University of Nairobi

ISSN: 2410-1397

Master Project in Social Statistics

# Poverty-based classification of households using cluster analysis.

**Research Report in Mathematics, Number 36, 2020**

Faith Mueni Musili

December 2020

School of Mathematics

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Social Statistics

**Master Project in Social Statistics** **University of Nairobi**

**December 2020**

# Poverty-based classification of households using cluster analysis.

**Research Report in Mathematics, Number 36, 2020**

Faith Mueni Musili

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Social Statistics

Submitted to: The Graduate School, University of Nairobi, Kenya

# Abstract

Poverty in rural areas is complex and multi-dimensional. Most of the poor households in Sub-Saharan Africa (SSA) rely on agriculture for livelihood. Agri-climatic shocks such as prolonged droughts, outbreak of animal and human diseases and crop and pest diseases make rural poor households in SSA vulnerable. Research gaps exist on poverty-based clusters in Kenya rural areas. The clusters would be fundamental in understanding the determinants of poverty.

This study uses K-means and K-medoid algorithms to identify poverty-based clusters in Kenya rural areas. The data used is collected from rural farming households. K-means and K-medoid algorithms are the most common clustering algorithms used and have been implemented by researchers.

The results show that rural poor households have low education level, high dependency ratio, low gender parity ratio, low income and low household diet diversity compared to rural non-poor households. Rural non-poor households own agricultural productive assets, seek extension services, are more aware of financial services and products available to farmers and access financial services more compared to rural poor households. Knowledge on the determinants of poverty in Kenya rural areas can be used by the government, institutions and partners, to formulate strategies and policies in an effort to reduce poverty. In future, research should be conducted on the role of land sizes and land tenure on poverty in rural Kenya.

**Master Thesis in Mathematics at the University of Nairobi, Kenya.**
**ISSN 2410-1397: Research Report in Mathematics**

**DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya**

# Declaration and approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

| | |
|---|---|
| | 23/11/2020 |
| Signature | Date |

Faith Mueni Musili
Reg No. I56/13195/2018

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

| | |
|---|---|
| | 24/11/2020 |
| Signature | Date |

Dr. Timothy Kamanu
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: tkamanu@uonbi.ac.ke

# Dedication

This dissertation is dedicated to my Son Nillan Mwangi who was barely six months old when I started this study programme. He was my inspiration throughout the journey.

# Contents

# List of Tables

# List of Figures

# Acknowledgments

I thank God for the health, academic ability and financial providence of my fees. A special thank you to Dr. Alex Mwaniki who shared the data used in this dissertation. Thank you IFAD and Government of Kenya (GOK) for funding the Kenya Cereal Enhancement Programme Climate Resillient Agricultural Livelihoods Window (KCEP-CRAL) programme.

I also thank my supervisor Prof. Timothy Kamanu for the guidance in this dissertation and the entire School of Mathematics fraternity.

To my husband Eddy and sisters Everlyne, Caroline and Mary, thank you for your daily support and words of encouragement during the entire duration of my masters. I can never thank you enough. To my dad and mum, thank you for your continued support, prayers and words of encouragement to keep soaring higher in the academic ladder.

To my classmates and friends, thank you for the support and motivation. A special thank you to my colleagues in the GeoScience Lab at ICRAF. Thank you for setting the academic standards higher.

To all others who have not been mentioned, I thank you for the role you played in the course of my masters programme.

Faith

Nairobi, 2020.

# 1 Introduction

## 1.1 Background

Poverty is the lack of essentials to satisfy one's basic needs (United Nations, n.d.). Poverty has an economic, political and social perspective. Economists define extreme poverty as living below $1.90 and moderate poverty as living below $3.10 per day (World Bank; International Monetary Fund, 2016). However, sociologists define poverty with focus to both individuals and households (Wagle, 2002). The social based definitions include: basic needs, assets, social exclusion, human capability and human rights aspects of poverty (Nge'the & Omosa, 2016).

A tenth of the global population live in extreme poverty with over 80% living in Africa and Asia (Johnston, 2016). On average, Sub-Saharan Africa (SSA) has a 41% poverty rate (World Bank group, 2017). As of 2016, over 36% of Kenyan population lived in poverty (World Bank; International Monetary Fund, 2016).

Poverty exists if, household members cannot afford food, parents cannot afford to cloth their children, household members cannot afford health care, parents cannot afford to educate their children or when young girls are married off because parents cannot afford basic needs (Ahmad & Ejaz, 2011). Poverty is manifested at household level but the effects are countrywide and global.

Determinants of poverty include household characteristics such as: education level, gender of household head, income, family size, dependency ratio and per capita expenditure. Poverty can be reduced by decreasing the household size and number of dependants (Ahmad & Ejaz (2011); Orbeta (2006)). Increasing households education level decreases chances of poverty (Chaudhry & Rahman (2009); Jamal (2005)). Income is key in poverty reduction (Pervez & Usman, 2011). In Nigeria, Akerele & Adewuti (2011) investigated socio-economic determinants of poverty in Ekiti state. They concluded that female-headed households are more susceptible to poverty compared to male-headed households. Bogale et al. (2005), found a strong association between poverty in rural areas and lack of human capital and assets. According to Geda et al. (2005), level of education, engagement in agricultural activity and household size are crucial in alleviating poverty in rural Kenya. This study classifies rural farming households into poor and non-poor using clustering algorithms.

## 1.2 Problem statement

Poverty reduction is a core global agenda with resources, policies and strategies being put in place. The major strategies include the Millennial Development Goals (MDGs) and Sustainable Development Goals (SDGs). MDG one aimed at halving the proportion of poor people in developing regions. This target was achieved between 1998 and 2010 but the change was uneven. The overall change was contributed by few Asian countries like China and India with low-income regions such as Sub-Saharan Africa (SSA) registering a small decline (United Nations, 2015). After 2015, SDG one was formulated with the aim of eradicating poverty by 2030.

Kenya has registered significant decline in poverty rates over the years including 10.7% decline between 2005-2006 and 2015-2016. Even with such decline, poverty is improbable to be eradicated by 2030 (World Bank Group, 2018). To reduce poverty in Kenyan rural areas, research on the determinants of poverty is core. This knowledge will inform the design and formulation of interventions and strategies.

## 1.3 Objectives

The main objective is to identify the household characteristics that differentiate rural poor and non-poor households using clustering algorithms. The specific objectives are:

1. To identify the household characteristics that differentiate rural poor and non-poor households.
2. To determine the direction of the characteristics in the rural clusters of poor and non-poor.
3. To evaluate performance of the clustering algorithms used.

## 1.4 Justification

Four out of six people living in extreme poverty live in rural areas (Sachs, 2014). 72% of Kenyan population live in rural areas with 36% being poor (Kenya National Bureau of Statistics, 2020). Most of the poor households in Kenyan rural areas rely on agriculture for livelihood and therefore, poverty reduction has been difficult to achieve with the increasing agro-climatic shocks.

Geda et al. (2005) examined the probable determinants of poverty in Kenya using binomial and multinomial logit models. Ahmad & Ejaz (2011) investigated the significant factors in clusters of poor and non-poor and their direction using clustering techniques. Mukherjee et al. (2011) suggested a fuzzy approach for identification of poor households.

This study was necessitated by the gaps in contextualized literature on household characteristics that determine poverty in rural Kenya. The gaps include the roles of; knowledge and access to financial subsidy programs to farmers, access to extension services, ownership of agricultural productivity assets and family size in poverty reduction. This study classifies households into poor and non-poor using clustering algorithms. The algorithms group similar households together based on the variables present. Clustering algorithms were used because the dependent variable was unknown.

# 2 Literature review

## 2.1 Dynamics of poverty

Poverty is multi-dimensional and complex. The multi-dimensional aspects include social, economic and political elements. Absolute poverty is the total lack of means to meet basic needs and is independent of location (United Nations, n.d.). Relative poverty is the lack of means to meet minimum living standards compared to other individuals in that location (Sabates, 2008).

Resources have been invested to reduce poverty across the globe in the form of commitments such as the MDGs and SDGs. Some causes of poverty in Africa include societal and political greed, poor governance structures, spatial inequality during distribution of resources, unemployment, poor infrastructure and political instability (Melake & Merhawi, 2018). Accumulation of debt from donor countries and institutions have contributed to the inability of African governments to invest in poverty reduction. However, corruption and misappropriation of funds and donations by some African leaders cannot be overlooked. Poverty reduction requires strong institutions, non-marginalization of communities during resource allocation and transparency in government systems (Tazoacha, 2001).

In Kenya, poverty reduction has been on the development agenda since independence (Tazoacha, 2001). Kenya has recorded remarkable decline in poverty over the years but poverty eradication is yet to be achieved. Pandemics like COVID-19 have highlighted instability and non-resilience of Kenyan households.

## 2.2 Determinants of poverty in rural areas

Poverty in rural areas is more pronounced in developing countries compared to developed countries (Jazaïry et al., 1992).

In Sub Saharan Africa, most of the rural poor rely on agriculture for livelihood. Agri-climatic shocks such as prolonged droughts, outbreak of animal and human diseases and crop and pest diseases make rural poor households vulnerable. Knowledge of the soil and water conservation methods and drought-resistance farm inputs is key in maximization of yield.

Bogale et al. (2005) studied the extent of poverty in rural Ethiopia using Foster–Greer–Thorbecke (FGT) poverty index and binary logit estimates. 40% of the households were rural poor. They found a linkage between lack of household assets and poverty in rural areas.

Etim & Udoh (2013) investigated the determinants of poverty in rural Nigeria using Tobit regression. Farm income, farm size, amount of agricultural loan, participation in farmer cooperatives, ownership of certain assets, access to extension services, use of modern farm inputs, education level of male household heads, dependency ratio, farming experience and income were significant determinants of poverty.

Owuor et al. (2007) investigated the key determinants of poverty in rural Kenya using a probit model. Access to credit, livestock assets, location, education level and participation in farmer seminars influenced the poverty status of the households. Female-headed households and distance to the market increased the probability of persistent chronic poverty.

Malik (1996) investigated the determinants of poverty in rural Pakistan using a logit model. Education, health status of household members, gender of household head, access to markets, person per room, sex ratio, distance of school from house and dependency ratio were significant determinants.

## 2.3 Methods in poverty-based classification of households

Mathematical and descriptive methods are used in poverty-based classification of households. Below are some mathematical methods used.

Tvedten & Nangulah (n.d.) classified households in Namibia as very poor, poor and non-poor. Standardized Consumption Level (SCL) and Composite poverty indices (CPI) methods were used. 53% of the households had SCL of less than N$7200 and were classified as poor. A household that spend 60-79% of total income on food was classified as poor while a household that spend 80-100% was very poor. 10% of the households in Namibia were very poor, 30% were poor while 60% were not poor.

Geda et al. (2005) investigated the determinants of poverty in Kenya using binomial and polychotomous logit models. The poor and non-poor were identified and their probability of being in extreme poverty calculated. They assumed that a response variable which captured the economic status of individual or household determined their placement into poor and non-poor categories. Engagement in agriculture, household size and education level were key determinants of poverty in Kenya.

Ahmad & Ejaz (2011) investigated the determinants of poverty in the clusters of non-poor and poor and their direction using two-step cluster analysis. Household income, family size, sex ratio, household education and dependency ratio were significant.

## 2.4 Support Vector Machines (SVMs)

Support Vector Machines are non-probabilistic, binary and linear supervised learning models. They were developed at AT&T Bell Laboratory by Boser et al. (1992), Guyon et al. (1993) and Cortes & Vapnik (1995). SVMs are used for both regression and classification of data.

Each training data point is placed into a category. The SVM algorithm builds a model that places new data points into either of the available categories. SVMs use kernel trick to perform linear and non-linear classifications.

Many possibilities of hyper planes exist. The SVM algorithm selects a plane with the largest margin distance for future points to be categorised with high confidence.

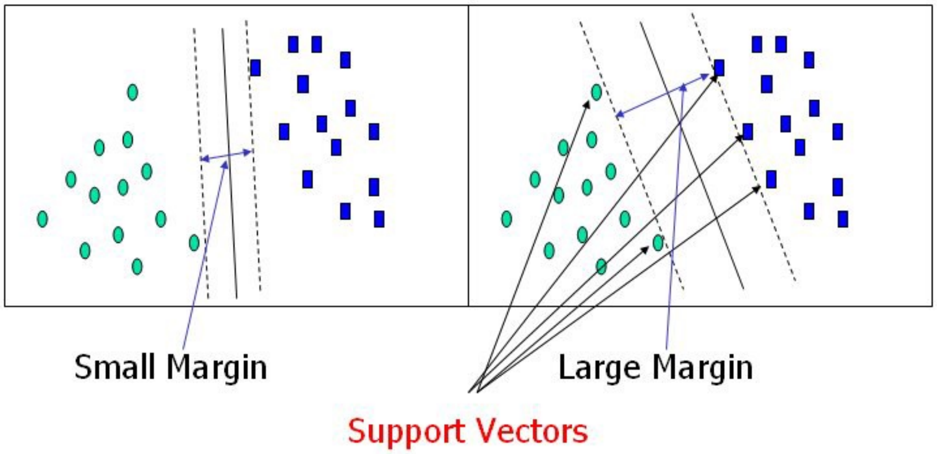

**Figure 1. Support vectors**

Support vectors determine position and orientation of the hyperplane.

**Advantages of Support Vector Machines**

1. Support Vector Machines are effective when the number of dimensions is greater than the number of samples.
2. Support Vector Machines work well when there is a clear margin of separation between classes.

3. Support Vector Machines are more effective in high dimensional spaces.
4. Support Vector Machines are suitable in extreme case binary classification.

**Disadvantages Support Vector Machines**

1. Implementation of Support Vector Machine algorithm is time consuming and therefore not suitable for large data sets.
2. Support Vector Machines do not perform well when the target classes overlap.
3. Support Vector Machines under perform when the number of features for each data point exceed the number of training data samples.
4. Support Vector Machines do not give a probabilistic explanation for the classification of data points.

