July 4–8, 2021, Évora, Portugal

New Approaches for Improving Nitrogen Efficiency based on Clustering Algorithms

Philipp Kastenhofer [a, *], Peter Prankl [a], Peter Riegler-Nurscher [a], Johann Prankl [a]

[a] Josephinum Research, Rottenhauser Straße 1, Wieselburg, Austria

* Corresponding author. Email: philipp.kastenhofer@josephinum.at

Abstract

Site-specific fertilization attempts to address the heterogeneities within a field by supplying plants with their respective nutrient requirements while also taking account of the heterogeneous yield capacity of the soils. This is intended to increase efficiency which, in addition to economic advantages, also brings ecological benefits.

To determine the actual nutrient requirements of plants, data from the local site and the vegetation condition must be interpreted and processed. Among other things, the use of satellite-based multispectral images (Sentinel-2) is being investigated and applied for the site description.

Fertilization models often assume a compensatory site which means that poorly developed areas of cereal fields in the spring can be compensated with a higher amount of fertilizer. In years with dry periods, especially in soils with low water holding capacity, the possibility of misinterpretation arises and thereby increases the ecological risks. These low-yield sites must be excluded from the existing system and treated separately. For this purpose, historical vegetation patterns and anomalies are examined, and corresponding areas are identified with the help of clustering algorithms like k-means clustering.

Low-yield areas could be detected in the fields of supporting farmers with an F1 score of 0.4. There were mainly differences in the expression of the areas, while the location was mostly well detected. This value was achieved by using the vegetation index NDVI with data from April to July, a modified, more robust distance function and with three clusters. In general, it has been shown that a smaller number of clusters leads to better results as the low-yield cluster here is very different from the other clusters.

Keywords: site-specific fertilization, precision farming, machine learning, remote sensing

1. Introduction

Nitrogen plays a major role in crop production from both a productivity and an environmental perspective. Nitrogen is an essential component of proteins and other compounds and is converted into large quantities in the agricultural nutrient cycle. These conversion processes can be associated with high losses. Nitrogen translocations to water bodies and outgassing into the air cannot often be avoided and contribute to environmental problems. Especially in areas where nitrogen fertilizers are used on a large scale (such as in America, China or Europe), these losses create an ecological risk (Basso *et al.*, 2019) Therefore, plants must be supplied with nutrients according to their needs and adapted to their location in order to keep these losses as low as possible and to keep the N balance at equilibrium (Spiess and Richner, 2005). One approach is site-specific fertilization which involves measuring the heterogeneities of a field and then adapting the amount of nutrients depending on the respective yield potential and the current nutrient requirements of the subfield (Isensee, Thiessen and Treue, 2003)

Sensor based crop measurements serve as the basis for nitrogen fertilization models. A correlation is estimated between the N uptake of the crop (determined via biomass measurements and N content analyses) and the measured sensor data (for example, from multispectral images from an UAV or a satellite)(Söderström *et al.*, 2017). Repeated measurements over time and regression analyses can be used to record the N uptake over time(Sharif, 2020). Depending on the yield expectations, a growth-adjusted amount of N can be administered through an idealized N uptake curve and a measured sensor value. An N uptake that is too high or too low can be detected and then compensated for with fertilization. The different yield potentials within a field must be taken into account by adjusting the uptake curves. Yield potentials can either be determined by the soil information and/or multi-year multispectral reflectance measurements(Yuzugullu *et al.*, 2020). It is especially important that low-yield areas be excluded.

In the existing fertilization model, it is assumed that the N uptake of the site can be compensated for, i.e. partial areas in winter wheat that are poorly developed in the spring (Spicker, 2016), by higher fertilization rates at the first application. In years with dry periods or at very sandy sites, there is however a risk of misinterpretation. Poorly developed sites might not be capable of being compensated with higher fertilization rates. To illustrate this, Figure 1 compares a compensable site with a non-compensable site. It can be seen that the non-compensable site contains sub-areas that are always under-developed and cannot be compensated for in the current growing season. However, the deficit in the sub-area of the compensable site is almost compensated for by May.

