INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Clustering Analysis of Food Security, Waste and Loss: Malaysia Agricultural Insights

Enoch Chen Sheng Hii [a], Siew Mooi Lim [a,*], Seng Xian Loo [a], Sheng Kit Yeap [a], Ching Yee Tan [a]

[a] *Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia*

Corresponding author: [*]siewmooi@tarc.edu.my

***Abstract*— As the world population grows rapidly nowadays, the demand for food has come to rise. The escalating demand for food has caused substantial wastage and loss, which not only hampers food security efforts but also aggravates greenhouse gas (GHG) emissions, intensifying the environmental crisis. Among numerous countries, Malaysia, with its diverse agricultural profile, emerges as a good fit for our case study. This study chooses the clustering technique to examine food sector data in Malaysia and investigate the link between the clustering results on food data and the data on GHG emissions. This case study aims to find crops depending on their production efficiency, underline those that match major waste, and estimate their contribution to greenhouse gas emissions. Three clustering techniques, Gaussian Mixture Modelling (GMM), Birch, and Density Peak clustering, are applied in the Production and Supply Utilisation Accounts (SUA) datasets, help to identify and cluster crops based on their similar traits to acquire uncovered patterns between the food sector and environmental issues. Using cutting-edge clustering algorithms and visualization tools, this study investigated in-depth the complex interactions among food production, waste, and greenhouse gas emissions in Malaysia. By addressing food production efficiency and waste reduction, the outcome will be a cascade of benefits that not only improve food security but also help to lessen negative environmental effects. This study illuminates the multifaceted dynamics of food production, waste, and environmental impact, offering valuable insights and pathways toward a more sustainable future for Malaysia and potentially other nations.**

***Keywords*—Unsupervised machine learning; clustering; Gaussian Mixture Modelling (GMM); birch clustering; density peak clustering.**

Manuscript received 5 May 2024; revised 9 Jul. 2024; accepted 19 Sep. 2024. Date of publication 30 Nov. 2024.

I. Introduction

In recent years, food security, food waste and loss, and GHG emissions have become crucial global issues that keep getting worse. Even though the food system and environmental issues seem like two different issues, they are profoundly connected and significantly affect each other. Globally, approximately one-third of the food produced, equivalent to 1.3 billion metric tons, is wasted or lost yearly, representing significant inefficiency in the food supply chain and a significant source of GHG emissions [1], [2]. These emissions, primarily methane from decomposing organic waste, contribute to the accelerating pace of climate change, further exacerbating the challenges of ensuring food security. Such vast amounts of food waste loss exacerbate food insecurity and contributes to environmental degradation.

A review of 78 articles reveals that the household-level calorie adequacy indicator is the most used single measure of food security, appearing in 22% of the studies [3]. Food waste loss is not just an issue common in developing countries but also in developed nations. In developing nations, inadequate infrastructure, poor harvesting techniques, and climatic conditions are the main reasons that cause food waste loss. In contrast, consumer behavior and retail practices in developed countries lead to higher per capita waste. This waste, especially when sent to landfills, becomes a significant source of GHG emissions [4].

Moreover, the food system's GHG emissions are not just why this issue exists. Agricultural practices, land-use changes, and transportation all contribute to the carbon footprint of the food people consume. As the global population continues to rise, the linear model of food production and consumption becomes

increasingly unsustainable in terms of meeting human needs and its impact on the planet [2].

Food waste is a significant problem worldwide. Major food loss and waste factors include poor management of perishables, stakeholder attitudes, buyer-supplier agreements, and supply chain interruptions. Livestock sector GHG emissions contribute to climate change and global warming [5], [6], [7]. Food systems worldwide struggle with GHG emissions and food waste, which harm food security, the environment, and economies. The cold chain for perishable goods in China accounts for 1–3% of total emissions, mainly from energy use and food losses. Food waste undermines ecosystems and food system resilience, with storage improvements potentially affecting food system stability. Models like BioBaM-GHG 2.0 show that afforestation can significantly reduce agricultural emissions. Regional differences in China's food system emissions point to the need for regional coordination. These findings emphasize the need for integrated approaches and detailed models to reduce food system emissions and waste [8], [9], [10], [11].

Researchers have devised several ideas to handle the problems of food security, food waste and loss, and greenhouse gas emissions. Among these is the change from linear to circular food systems, which give recycling and repurposing top priority over disposal. From organic farming to precision farming, sustainable agricultural methods, which range in impact on the environment, have been underlined as essential in guaranteeing food security while lowering environmental effects. Furthermore, under increasing focus is global cooperation with projects that share best practices, technologies, and policies to lower waste and emissions [1], [2].

Using both unsupervised and supervised machine learning, analysis was conducted to learn patterns and identify solutions depending on the outcomes of several machine learning techniques [12], [13], [14]. As the world grapples with these challenges, the collective insights from research underscore the urgency to act, with the well-being of current and future generations at stake. Understanding needs, family habits, and eating behaviors are critical to reducing food waste. Food loss and waste significantly impact food security, the environment, and the economy. Food insecurity, which is associated with global warming and poor health, was exacerbated by the COVID-19 pandemic. Food loss and waste in the Arab region can total more than 210 kg per person per year [15], [16], [17].

The food waste issue is severe in Malaysia. In addition, like many nations, Malaysia faces the dual challenge of ensuring food security while mitigating environmental impacts, particularly GHG emissions [18], [19]. A significant portion of GHG emissions can be attributed to the causes of massive food waste and utilization. This is proven by their previous research, where they showed statistics on the emission of CO_2 in Malaysia from 2000 to 2016, as shown in Fig. 2, which was mainly caused by food waste, as shown in Fig. 1. Current research suggests a strong correlation between food waste and GHG emissions [20]. However, a knowledge gap exists regarding the efficiency of food production across various regions in Malaysia.

Crucially, one must understand this efficiency and how it relates to food usage for common crops. Furthermore, the link between these crops' efficiency and their impact on greenhouse gas emissions is primarily unexplored. Therefore, this work attempts to close this gap by using clustering methods on datasets concerning food production efficiency and food consumption. By comparing and analyzing these datasets in the framework of GHG emissions, this study aims to offer a thorough knowledge of the interaction among food production, use, and environmental influence in Malaysia.

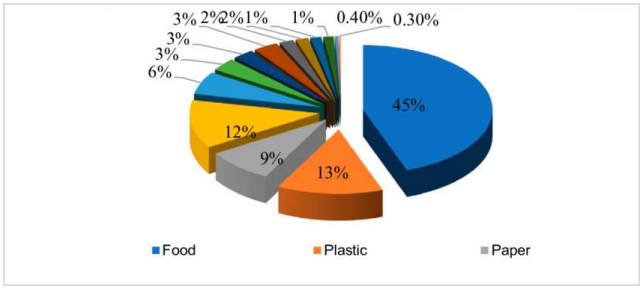

Fig. 1 Waste Pie Chart in Malaysia

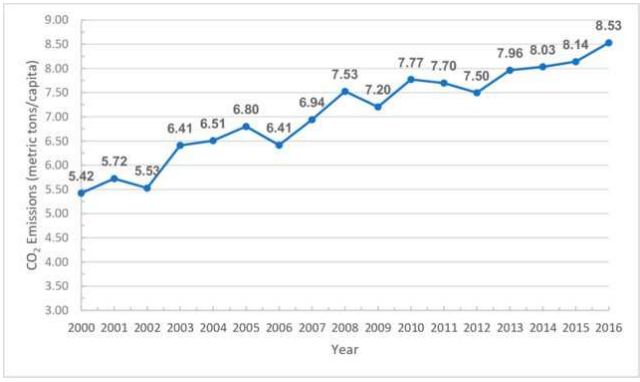

Fig. 2 CO_2 Emission Line Chart in Malaysia

A. Clustering methods in the Food sector

Advanced data collecting techniques cause many data to be acquired in several databases. Demand for organizing significant information from data and extracting insightful analysis has grown. Clustering is among the most basic issues in data mining and knowledge search. Based on their similarity, data points are split into several clusters; hence, data points in one cluster are more similar than those in another [21]. Mehrdad Rostami [22] developed a new hybrid food recommender system to address similar issues with precious systems. Based on the similarity of features, this work clustered users and food products using graph clustering and a deep-learning-based method.

B. Gaussian Mixture Model

The Gaussian mixture model (GMM) is a probabilistic clustering technique used to aggregate several Gaussian distributions to model complicated data distributions. Dealing with datasets with several underlying structures or when the data is not normally distributed will significantly benefit from it.