## 2.5 Unsupervised clustering algorithms

If a data set does not contains pre-existing labels, unsupervised clustering algorithms are used to group the data. Unsupervised classification require minimum human supervision. A distance measure is used to compute the dissimilarity matrix between the data points. Similar data points are placed in the same cluster and dissimilar data points are placed in different clusters.

K-means, K-medoids, hierarchical, model-based and density-based clustering algorithms are explained below.

### 2.5.1 K-means algorithm

K-means algorithm is the most common unsupervised classification algorithm. Hartigan & Wong (1979) is the standard K-means algorithm used. This algorithm states that total intra-cluster variation is equal to the sum of squared distances (Euclidean distances) between centroids and their items.

$$W\left(C_k\right) = \sum_{x_i \in C_k} \left(x_i - \mu_k\right)^2 \tag{1}$$

where:
- $x_i$ is a data point belonging to cluster $C_k$

- $\mu_k$ is the mean value of the data points assigned to cluster $C_k$

Each observation $x_i$ is assigned to a cluster such that the sum of squared distance is minimized.

$$\text{tot.withiness} = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{2}$$

K-means algorithm requires pre-specification of a distance measure and number of clusters. The specified distance measure is used to calculate similarity between data points. Euclidean and Manhattan distances are the most common distance measures used. Other distance measures include Pearson, Spearman and Kendall correlation distance measures.

**Euclidean distance**

$$d_{\text{euc}}(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

Where x and y are two vectors of length n.

**Manhattan distance**

$$d_{\text{man}}(x,y) = \sum_{i=1}^{n} |(x_i - y_i)| \tag{4}$$

Where x and y are two vectors of length n.

**Pearson correlation distance**

Pearson correlation measures the degree of a linear relationship between two profiles.

$$d_{cor}(x,y) = 1 - \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{5}$$

Where x and y are two vectors of length n.

### Spearman correlation distance

The spearman correlation method computes the correlation between the rank of x and the rank of y variables.

$$d_{\text{spear}}(x,y) = 1 - \frac{\sum_{i=1}^{n}\left(x'_i - \bar{x}'\right)\left(y'_i - \bar{y}'\right)}{\sqrt{\sum_{i=1}^{n}\left(x'_i - \overline{x'}\right)^2 \sum_{i=1}^{n}\left(y'_i - \overline{y'}\right)^2}} \tag{6}$$

Where $x'_i = \text{rank}\left(x_i\right)$ and $y'_i = \text{rank}(y)$

### Kendall correlation distance

Kendall correlation method measures the correspondence between the ranking of x and y variables.

$$d_{kend}(x,y) = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{7}$$

Where;
- $n_c$ : total number of concordant pairs
- $n_d$ : total number of discordant pairs
- $n$ : size of $x$ and $y$

### Optimal number of clusters

Elbow, Silhouette and Gap statistic are the common methods used to determine the optimal number of clusters.

### Elbow method

The Elbow method involves computing the within total sum of squares (WSS) for different clusters. The optimal number has minimal WSS.

WSS is plotted against the clusters, k. The position of the “bend knee” in the plot is the optimal number of clusters.

### Silhoutte method

The silhoutte method was developed by Kaufman & Rousseeuw (1990). This method involves computation of the average silhouette width ($Si$) for various clusters. High $Si$ implies that the data is well clustered.

$$\mathrm{Si} = (x-y)/\max(x,y) \tag{8}$$

where:
$x$ is the mean distance to the closest cluster.
$y$ is the mean-intra cluster distance

### Gap statistic

The Gap statistic method was developed by Tibshirani et al. (2001). This method compares total within intra-cluster variation for different number of clusters, k and their expected values. The value of k with the highest gap statistic is the optimal number.

$$\mathrm{Gapstatistic}(k) = \frac{1}{P}\sum_{b=1}^{P} \log\left(W_{kp}^{*}\right) - \log\left(W_k\right) \tag{9}$$

Select $k$ such that $\mathrm{k}+1 : \mathrm{Gap}(k) \geq Gaq(k+1) - s_{k+1}$

### Advantages of K-means algorithm

1. K-means algorithm is simple and fast to implement.
2. K-means algorithm works well with large data sets.
3. K-means algorithm generalizes clusters to different shapes and sizes.

### Disadvantages of K-means algorithm

1. K-means algorithm require pre-specification of the number of clusters.
2. K-means algorithm is sensitive to outliers and different ordering of data gives different results.
3. K-means algorithm does not perform well when clusters are of varying sizes and density.

### 2.5.2 K-medoids algorithm

K-medoid algorithm was first developed by Kaufmann & Rousseeuw (1987) and is a variant of K-means algorithm. The most common K-medoids algorithm is the Partitioning Around Medoids(PAM) algorithm (Kaufman & Rousseeuw, 1990). Unlike K-means algorithm, K-medoid algorithm uses actual data points as cluster centres. These data points are called cluster medoids. The medoid is selected such that the sum of distances to other points is minimized.

The distance measure and number of clusters are pre-specified.

**Advantages of K-medoids algorithm**

1. K-medoid algorithm is less sensitive to noise and outliers compared to K-means algorithm.
2. K-medoid algorithm is simple to understand and easy to implement.
3. K-Medoid algorithm is fast compared to other partitioning algorithms

**Disadvantages of K-medoids algorithm**

1. K-medoid algorithm is not suitable for clustering non-spherical data.
2. Different results may be obtained in re-runs of the algorithm as the medoids are randomly selected.
3. K-medoid algorithm requires pre-specification of the number of clusters.

### 2.5.3 Hierarchical clustering algorithm

Hierarchical clustering algorithm places data points into clusters using a hierarchy. The clusters are distinct from each other and the objects within each cluster are similar to each other.

The main output of the hierarchical clustering algorithm is a dendrogram, which shows the hierarchical relationship between the clusters.

The main types of hierarchical clustering algorithms are agglomerative and divisive. Agglomerative hierarchical clustering algorithm uses the top-down approach of clustering the data points. Divisive hierarchical clustering algorithm uses the bottom-top approach. Divisive hierarchical clustering algorithm is not common in practice.

### Advantages of hierarchical clustering algorithm

1. Hierarchical clustering algorithm does not require pre-specification of clusters used.
2. The output of hierarchical clustering algorithm is an attractive tree-like diagram called a dendogram.

### Disadvantages of hierarchical clustering algorithm

1. Hierarchical clustering algorithm involves many arbitrary decisions and is therefore considered unreliable.
2. Hierarchical clustering algorithm does not work with missing data.
3. Hierarchical clustering algorithm does not perform well when the data set has mixed data types.
4. Hierarchical clustering algorithm is not effective in large data sets.
5. Most people misinterpret the dendrogram.

## 2.5.4 Model-based clustering algorithm

Model-based clustering algorithm is based on the assumption that each cluster follows a parametric distribution. The data is modelled using normal or a mixture of Gaussian distributions where each data point is placed into a cluster (Fraley & Raftery (2002); Fraley et al. (2012)). All points have the probability of belonging to a cluster. A covariance matrix is used to determine geometric features of the cluster.

The mixture of Gaussians distribution equation is shown below:

$$\text{f(x)} = \sum_{k=1}^{K} \alpha_k f_k(x) \tag{10}$$

Where; $b_k$ is the contribution of the $k^{th}$ component in constructing f(x).

The model parameters are estimated using the Expectation-Maximization (EM) algorithm. The EM algorithm is initialized by hierarchical model-based clustering. Each cluster k is centered at the mean $\mu_k$.

### The Expectation-Maximization (EM) algorithm

EM algorithm assumes that the data is made up of multivariate normal distributions and can be used to maximize the likelihood function if some variables are unobserved.

#### Steps involved in EM algorithm

1. *E-step:* Solve for the distribution of the data using the model equation.
2. *M-step:* Maximize the expected likelihood function and estimate the model parameters.
3. Re-run E and M steps until convergence.

#### Advantages of Model-based clustering algorithm

1. Model-based clustering algorithm includes researched statistical inference methods and is therefore considered reliable.
2. Model-based clustering algorithm clusters are based on distributions.
3. Model-based clustering algorithm calculates a density estimation per cluster.

#### Disadvantages of Model-based clustering algorithm

1. Model-based clustering algorithm is slow when working with large data sets.

### 2.5.5 Density-based algorithm

Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996) is the common density-based clustering algorithmm used. DBSCAN algorithm is used to find arbitrary shaped clusters and to detect outliers. The algorithm works using a parametric approach. The clusters are constructed using density reachability and density connectivity approaches. A point q is considered density-reachable from a core point p if q is within the eps-neighborhood of p. Two points p and q are called density-connected if there is a third point t from which both p and q are density-reachable (Martin, 2009).

DBSCAN parameters are:

1. *e (eps)*: radius of the neighborhoods.
2. *minPts*: minimum number of data points that make up a cluster.

DBSCAN algorithm is sensitive to the choice of eps. A small eps gives sparse clusters while a large eps gives dense clusters which may be merged together.

### Advantages of Density-based clustering algorithm

1. Density-based clustering algorithm does not require pre-selection of the clusters.
2. Density-based clustering algorithm gives arbitrary-shaped clusters if present.
3. Density-based clustering algorithm is not affected by noise and outliers.

### Disadvantages of Density-based clustering algorithm

1. Variation in the density of data points makes noise points undetectable.
2. Density-based clustering algorithm is sensitive to selection of eps parameter.
3. The quality of the clusters depend on the distance measure used.
4. Density-based clustering algorithm does not work well with large data sets.

# 3 Methods

The data used in this study was collected as baseline for the Kenya Cereal Enhancement Programme Climate Resilient Agricultural Livelihoods Window (KCEP-CRAL) programme. KCEP-CRAL is a ten-year programme initiated in 2014 and is co-financed by the Government of Kenya (GoK), European Union (EU) and International Fund for Agricultural Development (IFAD). The overall development objectives of the programme are to contribute to national food security and smallholder income generation by supporting farmers to increase the productivity and profitability of key cereal commodities – maize, sorghum, and millet, and associated pulses.

The programme covers Embu, Kitui, Machakos, Tharaka Nithi, Kilifi, Kwale, Makueni and Taita taveta counties in Kenya. The location of the counties is shown in figure 2.

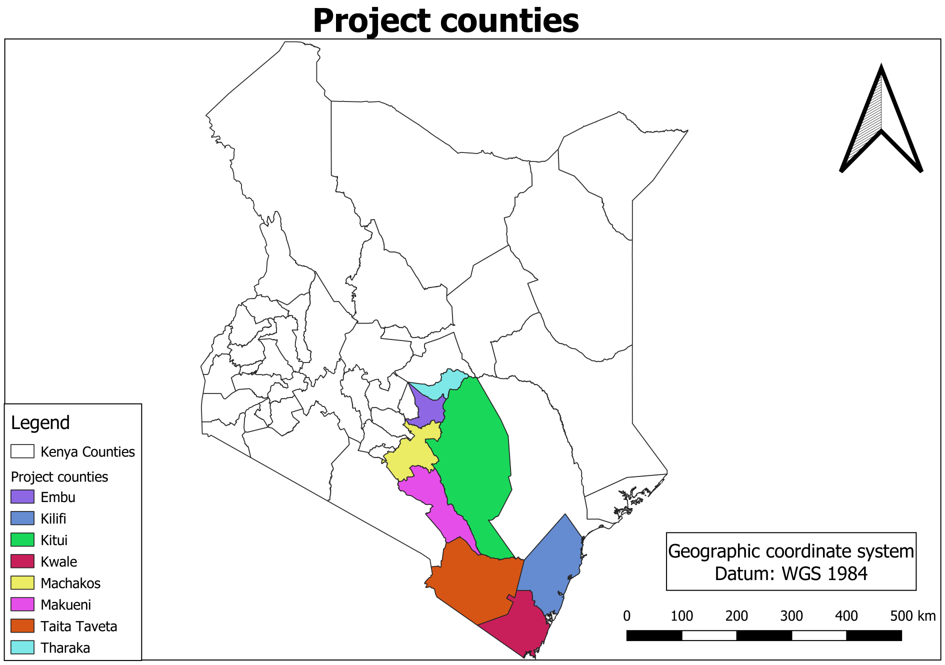

**Figure 2. Project counties**

## 3.1 Sampling design

The formulation of the baseline survey followed consultations with key stakeholders - GoK, IFAD and EU. Both quantitative and qualitative approaches were used to complement each other. Quantitative methods were used to collect and analyse household data while qualitative methods were used to draw insights from key informants and the farmers.

The sample size used was 1,050 potential beneficiary households and 1,050 control households. The sample size was determined based on Cochran's equation for N large at above 10,000 (Cochran, 1977).

$$n_0 = \frac{z^2pq}{e^2} \tag{11}$$

## 3.2 Data pre-processing

A mobile platform was used to collect the data. Therefore, no data entry was required. The data collected had 2100 observations and 2265 variables.

### 3.2.1 Removing missing data

The methods used in this study require complete data. 96% of the variables had missingness of above 3%. These variables were excluded from the study. The priority was to retain as many observations as possible. Variables with missingness of above 3% were excluded and then all observations with missing values excluded from the study. The resulting data had 2081 observations and 79 columns.

### 3.2.2 Calculated variables

Three variables were calculated and included in the data set. They are:

1. **Education level of a household**

   Education level of a household is the average of education levels of all household members. Education levels were categorised as follows: 0 points for a household member with no education; 5 points for a household member with up to primary school education or has attended youth or village Polytechnic; 10 points for a household member with up to secondary school education; and 15 points for a household member with up to college or university education.

2. **Sex ratio**
   Sex ratio is the ratio of males to females in a household.

3. **Dependency ratio**

Dependency ratio in a household is the proportion of dependents to non-dependents. Dependents are household members of below 15 years and those of above 64 years. Non-dependents are household members of between 16 and 64 years. Non-dependants are able to take care of their basic needs and are of the working age while dependants rely on non-dependents for their basic needs. Low dependency ratio in a household means that the non-dependent household members are able to suffice the needs of the household (Simon et al., 2012).

### 3.2.3 Data standardization

The data was standardized. Standardization is the conversion of all variables into a common data format so that each variable can contribute equally to the analysis.

