**Sustainability Research Institute**

SCHOOL OF EARTH AND ENVIRONMENT

UNIVERSITY OF LEEDS

---

# Asset-based poverty analysis in rural Bangladesh: A comparison of principal component analysis and fuzzy set theory

Sonia Ferdous Hoque

April, 2014

No. 59

**SRI PAPERS**

SRI Papers (Online) ISSN 1753-1330

First published in 2014 by the Sustainability Research Institute (SRI)

Sustainability Research Institute (SRI), School of Earth and Environment,

The University of Leeds, Leeds, LS2 9JT, United Kingdom

Tel: +44 (0)113 3436461

Fax: +44 (0)113 3436716

Email: SRI-papers@see.leeds.ac.uk

Web-site: http://www.see.leeds.ac.uk/sri

**About the Sustainability Research Institute**

The Sustainability Research Institute conducts internationally recognised, academically excellent and problem-oriented interdisciplinary research and teaching on environmental, social and economic aspects of sustainability. Our specialisms include: Business and organisations for sustainable societies; Economics and policy for sustainability; Environmental change and sustainable development; Social and political dimensions of sustainability.

**Disclaimer**

**Asset-based poverty analysis in rural Bangladesh: A comparison of principal component analysis and fuzzy set theory**



Email: eesfh@see.leeds.ac.uk

**Contents**

## Abstract

The development of Sen's capabilities approach has shifted the focus of poverty analysis from its traditional focus on measuring food intake, income or consumption expenditures to a broader multi-dimensional approach that uses a large number of indicators to assess human well-being or standard of living. This paper contributes to the literature on multi-dimensional poverty analysis, by investigating the extent to which results depend on the method of data analysis. Two distinct methods, principal component analysis (PCA) and fuzzy set theory (FST), have been applied to household survey data from rural Bangladesh. The study findings show that both PCA and FST can lead to reliable results in terms of poverty analysis. However, changes in procedures, such as the variables used, the number of factors extracted and the type of cluster analysis applied (in case of PCA) or the calibration and aggregation methods used (in case of FST) can lead to different results. Compared to PCA, which is totally data-driven, FST provides more flexibility, in terms of involving conceptual and theoretical inputs, and also leads to outputs that are easier to visualise and interpret.

**Key words:** multi-dimensional poverty assessment; principal component analysis; k-means cluster analysis; fuzzy set theory; Bangladesh.

**Submission date:** 03-02-2014; **Publication date**: 16-04-2014

## About the Author

Sonia Ferdous Hoque is currently a postgraduate research student (PhD) at the School of Earth and Environment at the University of Leeds, UK. She has completed her MSc in Environmental Management from the National University of Singapore (NUS) and BSc in Environmental Science from North South University, Bangladesh. She has received the Shell Gold Medal and Prize (2012) and the Chancellor's Gold Medal (2010) for being the most outstanding student of her masters and bachelors cohort respectively. Prior to joining her PhD program, Sonia worked as a Research Associate at the Institute of Water Policy, NUS (2011–2013), where her research focused on roles of private and public sector in urban water governance and water and wastewater tariffs in urban domestic and non-domestic sectors. She also served as a Research Officer at the Bangladesh Institute of Development Studies (2009–2010), where she was involved in research on climate change adaptation in Bangladesh.

**Asset-based poverty analysis in rural Bangladesh: A comparison of principal component analysis and fuzzy set theory**

## 1. The trend towards multi-dimensional poverty assessment

Poverty analysis has evolved from its traditional focus on measuring food intake, income or consumption expenditures to a multi-dimensional approach that uses a large number of indicators to assess human well-being or standard of living. This trend towards a broader asset-based method has been initiated by Amartya Sen's capabilities approach, which advocates analysing poverty or well-being in terms of 'functionings' and 'capabilities', rather than the maximization of utility through monetary income (Sen, 1985, Sen, 1993, Sen, 1999). Functionings are the states of 'beings or doings', such as being nourished and healthy, while capability refers to the freedom to choose between different functioning combinations. The ability of individuals to transform resources into valuable achievements (functionings) is also determined by the individual's own personal characteristics (e.g. age, gender, physical capacities) and the general environmental context.