GMM clustering is used in a similar case study conducted by [13] to find the possible causes of food insecurity in each area and to comprehend the features and structure of the food assistance network in a particular region. In crime cases, GMM is used to pinpoint groups of burglary and antisocial behavior criminal records. These clusters present Gaussian mixed model-like distributions (GMMs). While enabling probabilistic assignments [23], GMM efficiently manages complicated data structures. In weather forecasting, GMM is used to identify clusters corresponding to different distribution errors to correct ensemble forecast distributions [24].

C. Birch Clustering

BIRCH excels at clustering massive datasets by creating a compressed summary of the data points rather than processing the entire dataset directly. It is known for its efficiency and scalability, making it suitable for large datasets. BIRCH requires the maximum number of clusters for a better clustering quality performance [25]. It can incrementally incorporate new data points as they become available. This technique extracts load forms from large databases, clusters high, moderate, and minimal load types using cost functions, and determines the right number of clusters for global grouping [26], [27]. Similar research also proves that BIRCH's method for clustering complex large datasets first generates smaller and compact summaries from the original dataset that retain as much information as possible [28]. In the context of food security and waste management, BIRCH shows the emission of greenhouse gases in the current scenario and predicts the future environment [29].

D. Density Peak Clustering

Density Peak Clustering is a clustering algorithm that uses the concept of density peaks to identify clusters in the data. It is beneficial when dealing with datasets with varying densities and shapes. It can automatically ascertain the number of knee points in the decision graph according to the characteristics of different datasets and further determine the number of clustering centers without human intervention. On the other hand, the density peak clustering algorithm (DPC) requires manual determination of cluster centers and poor performance on complex datasets with varying densities or non-convexity. Hence, this study introduces a new DPC variant designed to be more resistant to noise in data.

Food waste data might have inherent noise due to variations or inconsistencies. This noise-resistant DPC variant could fit the proposed study well [30]. Besides, there is also a way to handle imbalanced data in clustering. The existing work by [31] applies clustering to minority class samples. The sub-clusters with sparse data (less dense sub-clusters), which are closer to the borderline of the majority class, are assigned higher weights to achieve a high probability of selection for generating a new synthetic sample.

II. Material and Method

A. Data Collection and Cleaning

The research's foundation was laid with the data collection process, where datasets for food production, food waste & loss relating to crops, and GHG emission in Malaysia were sourced from the Food and Agriculture Organization of the United Nations (FAOSTAT). All three of the datasets consisted of different features and different time ranges; however, the differences in the features of all three datasets made a connection among each other. After data was collected, explorations and visualizations were carried out to understand each dataset better. The three datasets consist of null values, meaning no values are recorded for that specific element for that specific year; therefore, the null values are filled with zero values.

B. Data Feature Selection and Preprocessing

The preprocessing data step starts with the food production dataset, where the 'area' column is filtered to 'Malaysia,' and the 'element' column is filtered to 'Area harvested' and 'Production.' This process was carried out due to previous research carried out by Renard and Tilman. The food production dataset from FAOSTAT was implemented in the research to find out crop production efficiency, where rows remained that consisted of the 'Area harvested' and 'Production' in the FAOSTAT dataset [32].

The dataset is then transformed, where each row represents either the *Area harvested* or *Production* element for a crop; each row containing the quantities for consecutive years is summed up, and a new column containing the total quantities is created, as shown in Table 1. Afterward, the dataset is transformed where it only has three columns, which are the *Item*, *Area harvested*, containing the total value of all the years each crop for area harvested, and *Production*, containing the total value of all the years for each crop for production as shown in Table 2. As *Area harvested* and *Production* are needed for clustering uses, dimensionality reduction is unnecessary as the dataset is already in 2-dimensional format.

As the values in the dataset vary significantly, standardization using *StandardScaler* becomes imperative. This method centers the data around a mean of zero with a standard deviation of one, ensuring that features with larger scales don't disproportionately influence the outcome of algorithms sensitive to feature magnitude. By transforming the data this way, a more balanced and fair comparison between the features can be achieved, facilitating better performance and interpretability for the three clustering algorithms.

TABLE I
FOOD PRODUCTION DATASET BEFORE TRANSPOSING

	Item	Element	Y1961	Y1962	Y1963	Y1964	Y1965	Y1966	Y1967	Y1968	…
0	Abaca, manila hemp, raw	Area harvested	1836	1821	2023	2064	2289	2198	2023	1214	
1	Abaca, manila hemp, raw	Production	4400	4000	4000	4400	3200	3259	3281	3519	
2	Areca nuts	Area harvested	6000	3800	5500	4500	5000	4500	5000	3000	
3	Areca nuts	Production	6500	4000	5000	4000	6000	4000	5000	3500	

TABLE II
FOOD PRODUCTION DATASET AFTER TRANSPOSING

	Item	Area harvested	Production
0	Abaca, manila hemp, raw	17654	33218
1	Areca nuts	110678	152003.48
2	Avocados	361	824
3	Bananas	1521709	25172384.11

TABLE III
FOOD WASTE & LOSS DATASET BEFORE TRANSPOSING

	Item	Element	Y2010	Y2011	…	Y2019	Y2020	Total
0	Almonds, in shell	Calories/Year	611	680	…	488.38	468.71	4048.09
1	Almonds, in shell	Export Quantity	0.0	14	…	63.27	5.13	129.4
2	Almonds, in shell	Fat supply quantity (g/capita/day	0.0	0	…	0	0	0
3	Almonds, in shell	Fats/Year	54	60	…	43.25	44.16	357.41
4	Almonds, in shell	Food supply (kcal/capital/day)	0	0	…	0.04	0.04	0.08

TABLE IV
FOOD WASTE & LOSS DATASET AFTER TRANSPOSING

	Item	Export quantity	Import quantity	Stock variation
0	Areca nuts	33029.46	24833.88	1007.52
1	Avocados	615.28	19160.19	0
2	Bananas	253421.23	175040.37	0
3	Cabbages	169715.52	1372403.27	0
4	Cashew nuts, in shell	49.03	923.37	-5263
5	Cassava, fresh	0	0	0

For the preprocessing of the food waste and loss dataset, the data was filtered based on the area where the area is 'Malaysia' and the element column where the values were 'Feed', 'Import Quantity', 'Loss', 'Other uses (non-food)', 'Processed', 'Residuals' and 'Stock Variation'. This step is carried out after the clustering on the food production dataset, as comparisons are made to find the common crops within the clusters produced by the food production dataset with the crops in the food waste and loss dataset. Each row represents a specific element for a crop, with columns detailing the quantities for consecutive years. A two-step data transformation process is carried out; for each crop-element combination, the total quantity across the years is calculated as shown in Table 3.

The dataset is transposed to reorient its structure. In this transformed format, each row represents a unique crop, while the columns correspond to the total quantities for each element. As shown in Table 4, the columns for the transposed dataset in total are eight columns; therefore, dimensional reduction is applied using Principal Component Analysis (PCA) to convert the dataset to a two-dimensional dataset. Due to the large data scales in the dataset, standardization is also carried out on the food waste and loss dataset to make the three clustering algorithms better perform when training on the data.

For the GHG emission dataset, unwanted features, such as 'Domain Code', 'Area Code (M49)', 'Element Code', 'Item Code', and 'Year Code' are removed from the dataset. Afterward, the preprocessed dataset is applied for data visualization and analysis.

C. Data Visualization

Before the clustering phase of the research, extensive data visualization techniques were applied to the Food Production dataset to gain a comprehensive understanding of its structure and characteristics before the clustering process. To begin with, bar charts were utilized to visualize the top 10 crops in terms of both the total areas harvested shown in Fig. 3 and total production shown in Fig. 4, where we can observe that oil palm fruit was top 1 for both charts. This approach provided an immediate grasp of the most significant crops in the dataset.

Subsequently, a pair plot was generated in Fig. 5 to explore the relationships between area harvested and production, offering insights into their correlation and potential outliers.

Density plots were created to delve deeper into the data distribution for harvested and production areas, highlighting regions of high and low data density. Boxplots in Fig. 6 were also employed for these two variables to identify their spread, skewness, and potential outliers, informing the data cleaning and preprocessing stages. Finally, a parallel coordinate plot was used to visualize the multi-dimensional nature of the area harvested and production across different crops shown in Fig. 7. This ensemble of visualization techniques served as a diagnostic tool, setting the stage for the subsequent clustering and in-depth analysis.