## 3.3 Exploratory data analysis

Exploratory data analysis was done to assess the feasibility of cluster analysis on the data and the cluster structures. The methods used are outlined below.

### 3.3.1 Feasibility of cluster analysis

Assessment of clustering tendency of the data was done to check whether clustering was suitable for the data and if the data contained meaningful clusters. The methods used were: Principal component analysis (PCA), Hopkin's statistic and visualization of the distance matrices.

**Principal component Analysis (PCA)**

PCA is a dimension-reduction method which transforms a large set of variables into fewer variables containing as much information as possible. PCA was developed by Pearson (1901); Hotelling (1933) and Jolliffe (2002)). PCA explains most of the variability in the data using fewer variables. Dimension reduction reduces the accuracy of the data.

**What are principal components?**

Principal components are new variables constructed as linear combinations or mixtures of initial variables. The principal components are uncorrelated. The first principal component contain maximum information. Then, the maximum remaining information is contained in the second principal component and so forth.

The first Principal component $Z_1$ is given by:

$$Z_1 = \phi_{11}Y_1 + \phi_{21}Y_2 + \ldots + \phi_{q1}Y_q \quad (12)$$

where $\phi_1$ is the first PC loading vector, with elements $\phi_{12}, \phi_{22}, \ldots, \phi_{q2}$. The $\phi$ are normalized.

The second Principal component $Z_2$ is given by:

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \ldots + \phi_{q2}X_q \quad (13)$$

All the subsequent principal components are calculated. The elements $\phi_{11}, \ldots, \phi_{q1}$ in equation (12) are the loadings of the first principal component calculated by maximizing the variance.

### How to choose principal components

Proportion of explained variance, eigen values and elbow method are used to select the most appropriate number of principal components.

1. **Proportion of explained variance**

   The first $x$ principal components which explain more than 70% of the total variation in the data are selected.

2. **Eigen values greater than 1**

   Use of eigen values to select the number of principal components uses the Kaiser criteria (Kaiser, 1960). The criteria states that eigen values greater than 1 are stable and should be included in the analysis (Girden & Kabacoff, 2001).

3. **Elbow method**

   An ideal scree plot curve is steep and then bends at an "elbow". After the "elbow", the curve flattens out. The "elbow" is chosen as the cutting-off point (Cattell, 1966).

PCA require use of variables with variance greater than 1. One variable with zero variance was excluded from this study. Aggregate income was used as the dependent variable. PCA was implemented in R programming language using the stats package version 3.6.2.

#### Hopkin's statistic

Hopkin's statistic (H) is a is a statistical hypothesis test which measures the clustering tendency of data (Hopkins & Skellam, 1954). The null hypothesis is that the data is generated using a poisson point process and is therefore uniform and randomly distributed. A H value close to 1 indicates that the data is clustered, random data results in values around 0.5 while uniformly distributed data results in values close to 0 (Aggarwal, 2015).

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} \tag{14}$$

Hopkin's statistic was calculated in R programming language using clustertend package version 1.4.

#### Distance matrices computation

Visualization of the distance matrices is used as a visual approach of assessing clustering tendency of data.

The implementation of this method was done in R programming language using factoextra package version 1.0.7.

### 3.3.2 Cluster structures

Random forests and Classification and Regression Trees (CART) algorithms were used to access the cluster structures of the data.

#### Classification and Regression Trees (CART)

CART algorithm was introduced by Breiman et al. (1983) and refers to decision tree algorithms that can be used for classification or regression predictive modelling. CART visualises numeric data using regression trees and categorical data using classification trees. In R programming language, CART is implemented using the name RPART (Atkinson & Therneau, 2000).

Constructing a classification or regression tree involves successive partitioning of data into groups based on the value of a predictor variable. The first partition takes the entire training data set and divides it into 2 groups. Each group is further divided into 2 subgroups and each of those subgroups is divided again. Partitioning the data into subgroups continues until all of the observations in a particular subgroup are the same or until some other criterion is met.

When deciding how to partition a group of observations, binary decisions involving each predictor variable are considered. The partition selected is the one which minimizes the residual mean deviance. For a regression tree, the residual mean deviance is defined by:

$$RMD = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-k} \tag{15}$$

Where:

- $Y_i$ is the i-th observed value of the dependent variable
- $\hat{Y}_1$ is the predicted value for the subgroup to which it is assigned
- $n$ is the total number of observations in the training set
- $k$ is the total number of subgroups (leaves) in the tree.

For a classification tree, the residual mean deviance is defined by:

$$RMD = \frac{\sum_{i=1}^{n} -2\log\left(p_{i,j}\right)}{n-k} \tag{16}$$

Where $p_{i,j}$ is the estimated probability that observation i would be assigned to class j and j is the class predicted by the model.

***Pruning the tree***

Pruning is done to remove unnecessary nodes from the tree. To decide which nodes to retain, fitting is done and stops when some level of $\alpha$ is attained. Where;

$$R_\alpha(S) = R(S) + \alpha|S| \tag{17}$$

$R_\alpha(S)$ is the cost for the tree
$R\left(S_0\right)$ is risk a zero split tree
$S_\alpha$ is the sub-tree.

$S_0$ is the complete model
$S_\infty$ is a form of the model with no splits.

The assumptions during pruning are:

1. If $S_1$ and $S_2$ are sub trees of $S$ with $R_\alpha(S_1) = R_\alpha(S_2)$, then either $|S_1| < |S_2|$ or $|S_2| < |S_1|$
2. If $\alpha > \beta$ then either $S_\alpha = S_\beta$ or $S_\alpha$ is a sub tree of $S_\beta$
3. When given some set of numbers $\alpha_1, \alpha_2, \ldots, \alpha_m$; both $s_{\alpha_1}, \ldots, s_{\alpha_m}$ and $R(s_{\alpha_1}), \ldots, R(s_{\alpha_m})$ can be computed efficiently.

CART algorithms require a training and test data set. In this study, 70% of the data was used as the training data set while 30% was used as the test data set.

## Random forests

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees. A training data set is used to construct decision trees. For each class, the mean prediction (for regression) and mode (for classification) is given (Ho, 1995); (Ho, 1998).

### *Variable Importance*

Random forests can be used to rank the importance of variables in a regression or classification problem. The importance score of the variables is computed by averaging the difference of out-of-bag error before and after the permutation over all trees. The score is then normalized. Features which produce large values are ranked as more important than features which produce small values (Zhu R & D, 2015).

Variable importance computation was done using both Mean Decrease Accuracy (%IncMSE) and Mean Decrease Gini (IncNodePurity). %IncMSE gives the decrease in model accuracy when the variable is excluded from the model. IncNodePurity measures the variable importance based on Gini impurity index.

Random forests algorithm was implemented in R programming language using the randomForest package version 4.6.

## 3.4 Optimal number of clusters

Both K-means and K-medoids algorithms require pre-specification of the number of clusters. Elbow, Silhouette and Gap statistic methods were used to determine the optimal number of clusters. R programming language was used to implement these methods.

## 3.5 K-means algorithm

K-means algorithm was implemented in R programming language using factoextra package version 1.0.7.

## 3.6 K-medoid algorithm

Gower's distance was used as the distance measure. Gower's distance is used for clustering mixed data types. The steps used for the different variable types are:

- **Quantitative (interval) variables** - uses range-normalized Manhattan distance.

$$s_j(x_1, x_2) = 1 - \frac{|y_{1j} - y_{2j}|}{R_j} \quad (18)$$

- **Ordinal variables** – First ranks the variable and then uses Manhattan distance to adjust the ties.

- **Nominal variable** - Converts the variables to binary columns and then uses Dice coefficient. When the values are equal, Dice Distance = 0 and when they are not equal, Dice distance is calculated using the equation below.

$$\text{DiceDistance} = \text{NNEQ} / (\text{NTT} + \text{NNZ}) \quad (19)$$

Where:

– N : number of dimensions

– NTT : number of dims in which both values are True

– NTF : number of dims in which the first value is True, second is False

– NFT : number of dims in which the first value is False, second is True

– NFF : number of dims in which both values are False

– NNEQ : number of non-equal dimensions, NNEQ = NTF + NFT

– NNZ : number of nonzero dimensions, NNZ = NTF + NFT + NTT

Gower's distance is computed as the average of partial dissimilarities. The general form of the coefficient is:

$$D_{\text{Gower}}\left(x_1,x_2\right) = 1 - \left(\frac{1}{p}\sum_{j=1}^{p} s_j\left(x_1,x_2\right)\right) \tag{20}$$

Where, $\mathrm{sj}(x1,x2)$ is the partial similarity function computed separately for each descriptor.

K-medoids algorithm was implemented in R programming language using factoextra package version 1.0.7.

## 3.7 Validation and performance evaluation

Cluster validation is the measurement of the goodness of the clustering results. Cluster validation techniques are categorized into: internal, external and relative (Theodoridis & Koutroumbas, 2006; Brock et al., 2008; Charrad et al., 2014). Internal cluster validation techniques use internal information of the clustering process to evaluate the goodness of the cluster results without reference to external information. External cluster validation techniques compare the cluster results to an external result such as available class labels. Relative cluster validation techniques evaluate the cluster results by changing different parameter values for the same algorithm e.g. varying the number of clusters (Kassambara, 2017).

Internal cluster validation techniques - Dunn index and Silhouette coefficient, were used. These methods measure compactness, connectedness and separation of cluster partitions.

### Average Silhoutte width

Average Silhoutte width measure how well data is clustered and estimates the average distance between clusters. The result is a silhouette plot which shows how close each point in one cluster is to points in the neighbouring clusters.

Average Silhoutte width ($s_i$) was used to check the goodness of the clusters generated. Average Silhoutte width values range from -1 to 1.

Where:

- $s_i$ = *1*: means the clusters are distinguishable from one another.
- $s_i$ = *0*: means clusters are indifferent.
- $s_i$ = *-1*: means clusters are wrongly assigned.

### Dunn index

Dunn index is based on calculating the size or diameter of a cluster (Dunn, 1974). Equation 21 is used to calculate Dunn index.

$$DI_m = \frac{\min_{1 \leqslant i < j \leqslant m} \delta\left(C_i, C_j\right)}{\max_{1 \leqslant k \leqslant m} \Delta_k} \tag{21}$$

Where:

- m is the number of clusters.
- $\delta\left(C_i, C_j\right)$ is the inter-cluster distance metric, between clusters Ci and Cj.

If the data contains compact and well-separated clusters, the diameter of the clusters is expected to be small and the distance between the clusters is expected to be large. Therefore, Dunn index should be maximized.

# 4 Results

## 4.1 Exploratory data analysis

### 4.1.1 Feasibility of cluster analysis

**Principal component Analysis (PCA)**

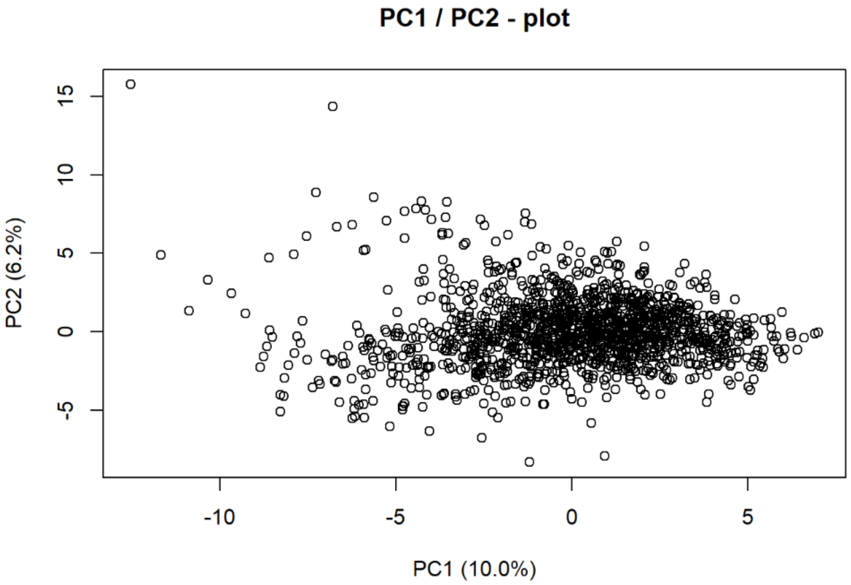

**Figure 3. Principal component analysis plot**

‘

**Hopkin’s statistic**

The Hopkins statistic was 0.219 which is less than the threshold 0.5. This means that the data is clusterable.

### Distance matrices computation

The ordered dissimilarity matrix (ODM) is shown below:

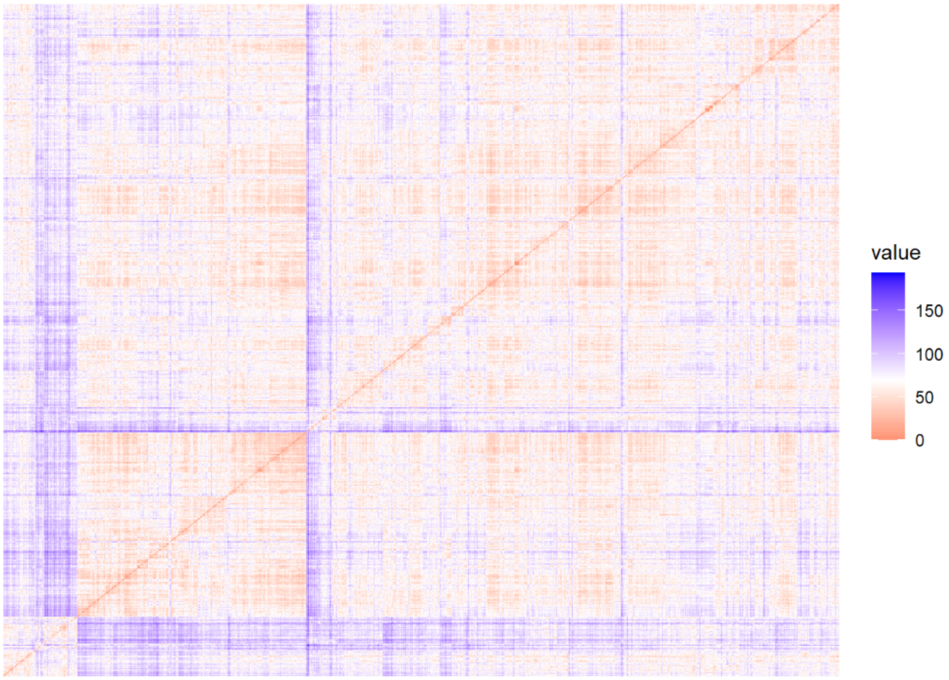

**Figure 4. Ordered dissimilarity matrix plot**

Red means low dissimilarity and blue means high dissimilarity. The Ordered Dissimilarity Matrix (ODM) in 4 confirmed that the data contained clusters.

