Mapping Recent Plantation Expansion in "Kampong Thom" and "Oddar Meanchey" Provinces, Cambodia

Innovative Tools to Support the Design and Monitoring of NBS and GHG Emission Reduction Projects (LoA/RAP/2023/48)

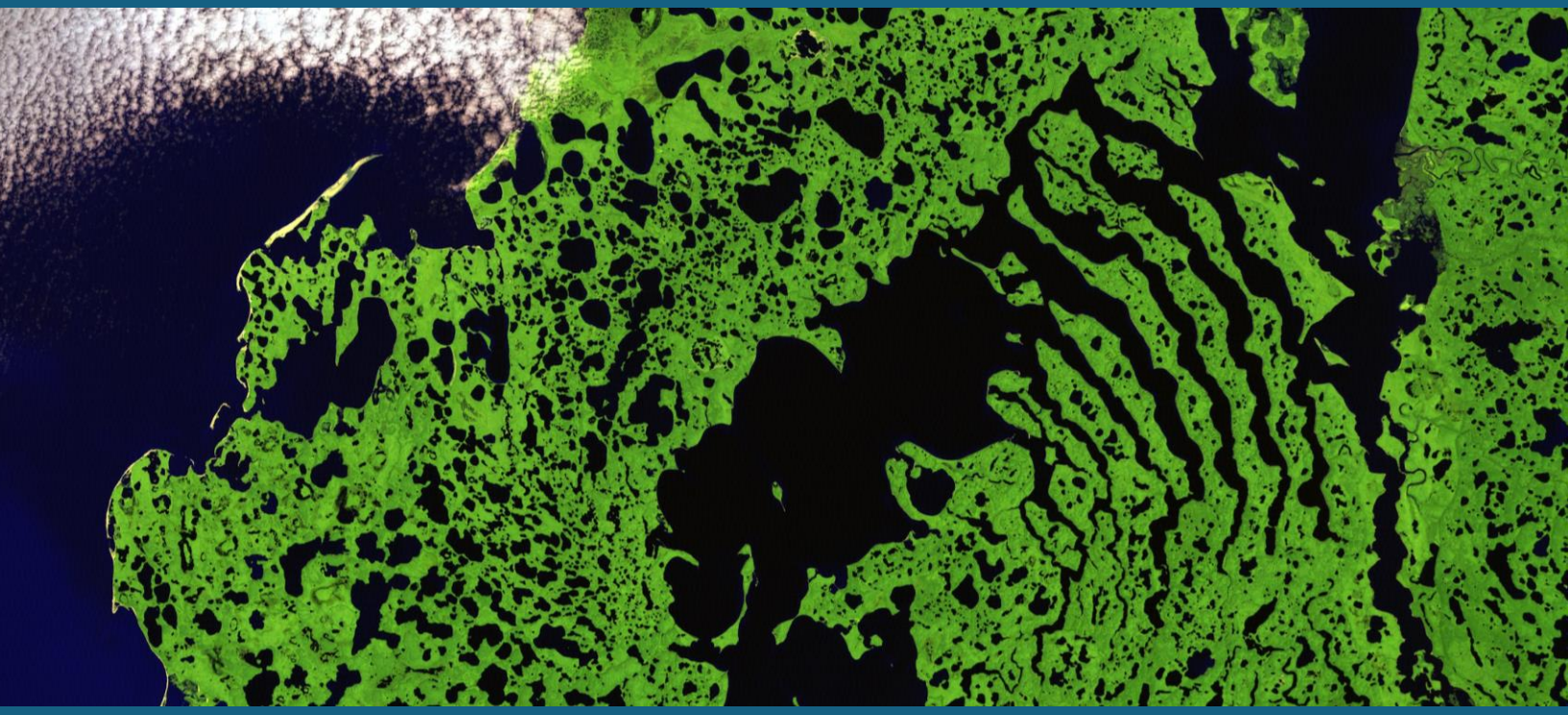

Photo Credit: USGS

PREPARED BY:
Asian Institute of Technology,
Pathum Thani, Thailand

Table of Contents

Chapter I: Introduction 3

Chapter II: Overview of the Approach 4

2.1 Algorithms Used in TSClusterer Tool 4

2.2 How to Use the "TSClusterer" Tool 6

Chapter III: Case Studies 9

3.1 Background 9

3.2 Methodology 10

3.3 Results 12

References 15

Chapter I: Introduction

Cambodia's agricultural sector has experienced significant transformation over the last few decades, positioning the country as a key player in the global agricultural market. Among the notable developments is the rapid expansion of plantation agriculture, particularly in crops such as cashew nuts and mangoes. These crops are not only vital for Cambodia's economy but are also seeing an increase in global demand, particularly in emerging markets. As agricultural practices in Cambodia evolve, it has become crucial to monitor and understand the dynamics of plantation expansion and land-use change.