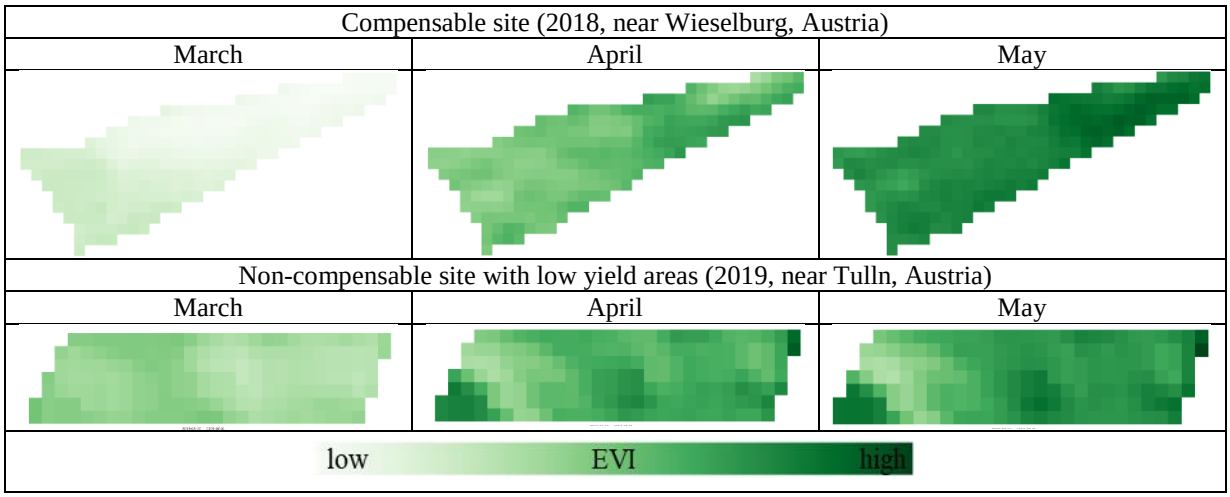


Figure 1. Comparison of the vegetation index EVI of a compensable site with a non-compensable site with winter wheat at three points in time

Low-yield sites must be detected and treated separately by excluding them from the existing fertilizer system regardless of the crop. In this research, only data from the winter wheat were used for a corresponding zonation. The mapping of low-yield sites can be used in crops other than winter wheat as well. The prerequisite for this is that a historical winter wheat vegetation course of the site must be used for zoning.

2. Materials and Methods

Unsupervised learning methods like clustering are used for the detection of low yield areas. An extensive dataset serves as the basis to derive sensor data over the vegetation periods from Sentinel-2 images. Clustering algorithms will be used to detect and interpret the patterns and the structures within the sensor data. Validation will be performed in pilot plots by comparing the model estimation with the farmers' assessments. Additionally, the consistency of the results over several years will be investigated.

2.1. Creating the dataset

In order to carry out a cluster analysis, a data set was created that represents the growth of plants as generally as possible and allows clustering into the main growth patterns. To ensure this, winter wheat fields in the main arable regions in Austria (Upper Austria, Lower Austria, and Burgenland) were observed from 2018 to 2020. The data set INVEKOS plots Austria 2018 to 2020 (Agrarmarkt Austria, 2020) was used as the basis for selecting the fields. To avoid border effects, a buffer of -20 m was first created for each observed field. From the resulting polygon, a random point was chosen which lies within the polygon. For the training data set, 95,235 random points in Austria were selected (see Figure 2 a). A similar approach was used for the test dataset. However, some of the fields were selected specifically in pilot farms. When possible, sensor data was selected from the same field over two years to assess the consistency of the cluster results. In addition, rather than selecting one random point in the field, multiple points were selected in a 20 x 20 m grid arrangement (see Figure 2 b). Vegetation trajectories in the growing season of winter wheat (February to July) were formed for these selected points. For this purpose, the vegetation indices EVI, MSAVI2 and NDVI were calculated from Sentinel-2 data. The atmospheric correction and cloud detection was done with the tool Sen2Cor. Only images with a cloud-probability of 0 were used. To obtain a uniform temporal resolution and to simplify the further processing steps with a uniform input vector, the data was linearly interpolated to a grid of 15 days (see Figure c).