## 4.1.2 Cluster structures

### Classification and Regression Trees (CART)

81 principal components were computed. The first 26 principal components explained 70.4% of the total variance in the data and had eigen values greater than 1. Therefore, 26 principal components were used in the CART algorithm. The proportion of explained variance and eigen values of the principal components plots are shown in figures 5 and 6.

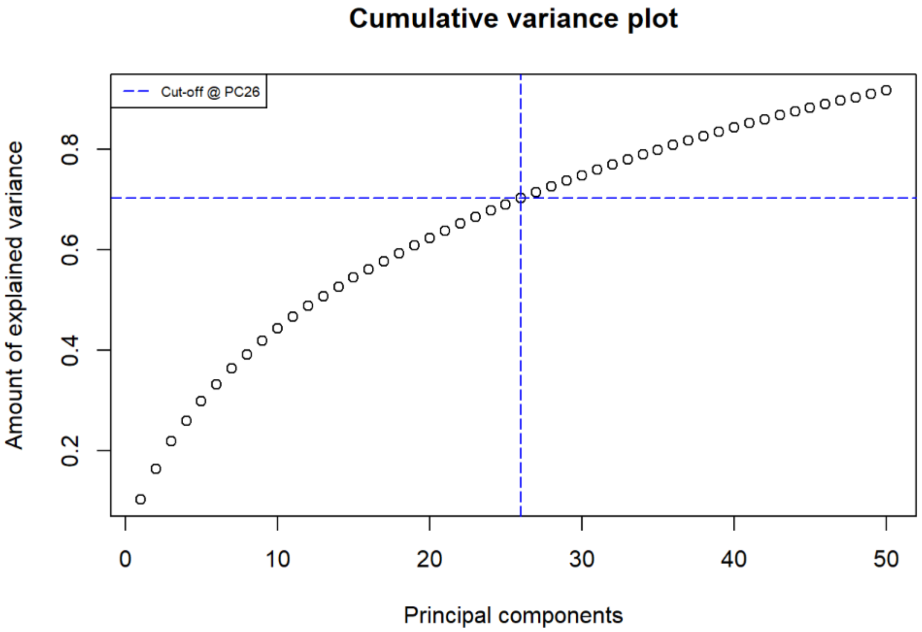

**Figure 5. Proportion of explained variance in the principal components**

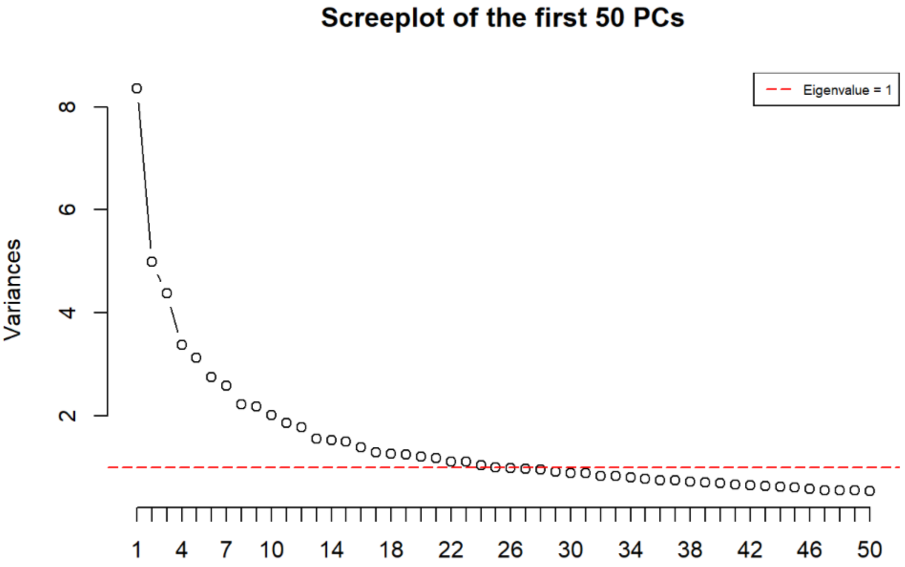

**Figure 6. Eigen values of the principal components**

The non-pruned Classification and Regression Tree is shown in figure 7.

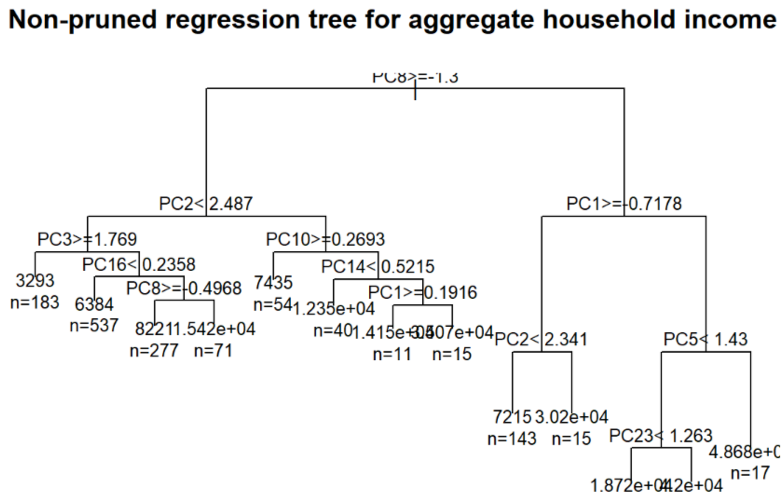

**Figure 7. Non-pruned Classification and Regression Tree (CART)**

The pruned Classification and Regression Tree (CART) is shown in figure 8..

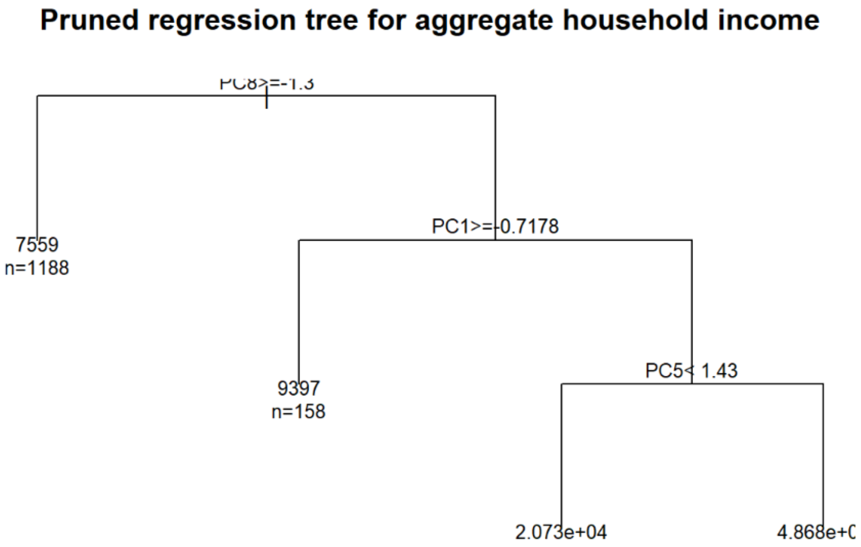

**Figure 8. Pruned Classification and Regression Tree (CART)**

Test dataset was used for prediction. The accuracy was 25.28%.

## Random forests

The percentage of explained variation was 46.7%. The classification accuracy was 0.16.

Using %IncMSE and IncNodePurity,the top 15 important variables are shown in table 1.

**Table 1. Importance of variables using mean decrease accuracy and mean decrease gini**

| %IncMSE | IncNodePurity | Variable |
|---|---|---|
| 44.262361 | 57757952260 | Primary household Income size |
| 10.977398 | 3379576995 | County |
| 10.663338 | 6698021106 | Sub-county |
| 7.231781 | 3995259222 | Ward |
| 6.574483 | 920340152 | Total non- food expenditure within the last year |
| 6.532458 | 5049673314 | Primary occupation of the household head |
| 6.459803 | 3886236193 | Age of household head |
| 6.189970 | 15016956701 | Household education level |
| 5.661327 | 12076160091 | Education level of the household head |
| 5.523263 | 4001942116 | Secondary occupation of household head |
| 5.473425 | 1089631562 | Did you grow any crops in the first season? |
| 5.384683 | 594063764 | Did any member of your household consume nuts/pulses within the last week |
| 5.243492 | 516135003 | Did you apply any crop management techniques in the second season? |
| 5.069130 | 4165793003 | What is your household's total land size in acres? |

## 4.2 Optimal number of clusters

### Elbow method

Elbow method gave two clusters as the optimal number for both K-means and K-medoids.

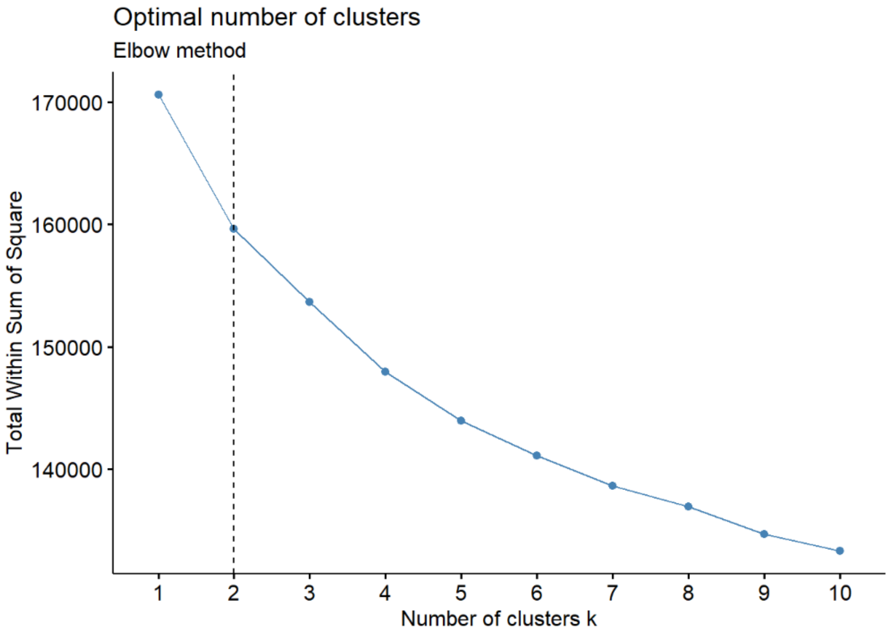

**Figure 9. Total within sum of squares against number of clusters plot**

## Silhoutte method

Silhouette method gave two clusters as the optimal number for K-means algorithm (figure 10) and three clusters as the optimal number for K-medoids algorithm (figure 11).

For K-means algorithm,

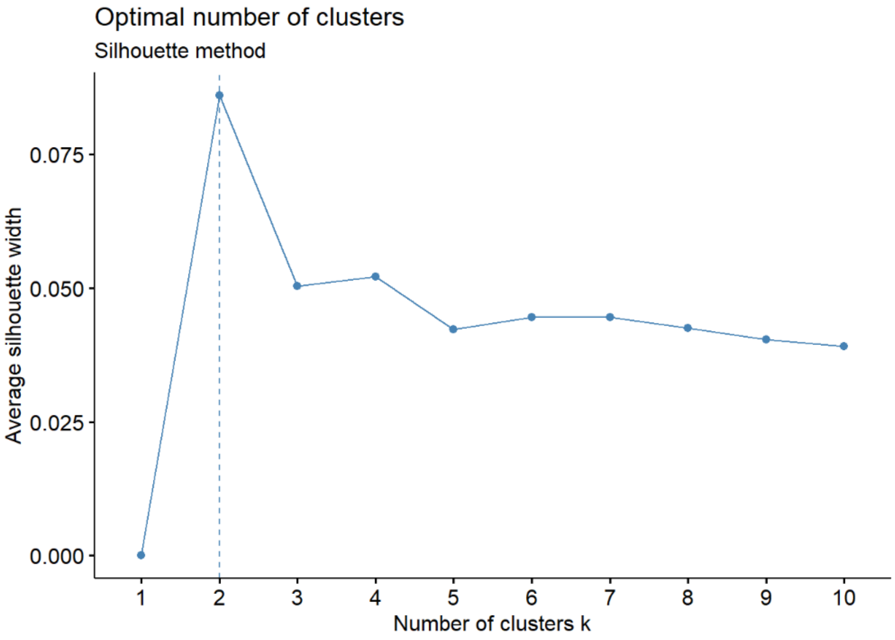

**Figure 10. Average Silhouette width plot for K-means algorithm**

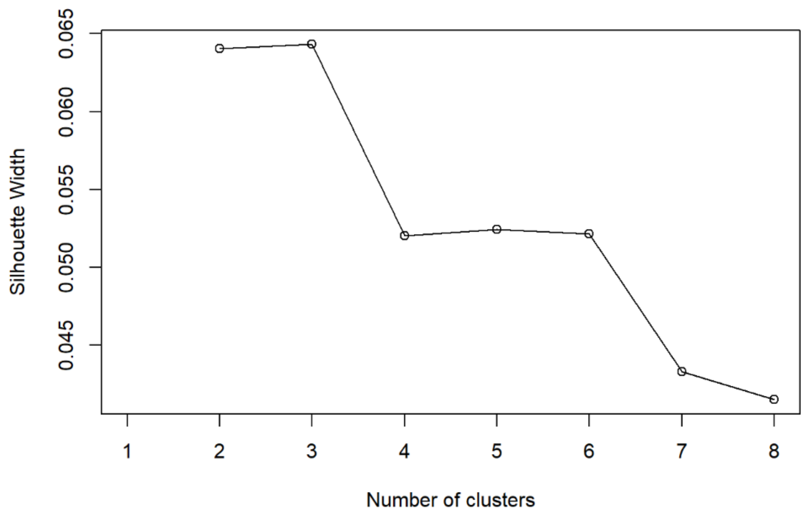

**Figure 11. Average Silhouette width plot for K-medoids algorithm**

## Gap statistic

Gap statistic method gave one cluster as the optimal number for K-means algorithm (figure 12) and three clusters as the optimal number for K-medoids algorithm (figure 13)).