This report focuses on the detection and analysis of plantation expansion in Cambodia, with a particular emphasis on two important provinces: Kampong Thom and Oddar Meanchey. Both provinces have seen substantial growth in plantation areas over the past decade, with Kampong Thom becoming a major hub for cashew nut production and Oddar Meanchey emerging as a key region for both mango and cashew nut cultivation. These shifts in land use are driven by a combination of factors, including government policies aimed at promoting agricultural growth, investments in infrastructure, and the growing demand for these products in both the domestic and international markets.

The study period is confined to the years 2013 to 2024, a time frame during which Landsat-8 imagery is available and provides valuable insights into vegetation and land cover changes. The use of remote sensing data, specifically the normalized difference vegetation index (NDVI), allows for a detailed assessment of the growth of vegetation, making it possible to monitor plantation expansion at a provincial scale.

A key objective of this report is to employ clustering algorithms to analyze and classify NDVI time-series data for each province. These algorithms enable the identification of distinct temporal patterns in vegetation growth, which are critical for understanding the dynamics of plantation expansion. With this case study, two clustering algorithms were implemented as a tool. They are K-means, a simpler yet effective clustering technique that groups data into clusters based on similarity, and VQ-VAE, a more computationally intensive method capable of identifying more complex patterns in large datasets.

Case studies in this report utilize Landsat-8 imagery to monitor the NDVI values over time, allowing for the detection of changes in vegetation and the identification of areas where plantation expansion has occurred. By analyzing NDVI time series data, the report identifies key temporal patterns that indicate recent plantation growth. These patterns are then used to isolate clusters corresponding to newly planted areas, offering valuable insights into the timing and extent of plantation expansion.

Through the application of remote sensing and clustering algorithms, this report offers an analysis of plantation expansion in Kampong Thom and Oddar Meanchey provinces. The subsequent chapters will explore the methodologies used in the analysis and present the clustering process results.

Chapter II: Overview of the Approach

2.1 Algorithms Used in TSClusterer Tool

The TSClusterer tool/class integrates two clustering algorithms for time-series data: classic K-means and VQ-VAE. Each of these algorithms brings its own strengths to the table, and understanding their inner workings and applications is essential for selecting the appropriate one for a given analysis. The selection depends on the data's complexity, the type of patterns expected, and the computational resources available.

K-means (Google Earth Engine API, 2025) Clustering is one of the most well-established and widely used clustering algorithms due to its simplicity, efficiency, and ease of implementation. It is a partition-based algorithm that divides a set of data points into a specified number of clusters, k, with the goal of minimizing the variance within each cluster. The process begins by selecting k initial centroids, either randomly or through some other initialization method. Each data point is then assigned to the nearest centroid, typically using the Euclidean distance metric. After all points are assigned to clusters, the centroids are recalculated as the mean of the points in each cluster. The algorithm then iterates between the assignment and centroid update steps until convergence, meaning the centroids no longer change significantly.

The key advantages of K-means are its simplicity, speed, and scalability. The algorithm is computationally efficient, making it an excellent choice for clustering large datasets, and it is easy to understand and implement. K-means is particularly suitable for problems where the data can be naturally divided into spherical clusters that are relatively equal in size and variance. For example, in land cover classification tasks, where time-series data of vegetation indices such as NDVI (Normalized Difference Vegetation Index) are clustered into distinct groups based on their temporal patterns, K-means can effectively identify homogeneous clusters.

However, K-means does come with its limitations. One of the most significant issues is that it is sensitive to the initial placement of centroids. Poor initialization can result in the algorithm converging to suboptimal solutions. Another limitation is that K-means assumes the clusters are spherical and equally sized, which may not always be the case. Despite these drawbacks, K-means is often a good starting point for clustering tasks due to its efficiency and simplicity, particularly when the underlying clusters are well-separated and the data is relatively simple.

On the other hand, VQ-VAE (Vector Quantized Variational Autoencoder) (Keras Documentation, 2025) is a more sophisticated approach combining deep learning with clustering techniques. VQ-VAE is an extension of the traditional autoencoder architecture, which is designed to learn a compressed representation of the input data. In VQ-VAE, the encoder maps the input data (in this case, time-series data) to a continuous latent space,

and the decoder attempts to reconstruct the original data from this compressed representation. However, instead of using the continuous latent space directly, VQ-VAE applies a vector quantization technique, which forces the encoder to map the data into a discrete set of vectors from a codebook. This encourages the model to learn distinct, meaningful clusters in the latent space, which can then be used for clustering.

