Vol.4.Issue.4.2016 (October -December)

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

http://www.bomsr.com

Email:editorbomsr@gmail.com

ISSN:2348-0580

RESEARCH ARTICLE

APPLICATION OF k-MEANS AND PARTITIONING AROUND MEDIODS (PAM) CLUSTERING TECHNIQUES ON MAIZE AND BEANS YIELD IN TANZANIA

JUSTINE NKUNDWE MBUKWA[1]., G.V.S.R. ANJANEYULU[2]

Department of Statistics, Acharya Nagarjuna University, Nagarjuna Nagar Guntur, Andhra Pradesh-India

JUSTINE NKUNDWE MBUKWA

ABSTRACT

This paper contributes to the application of k-means and k-mediods multivariate statistical methods for the purpose of revealing optimal clusters and assessing the consistency of individual districts within the group. Data used were extracted from united republic of Tanzania (Ministry of Agriculture, Livestock and Fisheries(MALF) (2003/12) consisted a total of (n=36 districts) with both maize and beans yield. The R-statistical computing version (3.1.1) was used. The study findings revealed that 36 districts were grouped into six clusters using k-means algorithm. Using the k-mediods,it was revealed that only 11 districts were found to very well structured since their silhouette width (s_i) is above 0.5. Nevertheless ,the clusters validation was done in such a way that individual district whose silhouette width (s_i) close to 1 was regarded as highly consistency clustered whereas districts with (s_i) greater or equal to 0.25 were said to be somehow well clustered and otherwise. The paper concludes that few districts that are very consistent given the threshold margin should to be monitored and evaluated effectively to ascertain productivity. The study recommends that the government should pay attention on allocating the scarce resources to the consistency clusters along with policy review in favour of smallholder farmers through access and timely for all important farm inputs in future.

Keywords: Clustering, k-means and PAM or k-mediods

1. INTRODUCTION

The main goal of this clustering multivariate statistical method is to find out the optimal groups of the observations vectors or objects which are homogeneous but dissimilar to each other. The optimal grouping for

observations or objects is done naturally (Johnson & Wichern (2007)). Clustering analysis methodology takes into account to identify the patterns/directional vectors from a dataset by grouping the (multivariate) observations into segments in order to be to provide the better proper cornerstone for data interpretations and decision making. This process is facilitated by statistical distance (Euclidean distance). Kalyankar & Alaspur (2013) pointed out that clustering technique has been divided into two parts, namely hierarchical and non hierarchical.

In Tanzania environment, the question of crop yield is still an ongoing debate among stakeholders that calls for comprehensive integrated in-depth solutions. The Agricultural sector contribute to about half of the GPD in Tanzania, nevertheless the average harvests are still very low (Mkapa, 2005).

Msuya et at.(2008) identified that there are variations of productivity among smallholder farmer in Tanzania. The research findings revealed that production and productivity are very low and slightly varied ranging from 0.01 tones/ha to 6.77 tones/ha. This low productivity remains to be a fundamental cause of stagnant economy and persistence poverty. Specifically, some scholars such as Amani et al.(2004; 2005), Skarstein (2005), Isinika et al.(2003), MAFC (2006) indicated that maize productivity is suffering because of insufficient practice of improved high yield farming as compared to subsistence.

On the other hand, the World Bank (2007) noted that Tanzania is among countries in Sub-Saharan Africa whose Agriculture is mainly done by small holder subsistence farmers. The output from this sector is very low and undermines economic growth, food insecurity as well as poverty persistence. The sector is not performing well as expected. Despite the number of policies and strategies that were in place post-independence yet productivity is very low. Although the sector employs about 67 per cent of labor force, yet the contribution to the Growth Domestic Product (GDP) is also very low. We are looking for such discovery knowledge clustering statistical tool in order to identify patterns relationships to enable all stakeholders to execute joint efforts to solve such complex problem.

2. Literature Review

Cluster analysis is sometimes refers to mining techniques which has gain its vital in many fields of studies. Its main goal is to split observations into distinct groups for knowledge discovering. It is concerned with searching for hidden patterns inside largely available data so that the information revealed can be transformed into usable fashion for decision making (Everitt & Dunn, 1991; Everitt et al., 2011).

There are numerous of studies have been done to the field of agricultural crop sciences in particular. Some of the works having been done include that of Kumar & Kannth (2013) who used data mining techniques to extract useful information from agricultural dataset of annual measurements of fertilizer nutrients consumed on wheat yield production in India. The research findings concluded that fertilizer nutrients were the most prominent factor for wheat yields. Another study by Veenadhari et al.(2011) was carry out to review of studies on how data mining techniques are useful in the field of agriculture. The research findings revealed that the techniques like ID3 algorithms, k-means, k-nearest Neighbors, Artificial Neural Networks and Machine vector were found to be vital to uncover the knowledge embedded in crop data. Tripathi & Kesswani (2012) carried the study to cluster KHARIF and RABI crops for ten years among districts in India with similar crop production using k-means (centroid criterion). The data analyzed were extracted from the Directorate of economics and Statistics Department of Agriculture and Cooperation Ministry of Agriculture in India. It was discovered that Indian states were classified into three similar groups. This indicated that some states were placed in identical group as per produce of KHARIF and others as per RAB crops produce. Again, Medar & Rajpurohit (2014) carried out a survey design to ascertain the application of data mining techniques in the field of agriculture. The research findings unveiled that the k-means, k-nearest neighbor (kNN) and Support Vector Machine (SVM) were substantial to find the unseen patterns from large amount of data. In a course of forming groups on the basis of area, production and productivity using some selected major crops in Karnataka state in India the study was done (Rathod et al., 2012). Data analysed were collected from various issues of 'Karnataka at a Glance' for the period of 1985-2005 for twenty years. The study was conducted under two segments whose one took into account the Pre WTO (World Trade Organization) era in between 1985-1995 whereas the second constituted the Post WTO era from 1995-2005. The data analysis was done by the Ward's hierarchical clustering method. The research findings revealed that the crops like sorghum, cotton, paddy, groundnut and

ragi were place in similar area. On the second segment, crops like paddy, maize, mango, ragi and groundnut were classified in the similar area. It was further indicated that, sugarcane and sorghum were similar in production for first period while sugarcane was classified in the second period. On the basis of productivity, major clusters were formed by horticulture crops for both periods. Furthermore, Narkhede & Adhiya (2013) carried out the study seeking to compare on the application of cluster techniques for crop prediction to solve the problem of noise and optimization via review of literatures in India. It was revealed that Beehive and improved K-means algorithms were excellent in solving the challenges in such a way that the good qualities of clusters were achieved to enhance crop prediction.