An empirical application of Sen's capabilities approach to poverty measurement requires identification of a set of indicators related to selected dimensions of well-being and adequate criteria to measure and aggregate them (Chiappero Martinetti, 2000). While a minimum set of basic functionings, such as health, nutrition, shelter, child mortality and education may be adequate for evaluating the extreme poor in developing countries, a wider range of indicators including social interactions and psychological well-being may also be required in developed country contexts (Chiappero Martinetti, 2000). While Sen has argued that the choice of relevant functionings and capabilities for any poverty measure is a value judgment rather than a technical exercise (Sen, 2008), many authors have subsequently suggested their own lists of basic functionings. The most prominent one is Nussbaum (2000)'s list of central human capabilities which defines characteristics of a full human life at a very general level and includes life, health, bodily integrity, senses, emotions, practical reasons, affiliation, other species, play and control.

Sen's theory of development as an expansion of human capabilities has provided the conceptual foundation for the human development paradigm, which aimed *"to shift the focus of development economics from national income accounting to people centred policies"* (Haq, 1995). This has led to the annual publication of UNDP's human development reports since 1990. These reports monitor progress through the human development index (HDI) - a composite measure comprising of indicators along three dimensions (with four indicators): life expectancy, educational attainment and command over resources for a decent living (UNDP, 2013). The HDI has been criticized on a number of grounds, including the choice of dimensions and indicators, the assignment of equal weights, the methods of calculation and data errors (Noorbakhsh, 1998, Wolff et al., 2011). In response, the UNDP argues that while the HDI is not enough to measure a country's level of development, it offers a broad proxy on some of the key issues of human development. More recently, the Oxford Poverty and Human Development Initiative (OPHI) has introduced the Multidimensional Poverty Index (MPI), which is composed of ten indicators corresponding to same three dimensions of the HDI (Alkire and Santos, 2010). The MPI aims to reflect on the overlapping deprivations that households face in relation to indicators emphasized in the Millennium Development Goals.

The operationalization of Sen's capabilities approach as well as other multidimensional measures of poverty has two inherent issues: the identification problem (setting up a break or poverty line to distinguish the poor and the non-poor) and the aggregation problem (summation of the dimensions of deprivation or well-being for each unit of analysis - individual/ household) (Neff, 2013). To address these issues, researchers conducting poverty assessments at micro levels have explored a wide variety of methods, of which principal component analysis (PCA) and fuzzy set theory (FST) have gained increased popularity in the last two decades (see sections 2.2 and 3.2 below for a review on the applications of these methods).

This paper contributes to the literature on multi-dimensional poverty analysis, by comparing results from two different methods (PCA and FST). These methods have been widely used in poverty assessments, yet there have been few comparisons of the two methods using the same data-set. The study by Lelli (2001) used factor analysis to validate the presence of seven main functionings amidst 54 indicators and then compared the factor scores and fuzzy membership scores for different age and social groups. In another study, Roche (2008) used PCA to identify three dimensions among eight housing adequacy indicators and then used FST to form synthetic indices. However, while these studies compared or combined the two methods, they provided little discussion on the strengths and limitations of the methods themselves.

The application of PCA is not only restricted to the poverty literature; it has been used to differentiate climate change vulnerabilities and adaptations based on socio-economic status (Islam, 2013, Berman et al., 2014, Sallu et al., 2010), as well as to generate weights for indicators in developing vulnerability indices. However, the methodological limitations of PCA in categorizing households by wealth groups or generating theoretically meaningful weights have often been overlooked in these studies. In comparison, FST has mainly been applied in poverty studies and rarely been used in vulnerability assessments. Given that FST is more theoretically grounded and not entirely data-driven like PCA, it can be used in a wider range of applications.

The main objectives of this paper are:

1. To elaborately illustrate the application of PCA (followed by cluster analysis) and FST in poverty assessment, with specific focus on how changes in certain methodological steps can lead to different results.
   a. To demonstrate how the number and type of variables and the number of factors extracted influence the 'factor loadings' in PCA.

b. To demonstrate how the calibration and aggregation approaches (using Ragin's and Cheli and Lemmi's methods) used influence the outcomes in FST [application of Ragin's method in poverty analysis and its comparison to Cheli and Lemmi's method has not yet been demonstrated in the literature].
2. To compare the results from these two methods and highlight the advantages and limitations of each method in different contexts.