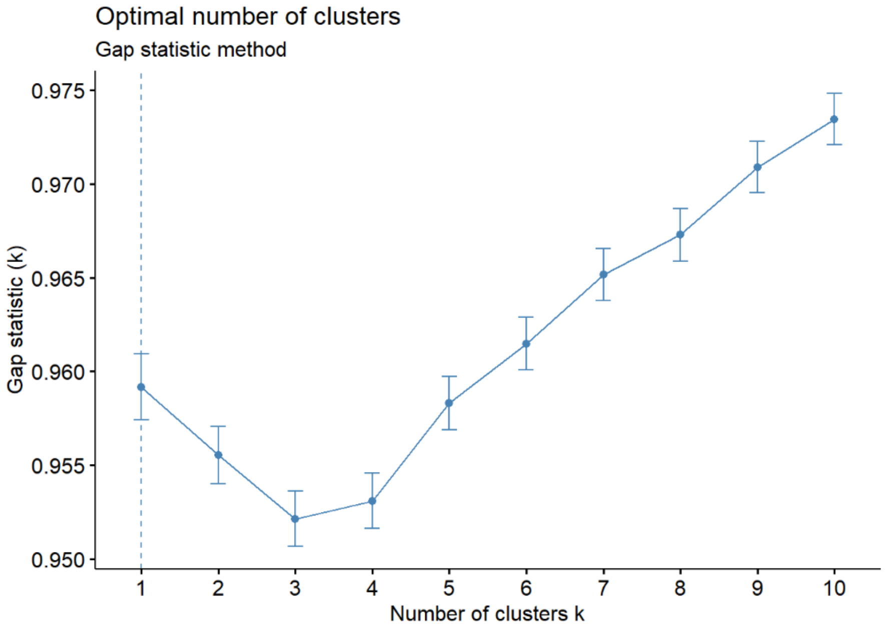

**Figure 12. Gap statistic plot for K-means algorithm**

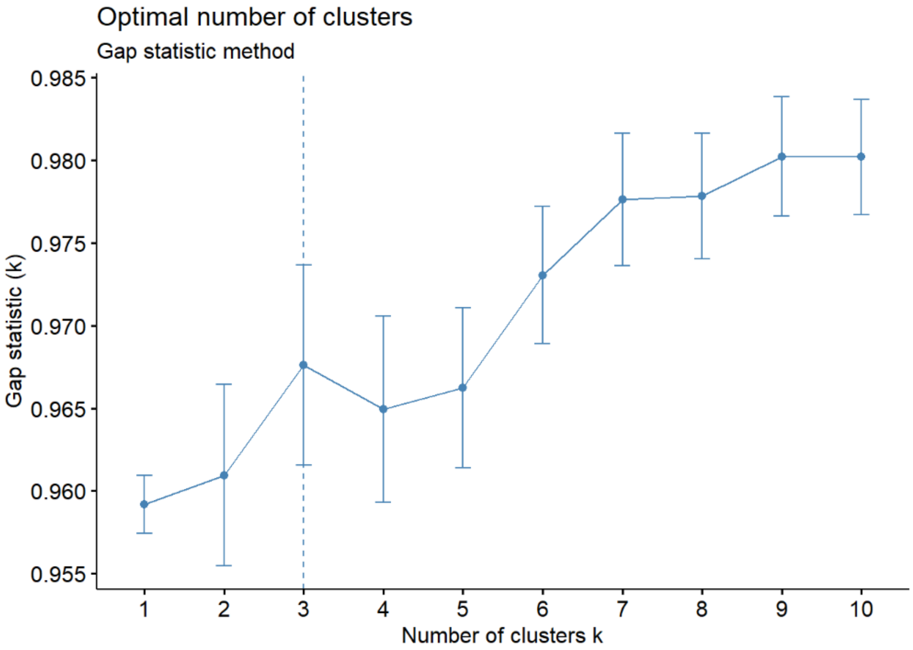

**Figure 13. Gap statistic plot for K-medoids algorithm**

## 4.3 K-means algorithm

The two K-means algorithm clusters were of sizes 707 and 1374. The within cluster sum of squares was 64,462 and 95,164.

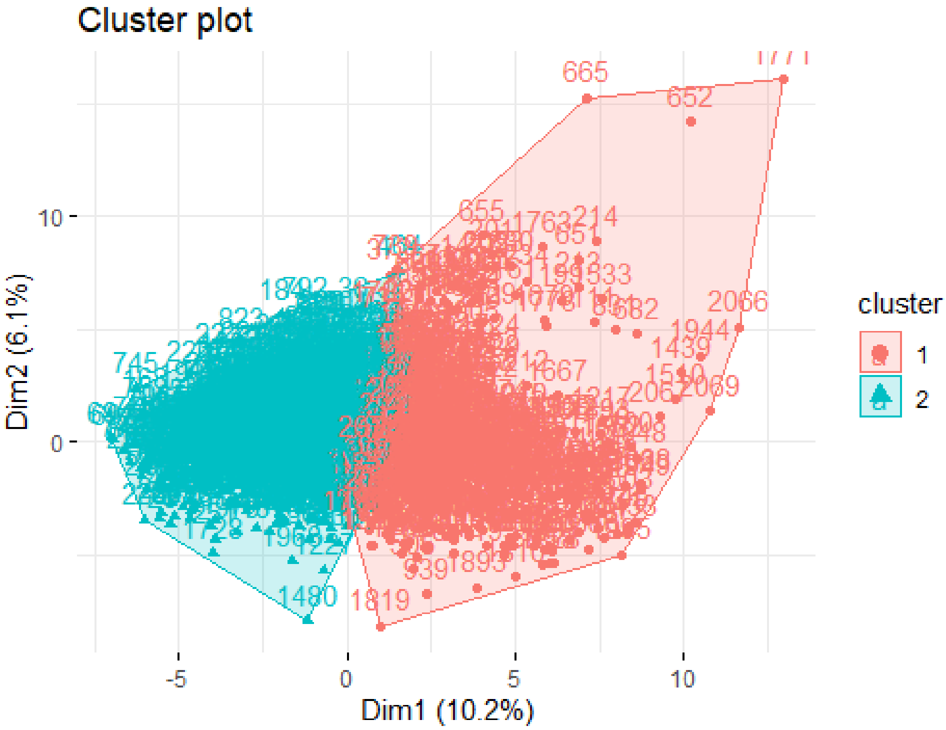

**Figure 14. Clusters generated using K-means algorithm**

The data used was collected from rural farming households and therefore the poverty levels were rural based. Cluster one contained rural non-poor households while cluster two contained rural poor households.

The means of the numeric variables are shown in table 2.

**Table 2. Means the numeric variables using K-means algorithm**

| | Rural non-poor | Rural poor |
|---|---|---|
| Aggregate monthly income | 11898.46 | 7346.79 |
| Monthly aggregate per capita income | 290.23 | 125.97 |
| Total land size in acres | 7.46 | 10.02 |
| Number of food groups consumed within the last week | 7.35 | 5.14 |
| Age of household head | 49.30 | 47.69 |
| Education level of household head | 7.55 | 6.71 |
| Average education level of household | 6.83 | 6.17 |
| Dollar Index | 0.71 | 0.29 |
| Household size | 4.75 | 4.99 |
| Number of children | 1.45 | 1.67 |
| Gender Parity index | 1.20 | 1.01 |
| Number of females in household | 2.35 | 2.51 |
| Number of meals taken by children in a day | 3.05 | 2.94 |
| Number of males | 2.40 | 2.48 |
| Number of meals taken excluding children in a day | 2.93 | 2.86 |
| Dependency ratio | 0.72 | 0.79 |
| Household access to shamba | 0.97 | 0.90 |
| Number of independent household members | 3.06 | 3.10 |
| Number of elderly | 0.24 | 0.21 |
| Household type | 0.82 | 0.80 |
| Sex ratio | 1.27 | 1.26 |

## Location

In Makueni, Taveta and Tharaka counties, the proportion of rural poor versus rural non-poor households was 50:40. The percentage of rural poor and rural non-poor households varied in Embu, Kilifi, Kitui, Kwale and Machakos counties. Over 80% of the households in Kilifi, Kwale and Kitui counties were poor.

**Table 3. County-based summary using K-means algorithm**

| County | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| **Embu** | 92 (39.0%) | 144 (61.0%) | 236 (100.0%) |
| **Kilifi** | 38 (15.9%) | 201 (84.1%) | 239 (100.0%) |
| **Kitui** | 66 (18.5%) | 291 (81.5%) | 357 (100.0%) |
| **Kwale** | 43 (17.9%) | 197 (82.1%) | 240 (100.0%) |
| **Machakos** | 82 (35.2%) | 151 (64.8%) | 233 (100.0%) |
| **Makueni** | 104 (43.9%) | 133 (56.1%) | 237 (100.0%) |
| **Taveta** | 123 (51.2%) | 117 (48.8%) | 240 (100.0%) |
| **Tharaka** | 159 (53.2%) | 140 (46.8%) | 299 (100.0%) |
| **Total** | 707 (34.0%) | 1374 (66.0%) | 2081 (100.0%) |

## Household demographics

The average age of household heads was 47 years in rural poor households and 49 years in rural non-poor households.

The average education level of the household head was 7 in rural poor households and 8 in the rural non-poor households.

The average education level was 6 in rural poor households and 7 in the rural non-poor households.

The average household size in both rural poor and rural non-poor households was 5. The average sex ratio was 1.3 in both clusters.

The average gender parity ratio was 1 in rural poor households and 1.2 in the rural non-poor households.

The average dependency ratio was 0.79 in rural poor households and 0.72 in rural non-poor households.

There was no distinction of both gender and marital status of household heads within the poor and non-poor clusters.

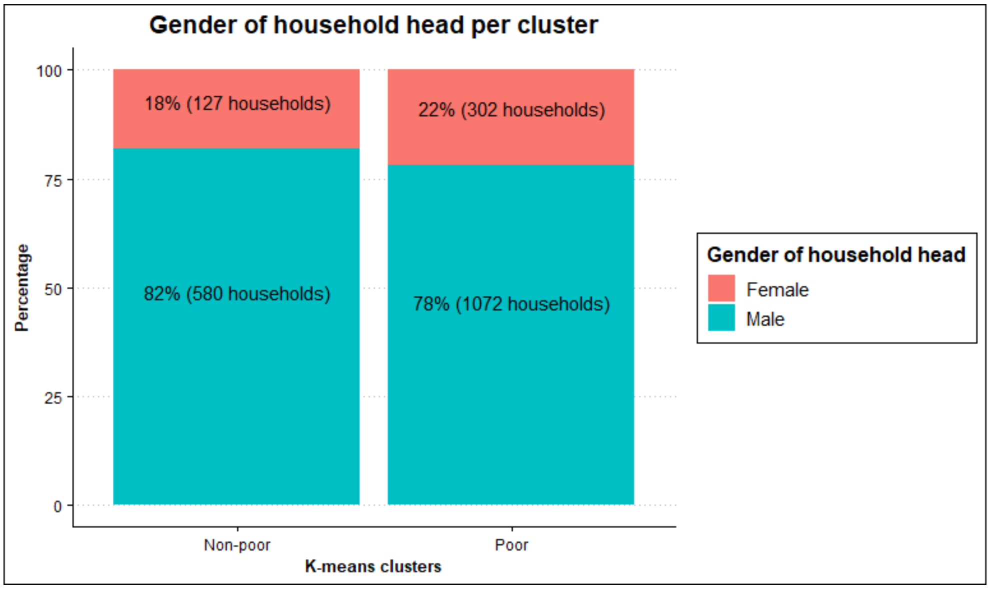

**Figure 15. Gender of household head**

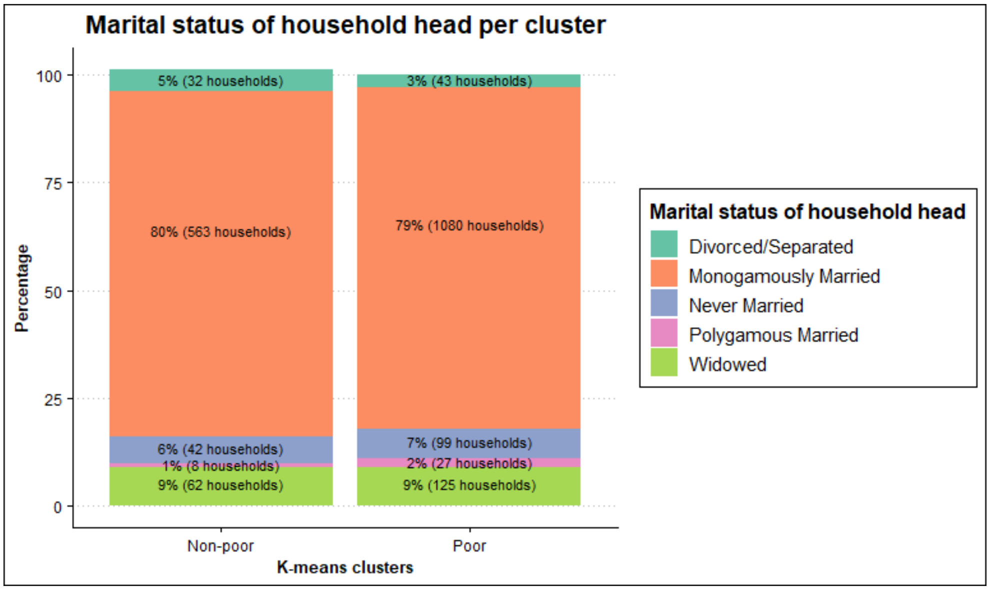

**Figure 16. Marital status of household head**

The results show that rural non-poor households have a high education level, low dependency ratio and high gender parity ratio compared to the rural poor households.