2.1 Contribution to the Existing Knowledge

In Tanzania, there are limited studies on an application of multivariate cluster analysis in the field agricultural crop science. Despite the contribution of this sector to Gross Domestic Product to our Nation, yet there required understanding the hidden structure embedded in dataset that might be used as base for decision making among policy makers. Among studies having been surveyed on clustering analysis in crop field, there still existing gap in Tanzania.

Comprehensive efforts to raise crop productivity have been shown. This was envisaged during the implementation of the Agricultural Sector Development Program (ASDP). This was in line with fertilizers subsidization among smallholder farmers via National Agricultural Input Voucher system (NAIVS) to raised productivity (Hepelwa et al., 2013). In this paper, two datasets were used. It comprised of the panel data from National Bureau of Statistics in 2007 before NAVIS and cross-sectional data from householder farmers (327) collected Tabora and Ruvuma regions after NAVIS in 2012. The target farmers were with access to fertilizer subsidy via that programme. The findings indicated that the average crop yield per acre were relatively higher (changed from 1,526.5 kg to 3,806 kg in 2007 and 2012 correspondingly). Therefore, the study findings revealed that majority poor smallholder farmers do not access the fertilizers.

There are still existing 39.6 per cent of technical inefficiency for maize production among smallholder farmers in Tanzania as results of education, lack of extension services, limited capital, land fragmentation, and unavailability and high input prices (Msuya et al., 2008). Baha et al. (2013) indicated a technical inefficiency of 37.7 per cent of maize among smallholders in Babati district. Thus, inputs such as farm size, formal education, number of plots owned by a farmer, number of times a farmer contacts with extension officers, use of insecticides and use of hand hoes should be taken seriously.

Meanwhile, Tanzania in particular is being collecting crop yield data from various districts and stored to the established database of the Ministry of Agriculture and Food Security Cooperatives in collaboration with National of Statistics. These data are available and can be accessed by anybody for the research purpose as well planning using legal procedures. Despite these efforts done to such an extent none of the research work has been shown using clustering multivariate methodology to establish optimal homogeneous clusters of districts having closely related in terms of maize and beans yield to enhance the effectiveness decision in improving the productivity.

3. Study Area and Methods

The study was carried in Tanzania. It is located in the eastern part of Africa. It is found along the Latitude 6^0 00' S and Longitude of 35^0 00' East of Greenwich. This location has effects on its climate. In the coast, it is noticed with tropical characterized by bimodal rainfall distributions and along the mountainous or southern highland in nature the weather or bimodal rainfall patterns is dominant. It bordered by Kenya, Uganda, Rwanda, Burundi, Democratic Republic of Congo (DRC), Zambia, Malawi, Mozambique and Indian Ocean.

The quantitative scale measured data on crop yields were collected from the Ministry of Agriculture,Livestock and Fisheries (MALF) in collaboration with National Bureau of Statistics maize and beans yield (2003/12). The data originated from 36 districts found in major five prominent for staple food production including Morogoro, Iringa, Mbeya, Ruvuma and Rukwa. In fulfilling the existing gap the k-means and PAM clustering algorithm were used to reveal the useful information embedded in data. The statistical data analysis was carried out using the R-statistical computing (3.1.1) version.

3.1 The k-Means Algorithm

This technique was first coined by MacQueen (1967), whose interest was to classify n- cases into k pre-defined clusters. It splits the data set into k groups such that each of the n-data items belongs to a group with a closest centroid/mean. The aim of this method is to minimize the objective function (the square error function).

$$z = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad \text{... (3.1.1)}$$

Where

$\left\| x_i^{(j)} - c_j \right\|^2$,is a chosen statistical distance measure between a data point $x_i^{(j)}$ representing an object in a cluster and the cluster centre c_j is the distance of the *n* data points from their respective cluster centroids.

z , is the sum of the squared error for all data objects within the data set.

3.2 Partioning Around Mediods (PAM) Aligorthm or k-Medoids

The k-mediods/median was first introduced by Kaufman and Rousseeuw (1987). It chooses data points as centers (medoids) to allocate the observations. It calculates in such a way that the total dissimilarity of all objects to their nearest medoid is minimal. It is said to be robust against k-means (centroids) because it is capable of handling outliers or extreme values.

The k-means method produces an elbow graph which is somehow ambiguous to comprehend. One of the weaknesses of the k-means is to fails to address some of the following questions such as an issue of the quality of clusters, if the within dissimilarities are less than the between similarities, which objects appear to be well classified? Which one is misclassified? Which one lies in between clusters? What is the overall structure of the clusters? Nevertheless, an alternative approach known as the average silhouette width has been extended along with PAM method. This average width is responsible for providing tentative answers to the emerged questions. In mathematical sense, this average width index can be derived as follows:

Given the each i^{th} item; let $u(i)$ be the average dissimilarity distance from point i to all other items in its cluster (within the same cluster). The value $u(i)$ can be interpreted as how well i^{th} is places to its cluster. However, the low the value the better the assignment has been done. Again, we can define the average dissimilarity distance of point i^{th} to a cluster M as the average of the distance from i^{th} to all points in M .

