Comparative Analysis of Machine Learning Algorithms and Statistical Techniques for Data Analysis in Crop Growth Monitoring with NDVI *

Manasha Arunachalam[1], Siddhaarth Sekar[1], Annastasia M. Erdmann[2], Sajith Variyar V. V.[3], and Ramesh Sivanpillai[4]

[1]Department of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore
[2] Haub School of Environment & Natural Resources, University of Wyoming, Laramie, 82071 USA.
[3]Amrita Center for Computational Engineering and Networking (CEN), Coimbatore ORCID : 0000-0003-3944-8155
[4] School of Computing, University of Wyoming, Laramie, WY 82071 USA. ORCID: 0000-0003-3547-9464.

Keywords: NDVI Analysis, Remote Sensing, ISODATA, Affinity Propagation, Gaussian Mixture Model, Agricultural Monitoring

Abstract

We assessed the potential of Machine Learning (ML) for mapping crop growth in three flood irrigated fields. Results generated from ML algorithms were compared to the output generated by the ISODATA algorithm. Affinity Propagation (AP) identifies the number of clusters by considering all data points as potential exemplars and iteratively refine the set, while Gaussian Mixture Model (GMM) algorithm treats the data as a mixture of several Gaussian distributions, allowing for flexible cluster shapes. In contrast, ISODATA, a statistical clustering method, requires an analyst to specify the number of output clusters followed by iterative splitting and merging of clusters based on variance and distance criteria. We acquired Landsat derived NDVI images for three flood-irrigated fields over a span of four years. These images were collected at the start of the growing season to ensure consistency. Initially we clustered the pixels in these images for each field using AP and determine the number of clusters. Next, we applied GMM to identify and define the clusters. Finally, we plotted the mean value of all the pixels in each cluster for every year and assigned the clusters into six thematic classes: the first three classes for consistent growth (good, average, or poor) across all four years, and the other three for mixed growth patterns (e.g., good in three years and average in one). Output maps generated from these methods were compared using IoU scores. ML methods had greater efficiency in terms of replicating the steps for other fields, whereas ISODATA requires analyst intervention and interpretation.

1. Introduction

Monitoring crop growth in irrigated fields is critical in agricultural practices and optimizing yields. One effective way to achieve this is by analyzing the Normalized Difference Vegetation Index (NDVI) (Singh et al., 2020), which serves as an indicator of vegetation health and land cover(Sasidhar et al., 2019). In this study, we explore the efficacy of machine learning clustering algorithms—Affinity Propagation and Gaussian Mixture Model (GMM)—in comparison with the ISODATA method for monitoring crop growth based on NDVI values.

Affinity Propagation identifies the optimal number of clusters by treating all data points as potential exemplars and iteratively refining this set, providing a robust framework for cluster determination. The GMM approach, on the other hand, models the data as a mixture of several Gaussian distributions, allowing for flexible and adaptable cluster shapes. In contrast, ISODATA (Venkateswarlu and Raju, 1992), a statistical clustering method, requires manual intervention, as it involves iterative splitting and merging of clusters based on variance and distance criteria, making it a more labor-intensive process.

To conduct our study, we collected NDVI values across three different fields over a span of four years. Each field contains one image per year, all taken in the same month and on nearly the same date each year to ensure consistency. We first used Affinity Propagation to determine the number of clusters and then applied GMM to identify and define these clusters. Following the application of these machine learning methods and ISODATA, we plotted the mean NDVI value of all pixels in each cluster for every year.

The clusters were categorized into six classes: three classes representing consistent growth (good, average, or bad) across all four years, and three classes representing mixed growth patterns (e.g., good in three years and average in one). Additionally, we created detailed plots to monitor the number of pixels in each cluster over time. These comprehensive plots reveal intricate patterns in the vegetative index of the fields.

Our findings demonstrate that machine learning methods, particularly Affinity Propagation and GMM, offer greater efficiency and accuracy in analyzing field patterns compared to ISODATA, which demands more time and manual intervention. This study highlights the potential of advanced clustering algorithms to enhance the monitoring and management of crop growth in irrigated fields.

Iterative Self-Organizing Data Analysis Technique is an unsupervised clustering algorithm that iteratively groups data points based on their similarity. It dynamically adjusts the number of clusters by splitting or merging them during the process, and makes it flexible for analyzing complex datasets.