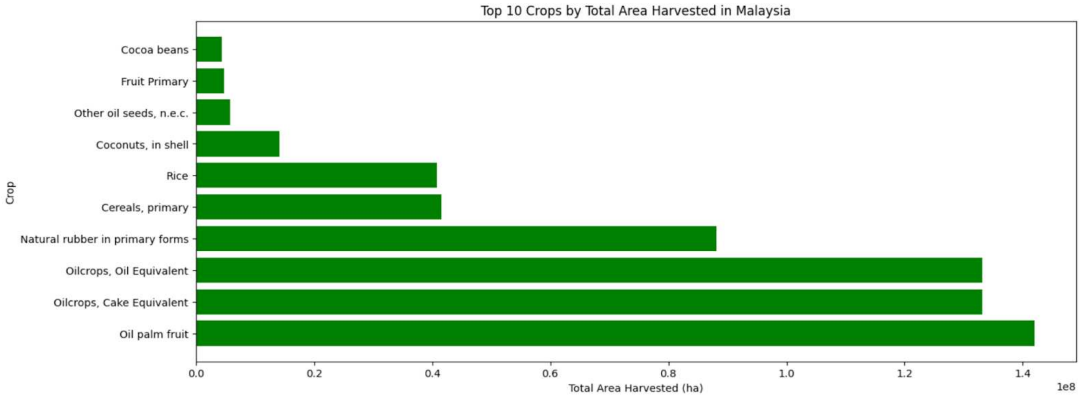

Fig. 3 Top 10 Total Area Harvested Bar Chart

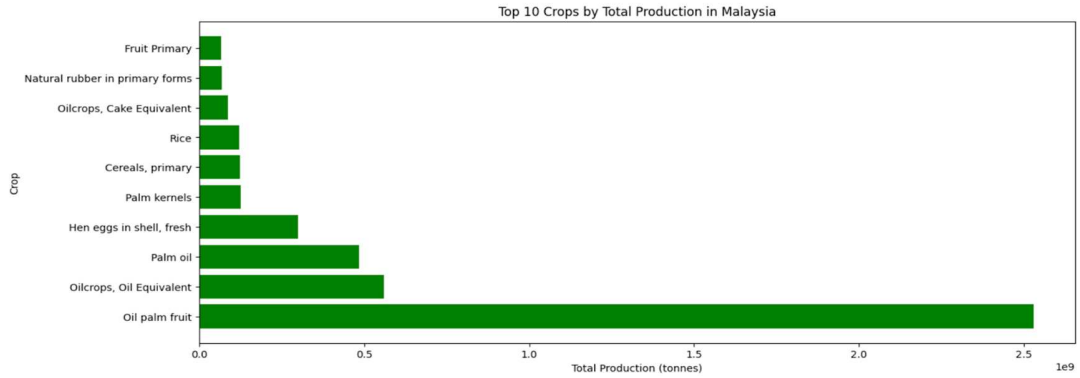

Fig. 4 Top 10 Production Bar Chart

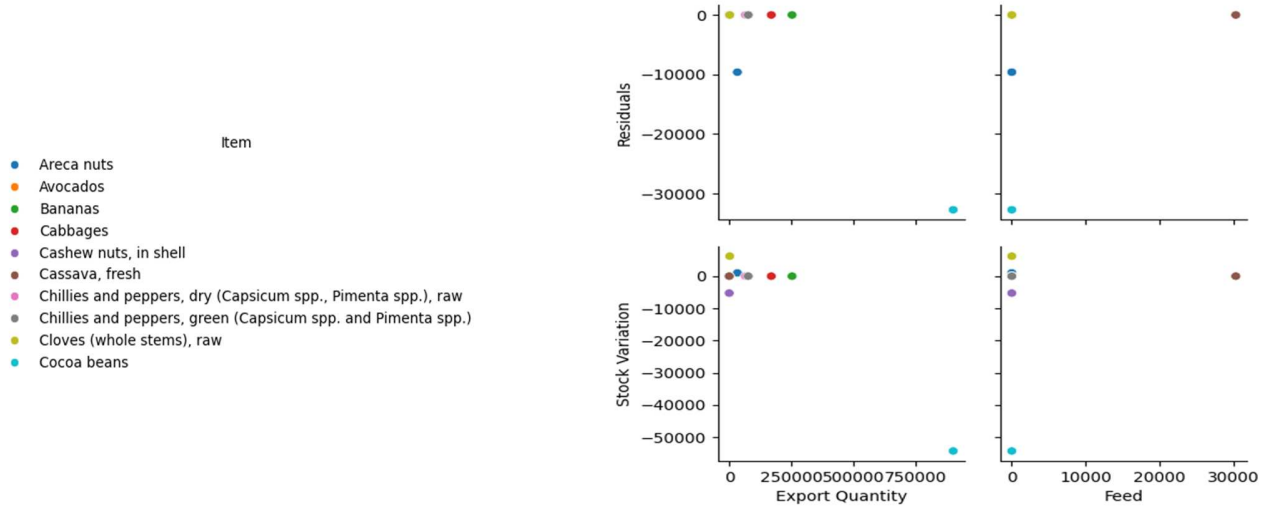

Fig. 5 Subset of Total Area Harvested & Production PairPlot Graph

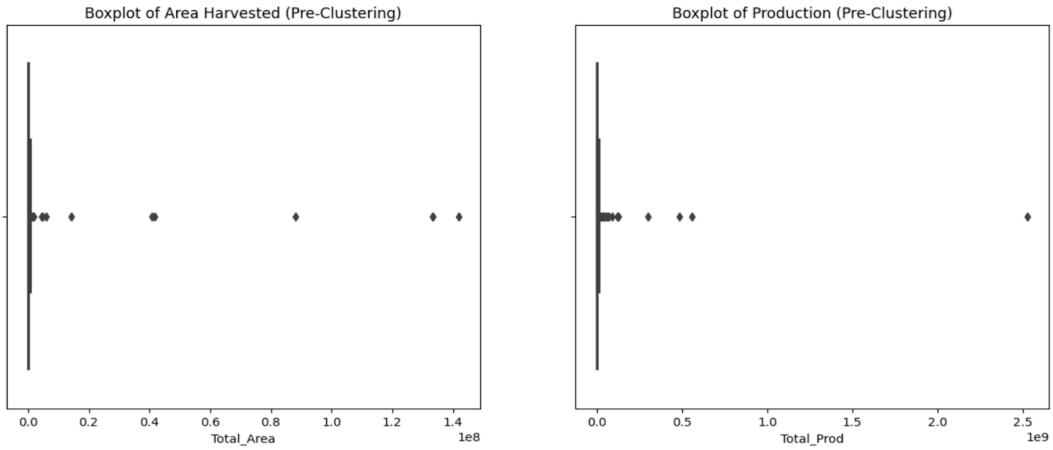

Fig. 6 Total Area Harvested and Production Boxplot

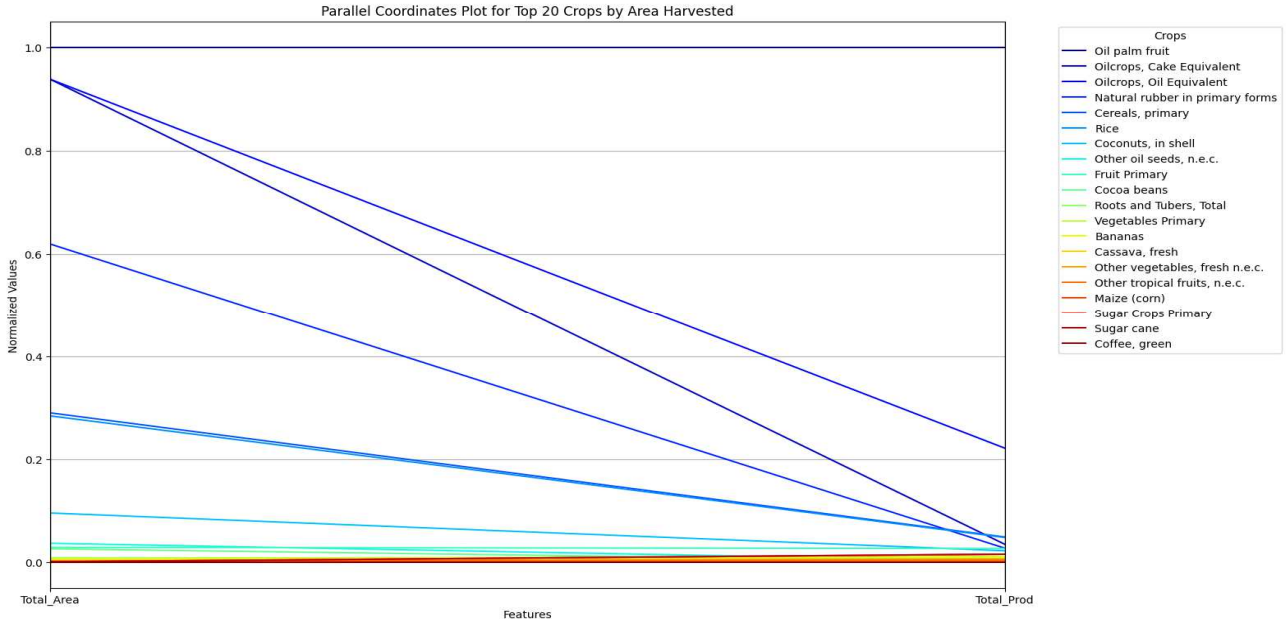

Fig. 7 Parallel Coordinate Plot

In the data visualization analysis of food waste and loss for crops in Malaysia, a series of plots yield critical insights for stakeholders. From Fig. 8, a bar plot ranking the top 10 items by total loss reveals that "Oil palm fruit" incurs the highest loss, trailed by "Maize (corn)" and "Rice (Paddy Equivalent)," pinpointing key areas for targeted waste reduction interventions.