Let, $v(i)$ be the smallest average dissimilarity distance to any other cluster i to all points in another (where i is not a member) or neighboring cluster $d(i,\mathrm{M})$. The $d(i,\mathrm{M})$, is the mean to all objects in any other cluster M.

The cluster whose average dissimilarity is lowest is said to be the neighboring cluster of i since is the next best fit cluster for point i . Now, the silhouette width index of the i^{th} is computed by:

$$S(\mathrm{i}) = \frac{\mathrm{v}(i) - u(i)}{\max\{\mathrm{v}(i), \mathrm{u}(i)\}} \quad \text{... (3.2.2)}$$

Thus, $S(\mathrm{i})$ can be calculated under the combination of $\mathrm{u}(\mathrm{i})$ and $\mathrm{v}(\mathrm{i})$ such that:

$$S(\mathrm{i}) = \begin{cases} 1 - \dfrac{u(i)}{\mathrm{v}(\mathrm{i})} & if \quad u(i) < v(i) \\ 0 & if \quad a(i) = b(i) \\ \dfrac{v(i)}{u(i)} - 1 & if \quad u(i) > v(i) \end{cases} \quad \text{... (3.3.3)}$$

Thus the above notation concludes that:

$$-1 \le S(i) \le 1 \quad \text{... (3.3.4)}$$

The interpretations of the average $S(i)$ were categorized into four groups [excellent/ strong structure (0.7-1.0), very good structure (0.51-0.7), weak structure (0.26-0.50) and no substantial structure (<0.25) (Kaufman & Rousseeuw ,1987).

The overall mean silhouette width refers to the average value of $S(i)$ for all specified clusters. Thus, the objects whose average silhouette width values are close to 1 are said to be well clustered (samples are away from the neighboring cluster) and those with small $S(i)$ values are likely to lie between clusters. Any value of $S(i)$ close to zero (0), indicates that the sample is on or very close to the decision boundary between two neighboring clusters while the negative value indicates that a sample might have been located to the wrong cluster. The procedures for undertaken by PAM is similar to the elbow k-means method and can be computed through observing the following steps: (1) Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters (2) for each value of k and estimate the average silhouette of observations (avg.sil) (3) plot the curve of average silhouette according to the number of clusters k defined (4) therefore, the location of the maximum is considered as the appropriate number of clusters.

Kaufman & Rousseeuw (1990) introduced the so called CLARA algorithm. This is an extension of the k-medoids technique for handling large number of objects (several thousand observations) than small one. In this paper, PAM algorithm is appropriate since the number of observations is small and the large clustering technique does not hold true.

4. Results and Discussions

As a basic rule of thumb for both two clustering algorithm, the set of maize and beans yield measurements were standardized first before embarking to partitioning process using the scale() function. The combined vector for maize and beans was given by w=cbind (maize, beans) function**.** On the other hand, it is recommended to determine the optimal number of clusters (k) before starting using the k-means algorithm. In this particular paper two methods were used. These are:

#The R code elbow method for k-means () with package “factoextra”:

```
>fviz_nbclust(z, kmeans, method = "wss") +
 geom_vline(xintercept = 6, linetype = 2)
```

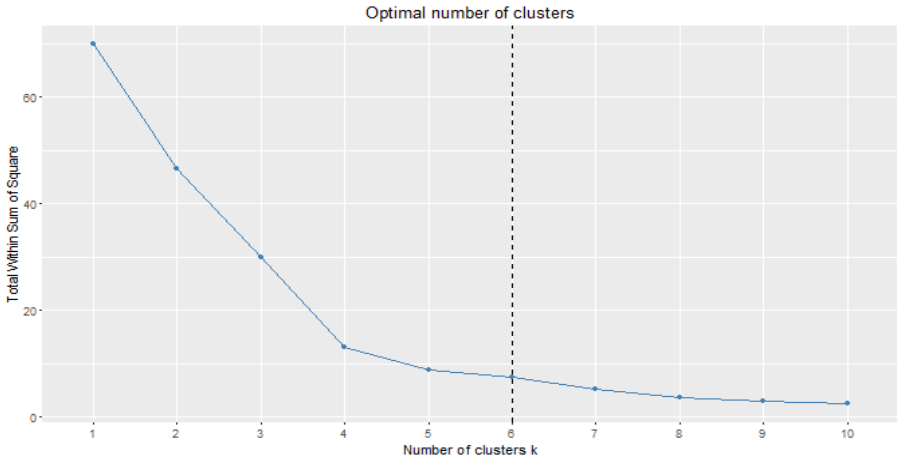

Figure (4.1): Plot of within-groups sum of squares against number of clusters.

Figure 4.1.1, indicates that the results of the within-groups sum of squares for 1 to 10 groups using the k-means. The solutions were plotted to see if there will be any indication of the number of groups. It can be visualized that after 6 clusters the observed difference in the within-cluster dissimilarity is not significant. Therefore, there are some absolute confidence that the optimal number of clusters to be used is should not go beyond six (6).

4.1 The k-Means Clustering Maize and Beans among 36 Districts in Tanzania

Based on the data analysed with pre-defined k=6 clusters, the findings revealed that the crop yield for maize and beans collected from thirty six (36) districts were grouped into six clusters on the basis of nearest centroid. It was also indicated that districts placed together produce the similar maize and beans crop yield on

average. About 91.0 per cent of the within sum of square (wss) is explained by all six clusters (table 2). It was further revealed that, the first cluster consisted of six districts, second cluster (2 districts), third cluster (3 districts), fourth (11 districts), fifth cluster (one district) and sixth cluster (13 districts). These research results having been revealed in the context of observed in crop yield of Tanzania are in line with the study done by Kumar and Kannth (2013); Veenadhari et al (2011); Narkhede & Adhiya (2013); Medar & Rajpurohit (2014).