Affinity Propagation (Dueck, 2009) is a widely used machine learning algorithm and unlike other clustering methods, Affinity Propagation autonomously identifies cluster centers and assigns data points to clusters. This makes it particularly useful for our dataset with unknown cluster numbers or non-spherical cluster shapes. Affinity Propagation is based on ”message-passing” between data points to identify cluster centers, known as exemplars, and assign data points to these centers. The algorithm

* This document includes content generated with the assistance of AI tools like ChatGPT.

https://doi.org/10.5194/isprs-archives-XLVIII-M-5-2024-15-2025 |

aims to find the most representative exemplars to cluster data into meaningful groups. It is particularly effective for datasets with numerous clusters or complex, non-linear distributions.

The algorithm uses three key matrices namely, Similarity matrix, responsibility matrix and the Availability matrix. Similarity Matrix (S): Measures similarity between data points based on features rather than visual attributes. The similarity score is the negative squared distance between points:

$$S(i,k) = -\|x(i) - x(k)\|^2$$

Where $-\|x(i) - x(k)\|^2$ is the squared Euclidean distance.

Responsibility Matrix (R): Indicates how well-suited a data point is to be the exemplar for another data point. It is updated as:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}[a(i,k') + s(i,k')]$$

Availability Matrix (A): Reflects the suitability of a data point to serve as an exemplar for others. It is updated as:

$$a(i,k) \leftarrow \min\left(0, r(k,k) + \sum_{i' \neq i} \max(0, r(i',k))\right)$$

The algorithm iteratively updates these matrices until convergence, where the values stabilize. The final matrices determine cluster assignments. It calculates the similarity matrix based on a chosen metric, such as negative squared Euclidean distance. Then initialize and iteratively update the responsibility matrix and availability matrix. The net responsibility is computed by summing the responsibility and availability for each data point. The data points with high net responsibility are identified as exemplars. Each data point is assigned to the nearest exemplar based on similarity. After determining the number of clusters using Affinity Propagation, we trained a Gaussian Mixture Model with the identified number of clusters to refine the clustering results.

Gaussian Mixture Model (Reynolds et al., 2009) is a probabilistic model that assumes all the data points are generated from a mixture of several Gaussian distributions, each represented by three parameters: a mean (μ), a covariance (Σ), and a mixing coefficient (π). The mean (μ) defines the center of the Gaussian distribution. The covariance (Σ) defines the spread or width of the Gaussian. The mixing coefficient (π) defines the proportion of the population represented by each Gaussian.

The steps involved in GMM are as follows: The three parameters are initialized, and the probability of each data point belonging to each cluster is calculated using the current parameters. This involves computing the value of the Gaussian probability density function for each data point and each cluster, then normalizing these values across clusters to get probabilities. The Gaussian probability density function is given by:

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

where x is a data point, D is the number of dimensions, μ is the mean, and Σ is the covariance matrix. The parameters are updated after calculating the probabilities which involves calculating new means, covariances, and mixing coefficients that maximize the likelihood of the observed data given these probabilities. It has application in diverse fields (Menon et al., 2022, Sinith et al., 2010, Vekkot and Gupta, 2019)

2. Methodology

2.1 Study Area

3 irrigated fields located in Albany County (Wyoming) were selected for this study. Water from the Laramie River was the primary source of irrigation for these fields. Winter precipitation in the form for snow was the secondary source of moisture for these fields. These fields were mostly flat with a few relatively low and high spots. Water tend to accumulate in these dips while it would not reach the high spots.

2.2 Data

Normalized Difference Vegetation Index (NDVI) images acquired by Landsat 8 and 9 satellites were downloaded from US Geological Survey (USGS). These images were acquired as close to mid-July as possible to align with peak crop growth stages while considering satellite temporal resolution and cloud cover. Acquisition dates included 7/16/2019, 7/14/2020, 7/6/2021, 7/19/2022, and 7/11/2023.

The original data was multiplied by 10000 and stored as integers. We use python for data pre-processing and analysis. Firstly, we opened the NDVI index images using the PILLOW library (Clark, 2015). The loaded images were converted into Numpy arrays (Harris et al., 2020) and each pixel's value is divided by 10000. The images obtained for field 1 have dimensions of 57 by 57 pixels, with each pixel representing an area of 900 square meters. Each field has 485 background pixels that are denoted by the value -3.2768. These 485 pixels are substituted with None. The total area of the field is 2487600 m2 (or 615 acres). Each pixel is considered as a feature and since we need to perform analysis across all years, the pixel values for four years are stacked to form a feature vector. We have 2764 pixels after removing the background valued pixels.