Another bar plot in Fig. 9 focusing on feed contributions shows that "Maize" leads the list, followed by "Soya beans" and "Sweet Potatoes," suggesting these crops are grown mainly for animal feed. The strong correlation of 0.98 between 'Feed' and 'Import Quantity' implies that a rise in feed correlates with an increase in imports, potentially indicating that most imports serve animal feed needs. The heatmap of the correlation matrix in Fig. 10 further indicates strong positive correlations between 'Loss' and 'Processed' (0.86) and 'Loss' and 'Stock Variation' (0.86), hinting at possible inefficiencies in food processing and storage.

In Fig. 11, the 2D density plot between 'Feed' and 'Loss' shows a high density of points at lower values. This suggests that most items have low feed and loss quantities, making them ideal candidates for redistribution efforts. These visualizations collectively provide a comprehensive understanding of Malaysia's food waste and loss landscape, offering actionable insights for optimizing the agricultural supply chain.

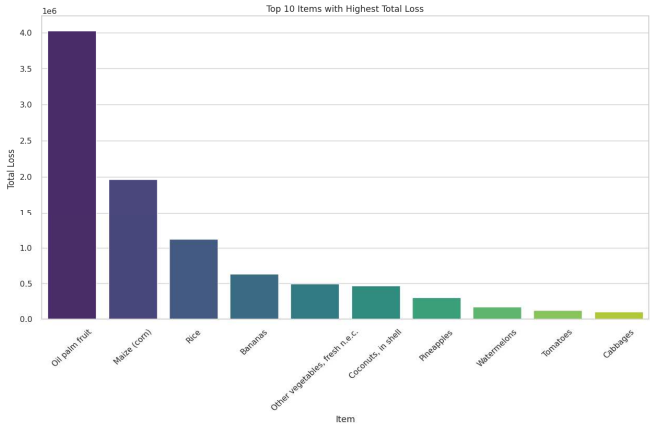

Fig. 8 Bar Plot of Total Loss by Item

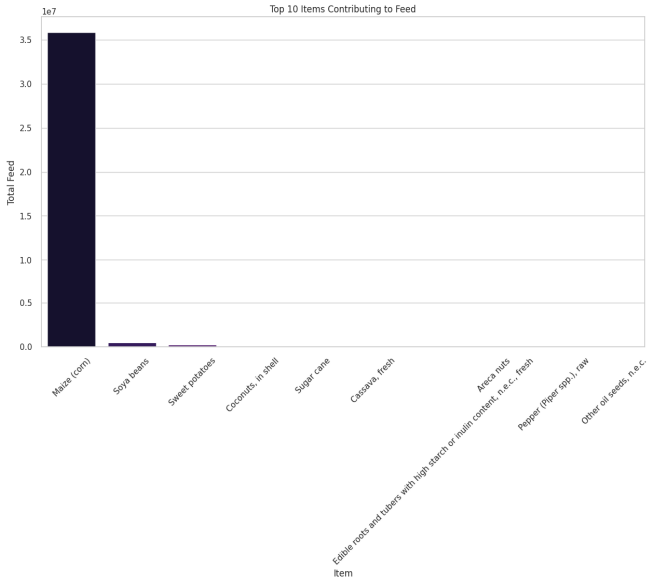

Fig. 9 Bar Plot of Top 10 Items Contributing to Feed

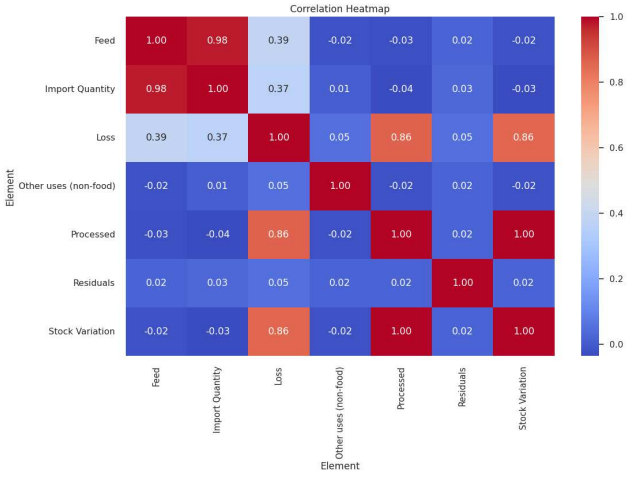

Fig. 10 Heatmap of Correlation Matrix

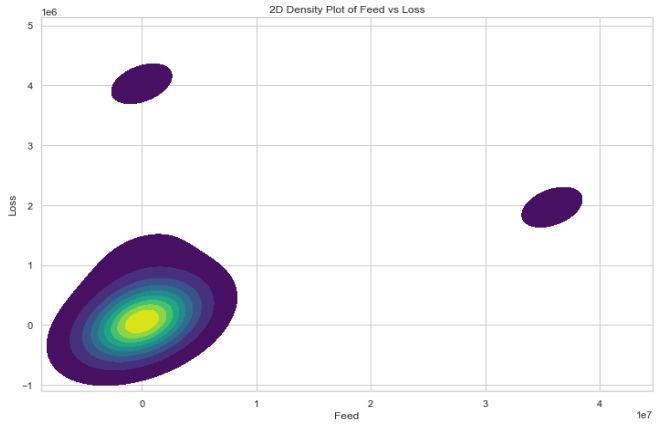

Fig. 11 2D Density Plot of Feed vs Loss

Applying visualization techniques in the GHG dataset helped one to acquire a better knowledge of GHG emissions in Malaysia. First plotted to show the trend of total GHG emissions over the years, a line graph offered a temporal viewpoint on how emissions have changed. This visualization was essential for spotting any trends, spikes, or emissions drop, providing the background for more investigation. A bar graph and pie chart were also produced to match the line graph and show the overall GHG emissions over the years. The pie chart gave a proportional view of yearly emissions that emphasizes the years with the most important contributions; the bar plot presented another angle on the temporal trends, allowing a year-by-year comparison of total emissions.

For 2020, extra bar graphs and pie charts were created to divide GHG emissions by element to highlight the most current statistics. The bar graph gave a comprehensive picture of the emissions for every element, enabling a simple comparison of their magnitudes. By contrast, the pie chart gave a whole picture of the share of every element in the total GHG emissions for that particular year. These visualizations, providing a whole picture of Malaysia's present GHG emissions, will serve as a fundamental basis for future studies and interpretations.

D. Clustering Algorithm

Three clustering algorithms, the Gaussian Mixture Model (GMM), BIRCH, and Density Peak Clustering, are applied in preprocessed food production and food waste and loss datasets. The GMM is a commonly used soft clustering method that approximates complex probability distributions by combining multiple weighted Gaussian distributions. In our analysis, each Gaussian distribution represents a distinct coverage class. In the context of our data, where we have observation vectors denoted by Y with specific attributes, the GMM aims to find κ mixture models, each with its mixture weight πi, mean vector μi, and covariance matrix Σi. GMM can then effectively capture multiple underlying patterns or clusters within the data, which makes it a versatile tool for various analytical applications as those components collectively model the observed data distribution. GMM's formula is shown below:

$$p(Y \mid \theta) = \sum_{k=1}^{K} \alpha_k \varphi_k \left(Y \mid \mu_k \sum_{k} \right) \quad (1)$$

$$\varphi_k(Y \mid \mu_k, \dot{a}_k) = (2\pi)^{-\left(\frac{d}{2}\right)} |\dot{a}_k|^{-\left(\frac{1}{2}\right)} \exp\left\{ -\frac{1}{2} (Y - \mu_k)^T \dot{a}_k^{-1} (Y - \mu_k) \right\} \quad (2)$$

One of GMM's primary advantages is its capability to model elliptical clusters, which allows greater flexibility in capturing the underlying patterns in the data. Applying GMM to the food production and waste and loss datasets could identify the distinct Gaussian distributions representing different clusters.