The primary advantage of VQ-VAE over K-means is its ability to capture complex, non-linear patterns in the data. Since it uses deep learning techniques, VQ-VAE can model detailed relationships in high-dimensional data that might not be effectively captured by K-means. It can also handle data with more complex structures, such as time-series data that exhibit non-linear temporal dynamics. For instance, in the case of land cover classification based on NDVI time-series data, where different land types may show overlapping temporal patterns, VQ-VAE can learn more complex clusters by leveraging its ability to capture non-linear dependencies within the data.

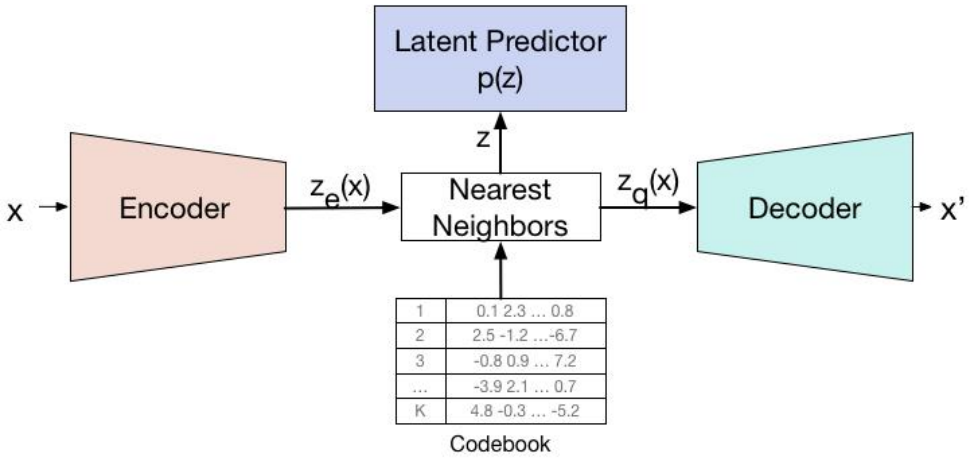

Figure 2.1: Architecture of VQ-VAE (Reprinted from Roy et al., 2018, https://ar5iv.labs.arxiv.org/html/1805.11063).

However, VQ-VAE is not without its challenges. The algorithm is computationally more intensive than K-means, particularly when dealing with large datasets. Training a VQ-VAE model also takes longer, as it involves both the encoder-decoder architecture and the quantization process. Furthermore, while the method can effectively handle more complex data structures, it requires a larger amount of training data to perform optimally. VQ-VAE also has more parameters to adjust, such as the size of the latent space, weights for loss components, and number of epochs/training iterations etc. , which can significantly impact the model's performance. Fine-tuning these parameters often requires multiple iterations of experimentation, making the process more time-consuming. Additionally, since the algorithm involves deep learning components, it may also be more prone to overfitting if not carefully regularized, especially when the dataset is small or lacks sufficient diversity. For smaller datasets or when computational efficiency is a priority, K-means might be a better choice, as it is easier to implement, faster to run, and less sensitive to overfitting.

While both algorithms are effective, the choice between K-means and VQ-VAE depends on the characteristics of the data and the specific needs of the analysis. K-means remains a powerful tool for simpler clustering tasks with well-separated, spherical clusters, especially when computational efficiency is a key consideration. VQ-VAE, however, is better suited for more complex clustering tasks where the data is high-dimensional, non-linear, and involves more complex temporal dynamics, as in the case of time-series data with overlapping patterns.

2.2 How to Use the "TSClusterer" Tool

The "TSClusterer" class is the main class for performing clustering. The useful functions of this class are summarized below.

```
khm_clusterer = TSClusterer.TSClusterer(aoi, startYear, endYear,
cloudTH, percentReductionTH, verbose=True)
```

While initializing the "TSClusterer" class, Landsat-8 NDVI will be calculated for each year based on the input parameters.

- "aoi": The study area feature collection. If the study area is in a shapefile, it can be read using "geemap.shp_to_ee" from the "geemap" library.
- "startYear": The start year of the analysis.
- "endYear": The end year of the analysis.
- "cloudTH": The acceptable cloud cover percentage to include while filtering image collections for each year. The value should be between 0 and 100.
- "percentReductionTH": The percentage reduction to be applied to reduce the image collection for each year. This parameter defines the percentage reduction. The value should be between 0 and 100.
- "verbose": Whether to print or echo method descriptions while running. Pass either True or False. The default value is True.

```
imageTemp = khm_clusterer.getRGBImageByYear(year)
```

This method can be used to get the RGB image for each year. By visualizing the RGB image for each year, we can assess its quality (e.g., whether there is cloud coverage in the image). If the quality of the image for a particular year is not satisfactory, it can either be removed using the "removeNDVIImageByYear" method, or the input parameters of the "TSClusterer" class can be adjusted and reinitialized.