a) b)

year	id	idp	02-16	03-03	03-18	04-02	04-17	05-02	05-17	06-01	06-16	07-01	07-16	07-31	geometry
2019	540	688	0.261897	0.276435	0.329420	0.401629	0.563764	0.675833	0.732078	0.822661	0.821004	0.670011	0.237132	0.233410	MULTIPOINT (446379.468 476098.537)
2019	632	816	0.322201	0.275323	0.375910	0.490433	0.632544	0.721569	0.784494	0.847419	0.763430	0.455773	0.146230	0.173269	MULTIPOINT (583564.253 518553.527)
2019	1087	1540	0.295098	0.328827	0.421391	0.560957	0.728458	0.853641	0.847284	0.851555	0.685834	0.339712	0.217985	0.162311	MULTIPOINT (628345.275 443537.261)
2019	1283	1808	0.340247	0.330125	0.490328	0.650531	0.790783	0.813364	0.864623	0.915883	0.653522	0.254515	0.167053	0.147356	MULTIPOINT (629309.092 477505.659)
2019	1353	1899	0.397068	0.471790	0.546512	0.621234	0.792974	0.795177	0.768310	0.741442	0.832977	0.604610	0.219660	0.178343	MULTIPOINT (463565.995 474953.295)

c)

Figure 2. a) Random section of the point selection for the training dataset with a buffer of -20 m; b) gridded field from the test data set with a buffer of -20 m; c) ready dataset excerpt

2.2. Filtering incorrect vegetation trends

Although all satellite images with a cloud probability > 0 were sorted out during the creation of the dataset, there are still erroneous vegetation trends (especially in the year 2020). These erroneous trends must be removed in order to achieve a good clustering result. For this purpose, erroneous courses were manually annotated in both the training data set and the test data set (training data set annotation n = 162). Subsequently, a support vector classifier was trained (cost factor = 5, gamma = 0.6, kernel = radial basis function). Inaccurate trajectories could be detected reliably with an accuracy of >90%.

After filtering out the incorrect vegetation trends with the support vector classifier, 49,613 to 50,460 instances remained depending on the vegetation index. Especially in 2020, many instances were discarded due to the frequent occurrence of bad weather periods with a high cloud coverage.

2.3. Cluster analysis

The k-Means method, implemented in scikit learn, was used for clustering. Several indicators were decisive for the result, including the hyper-parameters cluster number, vegetation index, number of dimensions (= number of days) and the distance function for the later assignment of the test instances to a cluster. The Euclidean distance is usually used as distance function for k-Means clustering. A new distance function was implemented to get a more robust distance function with respect to the outliers. For this purpose, the L1 distance from the cluster centroid to the instance of the test data set was calculated for each dimension and the maximum distance was discarded. The remaining dimensions are used for the distance calculations to the clusters where the test instances are assigned to the clusters based on the lowest Euclidean distance.

2.4. Cluster validation

For the validation of the clustering results, we introduced two approaches called "Permanence" and "Farm." In the "Permanence" approach, the clustering results of a point from 2018 and 2019 were compared. The resulting clusters were thereby converted into a binary categorization (0...no low-yield point, 1...low-yield point). Finally, the confusion matrix was formed to calculate the Precision, Recall, F1 score and Accuracy. In the "Farm" approach, three farmers from Lower Austria were asked to annotate their areas for low-yield sites. Again, the cluster result was converted into a binary categorization (0...no low-yield site, 1...low-yield site). The annotated areas were compared with the cluster result. From this a confusion matrix was formed to calculate the metrics Precision, Recall, F1 score and Accuracy.

The number of instances of the datasets was 32,108 for the permanence approach and 7,212 for the farm approach after filtering out the incorrect instances.

2.5. Grid search hyper-parameter optimization

All possible variants were calculated in a grid-search analysis to evaluate the effects of the hyper-parameters on the cluster result and to finally find the best variant. The performance metrics of the “Farm” evaluation and the “Permanence” evaluation were determined in each case. The hyper-parameters vegetation index (NDVI, EVI, MSAVI2, NDVI+EVI), distance function (normal, robust), temporal dimensions (April to July, May to July) and the cluster number (3, 4, 5, 6) thus result in 64 variants.

3. Results and Discussion

The following chapter discusses the results of the grid search based hyper-parameter optimization of the “Farm” and “Permanence” evaluation.