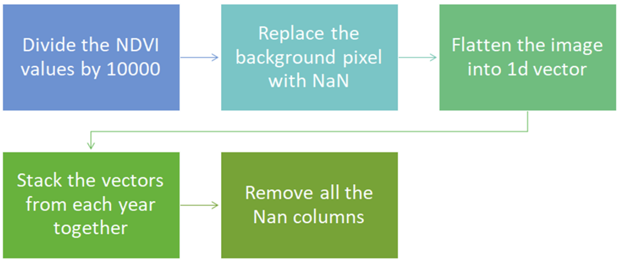
Divide the NDVI values by 10000
Replace the background pixel with NaN
Flatten the image into 1d vector
Stack the vectors from each year together
Remove all the Nan columns

Figure 1. Pre-processing NDVI data: Normalised NDVI, Substituting the garbage values in background pixels with NAN (Not a number in numpy), Flattening and stacking the images

To identify pixels with high and low NDVI values, we conducted a cluster analysis on the feature vector. Initially, Affinity Propagation was employed to determine the number of clusters in the field. Subsequently, a Gaussian Mixture Model was applied using the determined number of clusters. The overall workflow is shown in below in figure 1.

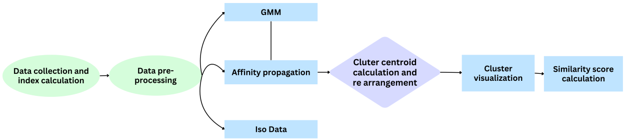
Data collection and index calculation
Data pre-processing
GMM
Affinity propagation
Iso Data
Cluter centroid calculation and re arrangement
Cluster visualization
Similarity score calculation

Figure 2. Overview of methodology: Steps involved in clustering using ISODATA, AP, and GMM

2.3 ISODATA

Individual NDVI images were stacked to create a multitemporal stack in ERDAS Imagine software (Hexagon Geospatial, Sweden). The multitemporal stack was then classified using the Iterative Self-Organizing Data Analysis Technique (ISODATA) unsupervised classification algorithm, initialized from statistics with 20 – 30 clusters, a minimum cluster size of 1% of pixels, a maximum standard deviation of 5.0, a minimum Euclidean distance of 4.0, and a maximum of 1.0 merges per iteration, with 100 iterations and a convergence threshold of 0.995. Since 2021 image was acquired earlier in the season, there was relatively less growth. Therefore, this image was excluded from the multitemporal stack and subsequent analysis. Image acquired in the next pass had cloud-cover over the study area. This entailed performing the entire classification process again on a revised multitemporal composite (4 images) with the parameters specified above.

The resulting clusters were then mapped into four classes based on the NDVI values shown in the multitemporal profile: Consistent High Growth, Consistent Low Growth, Expected Medium Growth, and Unexpected Growth. Consistent High Growth had NDVI values between 0.7 and 0.9. Consistent Low Growth had NDVI values between 0.15 and 0.35. Expected Medium Growth had NDVI values between 0.45 and 0.75, which were higher in wet years and lower in dry years. Unexpected Growth had NDVI values between 0.25 and 0.85, which were higher in dry years and lower in wet years.

2.4 Affinity Propagation

Affinity Propagation is used to determine the number of clusters for each feature set, It computes cluster assignments without needing to pre-specify the number of clusters.

The hyperparameters have been set to its default values during clustering, The damping factor of 0.5 ensures message updates are balanced, preventing oscillations for stable convergence. The preference parameter, set to the median of the similarity matrix by default, influences the number of clusters by indirectly controlling which points are chosen as exemplars. The convergence iteration parameter of 15 ensures that AP stops after 15 iterations without changes in cluster assignments. Similarly, the maximum iteration is set to 200 to cap the total iterations to avoid excessive computation. scikit-learn (Pedregosa et al., 2011) is used in python to perform affinity propagation

We use the number of clusters derived from AP as the number of components for GMM and perform futhur clustering.