- "year": The RGB image for this year will be returned for quality inspection.

```
khm_clusterer.removeNDVIImageByYear(year)
```

If any image quality is found to be unsatisfactory while inspecting using the "getRGBImageByYear" method, this method can be used to remove that image before analysis.

- “year”: The year to be removed from the set of NDVI images that will be used for analysis.

khm_clusterer.generateGTData(numRandPixels, additionalGTPoints, verbose=True)

This method generates ground-truth data to be used in the clustering algorithm. Both digitized points and random points within the study area can be generated using this method.

- “numRandPixels”: The number of random points within the study area to be used as ground-truth data.
- “additionalGTPoints”: A feature collection of additional ground-truth points. These points can be digitized separately and passed as a feature collection. If the dataset is in point shapefile format, it can be read using “geemap.shp_to_ee” from the “geemap” library.
- “verbose”: Whether to print or echo method descriptions while running. Pass either True or False. The default value is True.

`khm_clusterer.clusterViaKMean(nClusters, verbose=True)`

There are two clustering algorithms available: K-means (Google Earth Engine API, 2025) and VQ-VAE (Keras Documentation, 2025). This method clusters the data using the classic K-means algorithm.

- “nClusters”: The number of expected clusters.
- “verbose”: Whether to print or echo method descriptions while running. Pass either True or False. The default value is True.

`khm_clusterer.plotClusterCentersKMean()`

This method returns a time-series plot of the cluster centers clustered via the K-means algorithm.

`khm_clusterer.clusterViaVQVAE(dataCSVPath scalingPerc, rmseLossW, nClusters, epochs, verbose=True)`

There are two clustering algorithms available: K-means (Google Earth Engine API, 2025) and VQ-VAE (Keras Documentation, 2025). This method clusters the data using the VQ-VAE algorithm. After running this method, the loss curve of the training will be plotted at the end.

- “dataCSVPath”: The path to a temporary CSV file used to convert the GEE feature collection to CSV format. For example, '/content/TEMP1.csv'.

- “scalingPerc”: The percentage scaling of the data, which brings the data to a range between 0 and 1. Two values for percentage scaling can be provided as a list. For example, [2, 98] will scale the data between the 2nd and 98th percentiles.
- “rmseLossW”: The weight assigned to the reconstruction loss. If the cluster centers are too general, this weight can be increased; if they are too specific, it can be reduced. Typical values are 1, 10, 20, 50, or 100.
- “nClusters”: The number of expected clusters.
- “epochs”: The number of epochs (iterations) to train the algorithm.
- “verbose”: Whether to print or echo method descriptions while running. Pass either True or False. The default value is True.

```
khm_clusterer.plotClusterCentersVQVAE()
```

This method returns a time-series plot of the cluster centers clustered via the VQ-VAE algorithm.

```
khm_clusterer.mergeRelevantClusters(clusterIDList, verbose=True)
```

This method masks out non-relevant clusters from the output. For example, if the target class is recent cashew-nut plantations, it may be clustered into multiple clusters. This method isolates the relevant clusters and masks out the unnecessary ones. Relevant and unnecessary clusters can be identified by inspecting the plot of cluster centers.

- “clusterIDList”: A list of relevant cluster IDs. For example: [2, 3].
- “verbose”: Whether to print or echo method descriptions while running. Pass either True or False. The default value is True.

```
clusteredImg = khm_clusterer.getClusteredResult()
```

This method returns the clustered output image **before** applying the “mergeRelevantClusters” method.

```
finalImg = khm_clusterer.getFinalResult()
```

This method returns the clustered output image **after** applying the “mergeRelevantClusters” method.

```
khm_clusterer.exportOutput(outputFileName)
```

This method can be used to export the result to Google Drive.

- “outputFileName”: The name of the output GeoTIFF file. For example, 'result_kampong_thom.'

Chapter III: Case Studies

3.1 Background

In recent years, Cambodia has experienced significant growth in cashew nut cultivation, particularly in Kampong Thom province and also in Oddar Meanchey province.

Kampong Thom has emerged as a central hub for cashew nut production. In 2024, the province accounted for approximately 147,700 hectares of cashew plantations, making it the largest cashew-producing area in the country. Additionally, the Cambodian government has approved the establishment of the country's first cashew agro-industrial park in Kampong Thom, further enhancing the province's role in cashew processing and export (Khmer Times, 2024), (Phnom Penh Post, 2024).