3.1. F1 score

Figure 3 shows the variant with the best “Farm” F1 score. The variant is based on the vegetation index NDVI from April to July with the robust distance function and three clusters. In the line graph, the clusters centroids are shown. The low-yield cluster can be clearly differentiated from the other clusters. The low-yield points were drawn by the farmer with a red line in the selected fields. Some low-yield points were also classified as low-yield points by the estimation (=points). The sensitivity of the variant is slightly too low to fully detect even the small areas as shown in the right panel.

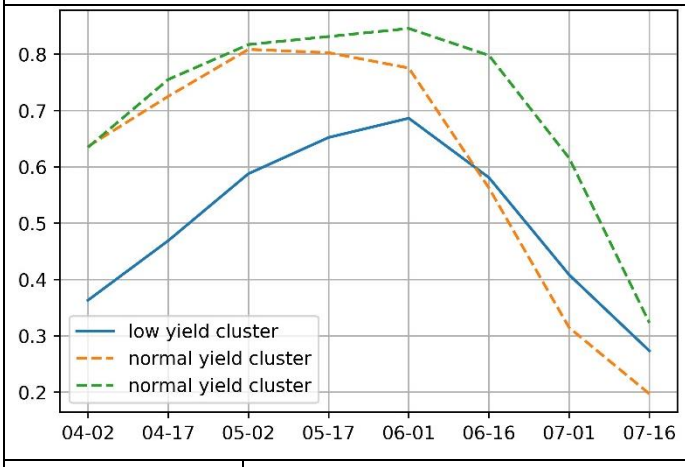
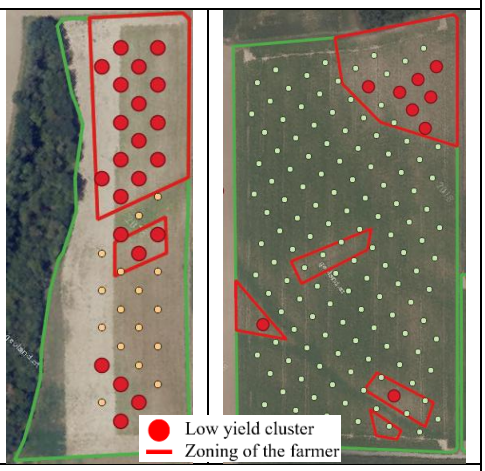
Vegetation-Index: NDVI Dimensions: April to July		Clusters: 3 Distance function: Robust
		
	Permanence	Farm
Precision	0,59	0,33
Recall	0,40	0,52
F1 score	0,48	0,40
Accuracy	0,84	0,71

Figure 3. Variant with best Farm F1 score: top) variant description; upper left) cluster centers; upper right) example fields of farm dataset; bottom) performance metrics of permanence and farm approach

Figure 4 shows the variant with the best "Permanence" F1 score. It is based on the MSAVI2 index from May to July with three clusters and the robust distance function. It can be seen that the lower yield cluster stands out less from the other clusters than the previously presented variants in the line plot. It can also be observed that in the sample plots the low-yield locations cannot be satisfactorily detected. It is also noticeable that the "Permanence" Precision is very close to 1 with 0.98 while the "Farm" Precision is very low with 0.17. The Farm F1 score is also rather low.