2.5 Gaussian Mixture Model

It is used to assign cluster labels for each feature set, by using the number of clusters identified by AP. All the hyperparameters of GMM are said to its delfault value during clustering. Scikit-learn (Pedregosa et al., 2011) is used in python to perform GMM

After cluster assignment, the clusters are rearranged based on their centroids, calculated as the mean of the feature values within each cluster. Clusters are evaluated against six thresholds: all years with values above 66%, exactly one, two, or three years above 66%, values below 33% across all years, and values falling between 33% and 66%. These thresholds allow for a detailed examination of changes and patterns over time. Finally, the results are visualized using plots that highlight how vegetative growth changes accross the 4 years.

2.6 Comparative Analysis

GMM is sensitive to initialization whereas AP and ISODATA do not. GMM, AP and ISODATA work well on tasks that entail probabilistic clustering, clustering of non convex data and exploratory clustering respectively. These methods have been applied to cluster vegetative cover of fields and the results are attached below.

3. Results

3.1 ISODATA

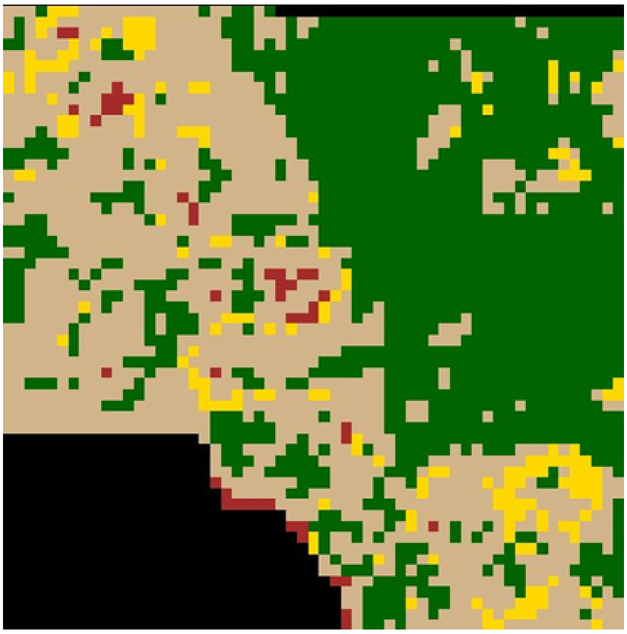

Figure 3. Classified image generated with the ISODATA classification showing a) Consistent high growth (green), b) Consistent low growth (brown), c) Expected medium growth (Tan), and d) Unexpected growth (Yellow) colors. Pixels assigned to unexpected growth class had high NDVI values in drought years but low values in wet years.

The classified multispectral image had the following class areas: Consistent High Growth, 292 acres; Consistent Low Growth, 10 acres; Expected Medium Growth, 273 acres; and Unexpected Growth, 40 acres. The spatial distribution of these classes illustrates the effects of irrigation efficiency and natural water availability. As expected, areas classified as Consistent High Growth corresponded to regions with consistent irrigation reach, demonstrating minimal variability in growth rates across wet and dry years. Areas classified as Expected Medium Growth corresponded to regions with poor irrigation reach, demonstrating variable growth rates dependent on natural water supplies. Conversely, some pixels from the Consistent Low Growth class appeared along the field's periphery, where crop growth diminished and bare ground encroached.

3.2 Affinity Propagation Clustering Results:

(Figure 1) illustrates the clustering results obtained for Field 1 using Affinity Propagation. The mean NDVI values for 2019, 2020, 2022, and 2023 across identified clusters are presented, revealing the temporal progression of vegetation health in the field. Thresholds of 0.33 and 0.66 were added as reference baselines to categorize pixels into low, medium, and high vegetation health clusters. The NDVI values consistently increase across clusters for all four years, indicating improved or stabilized vegetation health over time. Certain years, notably 2022, showed more variability in cluster NDVI values. An analysis of pixel distribution across clusters, as shown in Figure , revealed significant fluctuations in cluster sizes.