Cambodia's cashew nut production has expanded significantly in recent years, particularly in 2021, when exports to Vietnam surged by a large percentage (See Figure 3.1), (Venter, 2022) compared to the previous year. Given that cashew trees typically require about three years to reach maturity and begin producing nuts, it is likely that the plantations responsible for the 2021 export surge were established before 2018. This expansion in cashew cultivation should be detectable in NDVI time series data collected during that period. Since this study utilizes NDVI data spanning from 2013 to 2024, the analysis should capture these plantation expansions effectively.

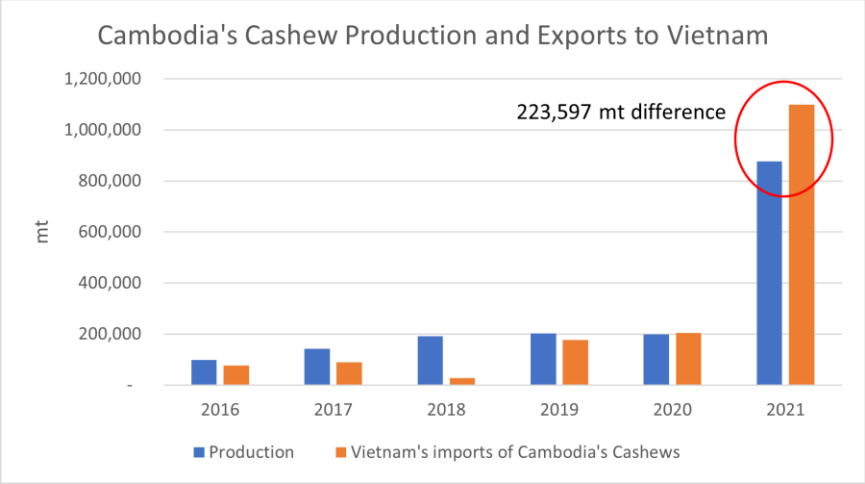

Figure 3.1: Cashew-nut production and export to Vietnam (Reprinted from Venter (2022)).

Recent data indicates a significant expansion in mango and cashew nut production in Oddar Meanchey province as well. Mango cultivation increased by 74.29% to 13,515 hectares, yielding 283,250 tons, a 28.40% annual growth. Cashew nut plantations expanded by 105.40% to 18,382 hectares, producing 14,600 tons, with a 22.59% annual growth (Khmer Times, 2023). These substantial changes in agricultural land use are detectable through NDVI time series analysis as well.

These developments in both Kampong Thom and Oddar Meanchey provinces underscore Cambodia's growing prominence in cashew nut and mango production, with significant investments aimed at enhancing processing capabilities and expanding export markets. This study tries to capture this recent expansion of plantations in Kampong Thom and Oddar Meanchey provinces using yearly Landsat-8 NDVI data from 2013 to 2024.

3.2 Methodology

In recent years, cashew nut and mango plantations have seen significant growth in Cambodia, particularly in the provinces of Kampong Thom and Oddar Meanchey. The increase in plantation areas in these regions has made them ideal candidates for remote sensing-based analysis to monitor land-use changes. To conduct this study, we have limited the analysis period to 2013–2024, as Landsat-8 imagery is available for this time frame, providing NDVI data suitable for analyzing land cover changes over time. The study areas, which correspond to the provincial boundaries of Kampong Thom and Oddar Meanchey, were obtained from The Humanitarian Data Exchange (https://data.humdata.org/).

Figure 3.2: A few examples of locations of plantations are seen on high-resolution base maps such as Bing and Google base maps.

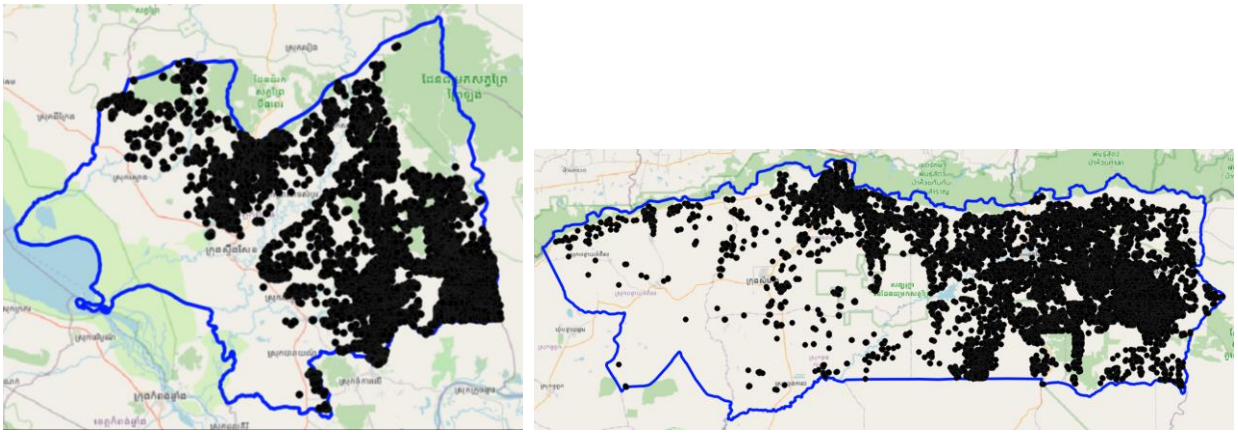

Figure 3.3: Manually digitize locations of plantations in Kampong Thom (left image) and Oddar Meanchey (right image) provinces using a base map.

