



CHAPTER THREE

THE DISTANCE FUNCTION APPROACH

3.1 INTRODUCTION

Poverty is a multi-dimensional phenomenon with several dimensions. Many dimensions are divided into several attributes. An example of a dimension of poverty is access to basic services. This dimension can comprise of the following attributes: access to water, toilet facilities, refuse removal, energy source for heating, lighting and cooking. Another dimension could be housing with the attributes: number of rooms, type of walls and roof, condition of dwelling, etc.

This chapter discusses the distance function techniques to combine attributes or dimensions of poverty of households using the Euclidean distance measure and the K-Means clustering technique.

The distance function is a concept widely used in Efficiency Analysis. It has however only rarely been applied to the analysis of household behaviour. Lovell *et al.* (1994) were the first to make such an attempt by taking a different approach to welfare measurement by employing distance functions. Deutsch and Silber (2005) employed these techniques in multivariate poverty analysis and their approach is applied in this section.

Considering the concept of distance functions in the literature, a distinction has been made between input and output distance functions. In this study the discussion is limited to input distance functions.

The distance function technique is borrowed from the production theory literature where it is used to measure efficiency. Consider a measure of the “distance” between a vector of the goods (functioning and capability) of a household and a comparison or yardstick vector. The distance function approach seeks to measure the amount by which the household’s set of attributes must be scaled up or down so that it has the same well-

being as the yardstick. This tool is called a *distance function* in economics literature or a *gauge function* in mathematics literature.

In mathematical notation the distance function is defined as follows:

$$D(x_i, W) \equiv \min d\{d : W(dx_i) = W^*, d > 0\} \quad (3.1)$$

where

x_i is a vector listing a number of features of the i^{th} household's circumstances,

W is the chosen weighting function,

W^* is the value of the weighting function for the yardstick, and

d is the distance measure which shows the minimum amount by which a household's circumstances would have to be scaled up or down so that it would be on a par with the yardstick.

The distance measure will depend on x_i , W and W^* . If the objective is a measure of relative welfare then it makes sense to choose the yardstick to be the household with either the lowest or highest well-being and to enquire about scaling back, or scaling up of the attributes of each household so that they have the same level of well-being as the yardstick?

To make it operational, a measure of well-being is required, essentially an aggregator function of the various household characteristics that represents the household's welfare. This is the analogue of the classic utility function. Deutsch and Silber (2005) use the translog function which is estimated by normalizing on one of the characteristics.

Let x_{ij} be the membership function of household i , ($i = 1, 2, \dots, n$), and attribute j , ($j = 1, 2, \dots, m$). Group the membership function for m attributes, (q_1, q_2, \dots, q_m), in columns and the membership function for n households, (p_1, p_2, \dots, p_n), in rows to obtain a data matrix X .

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} \quad (3.2)$$

Define the household with zero deprivation as follows:

$$BH = (\max_i x_{i1}, \max_i x_{i2}, \dots, \max_i x_{im}) \quad (3.3)$$

Define the household with the maximum deprivation as follows:

$$WH = (\min_i x_{i1}, \min_i x_{i2}, \dots, \min_i x_{im}) \quad (3.4)$$

The following three objective weights can be defined:

- Mean Weight Method,
- Entropy Weight Method, and
- Critic Method.

3.1.1 The Mean Weight Method

The Mean weight method assigns equal weight to each criterion. A neutral attitude is reflected and the objectivity of the performance evaluation process is guaranteed.

The Mean weight can be defined as follows:

$$MW_j = \frac{1}{m} \quad j = 1, 2, \dots, m \quad (3.5)$$

3.1.2 The Entropy Weight Method

Entropy is a measure of uncertainty in information and reflects the relative importance of its corresponding criterion in terms of the amount of the information it contains and it indicates the inherent contrast intensity of the corresponding criteria (Shannon and Weaver 1947).

The Entropy weight method is defined as follows:

$$EW_j = \frac{d_j}{\sum_{k=1}^m d_k} \quad j = 1, 2, \dots, m \quad (3.6)$$

where

$$d_j = - \sum_{i=1}^n (p_{ij}) \log_2 (p_{ij}) \quad \text{for } i = 1, 2, \dots, m,$$

$$p_{ij} = \frac{x_{ij}}{v_i}, \text{ and}$$

$$v_i = \sum_{j=1}^m x_{ij}.$$

3.1.3 The Critic Method

The Critic method was proposed by Diakoulaki *et al.* (1995), with the aim of determining the objective weights that incorporate the contrast intensity and conflict.

The Critic method is defined as follows:

$$CW_j = \frac{c_j}{\sum_{k=1}^m c_k} \quad j = 1, 2, \dots, m \quad (3.7)$$

where

$$c_j = s_j \sum_{k=1}^m (1 - r_{jk}),$$

s_j is the standard deviation of the sample proportion, and

r_{jk} is the linear correlation coefficient between vectors x_j and x_k .

The Minkowski metric weighted distances from the household with zero deprivation is defined as follows:

$$WD_{BH} = \left[\frac{\sum_{j=1}^m (|x_{ij} - \max x_{ij}|^\lambda w_j)}{\sum_{i=1}^n |x_{ij}|^\lambda} \right]^{\frac{1}{\lambda}}, \quad i = 1, 2, \dots, n. \quad (3.8)$$

where

w_j is the weighted coefficient, and

λ is the Minkowski factor for the norm.

The Minkowski metric weighted distances from the household with maximum deprivation is defined as follows:

$$WD_{WH} = \left[\frac{\sum_{j=1}^m (|x_{ij} - \min x_{ij}|^\lambda w_j)}{\sum_{i=1}^n |x_{ij}|^\lambda} \right]^{\frac{1}{\lambda}}, \quad i = 1, 2, \dots, n. \quad (3.9)$$

where

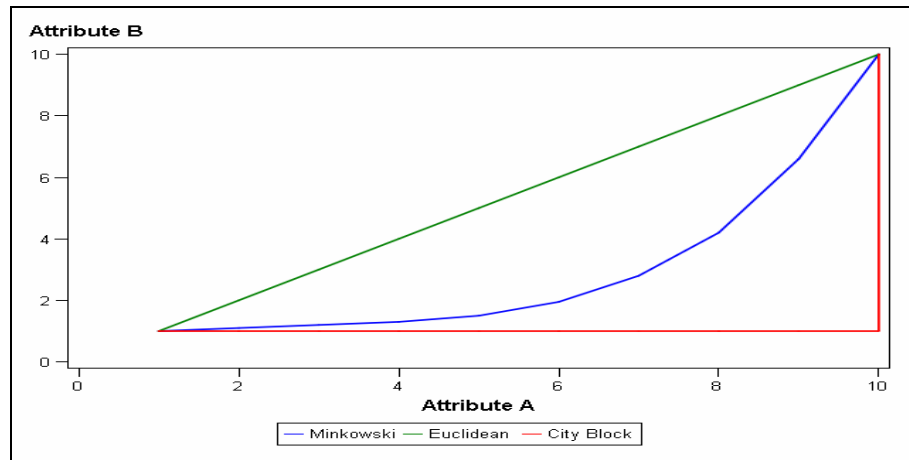
w_j is the weighted coefficient, and

λ is the Minkowski factor for the norm.

If $\lambda=1$, then the Minkowski distance is equal to the city block distance. If $\lambda=2$, then the Minkowski distance is equal to the Euclidean distance.

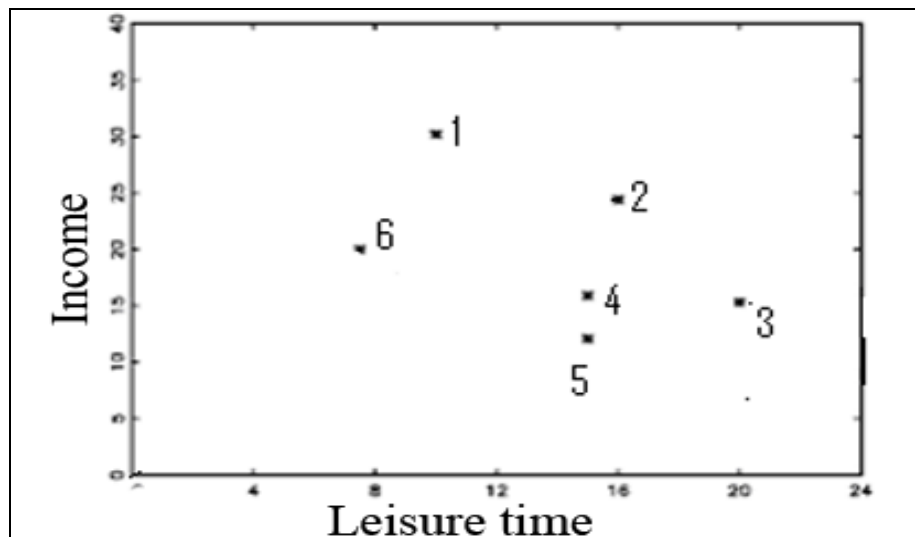
If $\lambda = \infty$ then the Minkowski distance is equal to the Tchebycheff distance. Figure 3.1.1 illustrates the Minkowski distance curves with different λ . A value for λ between the city block distance and the Euclidean distance is taken as $\lambda = 1.5$.

Figure 3.1.1: Distance curves for minkowski curves with different λ



Consider the following example in which a sample of 6 households are represented by 2 attributes, leisure time (X) and income (Y). Figure 3.1.2 shows the scatter plot of each household's attributes.

Figure 3.1.2: Scatter plot for attributes income and leisure time

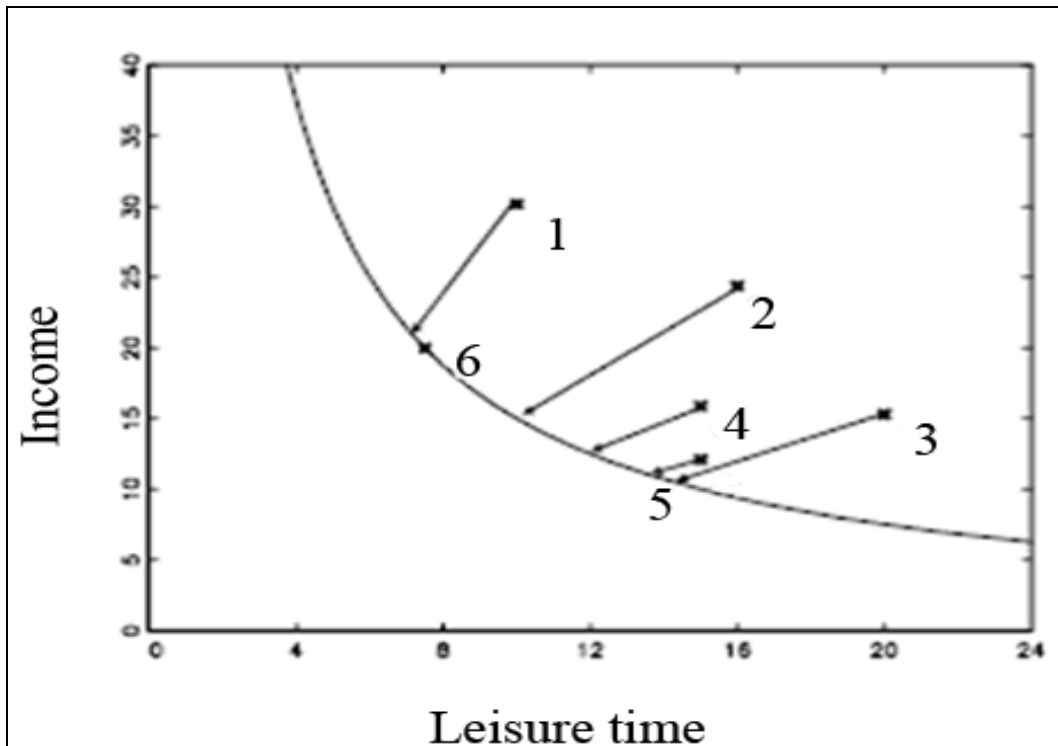


The attribute leisure time is plotted on the x-axis in hours; the attribute income is plotted on the y-axis in thousands of rands.

Let the aggregate measure of well-being be the geometric mean, $X^{0.25} Y^{0.25}$, then household 6 becomes the worst off household and the best off household is household 2.