Furthermore, in order to come up with the sense of useful information and interpretations, the similar districts converging to the pre-defined six clusters have been summarized (table 1.4.1). It is known geographically that some districts found in southern highland (unimodal rainfall distributions) and coast otherwise (bimodal rainfall distributions). The processing of putting districts in similar clusters did not take into account the issue of agro-ecological zone rather than centroid criterion. The findings indicated that more districts were placed to cluster six and four respectively. Though these research results are more exploratory in nature, however, they portray out the clear understanding on how government needs to be conscious in prioritizing the districts by committing more scarce funds to the cluster comprised of many districts with average maize and beans yield).

However, on the basis of these revealed, it calls for interventions from different stakeholders with joint efforts to come up with the comprehensive plans in order to raise the crop productivity. All favorable infrastructures are to be addressed concurrently to rise up crop yield so that the poverty and prolonged hunger come to an end. Along with these research findings, government interventions are required to in order to review the existing agricultural policy to centre on the equitable farm inputs, training, enhancing training of extension officers as well as enabling working environment to them , friendly cost of production, access and timely of farm inputs taking into consideration of smallholder farmers whose districts are highly related. Great attentions must be directed towards group four and six which seemed to be potential as have been compacted closely with similar yield. In this way it is possible to realize the economic efficiency in terms of allocation of financial scarce resources and raise productivity. The R-code and output for the fitted clusters using k-means have been shown as follows:

```
> maize.bean6<- kmeans(z, 6, nstart = 36)
> print(maize.bean6)
K-means clustering with 6 clusters of sizes 6, 2, 3, 11, 1, 13
Cluster means:
     [,maize]     [,beans]
1 -1.07697703 -1.24106902
2  1.66387357  0.06392372
3  0.04202657 -1.20275128
4 -0.24891115  0.73642494
5  4.28875149  3.51799426
6  0.11210050 -0.05321908
Clustering vector:
[1] 4 5 6 4 6 2 6 4 3 4 4 2 6 4 6 6 1 1 6 1 6 6 4 1 4 1 1 4 6 6 3 4 6 6 4
[36] 3
Within cluster sum of squares by cluster:
[1] 0.6449783 0.3262527 0.5118543 2.9491123 0.0000000 1.8598377
 (between_SS / total_SS =  91.0 %)
```

Table 4.1.1: The Convergence of the Districts to the Respective Clusters

```
> maize.bean<-data.frame(z, maize.bean$cluster)
> maize.bean
```

	maize	beans	maize.bean.cluster
1(Chunya)	-0.26428296	0.8817337675	4
2(Ileje)	4.28875149	3.5179942639	5
3(Kyela)	-0.12794866	-0.1977317016	6
4(Mbarali)	-0.40061726	0.4219208902	4

5(Mbeya Rural)	-0.29703860	-0.0313232317	6
6(Mbeya Urban)	1.26106768	0.0934831207	2
7(Mbozi)	0.09514383	0.0759664396	6
8(Rungwe)	0.18809904	1.2145507072	4
9(Kilombero)	0.32620391	-1.3209888732	3
10(Kilosa)	-0.38733794	0.5489168277	4
11(Morogoro Rural)	-0.69895920	1.7860324261	4
12(Morogoro Urban)	2.06667946	0.0343643222	2
13(Mvomero)	-0.18991879	-0.1583191693	6
14(Ulanga)	-0.05358449	0.6802919355	4
15(Iringa Rural)	-0.25011836	-0.1276649774	6
16(Iringa Urban)	0.16419627	-0.3728985120	6
17(Iringa DC)	-1.21154076	-1.4676910770	1
18(Iringa MC)	-1.38417185	-1.4020035231	1
19(Kilolo)	-0.31120320	0.0562601735	6
20(Kilolo DC)	-1.14071774	-1.3954347677	1
21(Ludewa)	0.34656553	-0.0006690399	6
22(Makete)	0.12701419	-0.4867569387	6
23(Mfindi)	-0.09165186	0.3606125065	4
24(Mfindi DC)	-1.18498213	-1.3363159692	1
25(Njombe Rural)	-0.23772433	0.5905189452	4
26(Mpanda)	-0.77686452	-0.8349009744	1
27(Mpanda Rural)	-0.76358520	-1.0100677848	1
28(Mpanda Urban)	-0.88664019	0.3387166552	4
29(Nkasi)	0.41296210	0.1766873556	6
30(Sumbawanga Rural)	0.42535613	0.3715604322	6
31(Sumbwanga Urban)	0.17481972	-1.4742598323	3
32(Mbinga)	-0.09076658	0.6124147965	4
33(Namtumbo)	0.56965802	0.1307060679	6
34(Songea Rural)	0.49263799	-0.1276649774	6
35(Songea Urban)	0.18544317	0.6649648396	4
36(Tunduru)	-0.37494392	-0.8130051231	3

The fitted clusters (table 4.1.1) have been further clarified in (table 4.1.2) indicating clusters, cluster sizes, sorted districts per clusters and geographical locations. The distribution of districts under the respective specified clusters has been shown as follows:

Table 4.1.2: Convergence of Districts by k-means Six Clusters

Clusters	**Cluster sizes**	**Sorted districts by respective cluster**	**Geographical locations**
1	6	Iringa DC, Iringa MC, Kilolo DC, Mfindi DC, Mpanda and Mpanda Rural	Southern Highland
2	2	Mbeya Urban, Morogoro Urban*	Southern Highland and coast*
3	3	Kilombero*, Sumbawanga Urban and Tunduru	Southern highland And coast*
4	11	Chunya, Mbarali, Rungwe, Kilosa, Morogoro Rural*, Ulanga*, Mfindi, Njombe rural, Mpanda urban, Mbinga and Songea urban	Southern highland and coast*
5	1	Ileje	Southern high land
6	13	Kyela, Mbeya Rural, Mbozi, Mvomero, Iringa Rural, Iringa Urban, Kilolo, Ludewa, Makete, Nkasi, Sumbwanga Rural, Namtumbo and Songea Rural	Southern high land And coast*
Total	**36**	Total within sum of squares (**wss**) by cluster **(91.0 percent)**	