To accurately identify plantation areas within these provinces, we utilized base maps from services such as Bing and Google Earth. These maps provided a clear visual reference to pinpoint plantation locations. For example, Figure 3.2 shows such example areas within the study region. These points were manually digitized on top of the base maps, with a total of 6,126 points digitized in Kampong Thom province and 9,125 points in Oddar Meanchey province, covering the entire area of the provinces. The locations of digitized points are shown in Figure 3.3. To account for other land cover types and dynamics in the study areas, we included an equal number of random points for each province. This means that 6,126 random points were generated for Kampong Thom and 9,125 random points for Oddar Meanchey. These random points were spread across the entire study area to provide a diverse sample of land cover types, ensuring that the analysis did not solely focus on plantations but also considered the broader landscape changes.

The next step involved creating a Landsat image collection for each year within the study period. During the creation of these collections, cloud cover was removed by utilizing the quality band in the Landsat images. Images with high cloud cover were filtered out to ensure that the analysis would focus on clear images. This cloud filtering process was critical, as cloud cover can significantly distort the results of NDVI (Normalized Difference Vegetation Index) analysis.

To reduce the yearly image collection to a single, we applied a percentile reduction technique. Percentile reduction was chosen over traditional min and max reduction methods. The min reduction tends to result in images with more cloud shadows, while the max reduction often leads to higher levels of cloud cover because clouds bring a high reflectance in the imagery. Percentile reduction strikes a balance, helping to mitigate the impact of both cloud cover and cloud shadows. For this analysis, we opted to use the 60th percentile to select the most representative image for each year. This was particularly important because we were modeling plantations, where it was crucial to remain as close as possible to the maximum NDVI value for each year. In contrast, areas like paddy fields,

which exhibit frequent periodic NDVI changes, might lead to misleading results. The 60th percentile reduction helped to avoid such fluctuations, ensuring that the resulting NDVI values reflected the stable growth patterns of plantations. This approach was empirically / visually verified to make sure that clouds are not affecting NDVI values as well.

Once the NDVI values were extracted for each year, they were extracted for each of the ground truth points. These NDVI time series data were then fed into the clustering algorithm. The clustering process aimed to identify distinct temporal patterns in the NDVI time series. One key pattern that emerged was a distinct dip in NDVI values, followed by a gradual increase in NDVI over approximately three years, after which the NDVI remained high. This temporal pattern is characteristic of recent plantation expansion, where new plantations initially show low vegetation cover and then gradually mature, leading to higher NDVI values over time.

By focusing on clusters that exhibited this specific temporal pattern, we were able to extract the areas corresponding to recent plantation expansion as the final result. These clusters were considered the most relevant for identifying the growth of cashew nut and mango plantations over the study period. This method allowed us to track the dynamics of plantation growth, providing valuable insights into the expansion of these agricultural areas in Kampong Thom and Oddar Meanchey provinces.

3.3 Results

The cluster centers generated by the clustering algorithm are shown in Figure 3.4. These cluster centers represent different temporal patterns of NDVI values over the study period. Among these clusters, those corresponding to recent plantation expansion exhibit a distinct pattern: a dip in NDVI values followed by a gradual increase in NDVI over approximately three years, after which the NDVI remains high. This pattern is indicative of newly established plantations that initially show low vegetation cover, followed by a steady increase as the plantation matures.

After isolating these clusters, we can see that cluster 2 and cluster 3 exhibited similar patterns of plantation growth in Kampong Thom province. However, there was a slight difference between the two: cluster 2 showed the early growth of plantations, while cluster 3 indicated a late plantation expansion. This suggests that cluster 2 corresponds to areas where plantations were established earlier in the study period, whereas cluster 3 corresponds to plantations that were established relatively more recently. In Oddar Meanchey province, only cluster 1 exhibited the characteristic pattern of recent plantation expansion. After isolating these clusters, the final results, which correspond to the areas of recent plantation expansion, are displayed in Figure 3.5.

