg as an area of interest with 2533 fields in the training tile and 2028 fields in the test tile with crop data and imagery from 2018. The South African fields are located east of Cape Town and reports are available for 2017 with 1715 fields in the train and 2436 fields in the test tile. Note that both of the tiles used in this paper were part of the training set in the AI4 Food Security challenge. We adjust the split because the actual test data is not yet public. The challenge dataset also limits the temporal coverage from April to November for South Africa whereas the

[1] https://platform.ai4eo.eu/ai4food-security-germany

[2] For more details, see: https://docs.sentinel-hub.com/api/latest/data/sentinel-2-l2a/

full year is available for Germany. This leads to a discrepancy of available scenes between the two areas of interest with 365 (GER) vs. 244 (SA) PF scenes and 144 (GER) vs. 76 (SA) S2 scenes respectively.

3 TASK AND METHODS

Crop type mapping algorithms often exploit temporal differences of vegetation activity by crop type throughout the season. Figure 1 shows an example of these differences. It plots the Normalized Difference Vegetation Index (NDVI) obtained from the Planet imagery for selected fields throughout the season for Germany (top) and South Africa (bottom). Note that the temporal x-axis is slightly different between top and bottom since the covered timeframes do not match exactly. In the top row, one wheat (thick blue) and meadow (thick orange) field are selected and plotted throughout the season. Other fields of the same crop type are shown in thinner lines in the background. Particularly from day of year (doy) 140 on, vegetation activity differs significantly. Wheat fields seem to be harvested between doy 170-220. On the other hand, meadows experience only a slight decrease in vegetation activity and rebound already before doy 220. These temporal differences can be exploited by machine learning models but as the bottom row for South Africa shows the patterns are often local. This is intuitive since the vegetation activity depends largely on the weather as well as the climate where spatial variation can be large. The bottom row is identical to the top row, just for South Africa and with lucerne instead of meadows. Wheat seems to be quite active between June and September whereas lucerne have low but constant vegetation activity.

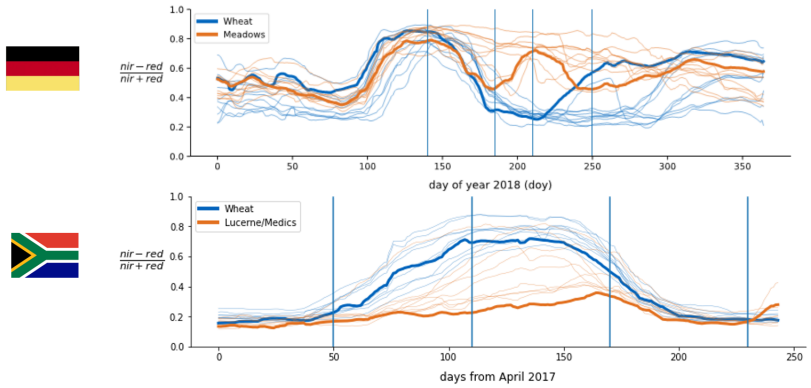

Figure 1: NDVI of selected fields over the season based on Planet Fusion imagery. Germany is in the top row, South Africa at the bottom. Clear differences in the temporal progressions between crops and both countries are visible. Blue markers underline different phases of the growing season.

Although the patterns vary widely between the two areas of interest, Figure 1 gives rise to the question of how much of the time series is actually needed to classify local fields properly. Classification early in the season can be advantageous since it is an essential input to anticipate an over- or undersupply of certain crops. We simulate the early classification task in the following way: On a day x in the season, the crop types for the train tile are known to the algorithm but the test tile is unknown. This is not an implausible scenario since practitioners often collect a small sample of field data in near-real time that can be used for modeling. All images up to the day x are used for training and inference. In the language of Figure 1, we are asking how much performance drops if we only include the time series up to a certain doy as input.

We pick different algorithms for the S2 and PF data since Kondmann et al. (2021) note that the number of fields is not large enough to train a deep learning model based on S2 alone on DENETHOR. Following Kondmann et al. (2021), we use Random Forests (RF) with sklearn defaults to estimate crop types with features derived from the S2 time-series. We compute the min,max,mean,median and standard deviation of each band and the NDVI across time as input. As this specification prioritizes spectral over spatial and temporal resolution, this approach favors the spectral depth of S2 (Kondmann et al., 2021). On the contrary, the Planet Fusion data is exploited best in conjunction with a deep learning model (Kondmann et al., 2021). We pick the Multi-Scale (MS) ResNet

model as temporal encoder which was originally proposed by Wang et al. (2018) for 1D pose estimation and used successfully as a temporal backbone in the BreizhCrops (Rußwurm et al., 2020) and DENETHOR repositories (Kondmann et al., 2021). The MSResNet is trained until convergence with a negative log-likelihood loss and a learning rate of 10^{-3} with weight decay of 10^{-6}. We compare accuracies and a macro-averaged F1 score as criteria since accuracies have limitations with imbalanced data. Macro-average F1 simply takes the unweighted average of F1 score per class. As a function of true positives (TP), false negatives (FN) and false positives (FP), F1 is defined as $F1 = TP/(TP + 1/2 * (FN + FP))$.