In this study, asset variables from Bangladesh Integrated Household Survey 2011-2012 have been used to analyse poverty using PCA and FST. Sections 2 and 3 introduce the mathematical and conceptual basis of the two methods and provide overviews of their use in poverty analysis, while section 4 provides a methodological comparison. Section 5 discusses the methods and results of applying these two methods in case of rural Bangladesh. A discussion of the results is provided in section 6.

## 2. Principal component analysis (PCA) and cluster analysis

### 2.1 The method

PCA is a multivariate statistical technique that reduces the number of variables in a dataset into a smaller number of dimensions or factors. Using the correlations between sets of variables, PCA extracts a number of factors that can be considered as salient unobserved variables capturing important aspects of the complete set. Each of these factors is a linear weighted combination of the initial variables and is uncorrelated to other factors. Mathematically, for n number of variables,

$$\begin{aligned} PC_1 &= w_1X_1 + w_2X_2 + w_3X_3 + \ldots + w_nX_n \\ PC_m &= w_{m1}X_1 + w_{m2}X_2 + w_{m3}X_3 + \ldots + w_{mn}X_n \end{aligned} \qquad [1]$$

where $w_{mn}$ is the weight assigned to the variable $X_n$ in the $m^{th}$ principal component. As the method relies on the correlations between sets of variables, correlations between individual variables should be greater than absolute 0.30 for the analysis to produce meaningful results (Mooi and Sarstedt, 2011). It is not problematic if single correlations are less; however, when all correlations tend to be around zero, the method stops being useful. The Kaiser-Meyer-Olkin (KMO) statistic, also called the measure of sampling adequacy, indicates whether the correlations between variables can be explained by other variables in the dataset and KMO values greater than 0.70 are usually considered as appropriate (Mooi and Sarstedt, 2011). Moreover, the Bartlett's test of sphericity can be used to test the null hypothesis that the correlation matrix is a diagonal matrix (that is, all non-diagonal elements are zero) in the sample. Since PCA requires high correlations, a small p-value will favour the rejection of the hypothesis (Mooi and Sarstedt, 2011).

The first principal component ($PC_1$) has the highest eigenvalue and accounts for the highest percentage of variance. The second component ($PC_2$) is completely uncorrelated with $PC_1$ and explains additional but less variation than $PC_1$. Eigenvalues describe how much variance is accounted for by a certain factor. If the number of components extracted is equal to the number of variables in the dataset, the cumulative variance will be 100%. On the other hand, communality indicates how much variance of each variable can be reproduced through factor extraction. Again, if the number of factors is equal to the number of variables communality of each variable will be 1.00. However, as the main aim of this process is data reduction, the Kaiser criterion of extracting factors with eigenvalues greater than one is frequently used and communalities should be at least 0.30 (Mooi and Sarstedt, 2011). This trade-off between simplicity and accuracy is an issue in any PCA.

In poverty analysis, the factors loadings of $PC_1$ are usually used to classify households into different socio-economic groups. As discussed below, the frequently used method has been to divide households based on quartiles or quintiles of the $PC_1$ scores. However, this method is somewhat arbitrary as some households in one quartile may have greater similarity to households in another quartile than to those in the quartile to which it has been designated. A better alternative is to use cluster analysis, a statistical procedure to identify homogenous groups of cases. There are three main forms of cluster analysis, namely hierarchical method, k-means clustering and two-step clustering, each of which uses a different approach of grouping. As the k-means clustering has been applied in this paper, only this method is explained. This procedure segments the data in such a way that the within-cluster variation is minimised. K-means clustering is thought to be superior to hierarchical methods as it is less affected by outliers and the presence of irrelevant variables. It is also suitable for applying to very large datasets, especially above sample size 500, as it is computationally less demanding (Mooi and Sarstedt, 2011). However, unlike the other methods, the researcher has to specify the number of clusters to retain, which sometimes makes it less attractive. In cluster analysis, the same variables used for PCA or the factor scores of $PC_1$ can be used as inputs.