BIRCH is a hierarchical clustering algorithm designed to process large datasets efficiently. Large datasets are efficiently processed by BIRCH, which incrementally and dynamically clusters incoming multi-dimensional metric data points. The algorithm's essence lies in using the CF-tree, a height-balanced tree data structure that summarizes the data's information. The nuances of food production and waste patterns are identified through BIRCH clustering on both datasets. It was especially beneficial given its ability to handle large datasets and produce a set of compact yet interpretable clusters. The clustering feature of a 3d vector is defined as CF = (n, LS, SS) where n is the number of objects in the cluster, and LS and SS are defined by the formula below:

$$LS = \sum_{i=1}^{n} x_i \text{ and } SS = \sum_{i=1}^{n} {x_i}^2 \quad (3)$$

Density Peak Clustering works on the principle of identifying cluster centers as density peaks, which are data points that neighbors with lower density surround. To determine these cluster centers, the algorithm leverages a decision graph that plots density against distance to determine them. Clusters are then formed based on the relative densities and distances of data points from these centers. Potential crops of focus for intervention and further analysis are highlighted when Density Peak Clustering is applied to our dataset, which provides insights into clusters based on the pattern of food production and food waste and loss. The Density Peak Clustering's formula consists of 3 focuses, which are:

1) Calculating local density:

$$\rho_i = \sum_{x_j \in X} \chi\left(\text{Eudist}\left(x_i, x_j\right) - \text{dist}_{\text{cutoff}} \right), \quad \chi(x) = \begin{cases} 1, & x \le 0 \\ 0, & x > 0 \end{cases} \quad (4)$$

2) Determine the cluster centroid via the decision graph:

$$\delta_i = \max_{x_j \in X, j \neq i} \text{Eudist}(x_i, x_j), \quad \delta_i = \min_{j:\rho_j > \rho_i} \text{Eudist}(x_i, x_j) \quad (5)$$

3) Forming clusters based on the centroid and its nearest high-density neighbor:

$$\text{For } \forall x_i \in X, \text{ the matrix } (\rho_i, \delta_i) \quad (6)$$

For the clustering algorithms that required a predefined number of clusters, specifically GMM and BIRCH, the cluster number selection was a pivotal step. For GMM, multiple criteria were employed to determine the optimal number of clusters. The Silhouette Score was used to ensure both cohesiveness within clusters and separation between them, as shown in Fig. 13. Additionally, both the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) were utilized as shown in Fig. 12. These criteria provided a balance between the goodness of fit of the model and the complexity of the model, with lower values indicating better models. By comparing the BIC and AIC values for different numbers of clusters, the optimal number of clusters was three clusters, so that the underlying patterns in the data were effectively captured without being overly complex.

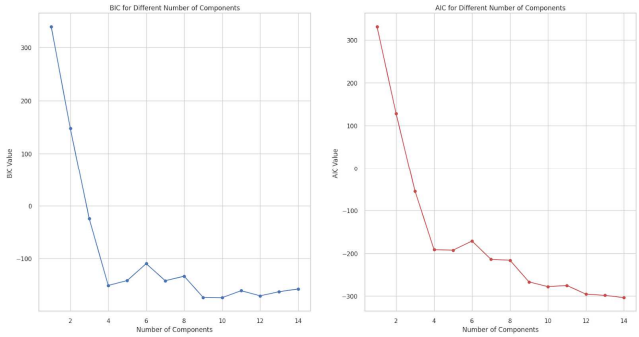

Fig. 12 BIC & AIC Plot

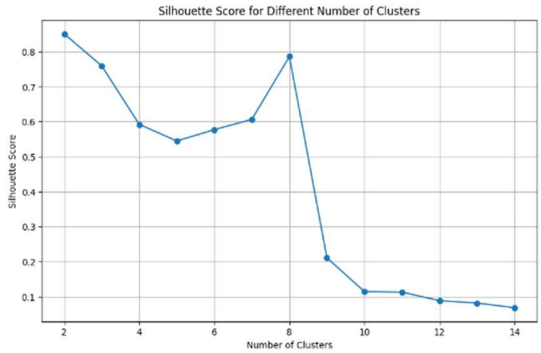

Fig. 13 Silhouette Plot

For BIRCH, the Distortion Elbow Method was applied. This method involves plotting the sum of squared distances against the number of clusters. By observing the point where the rate of decrease sharply changes, known as the 'elbow', the optimal cluster count identified was three clusters, where a balance between precision and computational efficiency was ensured. This is shown in Fig. 14.

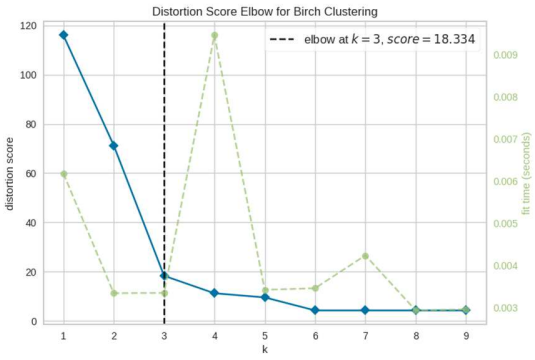

Fig. 14 Distortion Elbow Plot

Once the ideal number of clusters was decided upon and the data was grouped using the three techniques, each cluster was examined to grasp its unique qualities. Descriptive statistics and visualisers were produced for every cluster so it could spot trends, anomalies unique to every group. This stage was essential in comprehending the nature of food production and use inside every cluster, so enabling more focused recommendations and insights.

The next stage connected the results with the dataset of greenhouse gas emissions using the clusters found and described. By mapping the patterns of food production and waste from every cluster to their respective GHG emissions, one obtained a thorough integrated knowledge of the environmental impact of various food production and waste patterns. This linkage provided a foundation for identifying which clusters or patterns of food production and waste had the most notable environmental impact. The simplified overall methodology workflow can be referred to in Fig. 15 below.

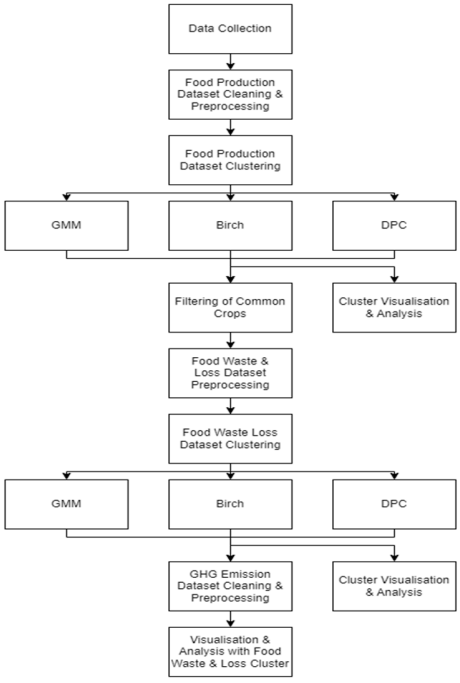

Fig. 15 Workflow of Methodology

III. Result and Discussion

In our analysis of the Food Production dataset, with the application of all three different clustering algorithms, Gaussian Mixture Model (GMM), BIRCH, and Density Peak Clustering, insightful outcomes were yielded. Specifically, GMM identified three distinct clusters, as shown in Fig. 16. It can be shown the cluster is differentiated into 3 colors, which are green, purple, and yellow, where each color cluster was characterized by different levels of area harvested and production. #Green represents crops with high production efficiency, where the production level is high, and the area harvested is low. Cluster

#Yellow represents crops that have low production efficiency, where the production level is low, and they are harvested in high. Cluster #Purple represents the crops within the cluster that have a consistent ratio of production to area harvested; where the area harvested increases, the production level also increases. Similar to GMM, Birch also gave a result of 3 clusters, as shown in Fig. 17. According to the plot, it could be seen that the clusters were classified by production efficiency, where Cluster #Purple was a crop with high production efficiency, Cluster #Green was a crop with low production efficiency, and Cluster #Red was a crop with consistent production to be harvested. However, for the cluster result found by DPC as shown in Fig. 18, the number of clusters found was only 1 cluster. It could be shown that DPC was unsuitable for clustering the food production dataset.