Field 1:

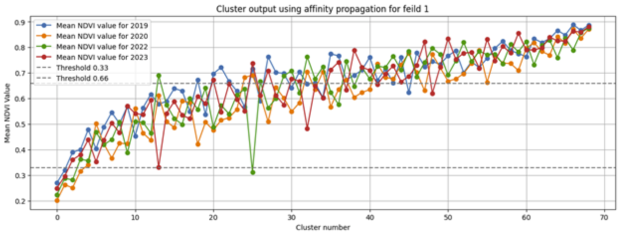

Figure 4. Average of NDVI value across 4 years for all the clusters generated by using Affinity Propagation for field 1

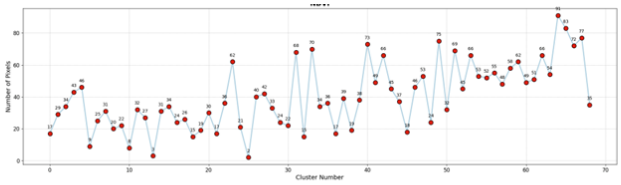

Figure 5. Number of pixels per cluster for field 1 using Affinity propagation

Field 2:

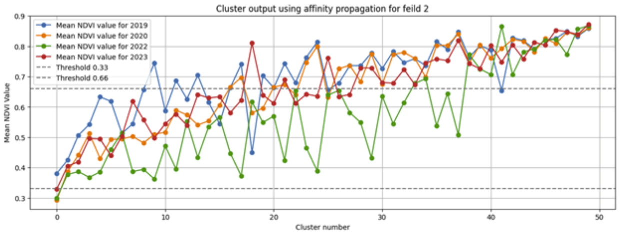

Figure 6. Average of NDVI value across 4 years for all the clusters generated by using Affinity Propagation for field 2

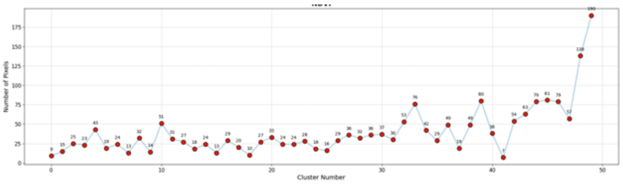

Figure 7. Number of pixels per cluster for field 2 using Affinity propagation

Field 3:

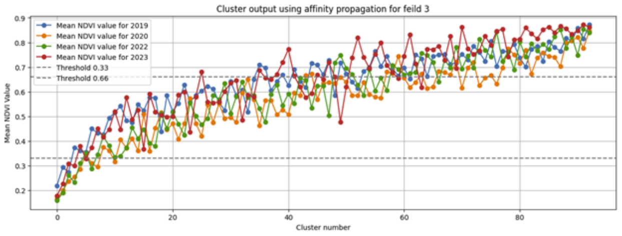

Figure 8. Average of NDVI value across 4 years for all the clusters generated by using Affinity Propagation for field 3

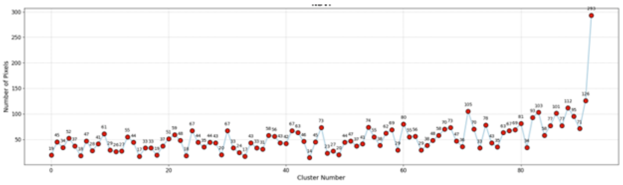

Figure 9. Number of pixels per cluster for field 3 using Affinity propagation

3.3 Gaussian Mixture Model Clustering Results:

After determining the cluster count from Affinity Propagation, a Gaussian Mixture Model (GMM) was applied to refine the clustering results further by fitting Gaussian distributions to each identified cluster. This method provided a probabilistic clustering approach, enabling the estimation of cluster overlap and further insights into the distribution of NDVI values within clusters.

The results from GMM for Field 1 are presented in (Figure 10), displaying mean NDVI values for each cluster across the four years. The GMM approach resulted in smoother transitions between clusters compared to Affinity Propagation. This is because of its ability to capture sub cluster variations within the primary clusters.

The Gaussian Mixture Model clustering for Field 1 shows a consistent upward NDVI trend across clusters, with most clusters in medium to high vegetation health zones. Yearly NDVI values are stable, though 2023 shows slight improvement.

This demonstrates the effectiveness of GMM in capturing finer details of NDVI variations, providing valuable insights for precision agriculture and vegetation health monitoring.

Field 1:

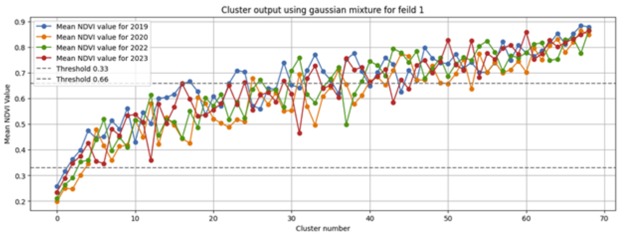

Figure 10. Average of NDVI value across 4 years for all the clusters generated by using Gaussian Mixture Model for field 1