4 RESULTS

For Germany, where the full year is available, we compare results after day of year (doy) 50,100,130,175,250 and after the full year. The growing season starts around day 100. In South Africa, we use days 10,15,30,75,150 and the full season for comparisons as the dataset only starts with the growing season in April. So day 10 would correspond roughly to doy 100. Figure 2 plots the respective accuracies and macro averaged F1 scores with Germany on the left and South Africa on the right. Because of their better temporal coverage, the PF data has a window at the beginning of the time period where no usable S2 observation is available. This is before doy 60 in GER and day of season 15 in SA mostly due to clouds. For Germany on the left, the MSResNet with PF reaches an accuracy of 30% and an F1 Score of 18% after 50 days which is fairly low. This is not surprising since the time window until mid-February is still before the start of the growing season. The MSResNet improves notably with the length of the time series included with the largest gains in the beginning of the season. Although the accuracy does not increase beyond doy 250, the F1 score still does and reaches its peak with the full season at 70%. This shows a better balance in the predictions per class compared to doy 250.

The RF model with S2 starts with 48% accuracy and 31% F1 at doy 100 which is comparable in accuracy to MSResNet and slightly lower in F1. At doy 130, both models still have similar accuracies at 60% but MSResNet is significantly better at handling the class imbalance with an F1 score of 52% vs 44%. MSResNet has an edge over RFs in both accuracy and F1 around doy 175. Afterwards, the S2 performance flattens and even slightly decreases towards the end of the season while the curve for MSResNet continues to increase. A reason for the decline for S2 might be that the temporal features such as mean NDVI over time stabilize after a certain point in the season. Hence, the marginal impact of new observations on the features is decreasing in the length of the time series. After the full season, there is a discrepancy in 18 percentage points (p.p.) in terms of F1 score which is significant. So MSResNet is available to make predictions earlier and has a slight edge early in the season that continues to grow throughout the year for the DENETHOR dataset.

The picture differs significantly, however, on the right for the SA dataset. Generally, performances are lower even though there are five instead of nine classes. As the SA dataset already starts in the growing season, performances start at a relatively high level without the slow climb on the left. The first model is again PF-based since a usable S2 observation is not yet available after 10 days because of clouds. This initial MSResNet model achieves 45% accuracy and 18% F1. After 15 days, the first comparison of RF and MSResNet is possible. RF have slightly higher accuracy with 48% vs 47% but a lower F1 score with 23% vs. 25%. MSResNet performance decreases slightly towards 30 days, however, with 44% accuracy and a constant F1 score. On the other hand, the scores of RF increase to 49% accuracy and 26% F1. From this point in time on, RF performance is superior to MSResNet with the largest gap in the middle of the season around day 150. The gap gets smaller towards the end of the season but remains significant at 3 p.p. accuracy and 7 p.p. F1 at day 250.

While MSResNet briefly has an edge over RF at the very beginning of the season in terms of F1, this advantage disappears quickly. It seems like the spatial and temporal depth of Planet Fusion data can not be fully exploited by the MSResNet model in the SA case. Compared to the fairly simple RF model the performance seems rather weak on the test set. One reason for this might be a shift in the class frequency between train and test tile which is comparably large for SA. Potentially the highly parameterized model may be particularly vulnerable to the class balance in the training set.

In summary, the MSResNet model did not manage to underline a significant advantage in early classification with PF data compared to a RF S2 baseline. In Germany, this is because the advantage was small early in the season but in fact became larger throughout the season. In South Africa,

performance was comparable in the beginning but fell off later in the season. Our results also show that even an architecture that is trained from scratch in two areas of interest may perform quite differently in those areas. Spatial generalization, even with labels and training from scratch, is therefore far from guaranteed in AI for Earth Observation.

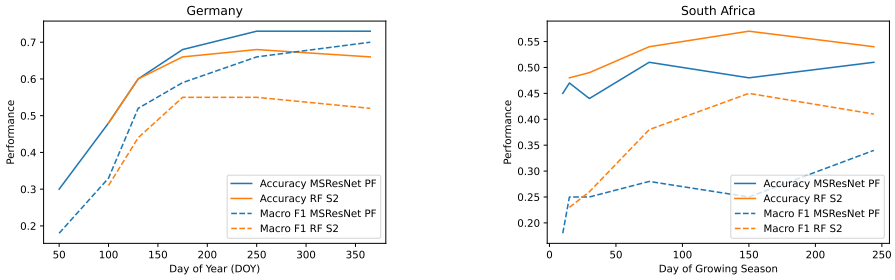

Figure 2: Accuracies and Macro-averaged F1 scores for Crop Type Classification in Germany (left) and South Africa (right). The x-axis shows until which day in the season satellite imagery was used for the respective score. MSResNet with PF data does better than the RF S2 baseline in Germany but is worse in South Africa. This is also the case early in the season but differences in performance tend to be smaller.