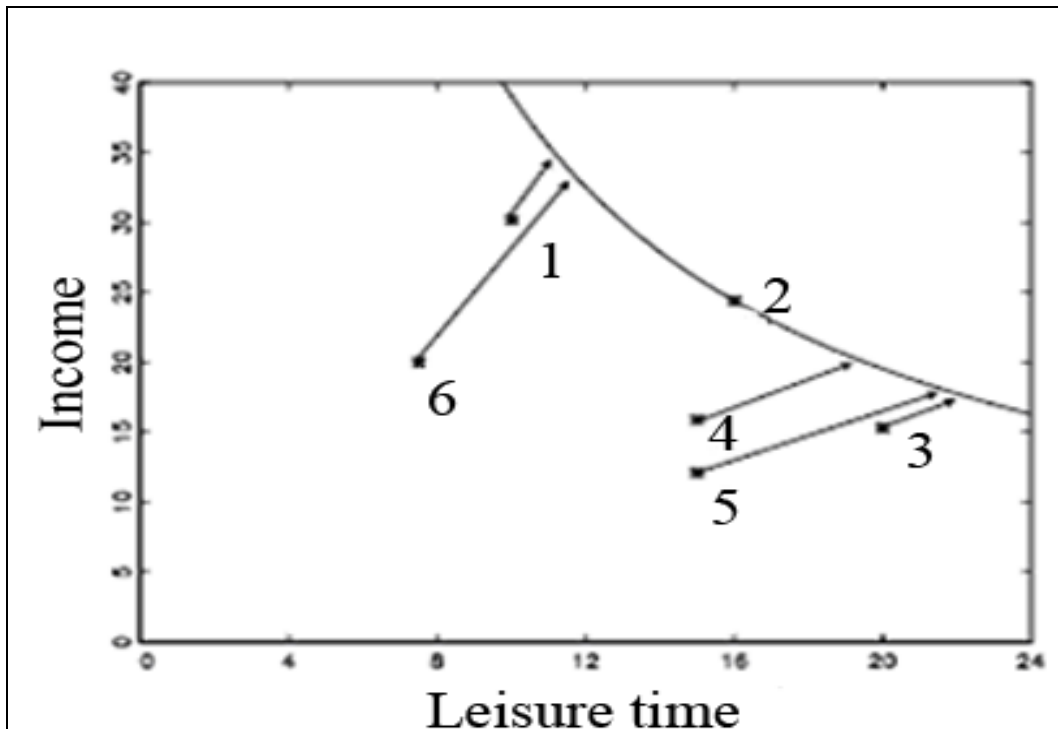
In figure 3.1.3 the aggregate measure passes through the point for household 6 and shows all of the combinations of the measured attributes which give exactly these levels of aggregate well-being. The distance measures of relative well-being are given by the length of the arrow which connect each of the rest of the households to the reference welfare value curve.

Figure 3.1.3: Scatter plot for attributes: worst aggregation curve



In figure 3.1.4 the aggregate measure passes through the point for household 2 and shows all of the combinations of the measured attributes which give exactly these levels of aggregate well-being. The distance measures of relative well-being are given by the length of the arrow which connects each of the rest of the households to the reference welfare value curve.

Figure 3.1.4: Scatter plot for attributes: aggregation curve



In table 3.1.1 the distance measures in the low reference column are those from figure 3.1.3, where the worst off household is the reference household. Household 6 is the worst off, so their circumstances need only be multiplied by 1 (that is, remain unchanged) for them to remain the worst off. Household 2 is the best off, their circumstances need to be scaled back by the most (multiplied by 0.62) to reduce them to the same welfare value as household 6.



Table 3.1.1: Distance measure for best and worst case aggregation curves

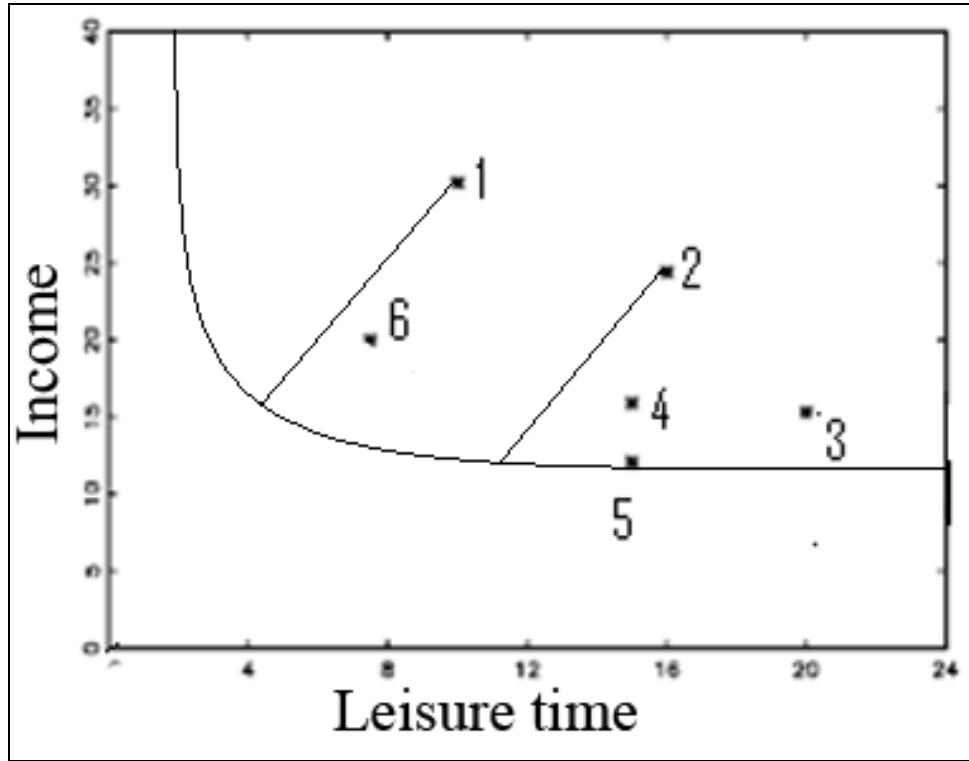
Household	D(x _i ,W)	
	Low Reference	High Reference
1	0.70	1.13
2	0.62	1.00
3	0.70	1.14
4	0.91	1.47
5	0.79	1.28
6	1.00	1.61

The distance measures in the high reference column are those from figure 3.1.4 which use the best off household as the reference. Household 6 is the worst off household and has to be scaled up by 61% in order to reach the reference level. Since the two columns are based on the same welfare measure they agree on the ranking of the households.

This approach is very easy to implement once an aggregating function is chosen. In this demonstration the aggregate curve, $X^{0.25}Y^{0.25}$, was chosen. What would have happen if another aggregate curve, $X^{0.75}Y^{0.25}$, had been chosen? Household 1 would have been the household with the highest standard of living and household 5 is the worst off household as shown in figure 3.1.5.

The distances and ranking of the other households will be altered. The results depend upon data on household circumstances and the weighting formula. The difficulty lies in the dependence of the answers upon the weighting formula. In standard models of consumer behaviour the weighting function is essentially the household's utility function rearranged in terms of income as a function of leisure for a given level of welfare.

Figure 3.1.5: Scatter plot for attributes: New aggregation curve



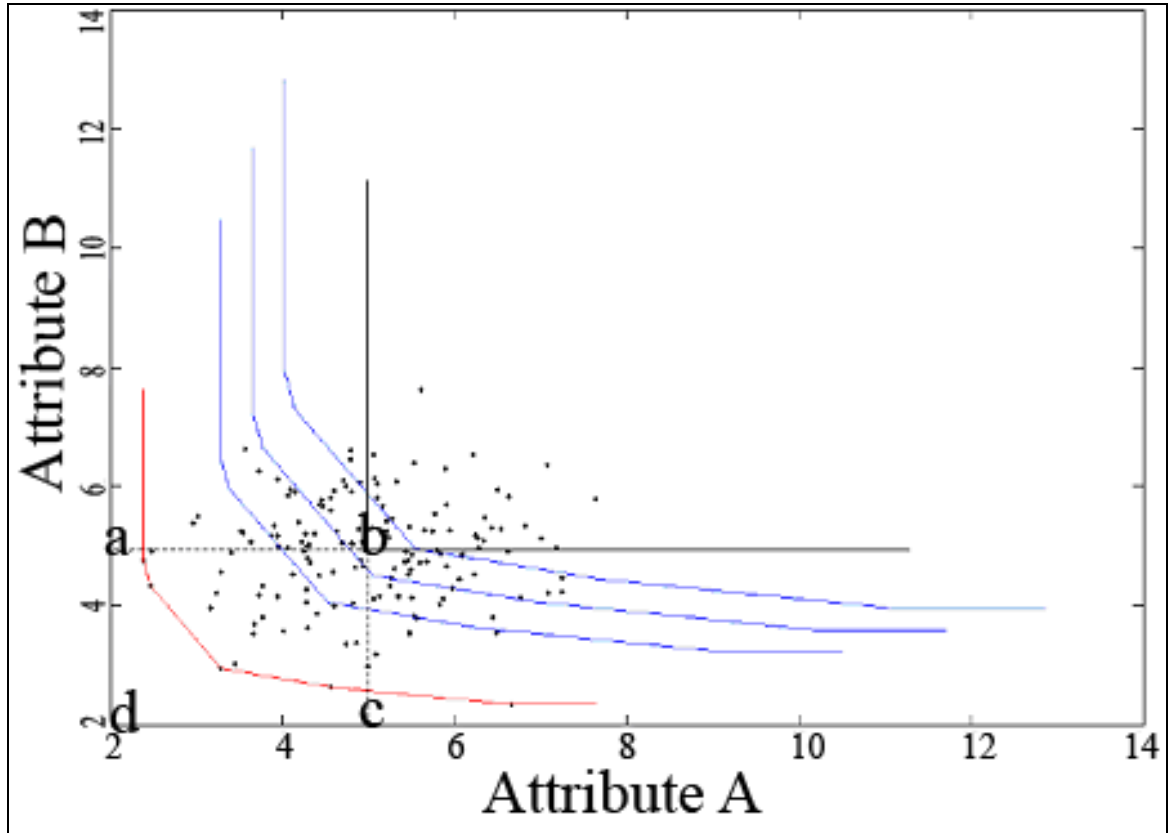
Anderson *et al.* (2005) avoids the need to choose aggregation functions and removes the dependence of the final index on the choice of aggregation functions by calculating a lower bound on the distance measure of relative well-being. The shared properties of the distance function are monotonicity and quasi-concavity. Monotonicity means that the measured attributes are such that it is reasonable to expect that if the household had more of any of them, then their well-being would not decrease. Quasi-concavity means that as the level of some measured attribute rises, well-being rises at a non-increasing rate which is closely related to inequality version.

The distance measure is defined as follows:

$$D(x_i) \equiv \min d\{d : W(dx_i) = W^*, d > 0\} \quad (3.10)$$

for all monotone, quasi-concave W .

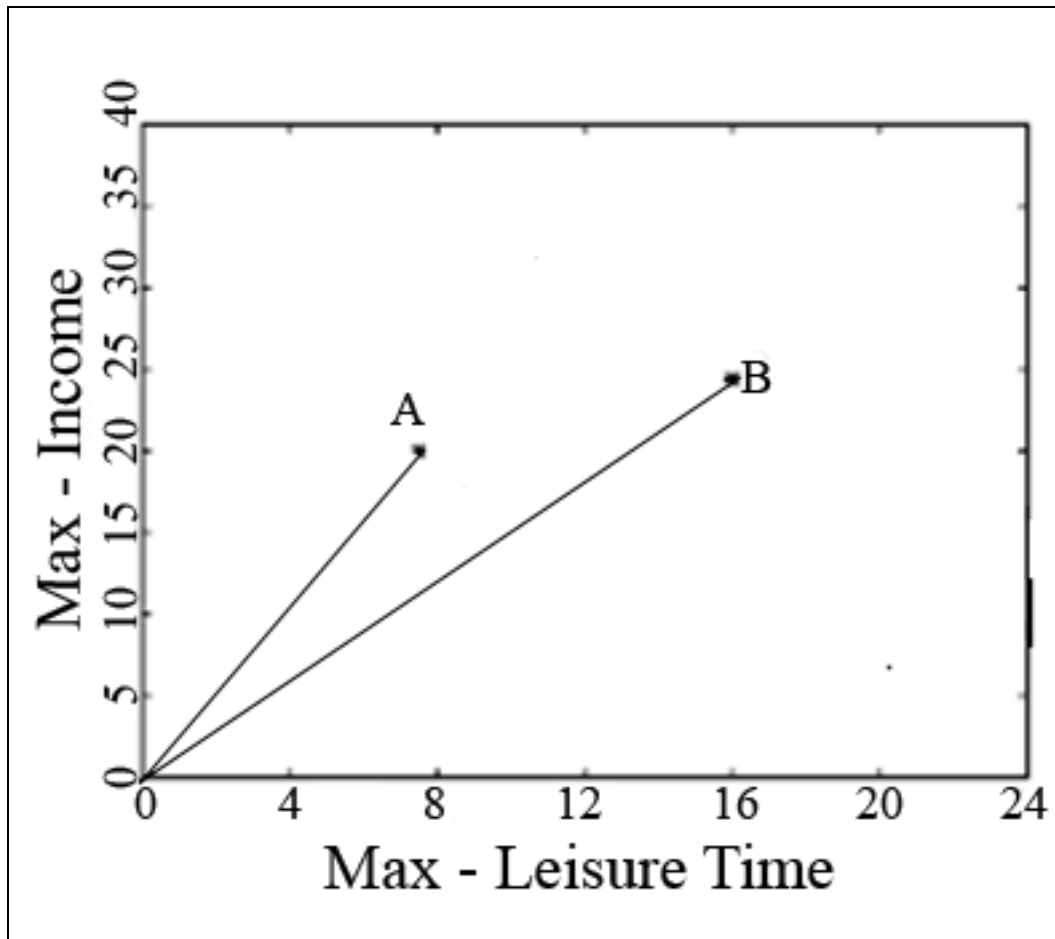
Figure 3.1.6: Welfare curves for attribute A and attribute B



The basic intuition is that welfare level sets, as shown in figure 3.1.6, of any aggregator with these properties are convex to the origin. A simple way of calculating bounds on the set of all possible curves in a finite dataset is proposed.