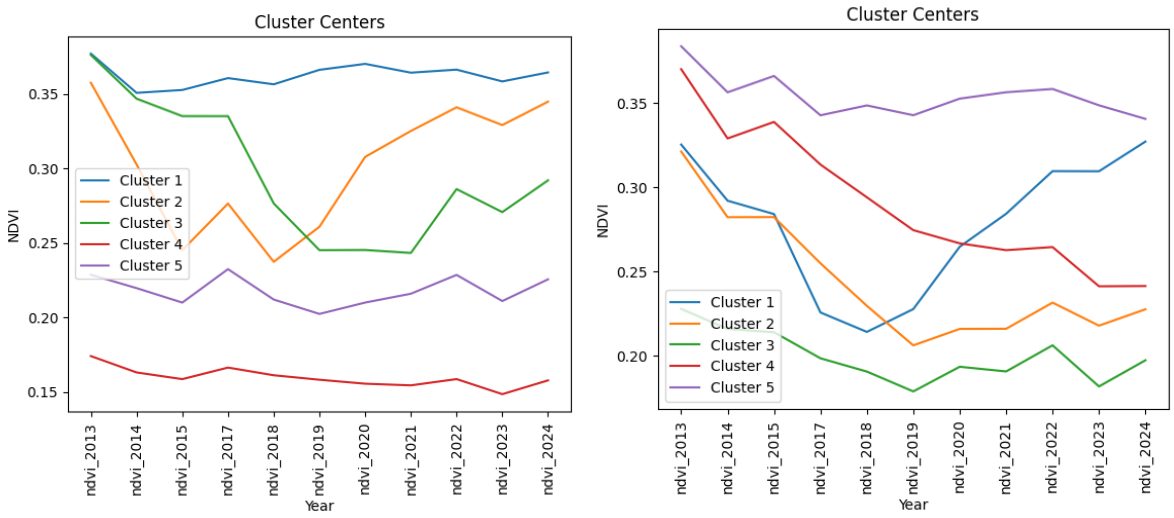

Figure 3.4: Resulting cluster centers of Kampong Thom (left image) and Oddar Meanchey (right image) provinces.

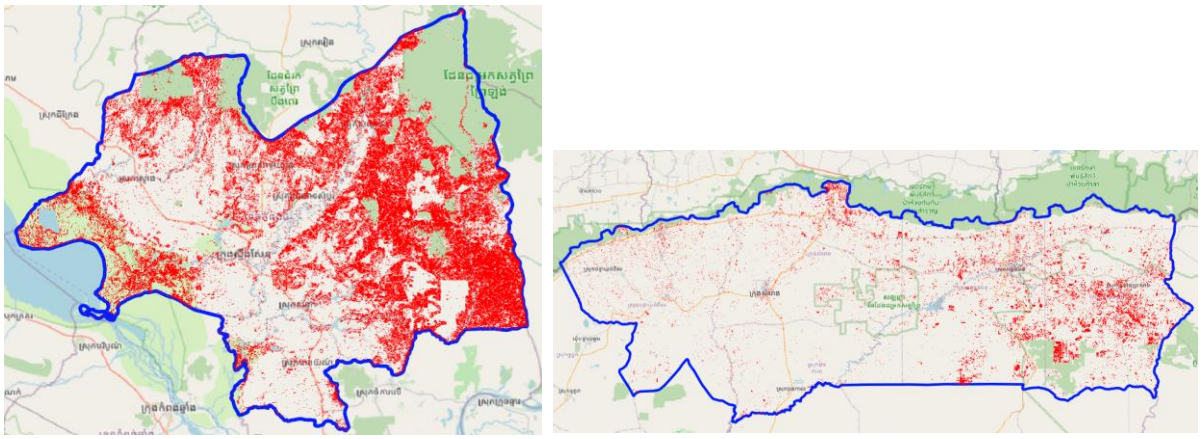

Figure 3.5: Results (recent plantation expansion) of Kampong Thom (left image) and Oddar Meanchey (right image) provinces.

The results for Kampong Thom province are particularly interesting, as the major plantation in this province is cashew nuts. There is substantial evidence from news media and academic literature indicating the recent expansion of cashew nut plantations in Kampong Thom. Based on this information, we can conclude that the areas identified in Kampong Thom province represent recently planted cashew-nut plantations. The timing of the plantation growth patterns observed in the NDVI time series aligns well with the reported expansion of cashew nut farming in the region, further validating the results.

In contrast, Oddar Meanchey province hosts a mix of mango and cashew nut plantations, both of which have seen significant recent expansion. Given the mixed nature of plantations in Oddar Meanchey, we can hypothesize that the mapped areas in this province correspond to recently planted mango and cashew nut mixed plantations. The NDVI time

series patterns of both crops would likely exhibit similar temporal characteristics of gradual NDVI increase after an initial dip as well.