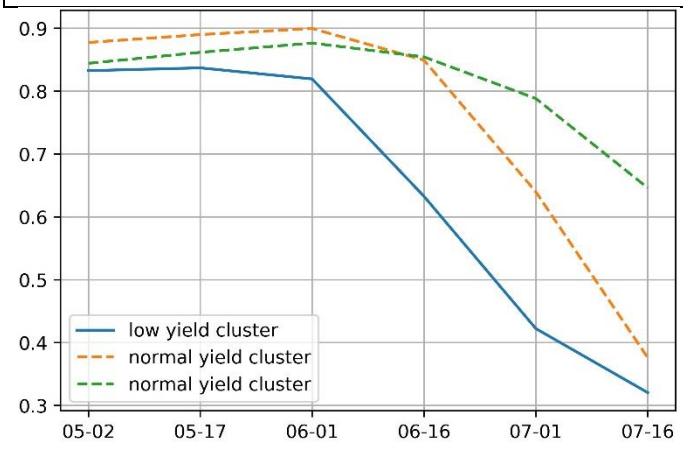
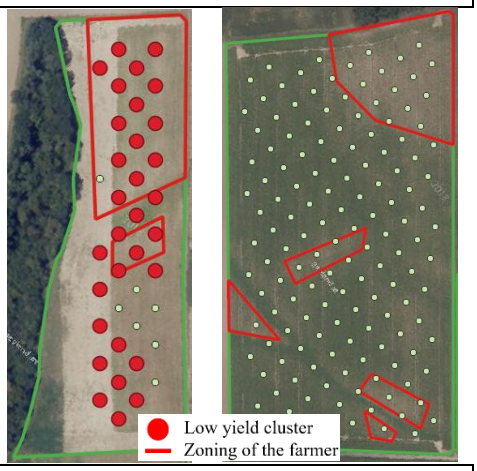
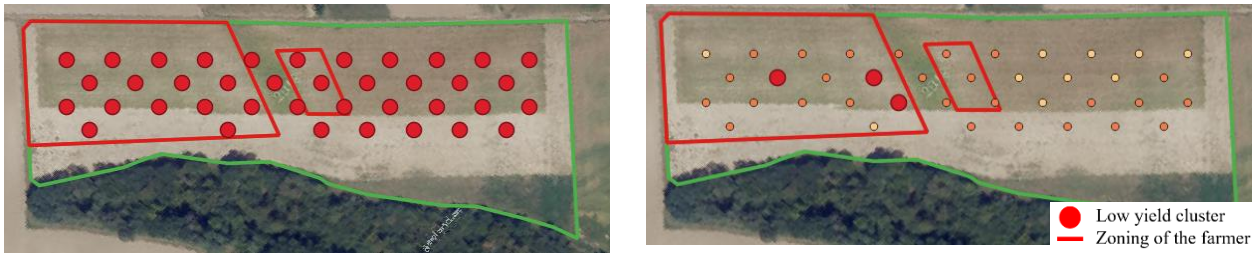
Vegetation-Index: MSAVI2 Dimensions: May to July		Clusters: 3 Distance function: Robust
		
	Permanence	Farm
Precision	0,98	0,17
Recall	0,37	0,35
F1 score	0,53	0,23
Accuracy	0,44	0,56

Figure 4. Variant with best Permanence F1 score: top) variant description; upper left) cluster centers; upper right) example fields of farm dataset; bottom) performance metrics of permanence and farm approach

3.2. Other metrics (Recall, Precision, Accuracy)

It turns out that the F1 score is possibly the best metric to rank the performance of the cluster analysis. Accuracy is not a good metric since the dataset is very unbalanced. Depending on the variant, the number of true negatives is about 5 to 10 times larger than the true positives. The recall would serve as a good metric if the low-yield areas could be reliably detected and if a classification of normal areas into a low-yield area is subordinate. However, variants with a high recall are somewhat too sensitive to low-yield points, and often entire fields are detected as such (see Figure 5 a). When optimizing for the precision metric, the exact opposite is the case. Hardly any areas are detected as low yield areas (see Figure 5 b) because the sensitivity of variants with high precision is too low.



a) b)

Figure 5. a) selected field with best farm recall (EVI, robust distance function, 3 cluster, May to July); b) selected field with best farm precision (NDVI, robust distance function, 6 cluster, April to July)

3.3. Influence of the individual hyperparameters on the result

The next step (in our research) involved grouping the results of the variants according to their vegetation index, distance function, temporal dimensions, and number of clusters to draw further conclusions on the influence of the parameters. The mean of the F1 scores was calculated from the resulting groups. There are no significant differences in the grouping according to the vegetation index and the distance function. The other two hyperparameters, dimension and number of clusters, influence the results. For the temporal dimensions, the variants using data from April to July generally perform better (mean “Permanence” F1 score: 0.43, mean “Farm” F1 score: 0.32). Those variants that use results from May produce lower F1 scores (mean “Permanence” F1 score: 0.35, mean “Farm” F1 score: 0.28). When grouped by the number of clusters, variants with a lower number of clusters perform better. Variants with 3 clusters produce the best results (mean “Permanence” F1 score: 0.43, mean “Farm” F1 score: 0.34). For example, variants with 6 clusters produce a mean “Permanence” F1 score of 0.35 and a mean “Farm” F1 score of 0.24. As the number of clusters increases, it is observed that the low-yield cluster no longer differentiates itself as strongly from the other clusters as in variants with fewer clusters.