The resulting distance measures reflect the minimum amount by which one would have to scale each household so that they shared equal ranking with the best and worst off household. They represent lower bounds on these measures for any way of choosing to weigh the various indicators as long as the weighting formula is monotone and quasi-concave. In figure 3.1.6 the welfare curves are convex to the origin, and the horizontal and vertical lines in the graph denote the median cut off points for the two attributes which define the intersection and union sets of poverty measurement. The intersection set of poverty is the square a b c d.

Figure 3.1.7: Euclidean distance measure of poverty



In this study it is proposed that the origin denotes zero poverty and the distance from the origin to the point on the scatter plot of the household can be considered a distance measure for poverty for the household as shown in figure 3.1.7. In the best situation, this distance measure should be zero, denoting no poverty or deprivation.

The distance measure can be used to compare the relative poverty between two households. To use this approach the values of the X-axis and Y-axis need to be changed. For the best case to be 0 on the X-axis, the household leisure time is subtracted from the max value. Similarly for the Y-axis, each household income is subtracted from the maximum value.

The household with the maximum income and maximum leisure time will be at the origin (0, 0). In figure 3.1.7 the distance measure from household A to the origin is shorter than the distance measure from household B to the origin thus implying that household B experiences more poverty than household A.

The fuzzy membership function allows categorical variables to be assigned a value between zero and one, therefore it can be treated as interval variables and a distance measure can be calculated for any household. The distance measure is calculated using the Euclidean distance and is discussed in the next section.

3.2 THE EUCLIDEAN DISTANCE MEASURE

3.2.1 Methodology

The fuzzy membership function that is applied to the attributes of poverty allows the Euclidean distance measure to be used to measure poverty within a single dimension consisting of several attributes.

The Euclidean distance measure will be explained using two attributes. The same explanation will apply to three attributes and similarly will apply to any number of attributes. The two attributes used in the explanation are “access to water” and “energy source for cooking”. The membership function is calculated according to the method proposed by Cheli and Lemmi (1995).

Table 3.2.1 shows the cross tabulation between the membership functions of the two attributes access to water and energy source for cooking, for the 905 748 households from the Republic of South Africa Census 2001. The value zero represents no deprivation in that attribute while the value one represents maximum deprivation in that attribute.

Table 3.2.1: Membership function frequencies for attributes: Water and Toilet

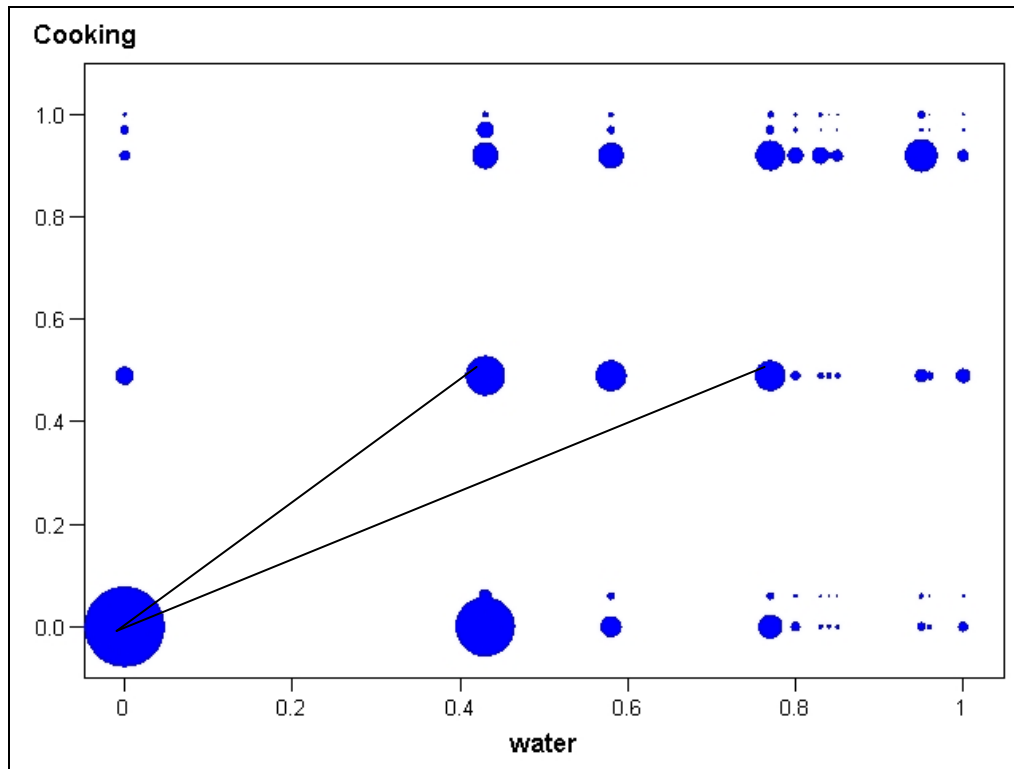
Water	Cooking						Total
	0	0.06	0.49	0.92	0.97	1	
0	263 005	5 993	14 487	4 713	3 384	846	292 428
0.43	144 435	7 692	65 937	30 478	12 680	1 701	262 923
0.58	19 324	2 599	40 418	28 579	2 987	1 514	95 421
0.77	24 980	2 792	39 734	38 627	3 324	1 942	111 399
0.8	4 151	780	4 169	11 333	1 054	736	22 223
0.83	1 045	342	2 284	12 727	227	906	17 531
0.84	951	452	1 646	2 174	100	165	5 488
0.85	987	218	1 388	6 039	297	284	9 213
0.95	3 410	1 273	7 013	45 624	797	2 655	60 772
0.96	1 323	162	2 930	2 092	145	135	6 787
1	4 536	617	9 144	6 182	715	369	21 563
Total	468 147	22 920	189 150	188 568	25 710	11 253	905 748

From table 3.2.1 it can be seen that there are 263 005 households that experience zero deprivation in both attributes and 369 households have no access to water and no energy for cooking. In between the worst case household and the best case household there are 64 different combinations of “access to water” and “energy source for heating”.

From the information in table 3.2.1 a scatter plot diagram (bubble plot) was drawn and the results are shown in figure 3.2.1. The ideal position for each household is to reach zero deprivation for each attribute. The points shown in the scatter point represent individual households. The household experiencing zero poverty or deprivation in each of the two attributes will be plotted on the origin (0, 0). The measure of the distance away from the origin for each household can be viewed as the measure of deprivation experienced by each household. This is only a relative measure to compare one household to another.

The Euclidean distance measure can be used to rank the households from the worst deprived to the least deprived. There are 66 points in figure 3.2.1 and 66 different distance functions can be calculated.

Figure 3.2.1: Bubble plot of membership function for attributes water and cooking



The general Euclidean distance formula can be reduced to the following equation for measuring relative deprivation because the Euclidean distance measure is from the household point back to the origin.

The distance measure d_i can be defined as follows:

$$d_i = \sqrt{u_{1i}^2 + u_{2i}^2} \tag{3.10}$$

where

u_1 is the membership function for the first attribute,

u_2 is the membership function for the second attribute.

3.2.2 Analysis

In table 3.2.2 the Euclidean distance measure for each household is calculated from a point plotted in a 6 dimensional space back to the origin. There are 222 577 households that have a Euclidean distance of zero and do not experience any deprivation in the six attributes, access to water, toilet facilities, energy source for heaters, energy source for lighting, energy source for cooking, and refuse removal. The membership function allows each household to be plotted on one of 94 325 points on a 6 dimensional space.

In table 3.2.2 the Euclidean distances measures are grouped into 19 categories. If a value is equal to the class limit then it is included with the upper class limit.

Table 3.2.2: Euclidean distance measures

Euclidean distance	Households
0	222 577
0.0-0.1	15 798
0.1-0.4	4 795
0.4-0.5	94 627
0.5-0.6	13 345
0.6-0.7	6 795
0.7-0.8	38 587
0.8-0.9	12 702
0.9-1.0	26 419
1.0-1.1	25 340
1.1-1.2	38 341
1.2-1.3	33 547
1.3-1.4	27 674
1.4-1.5	39 650
1.5-1.6	33 487
1.6-1.7	43 876
1.7-1.8	42 805
1.8+	185 382

Figure 3.2.2: Bar chart of frequency: Euclidean distance measures

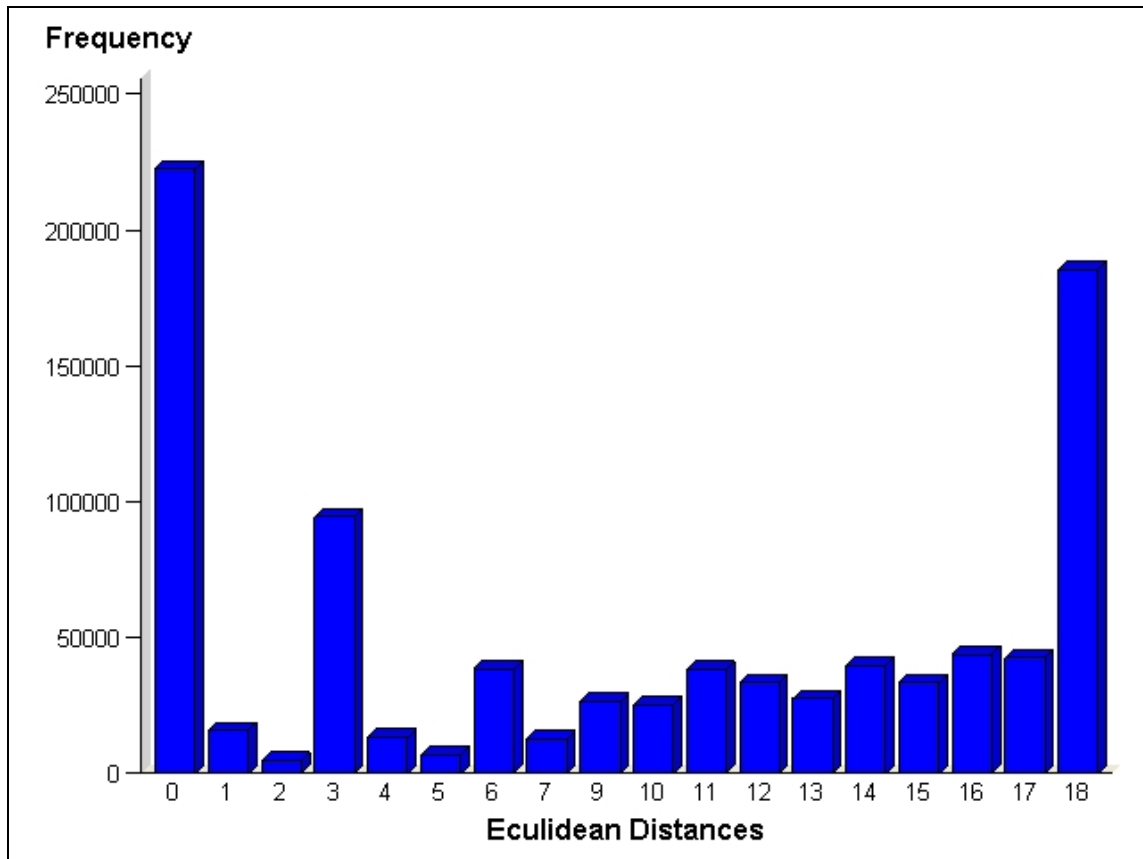


Figure 3.2.2 is a bar chart of the Euclidean distance measures and it clearly demonstrates the multidimensional measure of poverty. On the X-axis are the categories from table 3.2.2. There are 222 577 households that experience zero deprivation in basic services, while 185 382 households experience severe deprivation.

There are 17 categories in between clearly showing the different shades and grades of deprivation. This method can be used to measure the effectiveness of a poverty alleviation program for a particular city or town. The ideal situation is to get all the households into the zero category or as close to zero as possible. This measure can be calculated before a poverty alleviation program starts and then measured again to determine the effectiveness of the poverty relief measures.

3.3 K-MEANS CLUSTERING

Cluster analysis is the most widely known descriptive data mining method. Clustering is a very common approach used in a wide array of problems. The aim is to partition a data set into a set of clusters. In the poverty data analysis the matrix of n households (rows) and m attributes (columns) is clustered into groups that are internally homogeneous and heterogeneous from group to group.

Clustering is a general term that embraces various approaches, such as crisp clustering, fuzzy clustering, and mixture model-based clustering. In this analysis, the focus is only on K-Means cluster analysis. Although the general course of clustering is to maximize within-cluster similarity and/or between-cluster dissimilarity, various proximity measures (Euclidean, city-block, and Mahalanobis distances) and various distance criteria (within-cluster: average, nearest neighbor, and centroid distances; between-cluster: single, complete, average, and centroid linkages) exist, causing clustering results of the same data set to vary from one analysis to another.