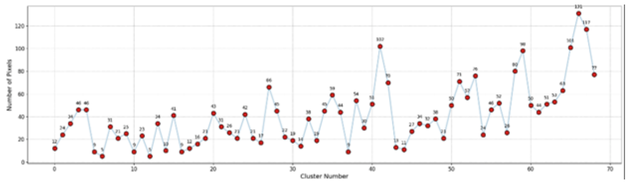

Figure 11. Number of pixels per cluster for field 1 using Gaussian Mixture Model

Field 2:

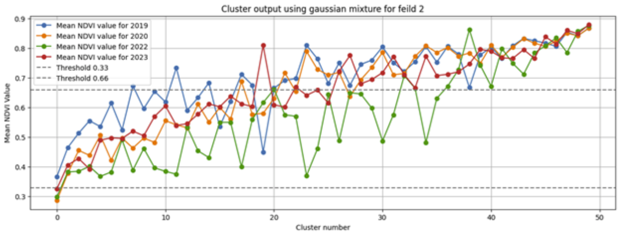

Figure 12. Average of NDVI value across 4 years for all the clusters generated by using Gaussian Mixture Model for field 2

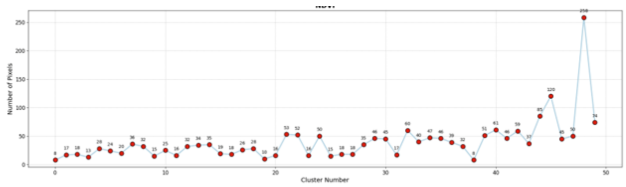

Figure 13. Number of pixels per cluster for field 2 using Gaussian Mixture Model

Field 3:

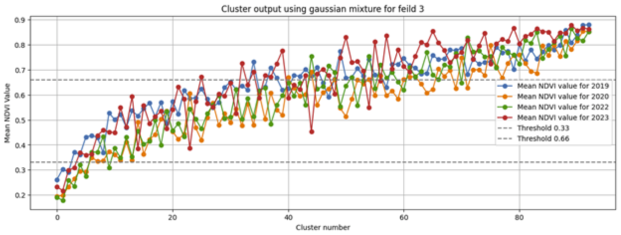

Figure 14. Average of NDVI value across 4 years for all the clusters generated by using Gaussian Mixture Model for field 3

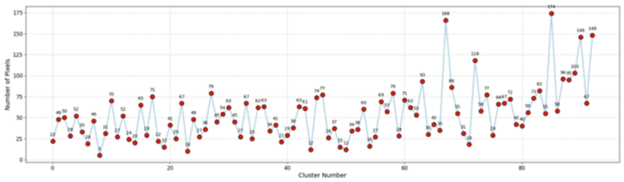

Figure 15. Number of pixels per cluster for field 3 using Gaussian Mixture Model

We can infer from the above plots, the number of clusters (sorted) to the number of pixels belonging to each cluster for both the clustering methods applied on the 3 fields across 4 years of data.

3.4 Comparitive Analysis

The Jaccard score (Ramli and Mohamad, 2009), also known as the Jaccard index or Intersection over Union (IoU), is a metric used to measure the similarity between two sets. It is particularly useful for evaluating clustering, classification, or segmentation tasks.

Before calculating the Jaccard score we need to reconstruct the image. This is due to the fact that we flattened the image before performing clustering. Image can be easily reconstructed by using the un-processed data as reference, substituting each pixel with the corresponding cluster number. The Jaccard score was calculated for each field between the clusters created by GMM and Affinity propagation. All of these values are very close to 1 as shown in (Table 1), indicating that there is no significant difference between the output given by GMM and Affinity propagation.

Field no.	Jaccard score
Field 1	0.9978
Field 2	0.9996
Field 3	0.9964

Table 1. Jaccard obtained when images classified with AP and GMM methods were compared to each other

4. Discussion

The clustering results highlights temporal changes in the fields over the course of 4 years. Unexpected patterns emerged within the Unexpected Growth class and the remaining pixels from the Consistent Low Growth class. These classes mapped areas that follow the path of a river that runs through the field, which could explain why the Unexpected growth class had such high growth rates in dry years and low growth rates in wet years. In dry years, the reduced river flow allowed vegetation to thrive in the riverbed, reflecting higher amounts of infrared radiation. In contrast, during wet years, higher water levels in the river led to reduced vegetation growth, absorbing more infrared radiation and lowering NDVI values. Similarly, pixels from the Consistent Low Growth class clustered on elevated terrain between river bends, where limited soil moisture further inhibited growth. These results help demonstrate how crop growth patterns are highly influenced by the relationship between irrigation practices, topography, and natural water sources.