3.4. Example fields

Figure 6 gives some overview of the zoning of the variant with highest F1 score (=0.4). It is observed that the location of the low yield points can be detected and that there is a partial correlation between the estimated zones and the zoning of the farmer. However, there are also the following error factors: low-yield points that are located in the border area of the field (e.g., in field 1 and field 3) cannot be detected due to the border area correction. The low-yield sites detected by the cluster analysis often have a different expansion than the areas classified by the farmer as indicated by a decrease in the F1 score. The areas classified by the farmer are a subjective assessment and may differ from the actual location and expansion under certain circumstances. In addition, there is also the issue that the 20 x 20 m data set has a low spatial resolution. Thus, there may be an influence by adjacent areas with higher yield potential in the border region of the low-yield area classified by the farmer. It is important to take note that each point is assigned to a cluster after only one year of observation which can lead to year-specific influences, such as weather, management practices, varietal influence, diseases, and nutrient deficiencies (not nitrogen-related).

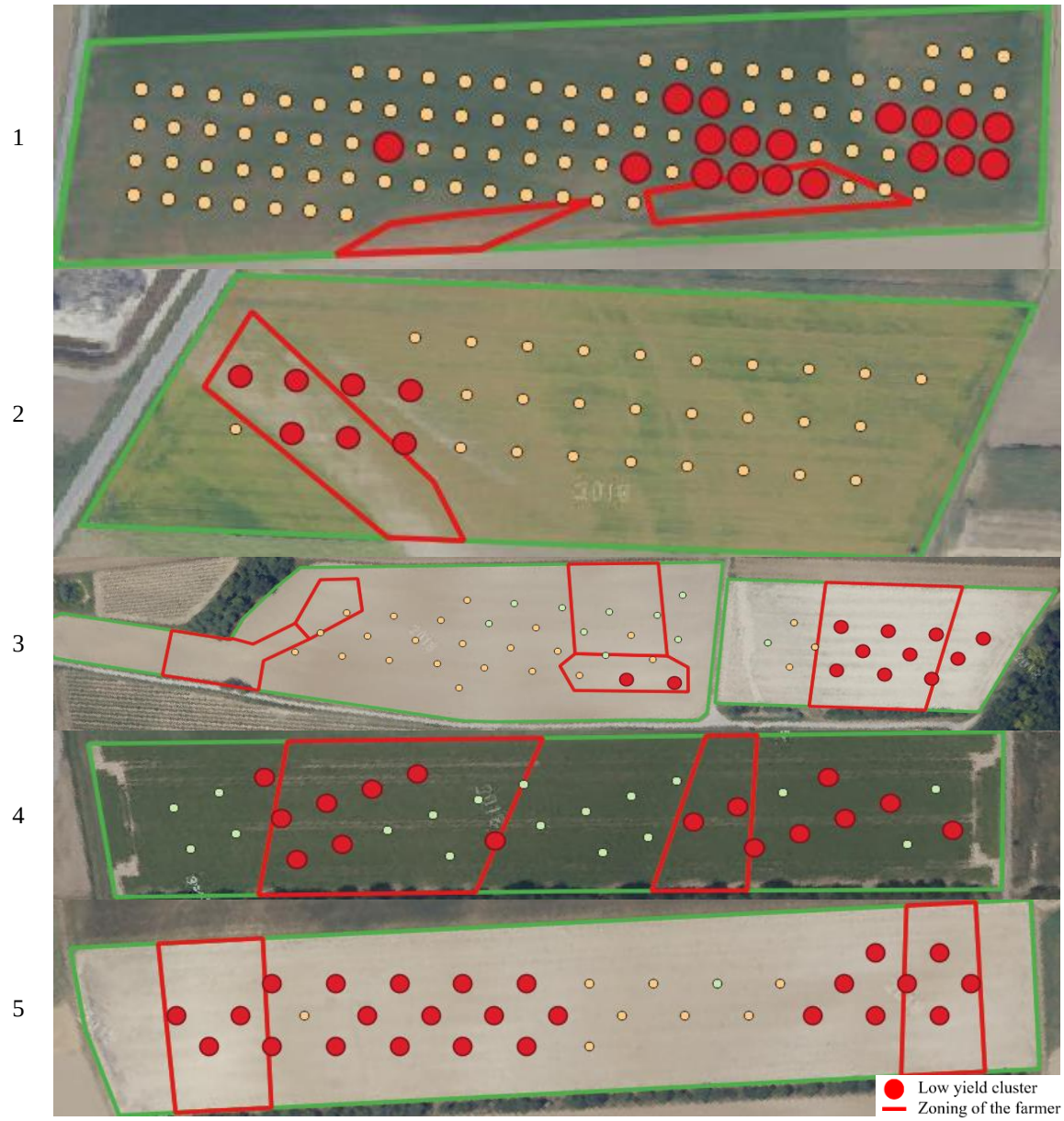


Figure 6. Example fields with best farm F1 score (=0.4), (NDVI, robust distance function, April to July, 3 cluster)

4. Conclusions

The detection and separate treatment of low-yield sites is an important ecological component of fertilizer models. The method used to detect low-yield sites is based on Unsupervised Learning methods. This involves examining historical vegetation patterns and anomalies and identifying corresponding areas using clustering algorithms such as k-means clustering. The advantages of this approach are that large amounts of data can be generated for a training and test data set, and it does not have to be labelled with soil samples or complex and cost-intensive analysis methods. It is difficult to validate the model, but initial comparisons with zoning or farmers' assessments are promising.