The purpose of cluster analysis is to place objects into groups or clusters suggested by the data, not defined a priori. The objects in a given cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar. Cluster analysis can also be used for summarizing data rather than for finding "natural" or "real" clusters; this use of clustering is sometimes called *dissection* (Everitt 1980).

Clustering analysis has the advantage of being intuitively simple and easily communicated. It can be used to detect similarity and/or abnormality in environmental conditions. It makes no assumptions about the statistical distribution of the indicators. However, Clustering analysis may be influenced by the covariance structure of the data set, especially when the Euclidean distance is used.

3.3.1 Methodology

Let x_{ij} be the membership function of household i , ($i = 1, 2, \dots, n$), for attribute j , ($j = 1, 2, \dots, m$). Group the membership function for m attributes q_1, q_2, \dots, q_m in columns and the membership function for n households p_1, p_2, \dots, p_n in rows to obtain a data matrix X .

$$X = \begin{pmatrix} X_{11} & X_{12} \dots & X_{1m} \\ X_{21} & X_{22} \dots & X_{2m} \\ \dots & \dots & \dots \\ X_{n1} & X_{n2} \dots & X_{nm} \end{pmatrix} \quad (3.11)$$

If there are two attributes, attribute X and attribute Y , with membership functions (x_1, y_1) and (x_2, y_2) then the bivariate Euclidean distance between the two households is define as follows:

$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (3.12)$$

If there are three attributes, attribute A , attribute B and attribute C . Suppose there are two households (x_1, y_1, z_1) and $x_2, y_2, z_2)$, then the Minkowski distance between the two households is defined as follows:

$$d_M = \sqrt[m]{|x_1 - x_2|^m + |y_1 - y_2|^m + |z_1 - z_2|^m} \quad (3.13)$$

where

m can be any positive integer, $(1, 2, 3, \dots)$.

When $m=2$ the Minkowski distance is the Euclidean distance and when $m=1$ the Minkowski distance is the city block distance.

There are two main ways to cluster data: partitive and hierarchical approaches. K-Means cluster analysis is a typical partitive clustering technique in which the data set is divided directly into a predefined number of clusters. This method implicitly assumes spherical shapes of the clusters. The main techniques of the non-hierarchical K-Means method are explained.

The basic idea of K-Means clustering is to introduce seeds, or centroids, around which units may be attracted, forming a cluster. The maximum number of clusters, G , can be determined in advance.

Non-hierarchical methods are fast, but they require the number of clusters to be chosen in advance. To avoid these disadvantages and to exploit the potential of both the methods, one can adopt two possible approaches. A sample of limited size is extracted from the data, and a hierarchical cluster analysis is carried out to determine G , the optimal number of clusters. Once a value for G is determined then the G means of the clusters are used as seeds in a non-hierarchical analysis of the whole data set using the number of clusters equal to G and allocating each observation to one the clusters.

Alternatively a non-hierarchical analysis can be carried out on the whole data set with a large value of G and then to consider a new data set, made up of the G group means, each endowed with two measurements, one indicating the cluster size and one the dispersion within the cluster. An hierarchical analysis is then carried out on this data set to see whether any groups can be merged. It is essential to indicate the frequency and the dispersion of each cluster. Otherwise the analysis will not take account of clusters having different numbers and variables.

The clustering node of SAS Enterprise Miner implements a mixture of both approaches in a three-stage procedure. Initially a non-hierarchical clustering procedure is run on all available observations. Then an interactive procedure is run; at each step of the procedure, temporary clusters are formed, allocating each observation to the cluster with

the seed nearest to it. Each time an observation is allocated to a cluster, the seed is substituted with the mean of the cluster, called the centroid. The process is repeated until convergence is achieved, namely, until there are no substantial changes in the cluster seeds. At the end of the procedure, a total of G clusters is available, with corresponding cluster centroids.

In the second stage a hierarchical clustering method is run on a sample of the data to find the optimal number of clusters. As the number of clusters cannot be greater than G , the procedure is agglomerative, starting at G and working downwards. The previous cluster means are used as seeds, and a non-hierarchical procedure is run to allocate the observations to the clusters. A peculiar aspect of this stage is that the optimal number of clusters is chosen with respect to a test statistic, a function of the R^2 index known as the cubic clustering criterion (CCC).

A Gaussian distribution for the observations to be clustered cannot always be assumed. To derive a statistical test, certain assumptions need to be made. Suppose that the significance of a number of clusters equal to G needs to be verified, then the general assumption is to assume that, under the null hypotheses, H_0 , the observations are distributed uniformly over a hypercube with dimension equal to the number of variables each cube representing a cluster, adjacent to the others. Under the alternative hypothesis, H_1 , clusters are distributed as a mixture of multivariate Gaussian distributions, centered at the mean of each cluster, and with equal variances.

The cubic clustering criterion is a function of the ratio between the observed R^2 and the expected R^2 under the null hypothesis. From empirical Monte Carlo studies, it turns out that a value of the cubic clustering criterion greater than 2 represents sufficient evidence against the null hypothesis and, therefore, for the validity of the chosen G clusters. Although it is approximate, the criterion tends to be conservative and it may have a bias towards a low number of clusters.

Once the optimal number of clusters has been chosen, the algorithm proceeds with non-hierarchical clustering to allocate the observations into the G chosen groups, whose initial seeds are the centroids obtained in the previous step. In this way, a final configuration of the observations is obtained.

The clustering algorithm repeats the following two steps until convergence:

- (1) Scan the data and assign each observation to the nearest seed (nearest using the Euclidean distance),
- (2) Replace each seed with the mean of the observations assigned to its cluster.

The distance function is the Euclidean distance, and Ward's method is used to recompute the distances as the clusters are formed.

The clustering methods that are discussed in this section are:

- Average Method,
- Centroid Method
- Ward Method.

In the Average method the distance between two clusters is the average distance between pairs of observations, one in each cluster. The average method tends to join clusters with small variances and is slightly biased towards producing clusters with the same variance.

The distance measure between the two clusters, C_K and C_L , is defined as follows:

$$D_{KL} = \frac{1}{N_K N_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j) \quad (3.14)$$

If

$$d(x,y) = |x-y|^2 ,$$

then the distance measure can be defined as follows:

$$D_{KL} = \|\bar{X}_K - \bar{X}_L\|^2 + \frac{W_K}{N_K} + \frac{W_L}{N_L} \quad (3.15)$$

Where

$$W_K = \sum_{i \in C_K} \|X_i - \bar{X}_K\|^2 ,$$

$$W_L = \sum_{i \in C_L} \|X_i - \bar{X}_L\|^2 ,$$

C_K is the K^{th} cluster subset (1, 2, ..., n),

N_K is the number of observations in C_K , and

\bar{X}_K is the mean vector for cluster C_K .

The combinatorial formula is

$$D_{JM} = \frac{N_K D_{JK} + N_L D_{JL}}{N_M} \quad (3.16)$$

In the Centroid cluster method the distance between two clusters is defined as the squared Euclidean distance between their centroids or means. The centroid method is more robust to outliers than most other methods but in other respects may not perform as well as the Ward's method or the average method.

The distance between the two clusters is defined as follows:

$$D_{KL} = \|\bar{X}_K - \bar{X}_L\|^2 \quad (3.17)$$

If the distance measure between observations x and y is

$$d(x,y) = |x-y|^2$$

then the combinatorial formula is

$$D_{JM} = \frac{(N_K D_{JK} + N_L D_{JL})}{N_M} - \frac{N_K N_L D_{KL}}{N_M^2} \quad (3.18)$$

In the Ward clustering method the distance between two clusters is the ANOVA sum of squares between the two clusters summed over all the variables. At each generation, the within cluster sum of squares is minimized over all partitions obtainable by merging two clusters from previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give proportions of variances. Wards method tends to join clusters with a small number of observations and it is strongly biased towards producing clusters with roughly the same number of observations. Ward's method joins clusters to maximize the likelihood at each cluster with equal spherical covariance matrices and equal sampling probabilities.

The distance between two clusters is defined as follows:

$$D_{KL} = B_{KL} = \frac{\|\bar{X}_K - \bar{X}_L\|^2}{\frac{1}{N_K} + \frac{1}{N_L}} \quad (3.19)$$

If

$$d(x,y) = (1/2) |x-y|^2$$

then, the combinatorial formula is

$$D_{JM} = \frac{(N_K + N_J)D_{JK} + (N_K + N_J) - N_J D_{KL}}{N_J + N_M} \quad (3.20)$$

3.3.2 Analysis

In this section the cluster node of Enterprise Miner is applied to the 10% sample data set of the Republic of South Africa 2001 census. There are 905 748 households in the sample and 6 attributes were selected for the analysis. The analysis was conducted using SAS Enterprise Miner's Cluster node. The clustering technique is illustrated using the following six attributes to measure the dimension of poverty: access to basic services.

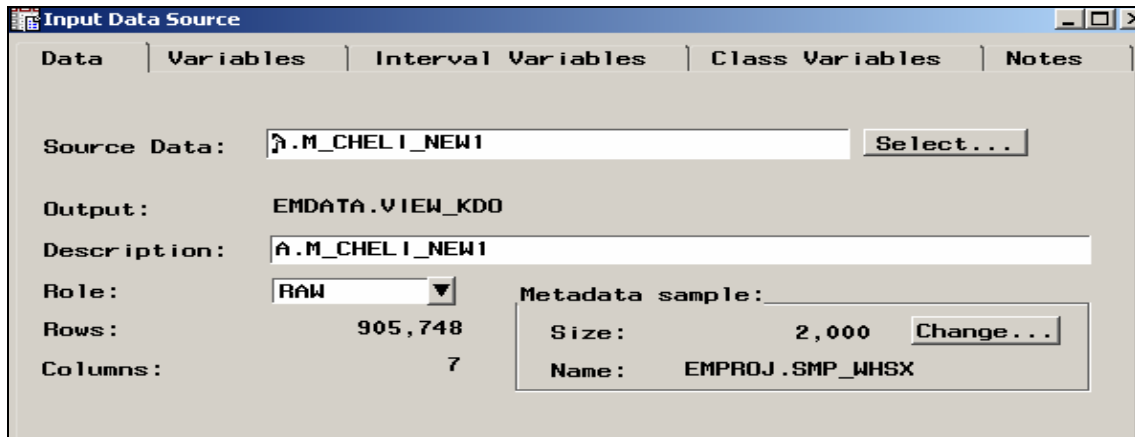
- access to water,
- toilet facility,
- energy source for cooking,
- energy source for heating,
- energy source for lighting, and
- refuse disposal.

In this section of the analysis two calculations are performed. In the first calculation the number of clusters is set to automatic and the clustering algorithm determines the number of clusters. In the second calculation the number of clusters are set to user specified, thus the number of clusters need to be determined a priori.

The automatic selection of the number of clusters works as a two step process. In the first step PROC DMVQ is run on the preliminary sample to create initial clusters, usually the maximum number of clusters as specified. In the second step PROC CLUSTER is run, using the means of the initial clusters as input. The smallest number of clusters that meet one of the following two criteria is selected. Firstly, the number of clusters must be greater than or equal to the minimum number of clusters specified in the selection criterion or alternatively, the cubic clustering criterion exceeds the set value.

The default value setting for the cubic clustering criterion is three.

Figure 3.3.1: Input data set: Data tab



In figure 3.3.1 the data set used in this calculation is shown to have 905 748 rows which represent the number of households and 7 columns which represent the six attributes and an identification variable called serial. The metadata sample is set at 2 000 and is used to identify categorical and interval variables. This data set is used for all the calculations in this chapter.

Figure 3.3.2: Input data set: Variables tab

Name	Model Role	Measurement	Type	F
SERIAL	id	ordinal	num	B
WATER	input	interval	num	B
REFUSE	input	interval	num	B
COOKING	input	interval	num	B
HEATING	input	interval	num	B
LIGHTING	input	interval	num	B
TOILET	input	interval	num	B

The names of the attributes are displayed under the variables tab in figure 3.3.2. The model role for the attributes is set to input, that is, they will be used in the clustering procedure. The model role for serial number of each household is set to id and will not

be used in the clustering process. For the column measurement all the attributes are set to interval, this allows the clustering algorithm to treat the attributes as continuous variables. Figure 3.3.2 also displays the SAS format and informat values.

Figure 3.3.3: Input data set: Interval variable tab

Data		Variables			Interval Variables			Cl
Name	Min	Max	Mean	Std Dev	Missing	Skewness	Kurtosis	
WATER	0	1	0.4144	0.3341	0%	0.0267	-1.273	
REFUSE	0	1	0.335	0.4134	0%	0.4723	-1.698	
COOKING	0	1	0.3213	0.3901	0%	0.6254	-1.301	
HEATING	0	1	0.3317	0.387	0%	0.5295	-1.496	
LIGHTING	0	1	0.2253	0.401	0%	1.3442	-0.118	
TOILET	0	1	0.3177	0.3819	0%	0.6138	-1.276	

In the data set the columns are the membership function for the attributes as proposed by Cheli and Lemmi (1995). As seen in figure 3.3.3 the membership function values for all attributes range from zero to one. A mean closer to zero indicates that many households do not suffer severe deprivation in that attribute. The standard deviation shows the spread of the membership values. The attributes “refuse removal” and “toilet facilities” have higher means and standard deviations than the other attributes, indicating that there are many households experiencing severe deprivation in these attributes.