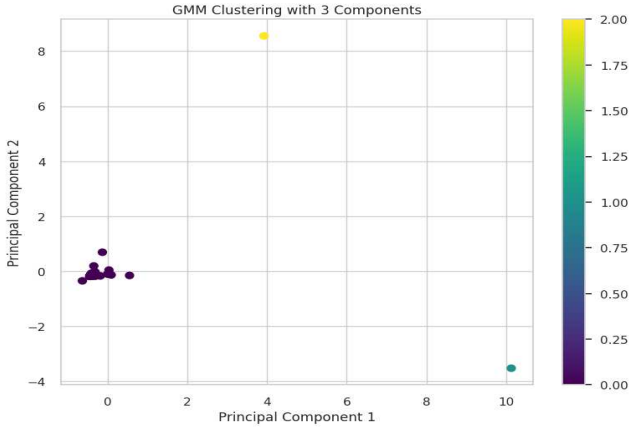

Fig. 16 Cluster Result of GMM

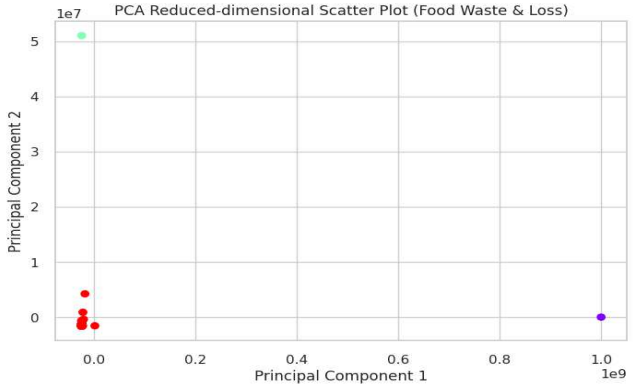

Fig. 17 Cluster Result of Birch

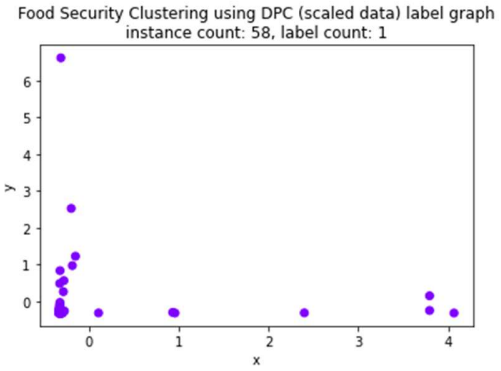

Fig.18 Cluster Result of DPC

In the Food Waste and Loss dataset, clustering algorithms unveiled unique crop utilization patterns. The clustering results offer valuable insights into key features and cluster distributions. Due to the small size of the dataset, the original scatter plot for clustering was challenging to interpret; therefore, a bar graph was used to represent cluster distributions. The analysis identified three significant clusters. Cluster #0 is characterized by high values in features like 'feed,' 'import quantity,' 'loss,' 'other uses (non-food),' 'processed,' and 'stock variation.' In contrast,

Cluster #1 primarily exhibits high values in 'import quantity' and 'loss.' Cluster #2 stands out with elevated values in 'feed,' 'import quantity,' 'loss,' and 'processed.' These findings corroborate the heatmap matrix, reinforcing import quantity and loss as pivotal indicators for clustering food waste and loss data. Our understanding of the dynamics in food waste and loss is enhanced by Cluster #2, which underscores the correlation between 'feed' and 'processed' features with 'import quantity' and 'loss', as beyond Cluster #0 shows generally high feature values.

In short, the clustering results provide a nuanced view of food waste and loss, highlighting the importance of 'import quantity' and 'loss' as key indicators. The clusters also suggest areas for targeted interventions and actionable insights to optimize the agricultural supply chain. As shown in Fig. 19, Fig. 20, and Fig. 21, it could be demonstrated that the GMM, Birch, and DPC found 3 cluster classes relating to food utilization.

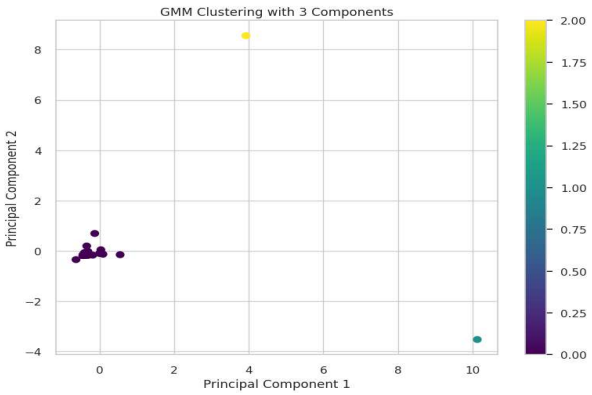

Fig. 19 Result of GMM Cluster on FWL Dataset

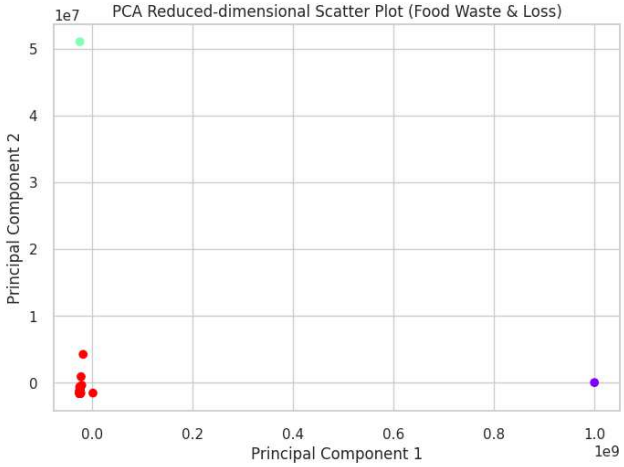

Fig. 20 Result of Birch Cluster on FWL Dataset

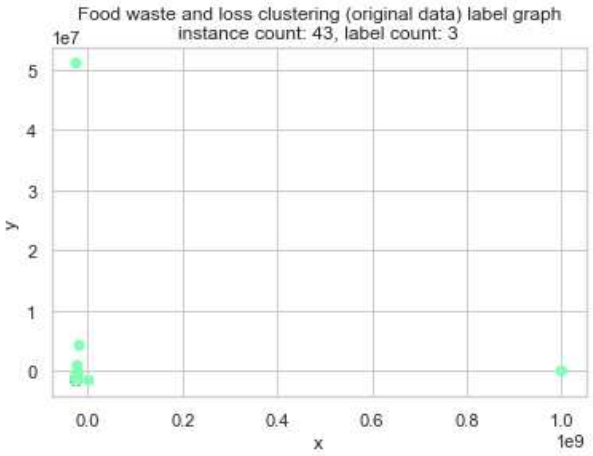

Fig. 21 Result of DPC Cluster on FWL Dataset

After the dataset was dimensionality reduced, the distribution of elements for each cluster was found, as shown in Fig. 22, Fig. 23, and Fig. 24, which shows which element each cluster contained.

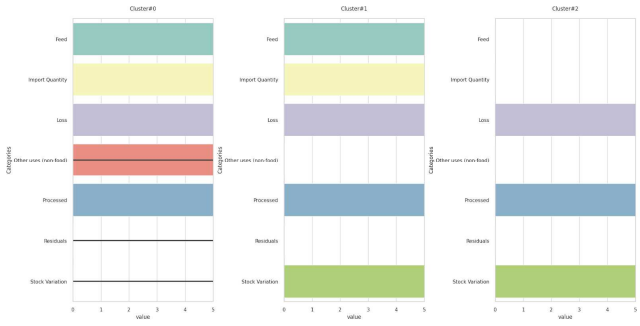

Fig. 22 Distribution of GMM Cluster on FWL Dataset

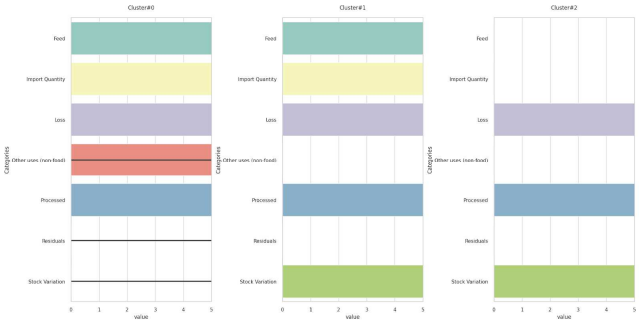

Fig. 23 Distribution of Birch Cluster on FWL Dataset

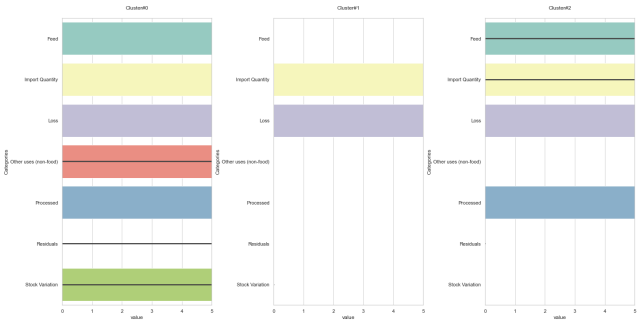

Fig. 24 Distribution of DPC Cluster on FWL Dataset

Performance score evaluations, which were Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index were performed on all three algorithms' clustering results on food production and food waste and loss datasets. A table was created to identify which algorithm performed the best, as shown in Table V and Table VI. Based on Table V, it could be demonstrated that Birch performed the best in clustering the food production dataset. In contrast, in Table VI, GMM and Birch got the same performance score and performed better in clustering the food waste and loss dataset. From this, it is shown that the cluster performed by both algorithms on the food waste and loss dataset was robust, meaning that there was likely a trustworthy representation of the actual patterns in the data.