Future research could improve this model through optimized validation like conducting and measuring soil or yield parameters. Multi-year observations of a site would weed out year-dependent factors (e.g. climate). The disadvantage here is the lack of area coverage of a crop (for example, winter wheat). This could be remedied by including other crops (e.g. maize) and adding more years to the data set (beginning in 2021). A further improvement can be achieved by expanding the scope of the data to include more data points which could lead to a higher generalization. More relevant characteristics could be extracted by increasing the bands used and calculating more indices. Additionally, a higher temporal resolution of the data, or the integration of earlier months, could lead to the extraction of even more relevant data. These suggestions must also be accompanied by improved pre-processing and improved handling of clouds.

References

Agrarmarkt Austria, 2020 *INVEKOS Schläge Österreich 2018, 2019, 2020*. Available at: https://www.data.gv.at/katalog/dataset/35e36014-ec69-439b-8629-389f52ffaa92

Basso, B., Shuai, G., Zhang, J. et al., 2019. Yield stability analysis reveals sources of large-scale nitrogen loss from the US Midwest. Sci Rep 9, 5774 https://doi.org/10.1038/s41598-019-42271-1

Isensee, E., Thiessen, E. and Treue, P.,2003. 'Mehrjährige Erfahrungen mit der teilflächenspezifischen Düngung und Ernte', *Agrartechnische Forschung*, 9

A. Sharifi, 2020. "Using Sentinel-2 Data to Predict Nitrogen Uptake in Maize Crop," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 2656-2662, doi: 10.1109/JSTARS.2020.2998638

Söderström, M., Piikki, K., Stenberg, M., Stadig, H., & Martinsson, J., 2017. Producing nitrogen (N) uptake maps in winter wheat by combining proximal crop measurements with Sentinel-2 and DMC satellite images in a decision support system for farmers. Acta Agriculturae Scandinavica Section B: Soil and Plant Science, 67(7), 637–650. https://doi.org/10.1080/09064710.2017.1324044

Spicker, A. B., 2016. Entwicklung von Verfahren der teilflächenspezifischen Stickstoffdüngung zu Wintergerste (Hordeum vulgare L.) und Winterraps (Brassica napus L.) auf Grundlage reflexionsoptischer Messungen (Doctoral dissertation, Technische Universität München)

Spiess, E. and Richner, W., 2005. 'Stickstoff in der Landwirtschaft', *Schriftenreihe der FAL*, 57.

Yuzugullu O, Lorenz F, Fröhlich P, Liebisch F., 2020. Understanding Fields by Remote Sensing: Soil Zoning and Property Mapping. Remote Sensing. 12(7):1116. https://doi.org/10.3390/rs12071116