Source: Research Findings (2016); (*) a district is along the coast

4.2 Partitioning Around Mediods (PAM) Algorithm /k-Medoids

In fact, the PAM algorithm came in operation since k-means fail to address some of the emerging questions (ibid). In this partitioning method, the validity of the goodness of cluster (internal Cluster validation measures) has been assessed using the silhouette width. The number of optimal clusters pre-specified has been identified by the elbow graph. It is plotted such that the within sum of the squares (wss) by cluster against number of optimal clusters. The figure 4.2.2, points out that beyond six there will be no addition wss substantially. However, the PAM method goes beyond by assessing the validity of that suggested clusters as well as individual objects whether have been sorted accordingly. Using R-statistical computing, with PAM Algorithm, it's possible to use the function fviz_nbclust () to present an elbow (wss) for 36 districts under six groups of maize and beans yield **(tons/ha)** as follow:

The R-code for Elbow method for PAM clustering

```
>fviz_nbclust(scaled(z), pam, method = "wss") +
 geom_vline(xintercept = 6, linetype = 2)
```

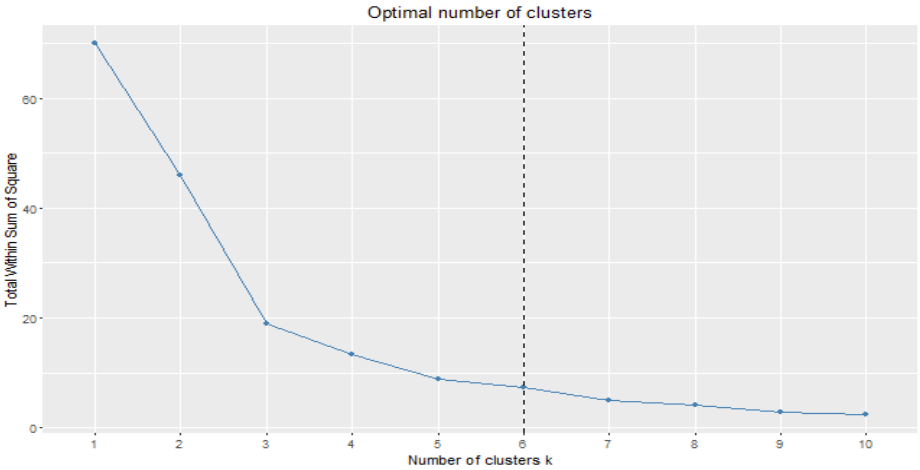

Figure (4.2.2): Plot of within-groups sum of squares against number of clusters using PAM algorithm.

In the above figure 4.2.2, six (06) clusters have been suggested by elbow method of PAM Clustering. At point indicating cluster 6 shows a sort of bend (knee). It should be noted that both approach for elbow graph (k-means and PAM) suggest six clusters to be used since beyond that value there is no substantial added value influence the results. However, under the PAM method silhouette width has been used to validate the established clusters. With respect to the silhouette width, the research findings revealed that on average, all districts have been grouped with an average rate of 0.4273818**.**

Using this particular index, the overall six clusters fit and individual district to be in its cluster have been assessed accordingly. With reference to the table 3 $s(i)$ width results per cluster indicated that:

Cluster I; $s(i)$ =0.35 ; comprised of eleven (11) districts namely Chunya, Ulanga, Njombe Urban, Mbinga, Kilosa, Rungwe, Morogoro Urban, Mbarali, Mfindi, Mpanda Urban and Songea Urban. The silhouette is very narrow and it indicates that the Structure is relatively weak. However, it has been revealed that chunya, Ulanga, Njombe Rural and Mbinga have been somehow structured well to its cluster since the values of $s(i)$ are above the average. However, Mfindi and Mpanda Urban have a $s(i)$ are 0.08446790 and 0.06956573. These indices are < 0.25 and close to zero correspondingly. In this regards, the results potray that both two districts hold the intermediate position between cluster 1 and 3 (cluster 3 is the neighbor of both two districts).

Cluster II; $s(i)$ =0.00 and consists of one district (Ileje). This distict is neither positioned to cluster 1 nor cluster 2. It is not recommended as a good cluster.

Cluster III, $s(i)$ =0.48 and consists of nine districts (Kyela, Mvomero, Iringa Rural, Mbeya Rural, Makete, Iringa Urban, Kilolo, Mbozi and Tunduru. In this particular cluster, the s(i) is good compared with the 1 and second clusters. Districts like Kyela, Mvomero, Iringa Rural and Mbeya Rural indicate to better clustered well with $s(i)$ (0.68554, 0.675916, 0.6469 and 0.5536) respectively. In this cluster, $s(i)$ =0.09960306 for Tunduru

district is less <0.25 and close to zero. Thus, it holds the intermediate position between cluster 3 and 6. In other words is not recommended as a good cluster.

Cluster IV; have $s(i)$ = 0.2331573. three districts were found to be somehow well structured (Mbeya Urban, Namtumbo, Morogoro Urban) while Sumbawanga Rural, Nkasi, Songea Urban were not properly clustered. That is their silhouette values are <0.25 (Kaufman and Rousseeu, 1987; 1990). However, the Ludewa district was indicated with $s(i)$ = -0.03793572. This depicted that, it is at an intermediate lying far from both cluster three and four. In other words it does not belong to any of the two clusters.