The results observed in this study are approximately consistent with findings from other academic literature (Chaya et al., 2024), which have also reported recent plantation expansion in these provinces. The mapped areas from the clustering algorithm in Kampong Thom and Oddar Meanchey provinces align with the observed patterns of plantation growth reported in other studies (see Figure 3.6), (Chaya et al., 2024).

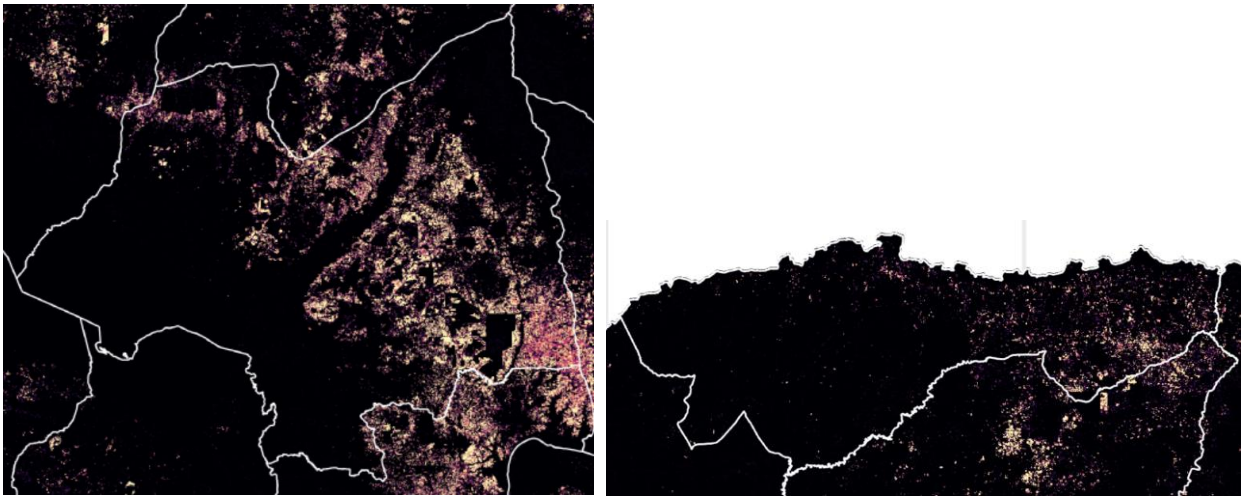

Figure 3.6: Results were found in another similar study (Chaya et al., 2024) for the Kampong Thom (left image) and Oddar Meanchey (right image) provinces.

References

1. Khmer Times. (2024, March 2). Kampong Thom to become cashew industrial hub. Khmer Times. https://www.khmertimeskh.com/501599066/kampong-thom-to-become-cashew-industrial-hub/
2. Phnom Penh Post. (2024, November 28). Cambodia approves first cashew agro-industrial park in Kampong Thom. Phnom Penh Post. https://www.phnompenhpost.com/business/cambodia-approves-first-cashew-agro-industrial-park-in-kampong-thom
3. Venter, T. (2022, January 24). 2021 Ends with more Questions than Answers in the Cashew Market. Tridge. Retrieved April 9, 2025, from https://www.tridge.com/stories/2021-ends-with-more-questions-than-answers-in-the-cashew-market
4. Khmer Times. (2023, March 21). Video: Oddar Meanchey Province has the potential to produce more than 2 million tons per year for domestic supply and export. Khmer Times. Retrieved April 9, 2025, from https://www.khmertimeskh.com/501259305/video-oddar-meanchey-province-has-the-potential-to-produce-more-than-2-million-tons-per-year-for-domestic-supply-and-export/
5. Roy, A., Vaswani, A., Neelakantan, A., & Parmar, N. (2018). Theory and experiments on vector quantized autoencoders. arXiv preprint arXiv:1805.11063. Retrieved April 9, 2025, from https://arxiv.org/abs/1805.11063
6. Google Earth Engine API. ee.Clusterer.wekaKMeans. Google Earth Engine API Documentation. Retrieved April 9, 2025, from https://developers.google.com/earth-engine/apidocs/ee-clusterer-wekakmeans
7. Keras Documentation. Vector-Quantized Variational Autoencoders. Keras Documentation. Retrieved April 9, 2025, from https://keras.io/examples/generative/vq_vae/
8. Chaya, V., Poortinga, A., Nimol, K., Sokleap, S., Sophorn, M., Chhin, P., McMahon, A., Nicolau, A. P., Tenneson, K., & Saah, D. (2024). Is Cambodia the World's Largest Cashew Producer? arXiv preprint arXiv:2405.16926.