### 2.2 Application of PCA in asset-based poverty analysis

As mentioned above, one of the challenges of estimating wealth using asset-based approach is the aggregation problem which involves assigning weights to different types of assets. While some early studies simply applied equal weights to all assets (Montgomery et al., 2000), it seems intuitively incorrect, for example, to treat ownership of agricultural land and a bicycle as of equal importance. PCA has been used extensively to address this problem and commonly the factor loadings from $PC_1$ are used as ‘weights’ for individual assets. However, this process is entirely data-driven and not theoretically grounded. Moreover, PCA has been used to identify the number of dimensions in large data-sets. Researchers have applied PCA on asset variables to estimate the relationship between household wealth and children’s school enrolment in India (Filmer and Pritchett, 2001), to investigate effect of wealth inequality on mortality and immunisation coverage (Houweling et al., 2003), to study inequality in living standards in Mexico (McKenzie, 2005) and to test the validity and limitations of the PCA method using Brazil and Ethiopia as examples (Vyas and Kumaranayake, 2006).

Filmer and Pritchett (2001) conducted PCA on asset variables obtained from Demographic and Health Survey (DHS) and used 21 indicators grouped into three broad categories (ownership of consumptive durable assets, dwelling conditions and housing materials, and ownership of land) to categorise households into the poorest 40, middle 40, and richest 20 percentiles. The authors established the empirical validity of PCA by comparing state-level averages of asset indices with headcount poverty rates and gross state product per capita in India. Moreover, they used national integrated household survey data from Indonesia, Nepal, and Pakistan to demonstrate correspondence between household classification based on asset indices and consumption expenditures.

McKenzie (2005) first highlighted the importance of using a broad class of indicators for assets to avoid the problems of 'clumping' and 'truncation' in PCA. Clumping or clustering occurs when households are grouped together in a small number of distinct clusters. Truncation implies a more even distribution of socio-economic status, but spread over a narrow range, making differentiating between socio-economic groups difficult (for example, not being able to distinguish between the poor and the very poor) (Vyas and Kumaranayake, 2006). This occurs when the same amount of certain assets are owned or not owned by most households in the sample. In order to avoid these issues, it is best to include as many variables as possible (McKenzie, 2005).

Vyas and Kumaranayake (2006) also employed DHS data to test the validity and challenges of using PCA to distinguish socio-economic status. Using the examples of urban and rural Brazil and Ethiopia, the authors found that due to problems of clumping (which is illustrated by the skewed distribution of factor scores), it was challenging to divide households in rural Ethiopia and urban Brazil into different socio-economic categories. As a result, in rural Ethiopia the mean difference in asset ownership was very less between the three poorest quintiles and in urban Brazil the difference was higher between the poorest two categories compared to adjoining ones. The authors further used cluster analysis to demonstrate that indeed 60% of all households in rural Ethiopia belonged to the poorest of the three categories used, while 46% in urban Brazil belonged to the richest group. The authors noted that the use of continuous variables (such as ownership of land) and a combination of other relevant asset variables can lead to better assessment of household wealth.

Houweling et al. (2003) used DHS data from 10 countries to study the impact of choice of selected variables on the observed economic status of households, which in turn, affects child mortality and immunisation rates. Using four alternative sets of variables in the PCA analysis, the authors tested the importance of including/excluding indicators that have direct impact on children's health (that is, water and sanitation), that have indirect impact through determining household wealth (that is, housing) and that are publicly provided at community level (that is, electricity). The results showed that the selection of variables greatly influenced the magnitude of observed inequalities in many countries. Reducing the number of variables to more homogenous items increased the common variance and led to odd results while grouping households by wealth status.

## 3. Fuzzy set theory (FST)

### 3.1 The method

The fuzzy set theory, developed by mathematician Zadeh (1965), has been one of the approaches used to address both identification and aggregation problems in multi-dimensional poverty analysis. Compared to a crisp set theory, that identifies an individual as either poor or non-poor based on a single precise poverty line, FST conceptualises poverty as the degree of membership to a poverty set, measured on a scale from 0 to 1, where 1 indicates full membership and 0 means full non-membership (Ragin, 2000). Mathematically, if X denotes a universal set, then the fuzzy set A is defined by the membership function $\mu_A$, as follows,

$$\mu_A:(X) \rightarrow [0,1]$$

where [0,1] is the interval of real numbers between 0 and 1, $\mu_A$ (x) = 0 if the element $x \in X$ does not belong to A, $\mu_A$ (x) = 1 if the element $x \in X$ belongs to A, and $0 < \mu_A (x) < 1$ if x partially belongs to A. The choice of membership function, $\mu_A$, takes several forms and can depend on the application context and the type of indicators.