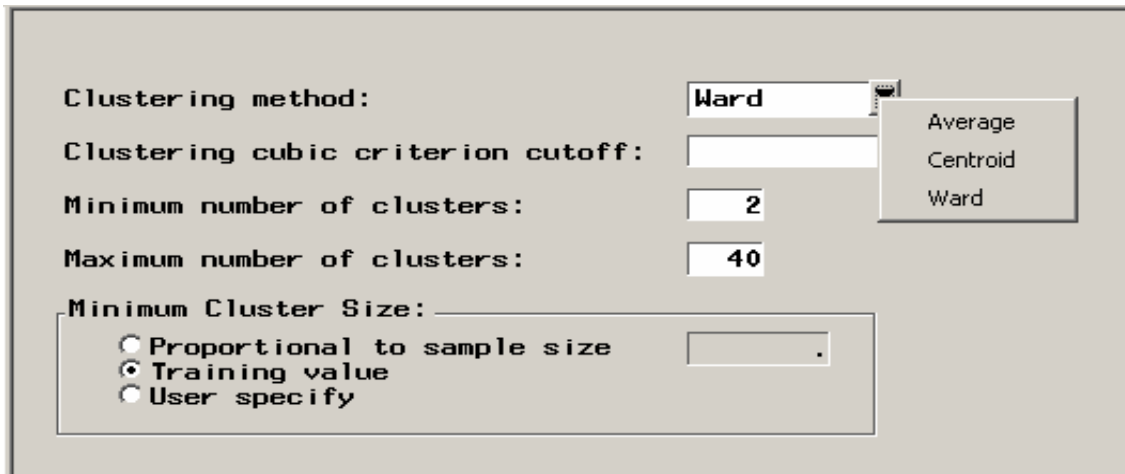
Figure 3.3.4: Cluster node: cluster tab

Data	Variables	Clusters	Seeds	Missing Values	Output	Notes
Segment Identifier:						
Variable name:		_SEGMNT_				
Variable label:		Cluster ID				
Role:		group ▼				
Number of Clusters:						
<input type="radio"/> User specify		3				
<input checked="" type="radio"/> Automatic		<input type="button" value="Selection Criterion..."/>				

Before the SAS Enterprise Miner clustering node can be run, certain options need to be selected. The first is the number of clusters, the second is the clustering criterion and the third is the clustering method. Many of the other settings are taken as default. In this first calculation the number of clusters is set to automatic as shown in figure 3.3.4.

Figure 3.3.5 shows the selection criteria tab of the seeds cluster in the cluster node. The clustering method must be selected. There are three different clustering methods, (Average, Centroid and Ward), that are available. For this calculation the Ward clustering method is selected. The maximum number of clusters is set to 40, the minimum number of clusters is set to 2 and the minimum cluster size is determined by the training value.

Figure 3.3.5 Cluster node: Cluster tab



Clustering method: Ward

Clustering cubic criterion cutoff:

Minimum number of clusters: 2

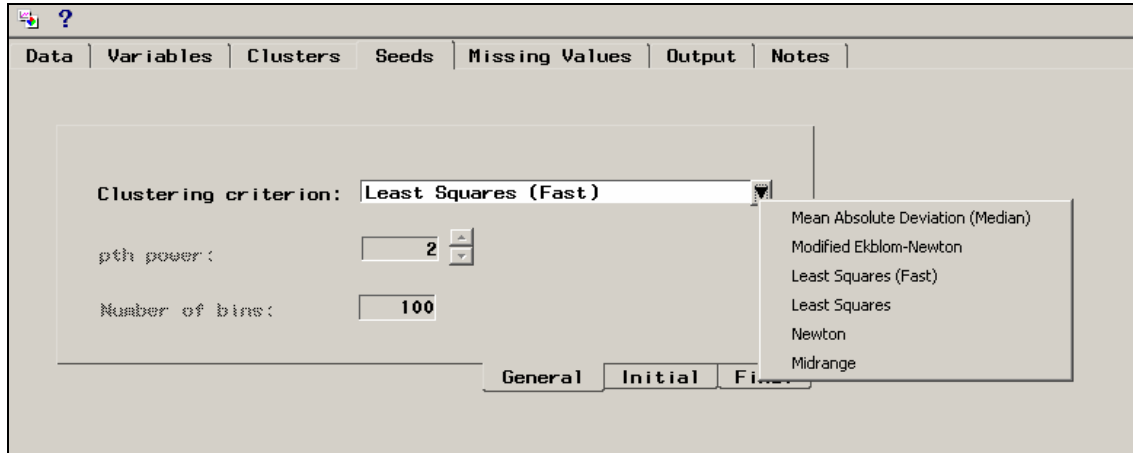
Maximum number of clusters: 40

Minimum Cluster Size:

- Proportional to sample size
- Training value
- User specify

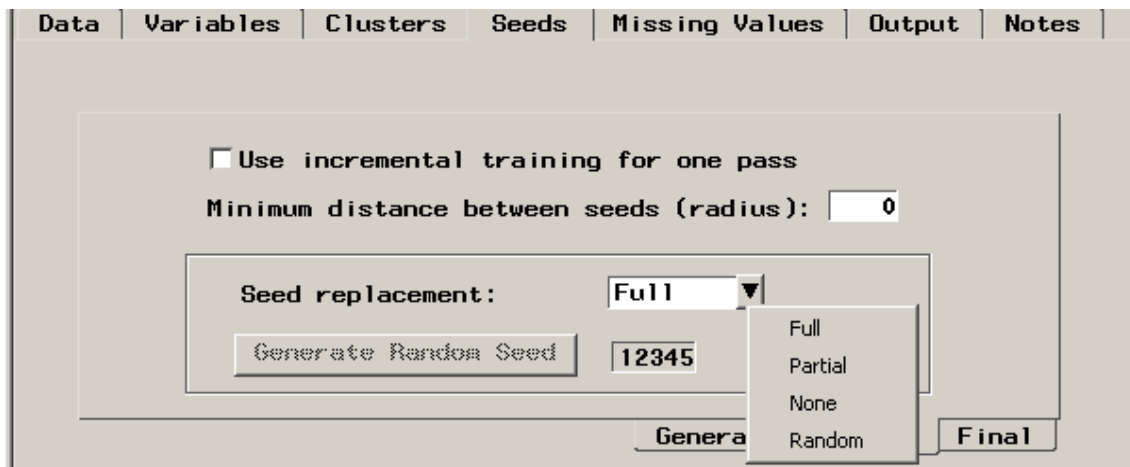
Under the seeds tab the clustering criterion needs to be selected. Figure 3.3.6 shows the different clustering criterion that can be used in the calculation. The mean absolute deviation requires the number of bins to be specified. (The default number is 100). The Modified Ekblom-Newton criteria require the p^{th} power to be specified. The p^{th} power can range between one and two with the default value of 1.5 and a maximum of 20 iterations.

Figure 3.3.6: Cluster node: Clustering criterion



The least squares criteria minimize the sum of squared distances between the data points and the cluster means by performing several iterations. The fast option in the least squares criteria limits the iterations to one. The midrange criterion minimizes the midrange distances between the data points and the cluster means. The least squares (fast) method was selected as the clustering criterion.

Figure 3.3.7: Cluster node: Seed replacement



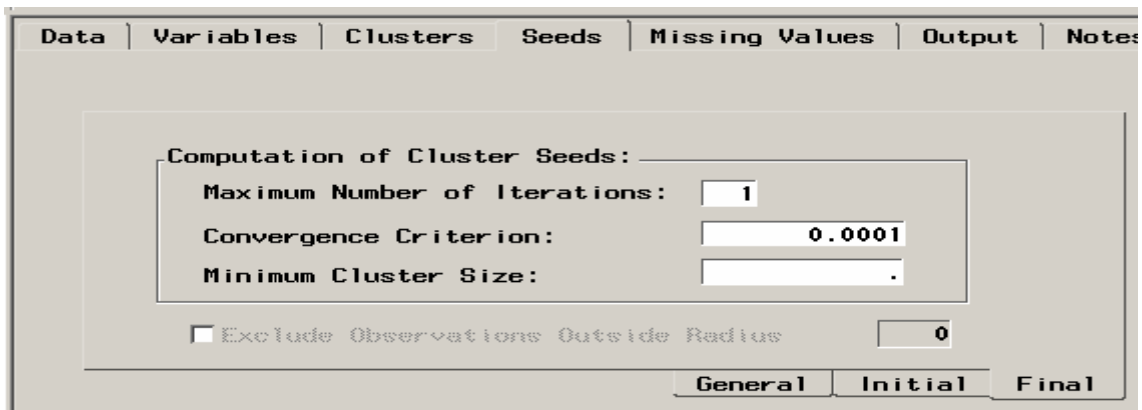
The initial sub tab of the seed tab in the cluster node as shown in figure 3.3.7 is used to specify how the cluster seed are initialized. If the incremental training for one pass is

selected then the seeds are allowed to drift as the algorithm selects initial seeds. The initial seeds must be complete cases, that is, no missing values in the training cases. The seeds are required to be separated by a Euclidean distance as specified by the minimum distance between seeds and are usually chosen as far apart as possible. To accomplish this, the seed replacement is set to full. If the seed replacement is selected as none then the initial seeds for the n clusters are the first n complete observations in the data set. While this option yields faster computation time, good clusters are not always obtained.

If partial is selected then only the seeds that do not meet the minimum distance requirement are replaced. In the random seed replacement the cluster seeds are randomly selected complete cases.

In this calculation the seed replacement is selected as Full with the minimum distance between seeds set as zero.

Figure 3.3.8: Cluster node: Computation of cluster seeds



In the final sub tab of the Seeds tab of the cluster node the stopping criteria for generating cluster seeds are stipulated as shown in figure 3.3.8. The maximum number of clustering iterations is set as 1 and the convergence criterion is set as 0.0001. No minimum cluster size is specified.

The SAS cluster node is run for the cluster analysis with the above mentioned settings and the following results are obtained:

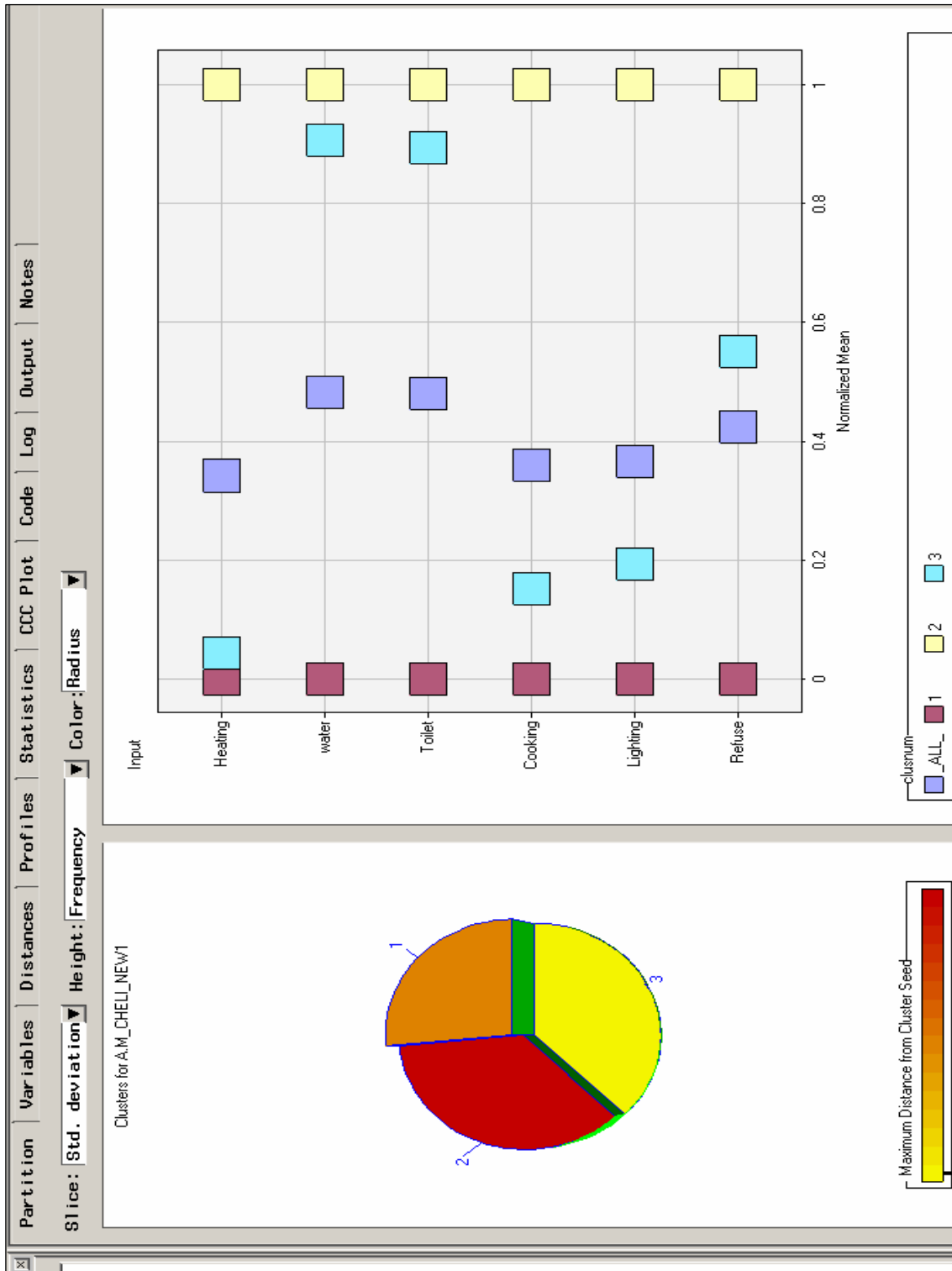
- The partition tab of the clustering results provides a graphical representation of key characteristics of the clusters from the training data.
- The variable tab lists all the input variables that are used in the clustering analysis.
- If there are more than three clusters the distance tab of the clustering results provides a graphical representation of the size of each cluster and the relationship among the clusters.
- The profiles tab displays a three dimensional bar chart of the interval input variables that were in the training sample data.
- The statistics tab displays a table of clustering statistics produced by PROC DMVQ.
- The cubic clustering criteria plot tab displays a graphic chart of the number of clusters against the training data set's cubic clustering criterion.
- The output tab displays the output obtained from running the SAS procedures.