Manual unsupervised classification for analyzing crop growth rates has many benefits and limitations. A human analyst can better understand the field conditions than an algorithm, granting the analyst more control over the mapping process to adjust parameters and thresholds based on real-world field conditions, which can enhance the relevance and interpretability of results. However, this understanding can be tainted by biases, resulting in conscious or unconscious manipulation of the results. Manual classification also requires significantly more time than machine-learning classification. Manually removing the 2021 image from the multi-temporal stack necessitated beginning the whole classification process over again. This can become tedious to the analyst, which can lead to more unconscious biases towards the results.

AP and GMM have several advantages over ISODATA, particularly in terms of consistency, repeatability, and ease of use. Both AP and GMM produce reliable and transferable results, making them ideal for consistent crop analysis across different fields and over multiple years. Unlike ISODATA, which can yield varying results due to its sensitivity to initial parameters and manual intervention, AP and GMM provide stable clustering outcomes. This stability allows new data from recent years to be added or older data to be removed seamlessly, as the algorithms can adapt to any dataset configuration.

One of the key benefits of AP is its ability to autonomously determine the optimal number of clusters, while GMM then takes these clusters and generates more flexible and adaptable shapes. In contrast, ISODATA often requires analysts to manually adjust parameters for each dataset, which can also lead to slight variations in results depending on the analyst's choices. AP and GMM, however, deliver identical results across different analysts due to their automated clustering processes.

An ideal solution would be a unified network that can both identify the optimal number of clusters and generate them autonomously.

5. Conclusion

The class map generated by ISODATA algorithm captured the growth classes. However, it heavily depends upon analyst's expertise and re-initialization when there are changes to the input dataset. The balance between analyst control and potential analyst bias underscores the need for caution when using manual methods. AP and GMM has proven to be consistent, easy to scale and easy to integrate with any type of data. To help reduce susceptibility to human error and streamline the mapping process, automated machine-learning techniques could be integrated, with future studies exploring hybrid approaches, combining the contextual awareness of manual analysis methods with the efficiency and objectivity of automated algorithms.

6. Acknowledgement

This work was conducted under the Memorandum of Agreement between Amrita Vishwa Vidyapeetham and the University of Wyoming, fostering collaboration in advanced research. We sincerely thank our faculty mentors, Dr Ramesh Sivanpillai and Dr Sajith Variyar V.V , for their invaluable guidance and support. A special thanks to Kuppala Nandini and Ch Navneet Krishna for their meaningful discussions and inputs, which greatly helped shape this work.

References

Clark, A., 2015. Pillow (pil fork) documentation.

Dueck, D., 2009. *Affinity propagation: clustering data by passing messages*. University of Toronto Toronto, ON, Canada.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T. E., 2020. Array programming with NumPy. *Nature*, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Menon, R. R., Kumar, S. D., Vismaya, C., 2022. Gmm-based document clustering of knowledge graph embeddings. Cited by: 5.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Ramli, N., Mohamad, D., 2009. On the Jaccard index similarity measure in ranking fuzzy numbers. *Matematika*, 157–165.

Reynolds, D. A. et al., 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).

Sasidhar, T. T., Sreelakshmi, K., Vyshnav, M., Sowmya, V., Soman, K., 2019. Land cover satellite image classification using ndvi and simplecnn. Cited by: 23.

Singh, B. M., Komal, C., Victorovich, K. A., 2020. Crop growth monitoring through Sentinel and Landsat data based NDVI timeseries. , 44(3), 409–419.

Sinith, M., Salim, A., Sankar K, G., Narayanan K V, S., Soman, V., 2010. A novel method for text-independent speaker identification using mfcc and gmm. 292 – 296. Cited by: 24.

Vekkot, S., Gupta, D., 2019. Emotion conversion in telugu using constrained variance gmm and continuous wavelet transform-f-0. 2019-October, 991 – 996. Cited by: 5.

Venkateswarlu, N., Raju, P., 1992. Fast isodata clustering algorithms. *Pattern Recognition*, 25(3), 335-342. https://www.sciencedirect.com/science/article/pii/003132039290114X.