**Calibration**

In order to solve the identification problem, conventional variables are transformed into membership scores, a process known as calibration. In the most rudimentary form, the researcher specifies the values of an interval-scale variable that correspond to three qualitative breakpoints that structure a fuzzy set: the threshold for full membership (fuzzy score = 0.95), the threshold for full non-membership (fuzzy score = 0.05), and the cross-over point (fuzzy score = 0.5) (Ragin, 2006). These break-points, called qualitative anchors, are based on theoretical criteria external to the data and accounts for the researcher's conceptualization, definition, and labelling of the set in question, which make fuzzy sets different from a mere ordinal ranking of cases. Ragin (2008) describes this calibration procedure as the direct method, which uses estimates of the log of the odds of full membership in a set as an intermediate step. In contrast, with the indirect method, the researcher initially sorts the cases into different levels of membership (with verbal labels and preliminary scores) and then refines these scores using the interval scale data (Ragin, 2008).

Other types of membership functions have also been widely used in the literature (Chiappero Martinetti, 2000, Lelli, 2001). Linear membership functions simply involve standardization of the raw indicator values and assume that the actual value of each case with respect to a given variable is directly proportional to its membership score [2]. In a quadratic sigmoid curve, the non- and full-membership values are defined in terms of certain thresholds ($\alpha$ and $\gamma$, respectively), while intermediate values between $\alpha$ and the cross-over point $\beta$ and between $\beta$ and $\gamma$ are based on quadratic interpolation [3]. Similar to the linear function, the trapezoidal function also standardises the intermediate values, except that in case of the former zero-membership and full-membership refers to the minimum and maximum values, whereas for the latter, they refer to values below and above certain thresholds [4]. Finally, in order to avoid the above arbitrary measures of membership scores, Cheli and Lemmi (1995) propose what they name as a 'totally fuzzy and totally relative' procedure, defining the membership to the fuzzy set on the basis of the distribution functions of the considered variables (referred to as CL method, henceforth). This type of calibration is exclusively based on empirical evidence and not on the judgment of the researcher [5a and 5b].

*Linear function*

$$\mu_A(x) = \begin{cases} 0 & \text{if} \quad x = x_{min} \\ \dfrac{(x - x_{min})}{(x_{max} - x_{min})} & \text{if} \quad x_{min} \leq x \leq x_{max} \\ 1 & \text{if} \quad x = x_{max} \end{cases} \qquad [2]$$

*Quadratic sigmoid function*

$$\mu_A(x) = \begin{cases} 0 & \text{if} \quad x \leq \alpha \\ 2[(x - \alpha)/(\gamma - \alpha)]^2 & \text{if} \quad \alpha \leq x \leq \beta \\ 1 - 2[(x - \alpha)/(\gamma - \alpha)]^2 & \text{if} \quad \beta \leq x \leq \gamma \\ 1 & \text{if} \quad x \geq \gamma \end{cases} \qquad [3]$$

*Trapezoidal function*

$$\mu_A(x) = \begin{cases} 0 & \text{if} \quad x \leq a_1 \\ \dfrac{(x - a_1)}{(a_2 - a_1)} & \text{if} \quad a_1 \leq x \leq a_2 \\ 1 & \text{if} \quad x \geq a_2 \end{cases} \qquad [4]$$

*CL frequency function*

$$\mu_A(x) = \begin{cases} 0 & \text{if} \quad x = x^1;\ k = 1 \\ \mu(x^{k-1}) + \dfrac{F(x^k) - F(x^{k-1})}{1 - F(x^1)} & \text{if} \quad x = x^k;\ k > 1 \\ 1 & \text{if} \quad x = x^K;\ k = K \end{cases} \qquad [5a]$$

where, F(x) is the cumulative function for variable x; k is the value taken by the variable x [k=1 means $x_{min}$, 1<k<K means $x_{min} \leq x \leq x_{max}$, k = K means $x = x_{max}$] and the values of x should be arranged in increasing order of deprivation. This equation can be simplified as:

$$\mu_A(x) = \begin{cases} 0 & \text{if} \quad x = x^1;\ k = 1 \\ \dfrac{F(x^k) - f(x^1)}{1 - f(x^1)} & \text{if} \quad x = x^k;\ k > 1 \\ 1 & \text{if} \quad x = x^K;\ k = K \end{cases} \qquad [5b]$$

where, $f(x^1)$ is the frequency associated with the minimum value for the variable x;