## Household income

On average, rural non-poor households had a monthly income of Kshs 11,898 while rural poor households had a monthly income of Kshs 7,346.

Crop farming was the main income source in 47% of the rural non-poor households and in 29% of the rural poor households. 32% of the rural poor households had no main household income.

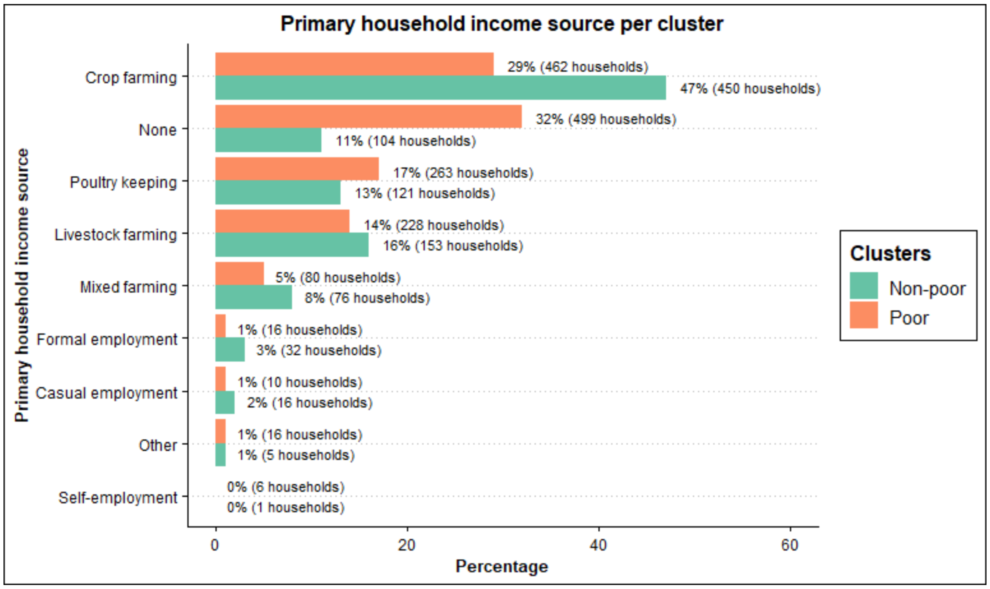

**Figure 17. Primary income of households**

## Crop Farming

83% of the rural poor households did not own any agricultural productive assets.

**Table 4. Ownership of agricultural productive assets**

| Ownership of agricultural productive assets | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| **No** | 56 (16.7%) | 279 (83.3%) | 335 (100.0%) |
| **Yes** | 651 (37.3%) | 1095 (62.7%) | 1746 (100.0%) |
| **Total** | 707 (34.0%) | 1374 (66.0%) | 2081 (100.0%) |

76% of the rural poor households had not accessed any extension service within the last year.

Table 5. Households access to extension services

| Access to extension services | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| No | 356 (23.6%) | 1153 (76.4%) | 1509 (100.0%) |
| Yes | 351 (61.4%) | 221 (38.6%) | 572 (100.0%) |
| Total | 707 (34.0%) | 1374 (66.0%) | 2081 (100.0%) |

The results show that rural poor households had larger land sizes than the rural non-poor but did not own agricultural assets and did not seek agricultural extension services compared to the rural non-poor.

### Access to financial services

80% of the rural poor households were not aware of any financial services or products available to farmers while only 20% of the rural non-poor households were not aware.

Table 6. Awareness of available financial services or products

| Aware of any financial services/products | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| No | 230 (19.8%) | 931 (80.2%) | 1161 (100.0%) |
| Yes | 477 (51.8%) | 443 (48.2%) | 920 (100.0%) |
| Total | 707 (34.0%) | 1374 (66.0%) | 2081 (100.0%) |

71% of the rural poor households had not accessed any financial services within the last year.

Table 7. Households access to financial services

| Accessed any financial services | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| No | 506 (28.9%) | 1242 (71.1%) | 1748 (100.0%) |
| Yes | 201 (60.4%) | 132 (39.6%) | 333 (100.0%) |
| Total | 707 (34.0%) | 1374 (66.0%) | 2081 (100.0%) |

23% of the rural non-poor household members did not own a bank account compared to 77% in the rural poor households.

**Table 8. Ownership of bank account by household members**

| Ownership of a bank account | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| **No** | 296 (23.2%) | 980 (76.8%) | 1276 (100.0%) |
| **Yes** | 411 (51.1%) | 394 (48.9%) | 805 (100.0%) |
| **Total** | 707 (34.0%) | 1374 (66.0%) | 2081 (100.0%) |

The results show that rural non-poor households were more financially aware and accessed financial services more compared to the rural poor households.

### Household Dietary Diversity (HDDS)

Ten categorized food groups were used in the survey ( see appendix).

Within the last week of the survey, rural poor households consumed 5 food groups on average while rural non-poor households consumed 7. This means that the rural non-poor households had a higher HDDS and better food consumption patterns than the rural poor households.

## 4.4 K- medoid algorithm

The two clusters in K-medoid algorithm were of size 950 and 1131. The build was 0.257 and swap was 0.253.

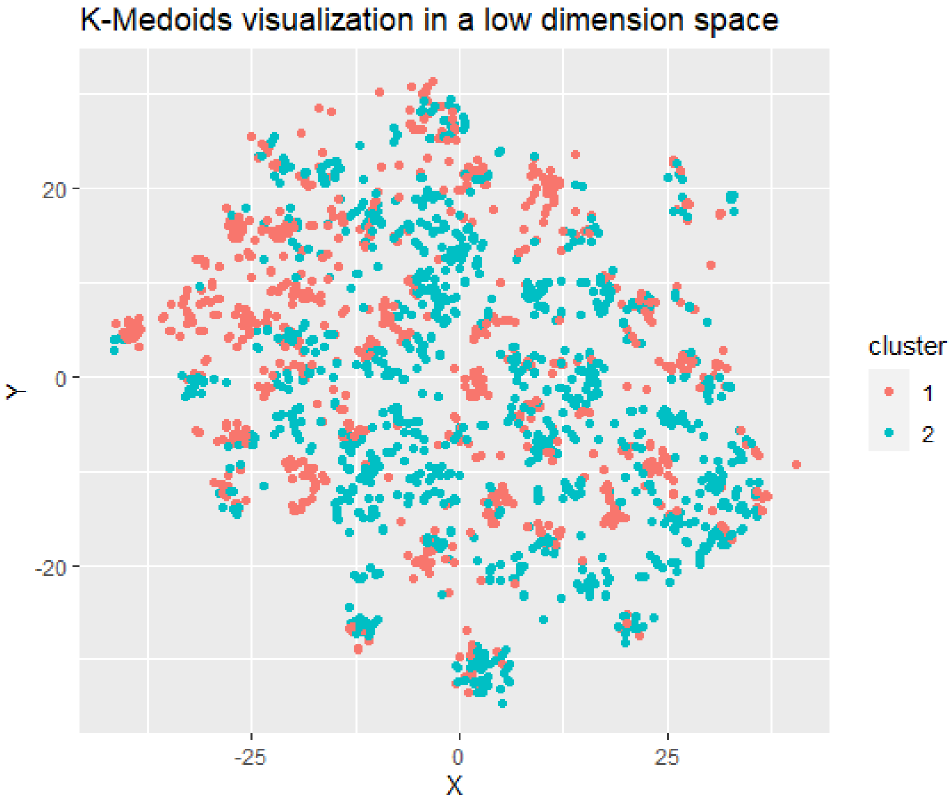

**Figure 18. Clusters generated using K-medoids algorithms**

Cluster one contained rural poor households while cluster two contained rural non-poor households.

The means of the numeric variables are shown in table 9.

Table 9. Means of numeric variables using K-medoids algorithm

| | Rural poor | Rural non-poor |
|---|---|---|
| Aggregate monthly income | 5032.49 | 12136.02 |
| Monthly aggregate per capita income | 110.96 | 241.26 |
| Total land size in acres | 10.69 | 7.86 |
| Number of food groups consumed within the last week | 4.46 | 7.10 |
| Education level of household head | 5.93 | 7.89 |
| Age of household head | 49.22 | 47.42 |
| Average education level of household | 5.84 | 6.86 |
| Dollar Index | 0.26 | 0.59 |
| Household size | 5.03 | 4.80 |
| Number of children | 1.72 | 1.50 |
| Depedancy ratio | 0.85 | 0.70 |
| Number of males | 2.52 | 2.39 |
| Gender Parity index | 1.02 | 1.12 |
| Number of females in household | 2.51 | 2.41 |
| Number of meals taken excluding children in a day | 2.85 | 2.92 |
| Number of meals taken by children in a day | 2.94 | 3.00 |
| Number of elderly | 0.25 | 0.19 |
| Sex ratio | 1.29 | 1.24 |
| Household access to shamba | 0.90 | 0.95 |
| Number of indepedent household members | 3.06 | 3.11 |
| Household type | 0.80 | 0.82 |

### Location

Embu, Kwale, Makueni, Taveta and Tharaka counties had a high percentage of rural non-poor households compared to rural poor. Kilifi, Kitui and Machakos counties had a high percentage of rural poor households compared to rural non-poor.

Table 10. County-based summary using K-medoids algorithm

| County | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| **Embu** | 173 (73.3%) | 63 (26.7%) | 236 (100.0%) |
| **Kilifi** | 117 (49.0%) | 122 (51.0%) | 239 (100.0%) |
| **Kitui** | 145 (40.6%) | 212 (59.4%) | 357 (100.0%) |
| **Kwale** | 150 (62.5%) | 90 (37.5%) | 240 (100.0%) |
| **Machakos** | 106 (45.5%) | 127 (54.5%) | 233 (100.0%) |
| **Makueni** | 130 (54.9%) | 107 (45.1%) | 237 (100.0%) |
| **Taveta** | 150 (62.5%) | 90 (37.5%) | 240 (100.0%) |
| **Tharaka** | 160 (53.5%) | 139 (46.5%) | 299 (100.0%) |
| **Total** | 1131 (54.3%) | 950 (45.7%) | 2081 (100.0%) |

### Household demographics

The average age of household heads in rural poor households was 47 years and 49 years in the rural non-poor households.

The average education level of the household head was 6 in rural poor households and 8 in the rural non-poor households.

The average education level was 6 in rural poor households and 7 in the rural non-poor households.

The average household size in both the rural poor and rural non-poor households was 5.

The average gender parity ratio was 1 in rural poor households and 1.1 in the rural non-poor households.

The average dependency ratio was 0.85 in rural poor households and 0.70 in rural non-poor households.

There was no distinction of gender and marital status of household heads within the two clusters.

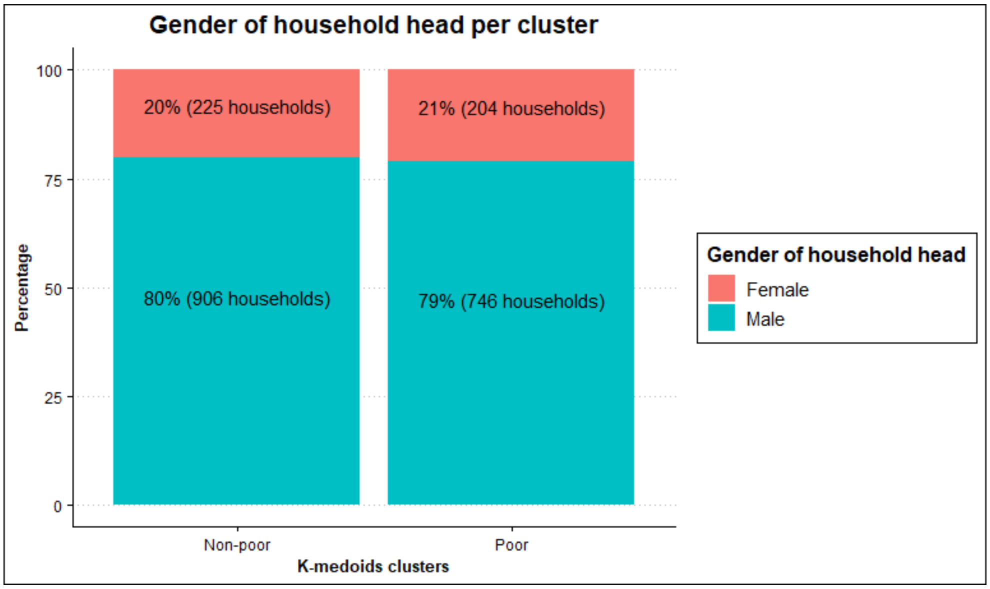

Figure 19. Gender of household head

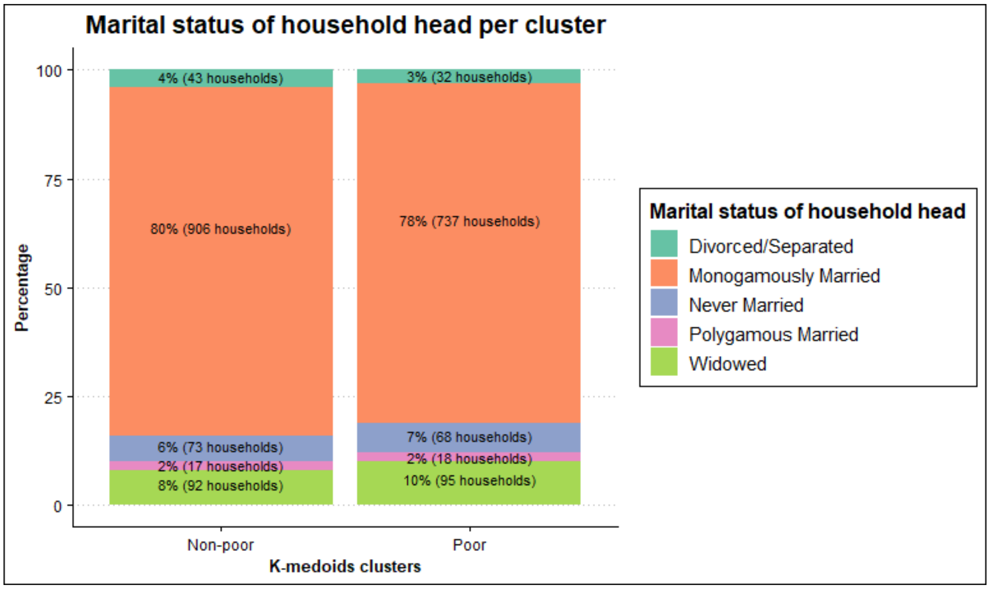

Figure 20. Marital status of household head

The results show that rural non-poor households had a high education level, low dependency ratio and a high gender parity ratio compared to the rural poor households.