Figure 3.3.9 shows the partition tab of the cluster results. On the left side of figure 3.3.9 is a three-dimensional pie chart with slice, colour and height with the following settings:

- Slice width is set to standard deviation, which is the root-mean-square standard deviation (root mean square distance) between cases in the cluster.
- Height is set to frequency.
- Colour is set to radius, which is the distance of the furthest cluster member from the cluster seed.

Each pie slice represents a cluster or segment. Each segment is labeled with a number, in this case from one to three. Cluster one has the highest frequency of 455 412 households and cluster three has the lowest frequency of 147 074 households.

Figure 3.3.9: Cluster Node: Cluster Tab Selection Criteria



A grid plot of the input means for the attributes that are used in the clustering analysis over all the cluster segments is displayed on the right hand side of the figure 3.3.9. The input means in the grid plot are normalized to fall within the range from 0 to 1. The normalized mean is the mean divided by the maximum value in the attributes.

The input means plots on the right of figure 3.3.9 display the input means for the variables that were used in the clustering analysis over all of the clusters. The input means are normalized using the following scale transformation function:

$$y = \frac{x - \min(x)}{\max(x) - \min(x)}$$

To explain the formula consider an example with five input variables

$$Y_i = Y_1, Y_2, \dots, Y_5$$

and three clusters

$$C_1, C_2, \text{ and } C_3.$$

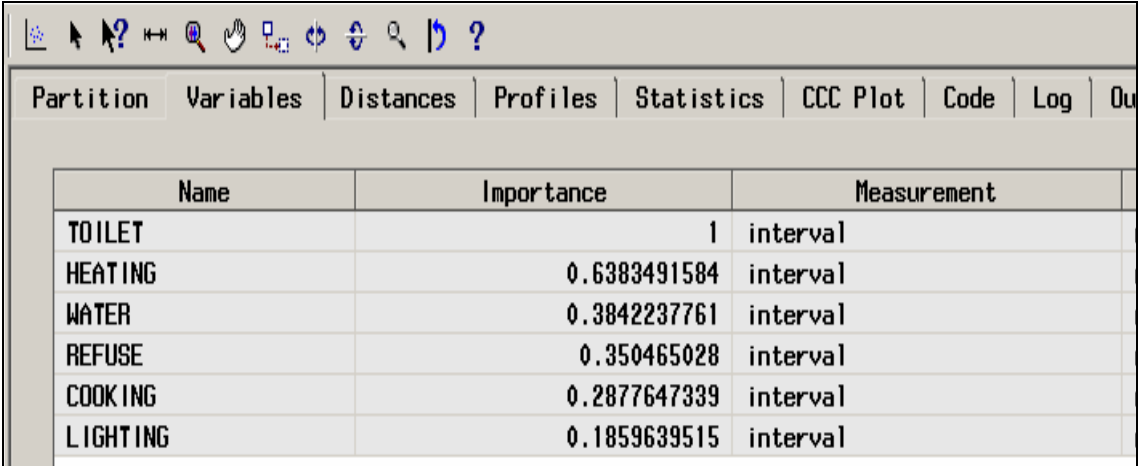
Let the input mean for variable Y_i in cluster C_j be represented by M_{ij} .

Then the normalized mean, or input mean, SM_{ij} is defined as follows:

$$SM_{ij} = \frac{M_{ij} - \min(M_{i1}, M_{i2}, M_{i3})}{\max(M_{i1}, M_{i2}, M_{i3}) - \min(M_{i1}, M_{i2}, M_{i3})} \quad (3.21)$$

The normalized means of the attributes as shown in figure 3.3.9 can only take on values between zero and one.

Figure 3.3.10: Cluster node results: Partition tab



Name	Importance	Measurement
TOILET	1	interval
HEATING	0.6383491584	interval
WATER	0.3842237761	interval
REFUSE	0.350465028	interval
COOKING	0.2877647339	interval
LIGHTING	0.1859639515	interval

The variable tab in the cluster results browser lists all the input variables that are used in the clustering analysis as shown in figure 3.3.10. For each input variable an importance value is calculated as a value between zero and one. If an input variable has an importance value of zero, this simply means that the input variable was not used as a splitting variable when the cluster analysis ran. It does not mean that this input variable should be dropped.

In figure 3.3.10 it can be seen that the attribute toilet has an importance value of one and none of the attributes have an importance of zero, that is, all the attributes were used in the cluster process.

In figure 3.3.11 the cubic clustering criterion is plotted on the Y-axis and the number of clusters plotted on the X-axis. In the cluster node the minimum number of clusters was set at 2 and the maximum number of clusters was set at 40 with the cubic clustering criterion cut-off value set at 3. In this analysis the cluster node automatically selected 3 as the number of clusters according to the cubic clustering criterion cut-off value. If cubic clustering criterion cut-off value is increased more clusters will be created.

Figure 3.3.11: Cluster node results: CCC plot

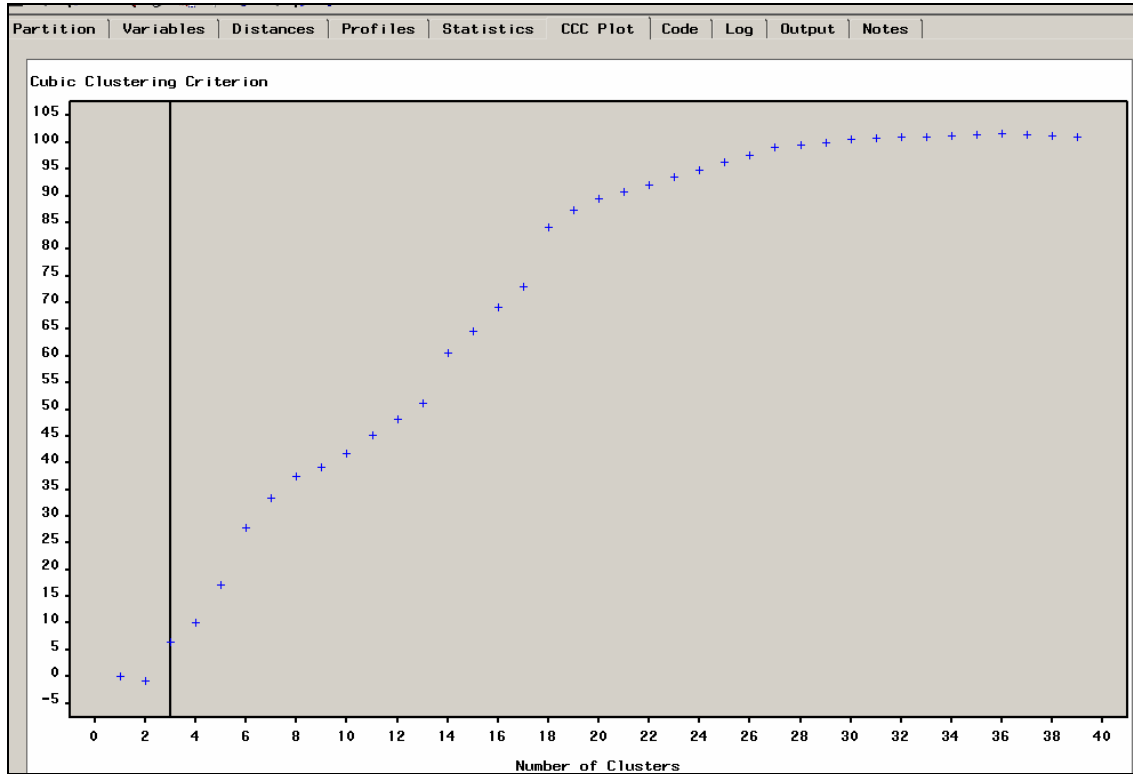


Table 3.3.1 displays information about each cluster obtained from the statistics tab of the cluster results in a tabular format. The cluster number and the frequency (number of households) of each cluster are given in columns one and two. For each cluster the mean of the input attribute is also given. The last column in table 3.3.1 is the Euclidean distance measure calculated from the cluster means of each attribute to the centre of origin. The three clusters were then ranked according to the Euclidean distance.

Table 3.3.1: Cluster node results: Statistics tab

Cluster	Frequency	Water	Refuse	Cooking	Heating	Lighting	Toilet	Distance
1	455 412	0.17	0.03	0.10	0.12	0.09	0.03	0.25
3	147 074	0.65	0.45	0.20	0.15	0.17	0.59	1.03
2	303 262	0.70	0.79	0.76	0.75	0.49	0.65	1.71

Figure 3.3.12 Bar chart: Three clusters.

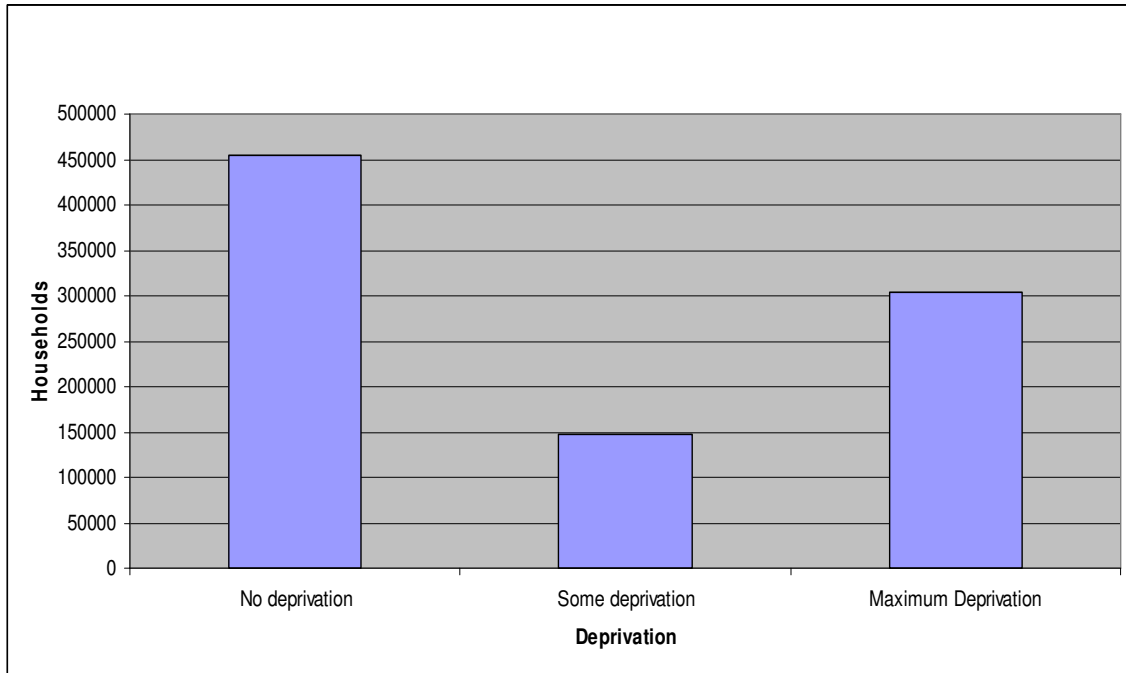


Figure 3.3.12 shows the frequencies of the three clusters created in the above analysis. Cluster 1 has 455 412 households and is labeled no deprivation with cluster 3 labelled some deprivation with 147 074 households. The worst off cluster is cluster 2 with 303 362 households and labeled maximum deprivation.