TABLE V
PERFORMANCE SCORE OF 3 ALGORITHMS FOR FOOD PRODUCTION DATASET

	GMM	Birch	DPC
Silhouette Score	0.76	0.85	-0.49
Calinski-Harabasz Index	80.86	146.50	0.57
Davies-Bouldin Index	0.50	0.19	1.35

TABLE VI
PERFORMANCE SCORE OF 3 ALGORITHMS FOR FOOD WASTE & LOSS DATASET

	GMM	Birch	DPC
Silhouette Score	0.86	0.86	-0.64
Calinski-Harabasz Index	45.49	45.49	0.12
Davies-Bouldin Index	0.07	0.07	1.88

Continuing with our GHG emission findings from the dataset, we have found that from Fig. 25, around 1990, there was a big spike in emissions, which can be explained by the rise in Malaysia's economy due to a rise in industrial operations. Besides that, from Fig. 26, the energy consumption is abnormally high compared to the others, which suggests that LULUCF-related operations have caused this effect.

Based on research done by Hassan et al., the rise in the oil palm industry, which involves LULUCF operations, contributes to the GHG emissions in the Energy Element from our visualization [21]. To add on, CO_2 occupies a fair amount of the pie chart in Fig. 27 and Fig. 28, which indicates a strong relation of CO_2 being emitted due to the energy item from Fig. 26. This could be a strong indicator that LULUCF is emitting most of the CO_2 and to reduce the emission, proper and strict policies must be imposed.

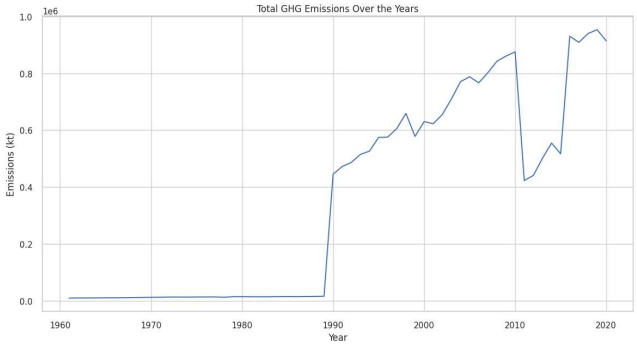

Fig. 25 Line Graph for GHG Emission

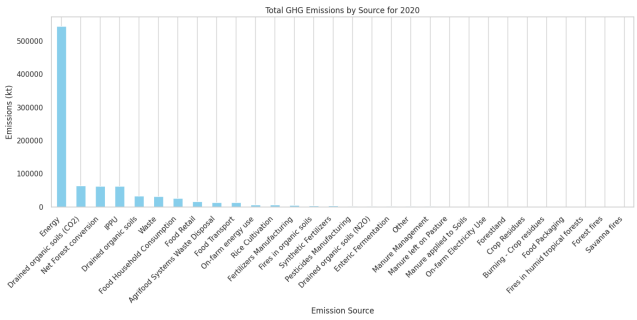

Fig. 26 GHG Emission Source Bar Chart for 2020

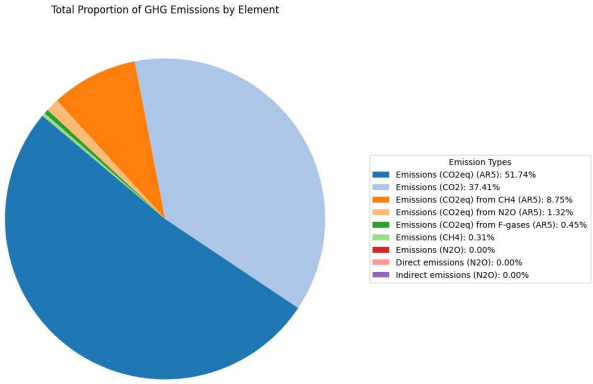

Fig. 27 Pie Chart for GHG Emission Source

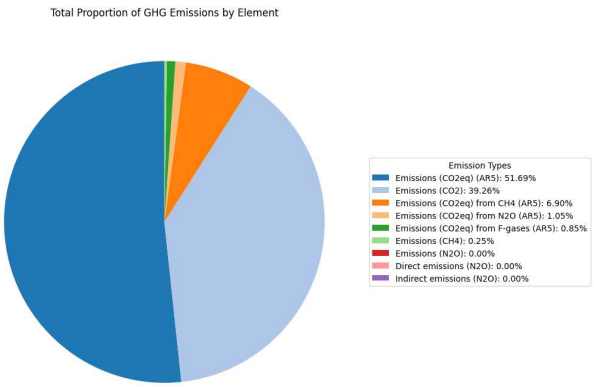

Fig. 28 Pie Chart for GHG Emission Source for 2020

The clustering results offer valuable insights into both food production and waste patterns. Identifying distinct clusters in the Food Production dataset suggests that targeted interventions could be more effective than a one-size-fits-all approach. Similarly, the clusters identified in the Food Waste and Loss dataset indicate areas where waste reduction efforts could be most impactful. For instance, the cluster characterized by high production but low utilization levels could be a key focus for waste reduction initiatives.

Malaysia's booming industrial operations, especially the rise of the oil palm industry and its LULUCF operations, have contributed to a significant emissions spike around 1990, as pinpointed by GHG analysis. CO_2, being a dominant emission in our findings, further underscores the impact of the Energy Item, suggesting LULUCF is a significant contributor. The results also pave the way for future research. The clustering algorithms could be expanded to other agricultural datasets or not limited to one nation, and the GHG emissions trends could be investigated in the framework of policy changes to evaluate their effectiveness. This research adds to the increasing corpus of knowledge meant to improve food security and slow climate change.

IV. Conclusion

Using advanced clustering algorithms and visualization tools, this study thoroughly investigated the complex interactions among food production, waste, and greenhouse gas emissions in Malaysia. Different clusters found by insights from the Food Production dataset underlined the need for focused interventions catered to particular production trends. The Food Waste and Loss dataset also revealed notable areas of waste that might be targeted for waste reduction initiatives.

Malaysia's industrial development, especially the expansion of the oil palm sector and related LULUCF activities, caused a notable increase in greenhouse gas emissions around 1990. The main concentration of CO2 emissions causes the notable environmental influence of these activities. Combining these results in many benefits when addressing food production efficiency and waste reduction, including improving food security and lowering negative environmental effects.

This study emphasizes the link between these problems and the need to discover all-encompassing answers since the world must simultaneously ensure food security and combat climate change. Future research could extend this analysis to other agricultural environments and areas, enhancing our knowledge and approaches to negotiating these worldwide issues. Through a data-driven approach, this study highlights the dynamics of food production, waste, and environmental impact. Therefore, it provides insightful analysis and roadmaps for Malaysia and other countries to have a more sustainable future.

Acknowledgment

The authors thank the reviewers for their insightful comments and suggestions, which improved this work.

References

[1] I. Vázquez-Rowe, K. Ziegler-Rodriguez, M. Margallo, R. Kahhat, and R. Aldaco, "Climate action and food security: Strategies to reduce GHG emissions from food loss and waste in emerging economies," Resources, Conservation and Recycling, vol. 170, p. 105562, Jul. 2021, doi:10.1016/j.resconrec.2021.105562.

[2] Y. Wang, Z. Yuan, and Y. Tang, "Enhancing food security and environmental sustainability: A critical review of food loss and waste management," Resources, Environment and Sustainability, vol. 4, p. 100023, Jun. 2021, doi: 10.1016/j.resenv.2021.100023.