Cluster V; had the average of $s(i)$ =0.8279901 width.It consists of two districts namely Sumbawanga Urban and Kilombero. The Districts placed in this cluster are well structured since its $s(i)$ is close to 1. Again there is no doubt about the two districts to its clusters $(s(i) = 0.8345 \; and \; 0.8215)$ are also close to one respectively. Hence, it is recommended to be excellent.

Cluster VI; $s(i) = 0.64$ and consists of six districts namely Mfindi DC, Kilolo DC, Iringa DC, Iringa MC, Mpanda Rural and Mpanda). It has been structured properly and recommended to be excellent. However, a doubt has been observed to Mpanda District whose $s(i) = 0.33079410$. This indicates to be relatively well clustered.

Contrast to k-means, the results from the PAM should be considered for recommendation for the silhouette width greater than 0.25 and not otherwise (ibid). With that threshold width, it gives us a confident that a district has been placed well to its cluster. The cluster fit validation and suitability assessment using silhouette width indices have been observed. However, the districts whose $s(i)$ width is indicated by (*) were not recommended as a good cluster (table 4.2.3):

Table 4.2.3: Convergence of the Districts by Silhouette Width

> summary(pam.maize.bean6)

Silhouette plot information:

	cluster	neighbor	sil_width
1 (chunya)	1	3	0.**54359761**
14(Ulanga)	1	3	0.49686317
25(Njombe Rural)	1	3	0.47286214
32(Mbinga)	1	3	0.47086966
10(Kilosa)	1	3	0.41353459
8(Rungwe)	1	4	0.38398070
11(Morogoro Rural)	1	3	0.37923803
35(Songea Urban)	1	4	0.31492676
4(Mbarali)	1	3	0.25157853
23(Mfindi)	1	3	0.08446790*
28(Mpanda Urban)	1	3	0.06956573*
2(Ileje)	2	4	0.00000000*
3(Kyela)	3	1	**0.68554323**
13(Mvomero)	3	1	**0.67591652**
15(Iringa Rural)	3	1	**0.64694436**
5(Mbeya Rural)	3	1	**0.55360007**
22(Makete)	3	5	0.46316669
16(Iringa Urban)	3	4	0.45749543
19(Kilolo)	3	1	0.43044805
7(Mbozi)	3	4	0.32086643
36(Tunduru)	3	6	0.09960306*
6(Mbeya Urban)	4	3	0.43110791
33(Namtumbo)	4	3	0.36085687
12(Morogoro Urban)	4	3	0.33061922
30(Sumbawanga Rural)	4	1	0.22339679*

29(Nkasi)	4	3	0.21827673*
34(Songea Rural)	4	3	0.10577906*
21(Ludewa)	4	3	-0.03793572*
31(Sumbwanga Urban)	5	3	**0.83450971**
9(Kilombero)	5	3	**0.82147057**
24(Mfindi DC)	6	5	**0.77803428**
20(Kilolo DC)	6	5	**0.76659963**
17(Iringa DC)	6	5	**0.75053811**
18(Iringa MC)	6	5	**0.73077571**
27(Mpanda Rural)	6	3	0.49251796
26(Mpanda)	6	3	0.33079410

Average silhouette width per cluster:

[1] 0.**3528623** 0.0000000 0.**4815093** 0.2331573 **0.8279901 0.6415433**

Average silhouette width of total data set:

[1] 0.4264558

Clustering vector:

[1] 1 2 3 1 3 4 3 1 5 1 1 4 3 1 3 3 6 6 3 6 4 3 1 6 1 6 6 1 4 4 5 1 4 4 1 3

Note: **DC=District council and MC= Municipal Council**

The figure 4.2.3, indicates the plot of six clusters with the respective silhouette width. The R-statistical code producing such a figure has been shown:

```
> plot(silhouette(pam.maize.bean6),  col = 2:5)
```

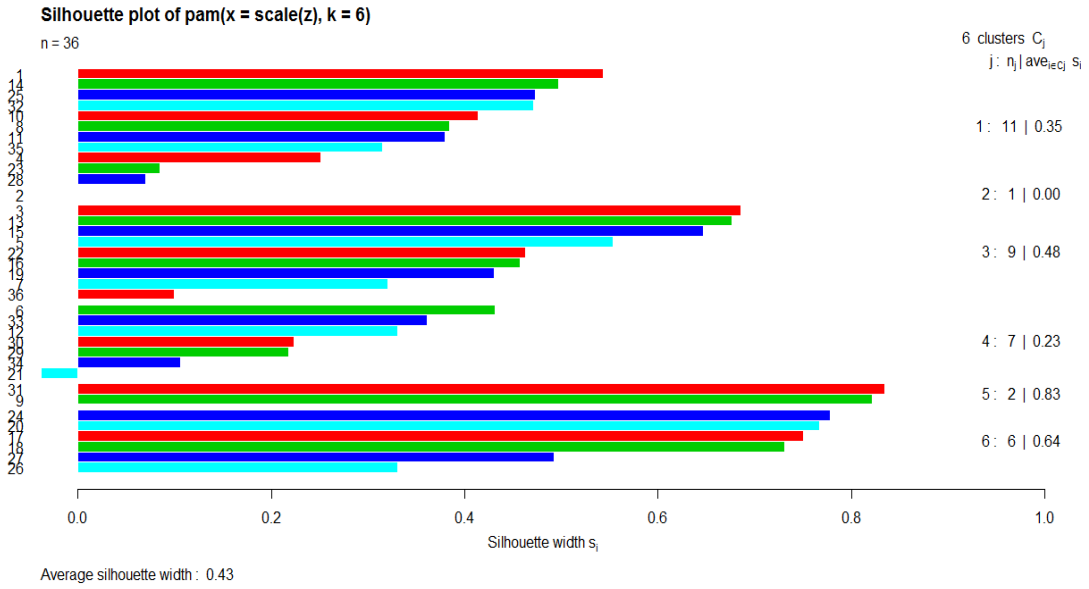

Figure (4.2.3): The Graph to Show the Number of Districts per Cluster

With reference to the figure 4.2.3 above, the number of elements (n_j) per cluster has been indicated. Each horizontal line corresponds to an element/district. The length of a lines corresponds to the silhouette width (s_i), which is the means similarity of each element to its own cluster minus the mean similarity to the next similar cluster. Also, it indicates the overall average silhouette width for six clusters validity. Alternatively, the breakdown of districts into number of districts per cluster and silhouette widths has been summarized below (table 4.2.4) and (figure 4.2.4) respectively:

Table 4.2.4: Average silhouette width for k=6 clusters

```
> fviz_silhouette(silhouette(pam.maize.bean6))
  cluster size ave.sil.width
1       1   11          0.35
2       2    1          0.00
3       3    9          0.48
4       4    7          0.23
5       5    2          0.83
6       6    6          0.64
Warning message:
Stacking not well defined when ymin != 0
```

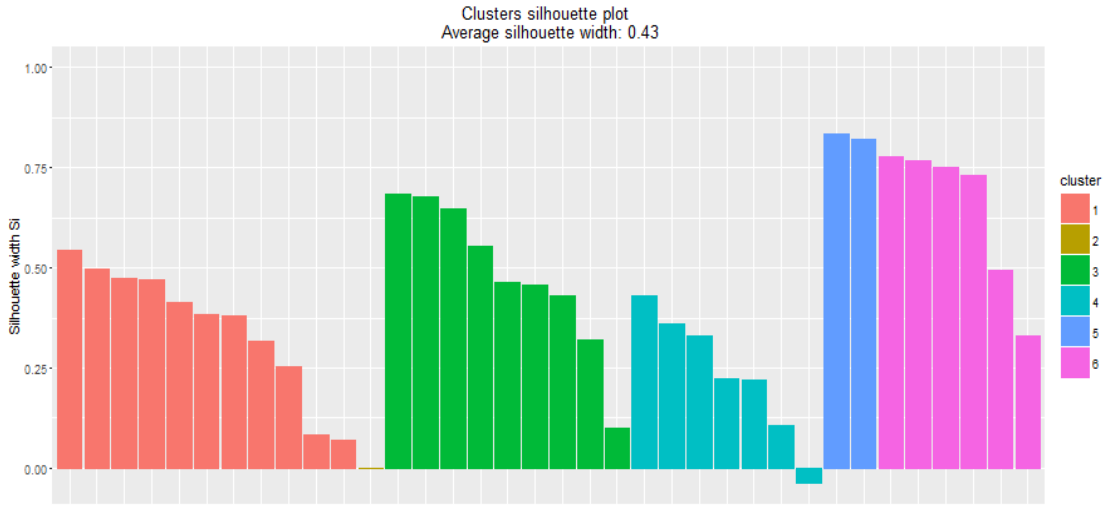

Figure (4.2.4): The histogram of the six clusters indicating the good clusters

It has been observed that one sample/district has a negative silhouette as shown with the figure 4 with blue bar crossing the x-axis. This means that they are not in the right cluster number 4. We can find the name of this sample/district and determine the closest cluster it belongs. Using R-code, it has observed that the Ludewa district is closer cluster three (03) and was wrongly located to cluster four (4) as shown below:

```
# Compute silhouette
> sil <- silhouette(pam.maize.bean)[, 1:3]
# Objects with negative silhouette
> neg_sil_index <- which(sil[, 'sil_width'] < 0)
> sil[neg_sil_index, , drop = FALSE]
      Cluster  neighbor   sil_width
21(Ludewa)    4        3  -0.03793572
```

5. Conclusion and Recommendation

The question of knowledge discovery in data remains as a most fundamental issue in defining an appropriate measure towards improving crop productivity. Using the cluster analysis the clear patterns have been identified with the respective to districts falling in both similar maize and beans yield. The k-means found to be good in identify similar groups, however it has been validated by the k-medoids (silhouette width). This study recommends the districts that have been clustered well on the basis of threshold quantity of not less than 0.25 silhouettes width. On the basis of these results, this paper recommends that when Tanzania government is planning to raise crop productivity, priority focus and attention should be on the clusters whose the silhouette width (s_i) greater than one though those close to one are the best. However, this should be in line with monitoring and control the crop yield in each district to realize the reasonable productivity. Again, this paper recommends that the issue raising crop productivity should be in collaborations with various

stakeholders such as individual farmers, government and non governmental agencies, policy makers, planning units and seed manufacturing firms.

Finally, it should be noted contextually that the classified groups cannot be stable over time. They might be subject to change depending on human psychology, government policies, geographical environment, climate change as well as technological innovation etc. However, in the short run, the results have a provided a stepping stone to understand the current trend/patterns such that effective monitoring and control can easily take place better prediction of the future yield given the realized benefits accrued from agricultural sector as a cross cutting among all development issues.

Acknowledgements

Heartily, we would like to recognize the financial support from the United Republic of Tanzania via Mzumbe University. This has played a significant role to carry out this part of the scholarly research work. Also, the word of thanks should go to the Ministry of Agriculture, Livestock and Fisheries of Tanzania where data were extracted. This paper has contributed to the robust optimal clusters among districts producing maize and beans yield in Tanzania using partitioning around mediods (PAM) techniques extended from the classical k-means.

References

[1]. Amani, H. K. R. 2004. "Agricultural Development and Food Security in Sub-Saharan Africa Tanzania Country Report". Retrieved on 21st March 2015 from http://www.fao.org/tc/Tca/work05/Tanzania.pdf

[2]. Baha, M., Temu, A., & Philip, D. (2013). Sources of Technical Efficiency Among Smallholders Maize Farmers in Babati District, Tanzania. *African Journal of Economic Review*, vol-*1, issue*-2, pp.1-14.