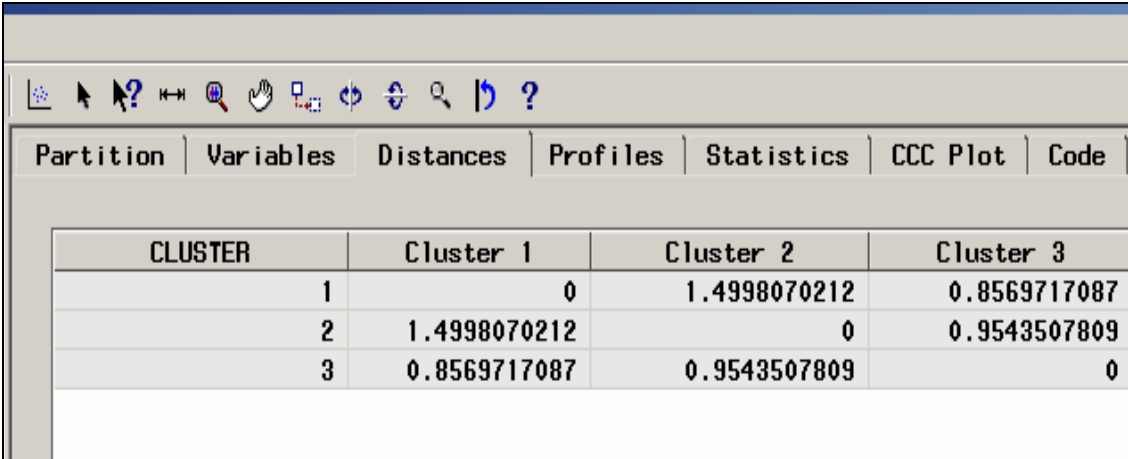
Figure 3.3.13: Cluster node results: Output tab

<u>Variable</u>	<u>Total STD</u>	<u>Within STD</u>	<u>R-Square</u>	<u>RSQ/(1-RSQ)</u>
water	0.33197	0.13705	0.832878	4.983643
Refuse	0.41374	0.04992	0.985729	69.071826
Cooking	0.39027	0.09287	0.944478	17.010833
Toilet	0.38643	0.11220	0.917335	11.097061
Heating	0.38394	0.07512	0.962465	25.642139
Lighting	0.41123	0.05062	0.985141	66.299323
OVER-ALL	0.38720	0.09192	0.944742	17.096881
Pseudo F Statistic =			859.23	

In figure 3.3.13 the statistics for the attributes obtained from the output tab of the cluster results are shown. The SAS procedure FASTCLUS is run and the pseudo F statistic is 859. The figure also shows the R Square value for each attribute. The R Square for all the attributes are fairly high, with the attribute water having the lowest R Square of 0.83.

The clustering algorithm created three clusters; therefore the distance tab results are in a table instead of a plot. Figure 3.3.14 shows the table of distances between the three clusters. Cluster 1 is furthest from cluster 2. If there were more than three clusters the Cluster Node results will produce a graphical representation for the distances between clusters.

Figure 3.3.14: Cluster node results: Distance tab

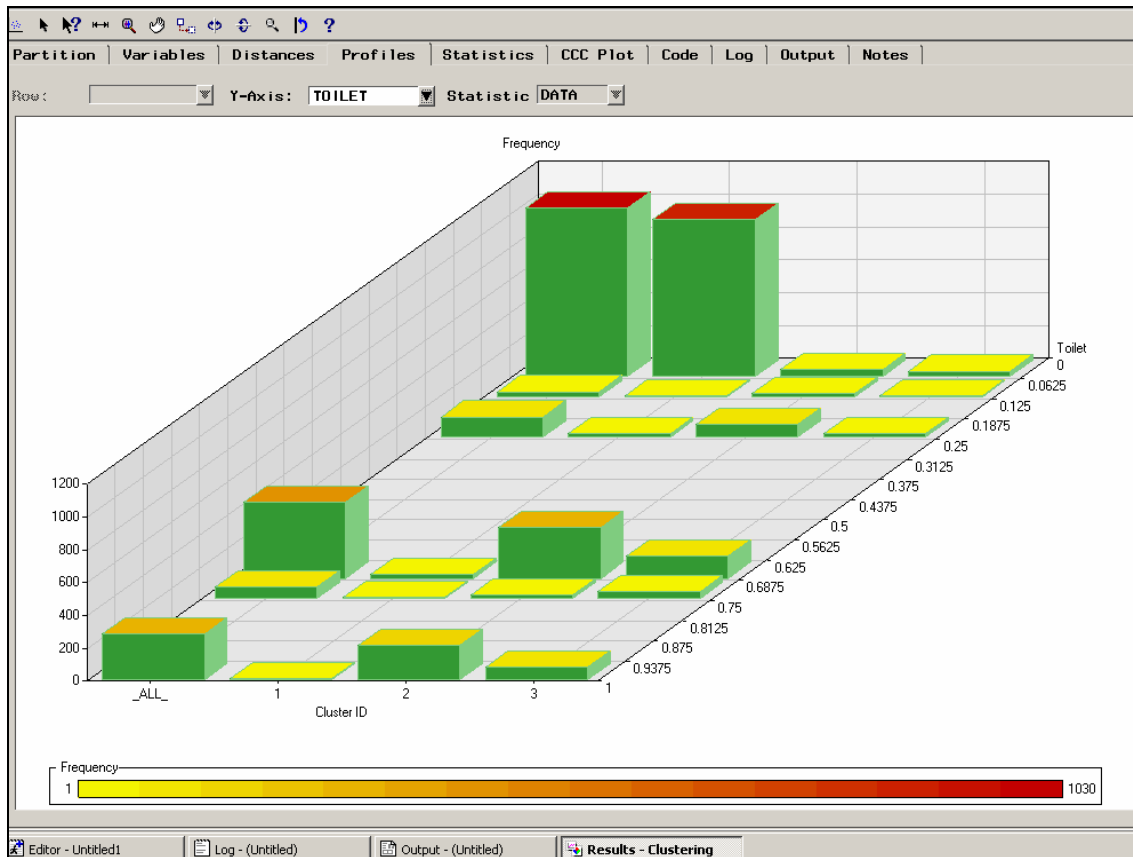


Partition	Variables	Distances	Profiles	Statistics	CCC Plot	Code
	CLUSTER	Cluster 1	Cluster 2	Cluster 3		
	1	0	1.4998070212	0.8569717087		
	2	1.4998070212	0	0.9543507809		
	3	0.8569717087	0.9543507809	0		

The three dimensional bar chart shown in figure 3.3.15 is for a random sample of 2000 households. The membership function for the attribute “toilet facilities” is shown on the X-axis and the numbers of the clusters are shown on the Y-axis with the height denoting the frequency. The ALL cluster shows the overall total.

The bar charts also show that cluster 1 consists of households that are least deprived in respect to the attribute “toilet facilities” while cluster 2 consists of households that are most deprived.

Figure 3.3.15: Cluster node results: Profiles tab



In the second calculation the number of clusters in the cluster node is set to nine as shown in figure 3.3.16. The data sets are the same that were used in the previous section, that is, 905 745 households with the following six attributes:

- Access to water,
- Toilet facilities,
- Energy source for heating,
- Energy source for cooking,
- Energy source for lighting, and
- Refuse disposal.

Figure 3.3.16: Cluster node results: Partition tab for 9 clusters

Data	Variables	Clusters	Seeds	Missing Values	Output	Notes
<p>Segment Identifier: _____</p> <p>Variable name: <input type="text" value="_SEGMNT_"/></p> <p>Variable label: <input type="text" value="Cluster ID"/></p> <p>Role: <input type="text" value="group"/> ▼</p> <hr/> <p>Number of Clusters: _____</p> <p> <input checked="" type="radio"/> User specify <input type="text" value="9"/> <input type="radio"/> Automatic <input type="button" value="Selection Criterion..."/> </p>						

When the number of clusters is set to user specified, the selection a criterion does not apply and a value for the cubic clustering criterion is not calculated.

The cluster node is run and the following results are obtained. Figure 3.3.17 shows the partition tab of the cluster results. The three dimensional pie chart on the left of figure 3.3.17 shows 9 clusters as specified. The grid plot of the input means, shown on the right hand side of figure 3.3.17 shows the overall input means as well as the input means for cluster 7 and cluster 3.

From figure 3.3.17 it can be seen that all households in cluster 7 have electricity, piped water, and flush toilets while the households in cluster 3 do not have electricity for lighting, do not have flush toilets and have no access to tap water.

A comparison of the input means is made for the best cluster which is cluster 7 and the cluster which has the most deprived households is cluster 3, and as observed before the best cluster has an input means of zero or very close to zero for all the attributes. In the comparison it can be seen that lighting is the variable that has the greatest spread and shown in figure 3.3.17 lighting is the first input means and heating has the smallest spread and is shown last.

Figure 3.3.17: Cluster node results: Partition tab for 9 clusters

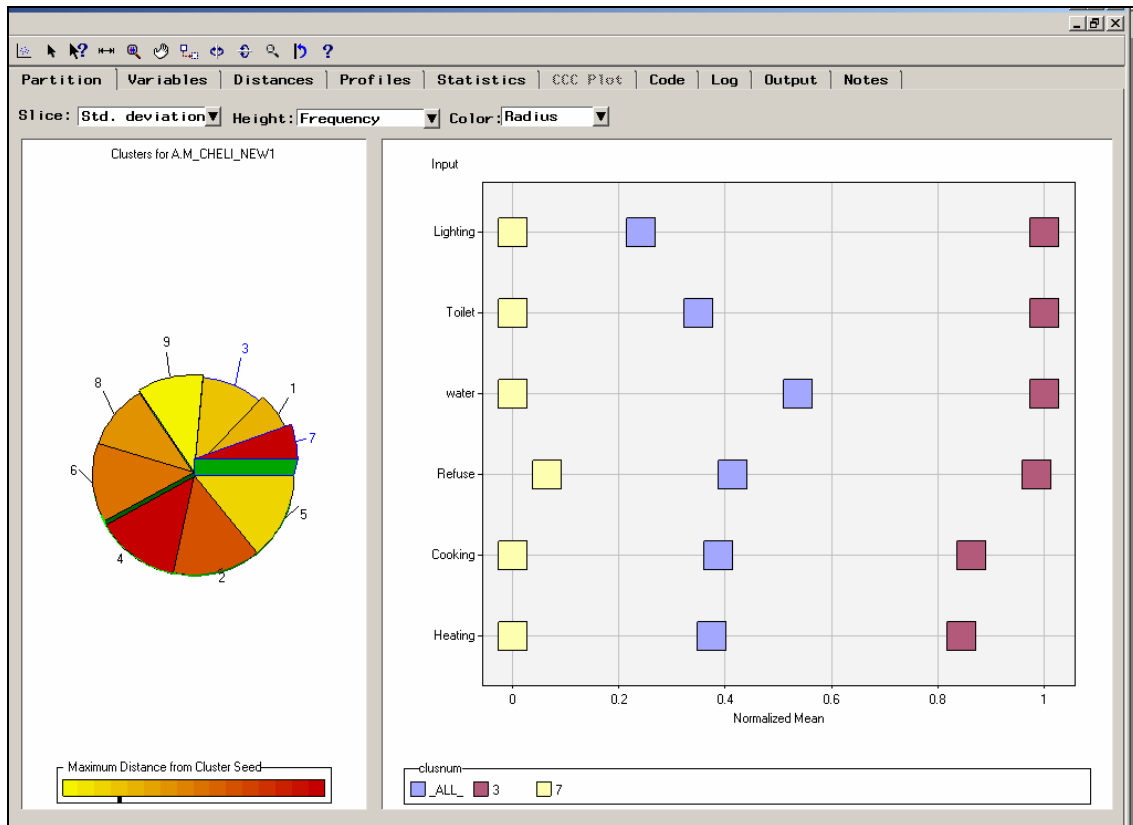


Figure 3.3.18 shows the variable tab in the cluster results browser, listing all the input variables that are used in the clustering analysis. The attribute “refuse removal” has the highest value of importance. The attributes “access to water”, “toilet facilities” and “energy source for heating” also have very high value of importance indicating that they contributed to the cluster formation.