5 Discussion

We notice a significant performance drop between DENETHOR and the SA dataset, particularly for the MSResNet model. This may be related to class imbalances between the training and test set in SA but other factors could also be relevant. It might be that the lower number of training fields and shorter overall time series limit the possibilities of the model. However, taking a shorter time series in GER still leads to notably higher scores so the length of the time series is likely not the deciding factor. Further, it may be that the choice of model fits particularly well to the temporal progression with stark seasonal contrast in GER. Potentially the vegetation signal can not be extracted in a similar capacity by the model in the fairly mild climate around the area of interest in SA.

Our study has several important limitations. First, our empirical comparison is limited by the methodological choices made. Particularly, other deep learning models such as TempCNN (Pelletier et al., 2019) or PseTae (Sainte Fare Garnot et al., 2020; Garnot & Landrieu, 2020) that also perform well on DENETHOR may improve scores in SA significantly. Second, one could assume that designated models for early classification (Mori et al., 2017; Rußwurm et al., 2019) may be better suited for this kind of analysis and we aim to explore this in future work. Third, the geographic scope of our analysis is small as the datasets we use prioritize temporal and spatial resolution over geographic extend. Therefore, the findings could be specific to the tiles in these datasets and may fail to generalize.

6 Conclusion

In this paper, we analyze the ability of selected methods to predict crops earlier in the season with PF and S2 data in Germany and South Africa. This is important in practice since it allows policy-makers rapid damage assessment in the wake of hazards. We hypothesize that daily PF data may be particularly helpful early in the season compared to S2 which has a lower revisit time. Our results with an MSResNet and PF show that performance early in the season is slightly better in Germany and slightly worse in South Africa compared to a RF S2 baseline. Over the season, these discrepancies become larger rather than smaller in both locations. Therefore, the early season classification abilities of certain methods and input combinations seem to be driven by a variety of factors that do not depend on temporal resolution alone. While potentially helpful in many cases, we conclude that satellite coverage in near-real time is not necessarily a silver bullet for early-season crop type mapping.

7 ACKNOWLEDGEMENTS

This work is jointly supported by the Helmholtz Association through the joint research school "Munich School for Data Science - MUDS" and the German Federal Ministry for Economic Affairs and Energy (BMWi) under the grant DynamicEarthNet (grant number: 50EE2005).

REFERENCES

Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012.

Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite image time series. 2020. URL http://arxiv.org/abs/2007.00586.

Lukas Kondmann, Aysim Toker, Marc Rußwurm, Andrés Camero, Devis Peressuti, Grega Milcinski, Pierre-Philippe Mathieu, Nicolas Longépé, Timothy Davis, Giovanni Marchisio, et al. Denethor: The dynamicearthnet dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Usue Mori, Alexander Mendiburu, Sanjoy Dasgupta, and Jose A Lozano. Early classification of time series by simultaneously optimizing the accuracy and earliness. *IEEE transactions on neural networks and learning systems*, 29(10):4569–4578, 2017.

Catherine Nakalembe, Inbal Becker-Reshef, Rogerio Bonifacio, Guangxiao Hu, Michael Laurence Humber, Christina Jade Justice, John Keniston, Kenneth Mwangi, Felix Rembold, Shraddhanand Shukla, Ferdinando Urbano, Alyssa Kathleen Whitcraft, Yanyun Li, Mario Zappacosta, Ian Jarvis, and Antonio Sanchez. A review of satellite-based global agricultural monitoring systems available for africa. *Global Food Security*, 29:100543, 2021. ISSN 2211-9124. doi: https://doi.org/10.1016/j.gfs.2021.100543. URL https://www.sciencedirect.com/science/article/pii/S2211912421000523.

Charlotte Pelletier, Geoffrey I Webb, and François Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5):523, 2019.

Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4):129, 2018.

Marc Rußwurm and Marco Körner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 11–19, 2017.

Marc Rußwurm, Sébastien Lefèvre, Nicolas Courty, Rémi Emonet, Marco Körner, and Romain Tavenard. End-to-end learning for early classification of time series. *arXiv preprint arXiv:1901.10681*, 2019.

Marc Rußwurm, Charlotte Pelletier, Maximilian Zollner, Sébastien Lefèvre, and Marco Körner. Breizhcrops: A time series dataset for crop type mapping. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences ISPRS (2020)*, 2020.

Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12322–12331. IEEE, 2020. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.01234. URL https://ieeexplore.ieee.org/document/9157055/.

Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. Cropharvest: A global dataset for crop-type classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Fei Wang, Jinsong Han, Shiyuan Zhang, Xu He, and Dong Huang. Csi-net: Unified human body characterization and pose recognition. *arXiv preprint arXiv:1810.03064*, 2018.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.