[3]. Everitt, B.S.,& Dunn, G.(1991). Applied Multivariate Data Analysis. John Wiley and Sons, New York.

[4]. Everitt, B.S., Landau, S., Leese, M.,& Stahl, D. (2011). Cluster Analysis, 5th edition. JohnWiley & Sons, Inc, New York.

[5]. FAOstat, "United Republic of Tanzania: Production and data," 2014, http:// http://faostat.fao.org

[6]. Hepelwa, A., Selejio. O., & Mduma, J. K. (2013). The voucher system and the agricultural production in Tanzania: is the model adopted effective? Evidence from the panel data analysis. *Environ. Dev. Initiat., Available from:http://www. efdinitiative. org/news/events/voucher-system-and- agricultural productiontanzania-model-adopted-effective-evidence(accessed June 2014)*.

[7]. Isinika, A., Ashimogo, G., & Mlangwa, J. (2003). AFRINT Country Report – *"Africa in Transition: Macro Study Tanzania"*, Dept of Agricultural Economics and Agribusiness, Sokoine University of Agriculture, Morogoro.

[8]. Johnson, R.A., & Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis* (6thedition). USA: Pearson Education, Inc., Publishing as Pearson prentice Hall.

[9]. Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. in 'Y. Dodge (editor) Statistical Data Analysis based on L1 Norm', 405-416.

[10] Kaufman, L., & Rousseeuw, P. (1990). *Finding groups in data: An introduction to Cluster analysis*. New York, NY: John Wiley & Sons.

[11] Khalid, M.N.(2011). Cluster Analysis -A Standard Setting Technique in Measurement and Testing. *Journal of Applied Quantitative Methods, vol-6, issue-2, pp.46-58.*

[12]. Kumar, A.V., & Kannth, T.V. (2013). Estimation of Influence of Fertilizer Nutrients Consumption of the Wheat Crop Yields in India: A Data Mining Approach. *International Journal of Engineering and Advanced Technology (IJEAT):ISSN:* ISSN: 2249 – 8958, *vol-3, issue-2, pp.316-320.*

[13]. Kalyankar, M. A., & Alaspurkar, S. J. (2013). Data Mining Technique to Analyse the Metrological Data. *International Journal of Advanced Research in Computer Science and Software Engineering*, vol-*3, issue*-2, pp.114-118.

[14]. MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol-1, issue-14, pp. 281-297).

[15] MAFC. 2006. Kilimo website http://www.kilimo.go.tz/Agr-Industry/Crops-grown-tz.htm

[16] Medar, R.A & Rajpurohit, V.S. (2014). A Survey on Data Mining Techniques for Crop Yields Prediction. *International Journal of Advanced Research in Computer Science and Management Studies, vol-2, issue-9, pp.59-64*

[17]. Mkapa, B.W.(2005).Towards Agricultural Growth in Tanzania: policies, priorities andstrategies. The 10th Sokoine Memorial Lecture by his excellence Benjamin William Mkapa, president of the United Republic of Tanzania, SokoineUniversity of Agriculture, Morogoro. 12th April, 2005.The Government Printer Dar es Salaam.

[18] Msuya, E. E., Hisano, S., & Nariu, T. (2008). Explaining productivity variation among smallholder maize farmers in Tanzania*. A presented in the XII World Congress of Rural Sociology of the International Rural Sociology Association, Goyang, Korea 2008.*

[19]. Narkhede, U.P., & Adhiya, K.P. (2013). A Study of Clustering Techniques for Crop Prediction- A Survey. American International Journal of Research Science, Technology, Engineering & Mathematics, vol-1, issue-5, pp.44-48

[20]. Rathod, S. A. N. T. O. S. H. A., Surendra, H. S., Munirajappa, R., & Chandrasekar, H. (2012). Cluster Analysis on Area, Production and Productivity for Majo Selected Crops in Karnataka. *Mysore Journal of Agricultural Sciences*, vol-46, issue-2, pp.293-299.

[21]. Skarstein, R. (2005). "Economic Liberalization and Smallholder Productivity in Tanzania; From Promised Success to Real Failure, 1985–1998, *Journal of Agrarian Change, Vol-5, issue-3, pp.334–362*

[22]. Tripathi, R.K., & Kesswani, N. (2012). Clustering Indian States on the Basis of Agricultural Produce of KHARIF and RAB crops. *International Journal of Electronics Communications and Computer Technology (IJECCT), vol-2, issue-2, pp.70-74*

[23]. Veenadhari, S., Misra, B.,& Singh, C.D. (2011). Data Mining Technique for predicting crop production: A Review Article. *International Journal of Computer Science & Technology, vol-2, issue-1, pp.98-100*

[24]. Verfaillie, E., Degraer, S., Schelfaut, K., Willems, W., & Van Lancker, V. (2009). Protocol for classifying ecologically relevant marine zones, a statistical approach. *Estuarine, Coastal and Shelf Science*, vol-*83, issue*-2, pp.175-185.

[25] World Bank. (2007). *World Development Report 2008*: *Agriculture for Development*. Washington, DC: World Bank.

Brief Biography of Corresponding Author

Mr. Justine Nkundwe Mbukwa is the Research Scholar at the Department of Statistics, University College of Sciences, Acharya Nagarjuna University, Guntur, Andhra Pradesh-India. He is doing a research work under the guidance of Prof. G.V.S.R. Anjaneyulu. He obtained his Bachelor of Science in Applied Statistics and Master's of Statistics from Mzumbe University and University of Dar es Salaam,Tanzania respectively. His research interest is covering the wide range of application of Multivariate Statistical Techniques particularly Machine Learning Algorithms/Statistical Pattern Recognition in natural sciences, engineering and social sciences.