[3] I. Manikas, B. M. Ali, and B. Sundarakani, "A systematic literature review of indicators measuring food security," Agriculture & Food Security, vol. 12, no. 1, May 2023, doi: 10.1186/s40066-023-00415-7.

[4] M. Kuiper and H. D. Cui, "Using food loss reduction to reach food security and environmental objectives – A search for promising leverage points," Food Policy, vol. 98, p. 101915, Jan. 2021, doi:10.1016/j.foodpol.2020.101915.

[5] C. Chauhan, A. Dhir, M. U. Akram, and J. Salo, "Food loss and waste in food supply chains. A systematic literature review and framework development approach," Journal of Cleaner Production, vol. 295, p. 126438, May 2021, doi: 10.1016/j.jclepro.2021.126438.

[6] C. L. Phooi, E. A. Azman, R. Ismail, J. Arif Shah, and E. S. R. Koay, "Food Waste Behaviour and Awareness of Malaysian," Scientifica, vol. 2022, pp. 1–11, Aug. 2022, doi: 10.1155/2022/6729248.

[7] M. A. Zubir, C. P. C. Bong, S. A. Ishak, W. S. Ho, and H. Hashim, "The trends and projections of greenhouse gas emission by the livestock sector in Malaysia," Clean Technologies and Environmental Policy, vol. 24, no. 1, pp. 363–377, Jul. 2021, doi: 10.1007/s10098-021-02156-2.

[8] Y. Dong and S. A. Miller, "Assessing the lifecycle greenhouse gas (GHG) emissions of perishable food products delivered by the cold chain in China," Journal of Cleaner Production, vol. 303, p. 126982, Jun. 2021, doi: 10.1016/j.jclepro.2021.126982.

[9] B. Bajželj, T. E. Quested, E. Röös, and R. P. J. Swannell, "The role of reducing food waste for resilient food systems," Ecosystem Services, vol. 45, p. 101140, Oct. 2020, doi: 10.1016/j.ecoser.2020.101140.

[10] G. Kalt et al., "Exploring the option space for land system futures at regional to global scales: The diagnostic agro-food, land use and greenhouse gas emission model BioBaM-GHG 2.0," Ecological Modelling, vol. 459, p. 109729, Nov. 2021, doi:10.1016/j.ecolmodel.2021.109729.

[11] G. Liu, F. Zhang, and X. Deng, "Half of the greenhouse gas emissions from China's food system occur during food production," Communications Earth & Environment, vol. 4, no. 1, May 2023, doi: 10.1038/s43247-023-00809-2.

[12] S. Wang, G. Azzari, and D. B. Lobell, "Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques," Remote Sensing of Environment, vol. 222, pp. 303–317, Mar. 2019, doi: 10.1016/j.rse.2018.12.026.

[13] R. S. Sucharitha and S. Lee, "GMM clustering for in-depth food accessibility pattern exploration and prediction model of food demand behavior," Socio-Economic Planning Sciences, vol. 83, p. 101351, Oct.

2022, doi: 10.1016/j.seps.2022.101351.
[14] R. S. Sucharitha and S. Lee, “Application of Clustering Analysis for Investigation of Food Accessibility,” Procedia Manufacturing, vol. 39, pp. 1809–1816, 2019, doi: 10.1016/j.promfg.2020.01.258.
[15] M. G. Abiad and L. I. Meho, “Food loss and food waste research in the Arab world: a systematic review,” Food Security, vol. 10, no. 2, pp. 311–322, Mar. 2018, doi: 10.1007/s12571-018-0782-7.
[16] R. Qasrawi et al., “Machine learning techniques for the identification of risk factors associated with food insecurity among adults in Arab countries during the COVID-19 pandemic,” BMC Public Health, vol. 23, no. 1, Sep. 2023, doi: 10.1186/s12889-023-16694-5.
[17] R. Diana, D. Martianto, Y. F. Baliwati, D. Sukandar, and A. Hendriadi, “Determinants of Household Food Waste in Southeast Asia: A Systematic Review,” Journal of Hunger & Environmental Nutrition, vol. 19, no. 5, pp. 792–803, Feb. 2023, doi:10.1080/19320248.2023.2174060.
[18] H. Jamaludin, H. S. E. Elmaky, and S. Sulaiman, “The future of food waste: Application of circular economy,” Energy Nexus, vol. 7, p. 100098, Sep. 2022, doi: 10.1016/j.nexus.2022.100098.
[19] N. Amirudin and T.-H. T. Gim, “Impact of perceived food accessibility on household food waste behaviors: A case of the Klang Valley, Malaysia,” Resources, Conservation and Recycling, vol. 151, p. 104335, Dec. 2019, doi: 10.1016/j.resconrec.2019.05.011.
[20] S. H. Vetter et al., “Greenhouse gas emissions from agricultural food production to supply Indian diets: Implications for climate change mitigation,” Agriculture, Ecosystems & Environment, vol. 237, pp. 234–241, Jan. 2017, doi: 10.1016/j.agee.2016.12.024.
[21] D. Renard and D. Tilman, “National food production stabilized by crop diversity,” Nature, vol. 571, no. 7764, pp. 257–260, Jun. 2019, doi:10.1038/s41586-019-1316-y.
[22] M. Rostami, M. Oussalah, and V. Farrahi, “A Novel Time-Aware Food Recommender-System Based on Deep Learning and Graph Clustering,” IEEE Access, vol. 10, pp. 52508–52524, 2022, doi:10.1109/access.2022.3175317.
[23] J. Q. Jerin, N. K. Khan, S. Biswas, and N. Sharmin, “ComparativeStudy of Clustering Algorithms: Scenario Based on Boston Crime Dataset,” 2023 26th Int. Conf. Comput. Inf. Technol. ICCIT 2023, pp. 1–6, 2023
[24] G. Jouan, A. Cuzol, V. Monbet, and G. Monnier, “Gaussian mixture models for clustering and calibration of ensemble weather forecasts,” Discrete and Continuous Dynamical Systems - S, vol. 16, no. 2, pp. 309–328, 2023, doi: 10.3934/dcdss.2022037.
[25] J. Lu, Y. Zhao, K.-L. Tan, and Z. Wang, “Distributed Density Peaks Clustering Revisited,” IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 8, pp. 3714–3726, Aug. 2022, doi:10.1109/tkde.2020.3034611.
[26] J. M. John, O. Shobayo, and B. Ogunleye, “An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market,” Analytics, vol. 2, no. 4, pp. 809–823, Oct. 2023, doi:10.3390/analytics2040042.
[27] N. Godcares, A. Sirsath, A. Bongale, P. Kadam, R. Jayawal, and S. Patil, “Exploring Customer Segmentation in the Context of Market Analysis,” 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), pp. 444–449, Oct. 2023, doi: 10.1109/r10-htc57504.2023.10461815.
[28] N. D. Sugiharto, D. Elbert, J. Arnold, I. S. Edbert, and D. Suhartono, “Mall Customer Clustering Using Gaussian Mixture Model, K-Means, and BIRCH Algorithm,” 2023 6th International Conference on Information and Communications Technology (ICOIACT), pp. 212–217, Nov. 2023, doi: 10.1109/icoiact59844.2023.10455950.
[29] N. Priya, K. Srinidhi, and T. Kousalya, “Carbon Footprint Monitoring System Using Machine Learning and Deep Learning Techniques,” 2023 12th International Conference on Advanced Computing (ICoAC), pp. 1–8, Aug. 2023, doi: 10.1109/icoac59537.2023.10250070.
[30] Z. Wang, H. Wang, H. Du, S. Chen, and X. Shi, “A novel density peaks clustering algorithm for automatic selection of clustering centers based on K-nearest neighbors,” Mathematical Biosciences and Engineering, vol. 20, no. 7, pp. 11875–11894, 2023, doi: 10.3934/mbe.2023528.
[31] A. S. Palli, J. Jaafar, M. A. Hashmani, H. M. Gomes, and A. R. Gilal, “A Hybrid Sampling Approach for Imbalanced Binary and Multi-Class Data Using Clustering Analysis,” IEEE Access, vol. 10, pp. 118639–118653, 2022, doi: 10.1109/access.2022.3218463.
[32] M. N. A. Hassan, P. Jaramillo, and W. M. Griffin, “Life cycle GHG emissions from Malaysian oil palm bioenergy development: The impact on transportation sector’s energy security,” Energy Policy, vol. 39, no. 5, pp. 2615–2625, May 2011, doi: 10.1016/j.enpol.2011.02.030.