**Aggregation**

In order to address the challenge of aggregating multi-dimensional measures, FST offers a range of aggregation functions, the most common ones being the union (logical OR) and intersection (logical AND) functions. The strong union operator focuses on the indicators that show the most favourable position by selecting the maximum membership function among the ones to be aggregated and thus reflect the worst achievements of each individual/household in depicting the considered dimension [6.1]. The strong intersection operator, in contrast, selects the minimum of the membership scores among the indicators to be aggregated and thus, emphasises one's best accomplishments [7.1] (Lelli, 2001). Thus, the strong union and intersection operators implicitly exclude any sort of compensation between indicators, and can used in case of indicators with a positive correlation. The weak union and intersection operators allow the possibility of compensation and can be used for aggregating independent indicators [6.2 and 7.2]. The bounded difference and bounded sum are used when indicators are negatively correlated [6.3 and 7.3] (Chiappero Martinetti, 2000).

*Fuzzy union operators*

| | | |
|---|---|---|
| Standard (or strong) union | $\mu_{A \cup B} = \max[\mu_A, \mu_B]$ | [6.1] |
| Weak union (or algebraic sum) | $\mu_{A+B} = [\mu_A + \mu_B - \mu_A.\mu_B]$ | [6.2] |
| Bounded sum | $\underline{\mu_{A \cup B}} = \min[1, \mu_A + \mu_B]$ | [6.3] |

where $\mu_{A \cup B} \leq \mu_{A+B} \leq \underline{\mu_{A \cup B}}$

*Fuzzy intersection operators*

| | | |
|---|---|---|
| Standard (or strong) intersection | $\mu_{A \cap B} = \min[\mu_A, \mu_B]$ | [7.1] |
| Weak intersection (or algebraic product) | $\mu_{A.B} = [\mu_A.\mu_B]$ | [7.2] |
| Bounded difference | $\underline{\mu_{A \cap B}} = \max[0, \mu_A + \mu_B - 1]$ | [7.3] |

where $\mu_{A \cap B} \leq \mu_{A.B} \leq \underline{\mu_{A \cap B}}$

Another frequently used aggregation function is the averaging operator, which allows the membership score of each element in the aggregated fuzzy set to lie between the minimum and the maximum. Mathematically, the unweighted averaging function is:

$$h_\alpha = (a_1,a_2,\ldots a_n) = [(a_1^\alpha + a_2^\alpha + \ldots + a_n^\alpha)/n]^{1/\alpha} \qquad [8]$$

where, $a_1$, $a_2$, ..., $a_n$ denote the membership grades of each element belonging to sets $A_1$; $A_2$.......$A_n$; min ($a_1$, $a_2$, ..., $a_n$) ≤ h ($a_1$, $a_2$, ..., $a_n$) ≤ max ($a_1$, $a_2$, ..., $a_n$); and $\alpha$=1 for the arithmetic mean, $\alpha$=-1 for the harmonic mean and $\alpha$= 0 for the geometric mean. The parameter $\alpha$ is related to the elasticity of substitution between different dimensions or indicators. A small value for $\alpha$ refers to a lower allowed substitutability between dimensions, meaning that one has to give up more of one dimension to get an extra unit of a second dimension while keeping the level of wellbeing constant (Decancq and Lugo, 2012).