## Household income

On average, rural non-poor households had a monthly income of Kshs 12,136 while the rural poor households had a monthly income of Kshs 5,032.

Crop farming was the main household income source in 41% of the rural non-poor households and in 29% for the rural poor households. 37% of the rural poor households had no major primary income source.

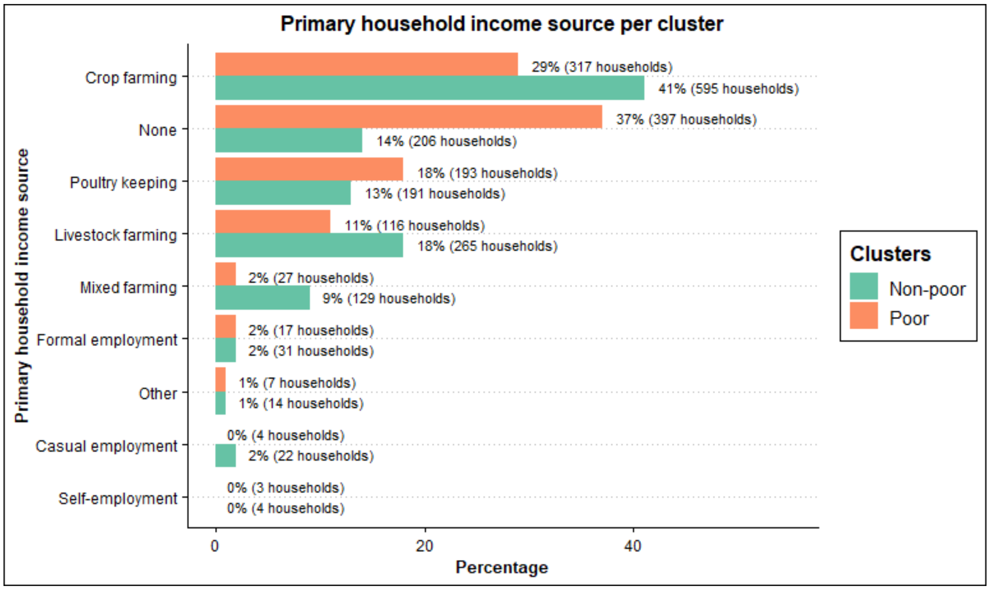

**Figure 21. Marital status of household head**

## Crop Farming

59% of the rural poor households did not own any agricultural productive assets.

**Table 11. Ownership of agricultural productive assets**

| Ownership of agricultural productive assets | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| **No** | 138 (41.2%) | 197 (58.8%) | 335 (100.0%) |
| **Yes** | 993 (56.9%) | 753 (43.1%) | 1746 (100.0%) |
| **Total** | 1131 (54.3%) | 950 (45.7%) | 2081 (100.0%) |

**Table 12. Households access to extension services**

| Accessed extension services | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| **No** | 735 (48.7%) | 774 (51.3%) | 1509 (100.0%) |
| **Yes** | 396 (69.2%) | 176 (30.8%) | 572 (100.0%) |
| **Total** | 1131 (54.3%) | 950 (45.7%) | 2081 (100.0%) |

The results show that rural poor households did not own agricultural assets and did not seek agricultural extension services compared to the rural non-poor.

### Access to financial services

53% of the rural poor households were not aware of any financial services or products available to farmers.

**Table 13. Awareness of available financial services or products**

| Aware of any financial services available | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| **No** | 541 (46.6%) | 620 (53.4%) | 1161 (100.0%) |
| **Yes** | 590 (64.1%) | 330 (35.9%) | 920 (100.0%) |
| **Total** | 1131 (54.3%) | 950 (45.7%) | 2081 (100.0%) |

**Table 14. Households access to financial services**

| Accessed financial services | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| **No** | 876 (50.1%) | 872 (49.9%) | 1748 (100.0%) |
| **Yes** | 255 (76.6%) | 78 (23.4%) | 333 (100.0%) |
| **Total** | 1131 (54.3%) | 950 (45.7%) | 2081 (100.0%) |

61% of the rural poor did not own a bank account.

**Table 15. Ownership of bank account by household members**

| Ownership of a bank account | Rural non-poor | Rural poor | Total |
|---|---|---|---|
| **No** | 496 (38.9%) | 780 (61.1%) | 1276 (100.0%) |
| **Yes** | 635 (78.9%) | 170 (21.1%) | 805 (100.0%) |
| **Total** | 1131 (54.3%) | 950 (45.7%) | 2081 (100.0%) |

The results show that rural non-poor households were more financially aware and accessed financial services more compared to the rural poor households.

### Household Dietary Diversity

Within the last week of the survey, rural poor households consumed 4 food groups on average while rural non-poor households consumed 7. This means that the rural non-poor had a higher HDDS and better food consumption patterns than the rural poor households.

## 4.5 Validation and performance evaluation

**Table 16. Performance of K-means and K-medoids algorithms**

| | K-means | | K-medoids | |
|---|---|---|---|---|
| **Number of clusters** | 2 | 3 | 2 | 3 |
| **Average silhoutte width** | 0.09 | 0.05 | 0.06 | 0.06 |
| **Dunn index** | 0.09 | 0.09 | 0.04 | 0.07 |

### Average silhoutte width

The average silhouette width was 0.09 for K-means algorithm (figure 22) and 0.06 for K-medoid algorithm (figure 23).

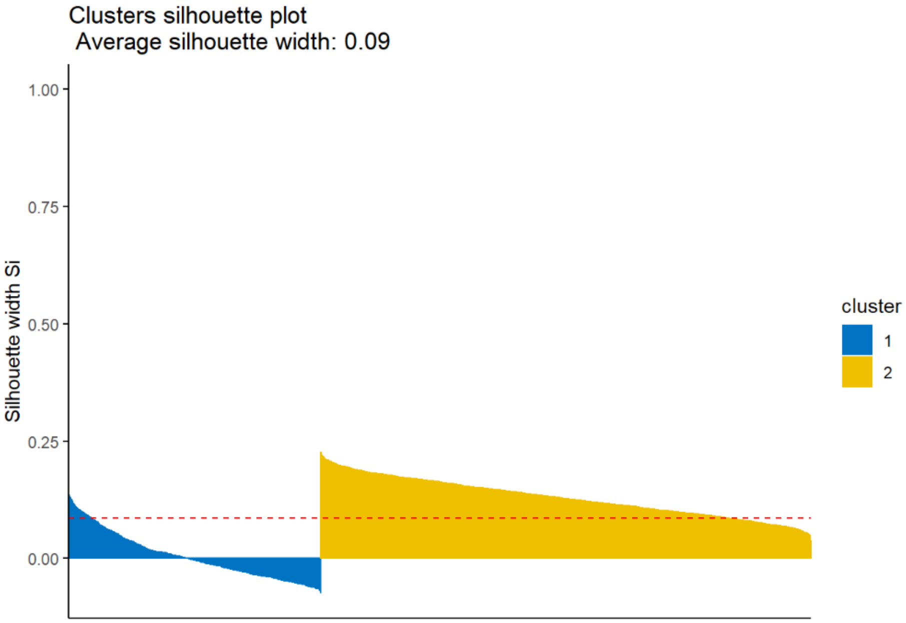

**Figure 22. Silhoutte width plot for K-means algorithm**

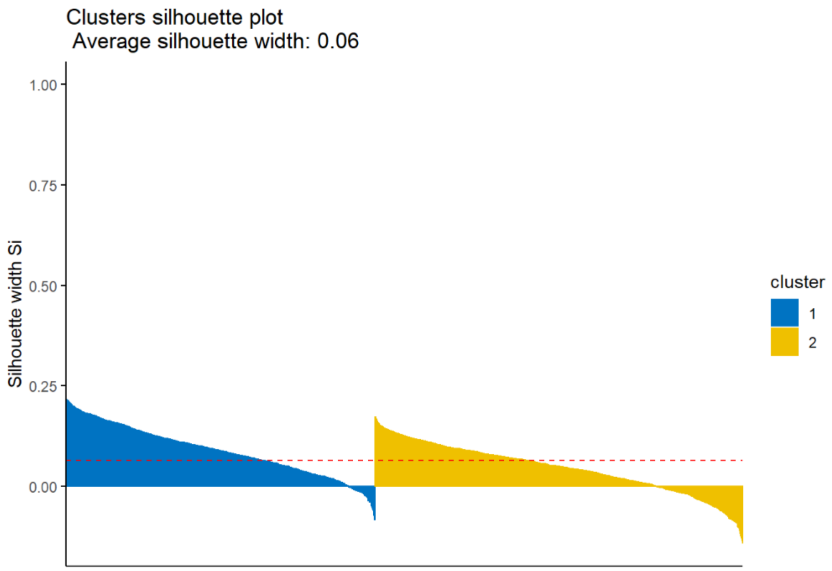

**Figure 23. Silhoutte width plot for K-medoid algorithm**

K-means algorithm had a high average silhouette width than K-medoids algorithm when two clusters were used. When three clusters were used, K-medoids algorithm had a high silhouette width than K-means algorithm.

### Dunn index

The Dunn index was 0.09 for K-means algorithm and 0.04 for K-medoid algorithm.

When both two and three clusters were used, K-means algorithm had a high Dunn index than K-medoid algorithm.

### Comparison of K-means and K-medoids clusters

**Table 17. Comparison of K-means and K-medoids clusters**

| K-Medoids clusters | K-Means clusters | | |
|---|---|---|---|
| | **Rural non-poor** | **Rural poor** | **Total** |
| **Rural non-poor** | 565 (50.0%) | 566 (50.0%) | 1131 (100.0%) |
| **Rural poor** | 142 (14.9%) | 808 (85.1%) | 950 (100.0%) |
| **Total** | 707 (34.0%) | 1374 (66.0%) | 2081 (100.0%) |

1,373 out of 2,081 households were placed in same clusters by both K-means and K-medoids algorithms. This accounts for 66% of the households.

Out of the 1,131 households classified as rural non-poor by K-medoids algorithm, 50% were classified as rural non-poor by K-means algorithm.

Out of the 950 households classified as rural poor by K-means algorithm, 15% were classified as rural non-poor and 85% classified as rural poor by K-medoids algorithm.

Out of the 707 households classified as rural non-poor by K-means algorithm, 50% were classified as rural non-poor and 15% classified as rural poor by K-medoids algorithm.

Out of the 1,374 households classified as rural poor by K-means algorithm, 50% were classified as rural non-poor and 85% classified as rural poor by K-medoids algorithm.

## 4.6 Summary of findings

1. Rural poor households have low education level, high dependency ratio and low gender parity ratio compared to rural non-poor households.

2. Rural non-poor households have high aggregate income compared to rural poor households.

3. Less rural poor households own agricultural productive assets and seek extension services compared to rural non-poor households.

4. Rural non-poor households are more aware of financial services or products available to farmers and access financial services more compared to the rural poor households.

5. Rural poor households have low dietary diversity score (HDDS) compared to rural non-poor households.

# 5 Discussion and conclusions

The main sources of income in rural poor households include agriculture, fishing and forestry among other small-scale industries. Agriculture suffers from major shocks such as drought, pest and diseases and floods. The objective of the study was to identify household characteristics that differentiate rural poor and non-poor households, their direction and compare performance of K-means and K-medoids algorithms.

More rural non-poor households owned agricultural productive assets and sought extension services compared to rural poor households. Rural poverty and lack of agricultural assets are highly correlated (Bogale et al., 2005). Access to extension services play a critical role in reducing poverty in rural areas (Danso-Abbeam et al., 2018).

A higher percentage of rural non-poor households were aware of financial services or products available to farmers and accessed financial services more compared to rural poor households. Access to formal credit services by farmers and membership in savings groups are important in alleviating poverty in rural areas (Abraham, 2018).

Rural poor households had low aggregate income, low education level and low gender parity ratio compared to rural non-poor households. Education is an important determinant of poverty in rural areas (Geda et al. (2005); Chaudhry & Rahman (2009); Jamal (2005)). Income is key in reducing poverty in rural areas (Pervez & Usman, 2011).

Rural non-poor households had low dependency ratio compared to rural poor households. Dependents increase the expenditure of a household while the income remain constant.

Rural non-poor households had higher dietary diversity score (HDDS) than rural poor households. This means that rural non-poor households consumed more diverse meals. Food consumption patterns and food security are synonymous with food availability, affordability and nutritional knowledge. According to Cordero Ahimán et al. (2017) and Cheteni et al. (2020), rural non-poor households have higher HDDS than non-poor households. In this study, rural poor households had low education level, low income and low dietary diversity compared to the rural non-poor households. This means that education and income are key in achieving food security in rural households.

The comparison of K-means and K-medoids algorithms in this study was inconclusive. Using Dunn index, when two and three clusters were used, K-means algorithm had higher Dunn Index than K-medoids algorithm. However, using Silhouette width, K-means algorithm had higher average silhouette width than K-medoids algorithm when two clusters were used. When three clusters were used, K-medoids algorithm had a higher Silhouette width than K-means algorithm.