Figure 3.3.18: Cluster node results: Variables tab for 9 clusters

Name	Importance	Measurement
REFUSE	1	interval
WATER	0.980077639	interval
TOILET	0.9075777739	interval
HEATING	0.8531164653	interval
LIGHTING	0.7128022343	interval
COOKING	0.4971513291	interval

Table 3.3.2: Cluster node results: Statistics tab for 9 clusters

	Cluster	Freq	Water	Refuse	Cooking	Heating	Lighting	Toilet	Dist
no deprivation	7	263 553	0.01	0.06	0	0.01	0	0.01	0.06
very little deprivation	1	148 046	0.45	0	0.11	0.07	0.02	0.01	0.47
little deprivation	5	52 898	0.36	0.07	0.35	0.9	0.01	0.16	1.05
below average deprivation	2	47 690	0.57	0.02	0.31	0.19	0.3	0.77	1.07
average deprivation	6	95 697	0.65	0.83	0.23	0.18	0.03	0.6	1.25
above average deprivation	4	36 343	0.45	0.02	0.53	0.64	0.98	0.21	1.39
extreme deprivation	9	106 131	0.68	0.81	0.86	0.81	0.06	0.67	1.72
very extreme deprivation	8	73 979	0.72	0.84	0.81	0.69	0.99	0.49	1.89
maximum deprivation	3	81 411	0.78	0.83	0.74	0.76	0.99	0.93	2.07

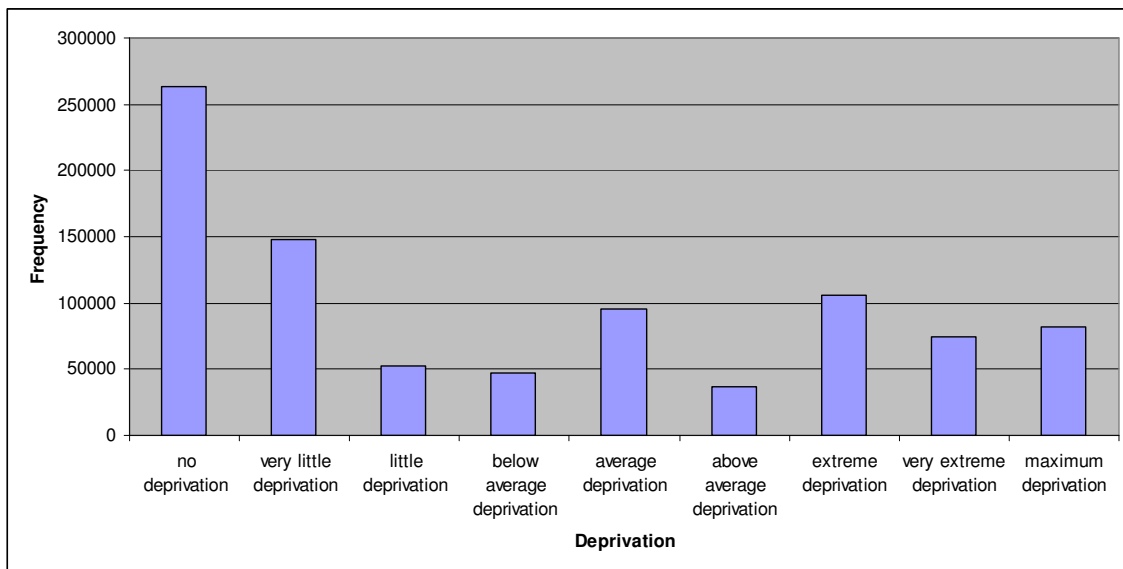
Table 3.3.2 displays information on the 9 clusters obtained from the statistics tab of the results browser in tabular format. The cluster number and the frequency (number of households) of each cluster are given in columns two and three. For each cluster the mean of the input attribute is also given. The last column in table 3.3.2 is the Euclidean distance measure calculated from the cluster centroids of each attribute to the centre of origin. The clusters are ranked according to the Euclidean distance. The cluster with the smallest Euclidean distance is categorized as the cluster with households that were the best off and the cluster with the largest Euclidean distance regarded as the cluster with households that are worst off in terms of deprivation of basic services.

Households that have a cluster mean of zero for any attribute experience zero deprivation in that attribute. The cluster means of all the attributes in cluster 1 are very close to zero. In table 3.3.2 the first column describes the clusters and cluster 7 is described as households experiencing zero deprivation. The maximum possible Euclidean distance measure is the square root of six, 2.45, (that is, when the cluster means for all the attributes are equal to one),

Cluster 3 has an Euclidean distance measure of 2.07 and all its households are described as experiencing maximum deprivation in basic services. Table 3.3.2 shows the multidimensional measure of deprivation from households experiencing no deprivation to households experiencing maximum deprivation. There are 263 553 households in cluster 7 that experience no deprivation of basic services. Cluster 3 has 81 411

households that experience maximum deprivation of basic services, this can be described as the union measure of poverty where the households experience deprivation in all attributes. The other seven clusters experience the union measure of poverty, i.e. deprivation in at least one attribute.

Figure 3.3.19: Bar chart: Nine clusters



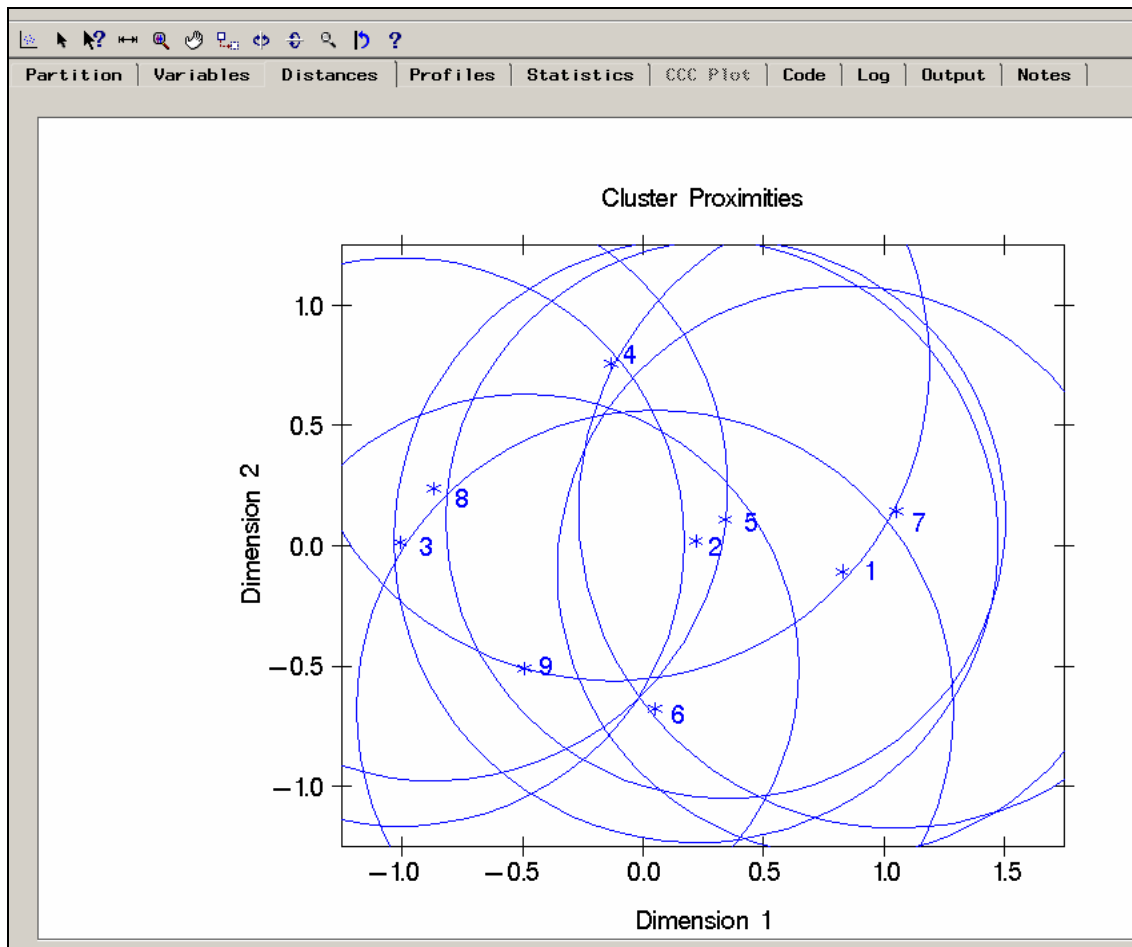
In figure 3.3.19 the frequencies for each category of deprivation is plotted as a bar chart. The first bar represents households that experience no deprivation. The middle seven clusters comprise of households that experience different degrees of deprivation.

If there are more than three clusters the distance tab in the clustering results browser provides a graphical representation of the size of each cluster and the relationship among the clusters as shown in figure 3.3.20

The graph axis is determined from multidimensional scaling analysis, using a matrix of distances between cluster means as input. The asterisks represent the cluster centre and the circles represent the cluster radii. A cluster that has only one case is represented as an asterisk. The radius of each cluster depends on the most distant case in that cluster

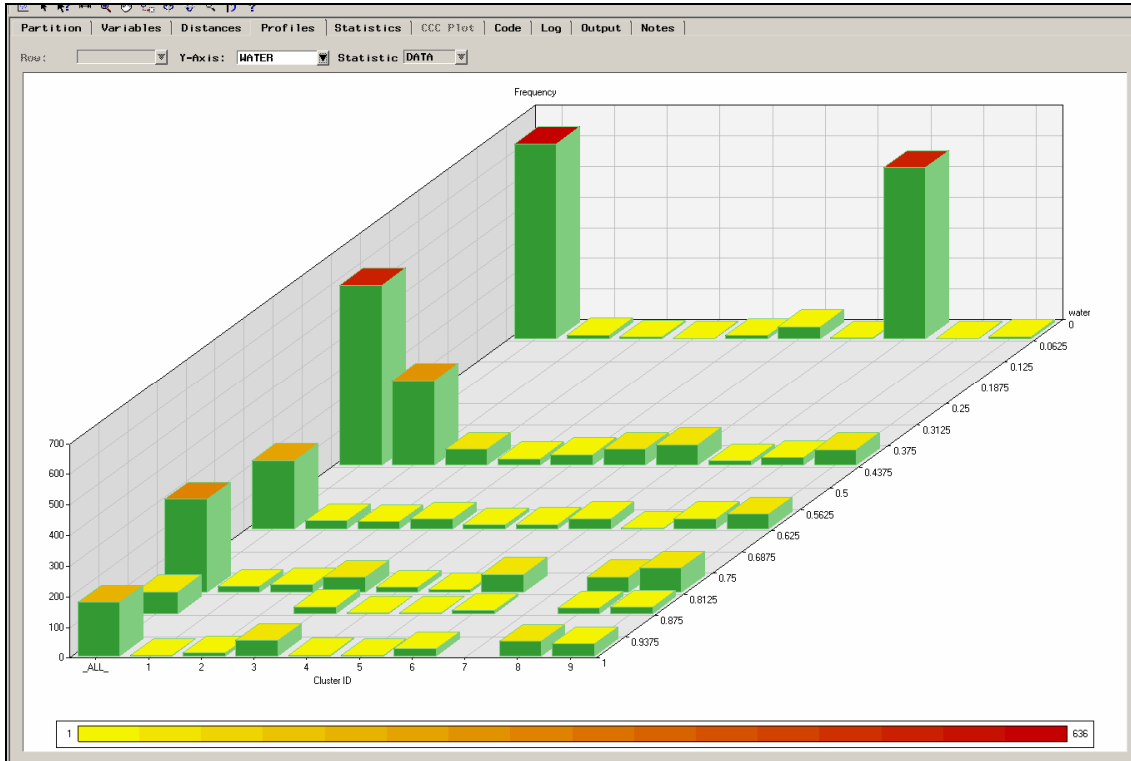
and cases may not be uniformly distributed within the clusters, thus it may appear that clusters overlap. This is in fact not true since each case is assigned to one cluster only. Figure 3.3.20 clearly shows that cluster 7 comprises of households that are least deprived while cluster 3 comprises of households that are most deprived in terms of basic services.

Figure 3.3.20: Cluster node results: Distance tab for 9 clusters



The three dimensional bar chart shown in figure 3.3.21 is for a random sample of 2 000 households. The membership function for the attribute “water” is shown on the X axis and the numbers of the clusters are shown on the Y axis with the height denoting the frequency.

Figure 3.3.21: Cluster node results: Profiles tab for 9 clusters



The ALL cluster shows the overall total. The bar charts also show that cluster 7 consists of households that are least deprived in respect to the attribute “water” while clusters 3 and 8 consists of households that are most deprived.

Table 3.3.3: Cluster node results: Output tab for 9 clusters

Attribute	Total STD	Within STD	R Square	RSQ/(1-RSQ)
water	0.33	0.17	0.73	2.73
cooking	0.39	0.21	0.71	2.41
heating	0.38	0.16	0.83	4.97
lighting	0.41	0.11	0.92	13.16
toilet	0.39	0.18	0.77	3.53
refuse	0.41	0.16	0.85	5.89
overall	0.38	0.16	0.81	4.31

In table 3.3.3 the statistics for the attributes obtained from the output tab of the 9 cluster results are shown. The SAS procedure FASTCLUS is run and some of the statistics that the cluster algorithm calculated for each attribute is shown.

The overall R Squared is 0.81 and the Pseudo F statistics is 488 879. The pseudo F statistics measures the difference between clusters. The number of clusters should be chosen such that the information loss is limited, that is, when the pseudo t^2 is maximum plus one and the pseudo F is maximized (Luzzi *et al.* 2005).

3.4 CONCLUSION

This chapter shows that the Euclidean distance measure removes the need for an aggregation function to measure and compare individual household poverty. The techniques derived can be used to rank households in respect of poverty measurement. The clustering algorithm generates clusters to demonstrate the multidimensionality of poverty measurement and combined the union approach and intersection approach to poverty measurement. The clusters that were created have no order in ranking the various depths and severity of poverty and deprivation experienced by households. This shortcoming is solved in the next chapter.