The union, intersection and averaging operators can be applied both in weighted and unweighted forms. Decancq and Lugo (2012) argue that weights are central in determining the trade-offs between different dimensions or indicators of poverty or well-being. Like the membership functions used for calibration, the weighting procedure used for aggregation also depends on the choice of the researcher. Data-driven weights are a function of the distribution of the achievements in the society and are not based on any value judgement about how the trade-offs between the dimensions should be (Decancq and Lugo, 2012). One of the common data-driven weighing approach is the frequency based system, in which the weight is determined as a function of the distribution of the membership scores or achievement levels of that indicator. It is based on Desai and Shah (1988)'s assumption that a larger weight should be assigned to an indicator for which a smaller proportion of people exhibit a low achievement. This reflects the fact that individuals attach a higher importance to the shortfalls in indicators where majority do not fall short. Cerioli and Zani (1990) suggested the weights to be equal to the inverse of the proportion of individuals who are deprived with respect to a given indicators, while Cheli and Lemmi (1995) generalised the latter weighing structure as:

$$w = \ln [(1/n).\sum\mu\ (x)] \qquad [9]$$

Another data-driven approach is to use the factor loadings generated from principal component analysis. However, critics (Nardo et al., 2005, Somarriba and Pena, 2009) have pointed out that as PCA puts more weight on dimensions that are strongly correlated to other dimensions, it fails to represent the real influence of the indicators on well-being or poverty levels.

Normative weighing approaches depend on the value judgements about the trade-offs and are not based on the actual distribution of the achievements in the society under analysis (Decancq and Lugo, 2012). The most common methods under this approach are equal weighting, which is simple and regard all dimensions as being equally important, and arbitrary weighting, in which researchers decide to give more weight on indicators that are considered important. Other methods include obtaining weights from surveys, participatory methods or from theoretical concepts.

### 3.2 Application of fuzzy set theory in asset-based poverty analysis

FST has been applied to a number of poverty assessment studies (Cerioli and Zani, 1990, Chiappero Martinetti, 1994, Cheli and Lemmi, 1995, Qizilbash, 2002, Qizilbash and Clark, 2005, Chiappero Martinetti, 2006, Neff, 2013), especially as a means of operationalizing Sen's capability approach.

Cerioli and Zani (1990) suggested to represent the individual's global deprivation as a weighted aggregate of the membership degrees to the fuzzy set of the deprived people, where the membership function takes the trapezoidal form [4] and the weighing structure equals the inverse of the proportion of individuals who are deprived with respect to the given item. As mentioned in section 3.1, Cheli and Lemmi (1995) defined the membership on the basis of the distribution functions of the considered variables [5] and coupled it with a weighted averaging operator where the weighing system [9], which in case of simple dichotomous variables coincides with the Cerioli and Zani's one.

Making use of the 1994 Italian household survey and mostly drawing on the CL approach, Chiappero Martinetti (2000) aggregated a number of indicators into five functionings (housing, health, education, social interactions and psychological conditions) and studied the inequality in deprivation index by population subgroups. Chiappero Martinetti (2006)Highlighting the different methods of constructing membership functions and the different classes of fuzzy aggregation operators, Chiappero Martinetti (2006) argues that FST is a useful and flexible tool for operationalizing the capability approach while preserving its richness and complexity.

In measuring poverty and deprivation using 1993 South African household survey data, Klasen (2000) selected 14 components of well-being and intuitively scored each component on a scale of 1-5, where 5 represented the best condition. The author derived a weighted aggregation of each component (using weights generated from PCA) as well as an unweighted average, and found high correlation between the two results. In order to distinguish the 'poor households' and the 'severely poor households', cut-off lines were selected at the 40$^{th}$ and 20$^{th}$ percentiles of the aggregate scores, respectively. While Klasen (2000) did not apply FST, his method or ranking various levels of deprivation provided useful indications of applying such methods in fuzzy analysis.

Also working with South African data, Qizilbash and Clark (2005) used questionnaire responses to identify cut-offs between the poor and non-poor. Cut-offs were admissible or treated as meaningful if they were endorsed by at least a small percentage of the sample households. For example, using the 5% criteria, 1-3 years of schooling was considered as the cut-off for education deprivation as it was endorsed by 6.06% of the households. The authors found that their questionnaire approach led to even tougher or lower cut-offs for some indicators compared to Klasen's approach.

Neff (2013) used household data from two villages in Andhra Pradesh (India) and identified three core dimensions of deprivation (housing, nutrition, and health) and two non-core dimensions (social interactions and subjective well-being). The author used crisp sets (with only 0 and 1 values), linear membership function as well as Ragin's direct method to calibrate different indicators under each dimension. Instead of using the weighing systems described above, the author used qualitative criteria to aggregate the five dimensions into a 10-value fuzzy scale, where households not deprived on any core or non-core dimensions were identified as fully non-poor and those deprived on all three core dimensions were classified as fully poor.