Some key conclusions that stood out from known facts include: diet diversity is a determinant of poverty in rural areas, farmers knowledge of available financial subsidy programs is a determinant of poverty in rural areas and household size is not a determinant of poverty in rural areas.

## 5.1 Limitations

The study limitations were:

- The research was confined to the variables present.
- In the original dataset, 2,183 variables had missingness of between 3% and 100% and therefore the number of variables used in the study were reduced from 2265 to 82 for the study.

## 5.2 Recommendations

To tackle poverty in Kenyan rural areas, government, donors and partners can:

- Invest in sensitization of the community on importance of education and reduce or remove school related fees for children to access education.
- Create awareness and provide subsidy programs for farmers to access financial services, farm inputs, farm productivity assets and access to better markets to improve yield.
- Provide farmers with more and easy access to extension services. This will educate them on farm inputs, soil management, coping with climate change and storage of produce to maximize yield, income and improve food security.
- Sensitize the community on nutrition, nutritious foods and importance of diet diversity.

## 5.3 Further research

Research should be conducted on the role of land sizes and land tenure on poverty in rural Kenya. Also, researchers should include more variables when collecting data on determinants of poverty in rural areas.

# References

Abraham, T. W. (2018). Estimating the effects of financial access on poor farmers in rural northern Nigeria. *Financial Innovation*, *Vol. 4, No. 1*.

Aggarwal, C. C. (2015). *Data Mining*. Springer International.

Ahmad, Z., & Ejaz, Z. (2011). Classification of households with respect to poverty by using cluster analysis. In *the 11th islamic countries confrence on statistical sciences* (Vol. 21, p. 369-381).

Akerele, D., & Adewuti, S. (2011). Analysis of poverty profiles and socioeconomic determinants of welfare among urban households of Ekiti State, Nigeria. *Current Research Journal of Social Sciences*, *Vol. 3*, 1-7.

Atkinson, E. J., & Therneau, T. M. (2000). An Introduction to Recursive Partitioning Using the RPART Routines. *Mayo clinic*, *Vol. 61*, 33.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms* (No. 0306406713). USA: Kluwer Academic Publishers.

Bogale, A., Hagedorn, K., & Korf, B. (2005). Determinants of poverty in rural Ethiopia. *Quarterly Journal of International Agriculture*, *Vol. 44, No. 2*, 101-120.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (p. 144–152).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1983). *Classification and Regression Trees*. Wadsworth, Belmont: Taylor & Francis publishers.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, *Vol. 25, No. 4*.

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *Vol. 1, No. 2*, 245-76.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, *Vol. 61, No. 6*, 1-36.

Chaudhry, I., & Rahman, S. (2009). The Impact of Gender Inequality in Education on Rural Poverty in Pakistan: An emphirical Analysis. *European Journal of Econiomics, Finance and Administrative Sciences*, 174-188.

Cheteni, P., Khamfula, Y., & Mah, G. (2020). Exploring Food Security and Household Dietary Diversity in the Eastern Cape Province, South Africa. *Sustainability*, *Vol. 12, No. 5*, 1851.

Cochran, W. (1977). *Sampling Techniques (Third edition).* New York: John Wiley & Sons.

Cordero Ahimán, O., Santellano Estrada, E., & Garrido Colmenero, A. (2017). Dietary Diversity in Rural Households: The Case of Indigenous Communities in Sierra Tarahumara, Mexico. *Journal of Food and Nutrition Research*, *Vol. 5, No. 2*, 86-94.

Cortes, C., & Vapnik, V. N. (1995). Support-vector networks. *Machine Learning*, *Vol. 20, No. 3*, 273-297.

Danso-Abbeam, G., Ehiakpor, D. S., & Aidoo, R. (2018). Agricultural extension and its effects on farm productivity and income: Insight from Northern Ghana. *Agriculture and Food Security*, *Vol. 7, No. 1*, 1-10.

Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, *Vol. 4, No. 1*, 95-104.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining* (p. 226–231). Portland, Oregon: AAAI Press.

Etim, N.-a. A., & Udoh, E. (2013). The Determinants of Rural Poverty in Nigeria. *The Determinants of Rural Poverty in Nigeria*, *Vol. 3, No. 2*, 141-151.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *Vol. 97, No. 458*, 611.

Fraley, C., Raftery, A. E., Murphy, T. B., & Murphy, S. (2012). *Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.* (Tech. Rep.). Department of Statistics, University of Washington.

Geda, A., de Jong, N., & Mwabu, M. S. K. G. (2005). *Determinants of Poverty in Kenya: A Household Level Analysis.*

Girden, E. R., & Kabacoff, R. (2001). *Evaluating research articles from start to finish (Third edition).* California: Sage Publications.

Guyon, I., Boser, B., & Vapnik, V. (1993). *Automatic capacity tuning of very large VC-dimension classifiers* (L. G. C. Jose Hanson S, Cowan JD, Ed.). San Mateo: California, USA: Morgan Kaufmann Publishers.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *Vol. 28, No. 1*, 100-108.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (p. 278-282).

Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Vol. 20, No. 8*, 832-844.

Hopkins, B., & Skellam, J. G. (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany*, *Vol. 18, No. 2*, 213-227.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, *Vol. 24, No. 6/ 7*, 417-441/ 498-520.

Jamal, H. (2005). *In Search of Poverty Predictors: The Case of Urban and Rural Pakistan* (Vols. 44, No. 1; Tech. Rep.).

Jazaïry, I., Alamgir, M., & Panuccio, T. (1992). The State of World Rural Poverty: An Inquiry into Its Causes and Consequences. New York: University Press.

Johnston, R. B. (2016). Arsenic and the 2030 Agenda for sustainable development. In *Arsenic research and global sustainability - proceedings of the 6th international congress on arsenic in the environment, as 2016* (p. 12-14).

Jolliffe, I. T. (2002). *Principal Component Analysis (Second edition).* New York: Springer Series in Statistics.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *Vol. 20, No. 1*, 141-151.

Kassambara, A. (2017). Practical guide to cluster analysis in R: unsupervised machine learning. *Journal of Computational and Graphical Statistics*, 187.

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley & Sons.

Kaufmann, L., & Rousseeuw, P. (1987, 01). Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, 405-416.

Kenya National Bureau of Statistics. (2020). *Comprehensive Poverty Report* (Tech. Rep.). Nairobi.

Malik, S. (1996). Determinants of rural poverty in Pakistan a micro study. *Pakistan Development Review*, *Vol. 35, No. 2*, 171-187. doi: 10.30541/v35i2pp.171-187

Martin, E. (2009). *Encyclopedia of Database Systems* (L. LIU & M. T. ZSU, Eds.). Boston, MA: Springer US.

Melake, T., & Merhawi, W. (2018). Drivers of poverty in Sub-Saharan Africa: Policy implications for achieving Agenda 2030 for Sustainable Development. *International Journal of Scientific and Research Publications (IJSRP)*, *Vol. 8*, 8462.

Mukherjee, S., Chatterjee, A., Bhattacharyya, R., & Kar, S. (2011). A Fuzzy Mathematics Based Approach for Poor Household Identification. *Journal of Development Economics*, *Vol. 1, No. 1*, 22-27.

Nge'the, N., & Omosa, M. (2016). *Drivers and Maintainers of Poverty in Kenya: A Research Agenda.* Nairobi: Institute for Development Studies, University of Nairobi.

Orbeta, A. C. (2006). Poverty, vulnerability and family size: Evidence from the Philippines. *Poverty Strategies in Asia: A Growth Plus Approach*(No. DP 2005-19), 171-193.

Owuor, G., Ngigi, M., Ouma, A. S., & Birachi, E. A. (2007). *Determinants of rural poverty in Africa: The case of small holder farmers in Kenya* (Vols. 7, No. 17).

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine Series 6*, *Vol. 2, No. 11*, 559-572.

Pervez, Z. J., & Usman, A. K. (2011). The Role of Education and Income in Poverty Alleviation: A Cross-Country Analysis. *The Lahore Journal of Economics*, *Vol. 16, No. 1*, 143-172.

Rice, J. (2007). *Mathematical Statistics and Data Analysis* (No. 978-0534-39942-9). Belmont, California: Brooks/ Cole Cengage Learning.

Sabates, R. (2008). The Impact of Lifelong Learning on Poverty Reduction. *IFLL Public Value Paper 1. Latimer Trend, Plymouth.*, 5-6.

Sachs, J. (2014). *Ending Extreme Poverty.* Washington, DC: World Bank.

Simon, C., Belyakov, A. O., & Feichtinger, G. (2012). Minimizing the dependency ratio in a population with below replacement fertility through immigration. *Theoretical Population Biology*, *Vol. 82, No. 3*, 158–69.

Tazoacha, F. (2001). The causes and impact of poverty on sustainable development in Africa. In *A paper presented at the conference held in bordeaux, france* (pp. 22–23).

Theodoridis, S., & Koutroumbas, K. (2006). *Pattern Recognition (Third Edition).* USA: Academic Press.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *Vol. 63, No. 2*, 411-423.

Tvedten, I., & Nangulah, S. (n.d.). *Social Relations of Poverty: A Case-Study from Owambo, Namibia.* USA: Chr. Michelsen Institute.

United Nations. (n.d.). *Poverty | United Nations Educational, Scientific and Cultural Organization.* Retrieved 2015-11-04, from `www.unesco.org.`

United Nations. (2015). The Millennium Development Goals Report. *United Nations*, 72.

Wagle, U. (2002, mar). Rethinking poverty: definition and measurement. *International Social Science Journal*, *Vol. 54, No. 171*, 155-165.

World Bank group. (2017). *Atlas of Sustainable Development Goals 2017: From World Development Indicators.*

World Bank Group. (2018). *Kenya Economic Update: Policy Options to Advance the Big 4* (Vol. 17; Tech. Rep.). Nairobi.

World Bank; International Monetary Fund. (2016). *Global Monitoring Report 2015/2016 : Development Goals in an Era of Demographic Change.* Washington, DC: World Bank.

Zhu R, Z., & D, K. M. (2015). Reinforcement Learning Trees. *Journal of the American Statistical Association*, *Vol. 110, No. 512*, 1770-1784.

# Appendix

### Implementation of Principal Component Analysis (PCA)

1. Standardize of the dataset.
2. Construct the covariance matrix.

   Covariance measures the relationship between two random variables (Rice, 2007). Covariance is calculated using the formulae:

$$\Sigma_{p,q} = \begin{bmatrix} \sigma^2_{pp} & \sigma^2_{pq} \\ \sigma^2_{qp} & \sigma^2_{qq} \end{bmatrix}, \quad q > p \tag{22}$$

$$\text{where } \operatorname{var}(p) = \frac{1}{n-1}\sum (p_i - \bar{x})^2 \text{ and } \operatorname{cov}(p,q) = \frac{1}{n-1}\sum (p_i - \bar{p})(q_i - \bar{y}) \tag{23}$$

3. Calculate eigenvectors and eigenvalues in equation 22 and order them.
4. Select m eigenvectors which correspond m largest eigenvalues.
5. Construct a projection matrix K using the largest m eigenvectors.
6. Obtain a new m-dimension subspace by using the projection matrix to transform the input dataset.

### Calculating Hopkin's statistic

1. Sample $n$ points from data set D.
2. For each observation, calculate the distance between it and its nearest neighbour.
3. For each subsequent point, calculate the distance between each point and its nearest neighbour.

4. Calculate the Hopkin's statistic (H) using the formulae below:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} \tag{24}$$

### Calculating gap statistic

1. Implement the algorithm using various number of clusters, $k$ and compute the total within intra-cluster variation $W_k$
2. Using random uniform distribution, generate P reference data sets. Then, cluster each of these reference data sets with varying number of clusters $\mathrm{k} = 1, \ldots, k_{\text{max,}}$ and compute the corresponding total within intra-cluster variation $W_{kp}$
3. Compute the gap statistic

### Calculating Silhouette coefficient

1. Implement the algorithm by using k (1 to 10) clusters.
2. Calculate Silhouette coefficient ($Si$) for each k.
3. Plot $Si$ against the clusters k.
4. The value of k with maximum $Si$ is chosen as the optimal number.

$$\mathrm{Si} = (p - q)/\max(p, q) \tag{25}$$

### Implementation of K-means algorithm

1. Select K clusters to be used.
2. Randomly select k objects as the initial cluster centres.
3. Compute the distance matrix between each pair of objects and place each near the closest centroid.
4. For each cluster, re-calculate the centroid using all data points.
5. Repeat the above two steps to minimize the total within sum of square until the maximum number of iterations is reached.

### Implementation of K-medoids algorithm

1. Select K clusters to be used.
2. Select k objects (to become medoids)
3. Compute the dissimilarity matrix.
4. Assign each object to the closest medoid.
5. For each cluster, check if any of the objects within the cluster decreases the average dissimilarity coefficient; if any such object(s) exist, select it as the medoid for that cluster;
6. If any medoid changes after step 4, repeat steps 3 and 4, else end the algorithm.

### Implementation of hierarchical clustering

1. Place each data point into its own cluster
2. Identify the clusters closest to each other
3. Merge the most similar clusters
4. Iterate step 2 and 3 above until all similar clusters are merged together.

### Calculating the Dunn's index

1. Compute the least pairwise separation distance (min.separation) between each data point in cluster $k1$ and data points in the other clusters.
2. Use the maximum intra-cluster distance to compute the maximum diameter between objects in each cluster.
3. Calculate the Dunn index ($D_i$):

$$D_i = \frac{\text{minimum separation}}{\text{maximum diameter}}$$

### Generating ordered dissimilarity matrix (ODM)

1. Compute the dissimilarity matrix (DM) using a distance measure. Manhattan distance (equation 4) was used in this study.
2. Create an ordered dissimilarity matrix (ODM) by ordering the DM for similar objects to be close (Bezdek, 1981).
3. Display the ODM using a visual output.

### Food groups used

1. Cereals, grains and cereals products
2. Roots and tubers
3. Nuts and pulses
4. Vegetables
5. Meat, fish and animal products
6. Fruits
7. Milk and milk products
8. Fats and oils
9. Sugar, sugar products and honey
10. Spices and condiments