## 4. Methodological comparison – PCA vs FST

Based on the discussions above in sections 2 and 3, Table 1 summarises the conceptual basis, technical issues and pros and cons of PCA and FST.

*Table 1. Comparison of poverty analysis using PCA and FST*

| | Principal component analysis | Fuzzy set theory |
|---|---|---|
| Theoretical basis | Form of multivariate statistical analysis (Quantitative) | Based on Set theory (Simultaneously qualitative and quantitative) |
| | It is a method of data reduction that relies on the correlation between a larger number of variables to construct a smaller number of latent variables (or factors). | It is a continuous set, ranging from 0 to 1, that is calibrated to indicate the degree of membership of each case in a given set (0 indicates non-membership, 1 indicates full membership) |
| Examples of applications in socio-environmental sciences | Poverty analysis – disaggregating households by socio-economic status; investigating inequalities in asset ownership by social group or regions<br>Vulnerability analysis – assessing climate change impacts and adaptation by wealth status | |
| | Regression analysis - generating weights for individual variables, and investigating the effect of each variable on the outcome, independently of other variables. | Truth tables - generating truth tables to investigate the combinations of variables related to a given outcome. |
| Technical considerations | Variables which have same values for most cases or do not exhibit certain degree of correlation with any other variable should be omitted. | Each variable can be individually calibrated using different membership functions and is not linked to others. |
| | Cut-off points for variables are entirely data-driven and have no theoretical basis. | Qualitative anchors can be used to identify key breakpoints on variables, thus, ensuring correspondence between theoretical concepts and measurement of set membership. |
| | Each principal component is a weighted | Variables can be aggregated into a |

| | | |
|---|---|---|
| | sum of the values for each of the variables, where the weights are data-driven. | single value using logical operators or qualitative judgment. |
| | Coding should be done with care, as any difference in the numerical codes used for ordinal variables will be interpreted as variance in the variable. | The direction of coding for each variable should be same for the method applied; for example, in CL method, variable should be coded in increasing order of deprivation because of the nature of the formula. |
| Strengths | Eliminates redundant variables with minimal data loss, by combining homogeneous variables into one component.<br>Can identify the number of dimensions in the data.<br>Useful in studies where variables under a given dimension show high correlations with each other.<br>More suitable for scale variables, although all types of variables may be included. | Provides flexibility; researcher can calibrate and aggregate variables based on theoretical or contextual understanding.<br>Impact of irrelevant variation within a variable can be reduced (see section 5.2)<br>Variables do not need to be correlated with each other; inclusion/exclusion of one variable does not affect another.<br>More suitable for ordinal variables, although all types of variables may be included.<br>Ragin's method is useful for making temporal and spatial comparisons. |
| Limitations | Concerned with the maximization of the variance in the variables, which might not be theoretically correct.<br>Retaining more homogenous variables (i.e. variables with higher correlations) produces more significant results in PCA, but very odd results in cluster analysis.<br>Inclusion/exclusion of only one additional variable, recoding even one variable and changing the number of factors extracted significantly affects the results.<br>In studies with large number of variables, often theoretically relevant variables may need to be excluded to produce good PCA results.<br>Temporal and spatial comparisons cannot be made using separate data-sets; in this case, all data may first be combined and then split after PCA is conducted. | Difficult to define cut-off lines between different socio-economic groups as the results show a gradation of deprivation.<br>Temporal and spatial comparisons cannot be made using CL method; in this case, all data may first be combined and then split after fuzzy analysis. |

## 5. Asset-based poverty analysis in Bangladesh

In this paper, data from Bangladesh Integrated Household Survey (BIHS) 2011-2012 (Ahmed, 2013) is used to conduct asset-based poverty analysis for rural Bangladesh using two distinct methods: 1) Principal component analysis followed by cluster analysis, and 2) Fuzzy set theory. The BIHS collected demographic and socio-economic data from 6503 households that are nationally representative of rural Bangladesh. It contains detailed information on households' access and ownership to a wide range of assets, including education, consumptive and productive assets, savings and loans, land and water bodies, livestock and poultry, water supply and sanitation facilities, food consumption and housing conditions. In both PCA and FST, 22 variables grouped under eight dimensions have primarily been used for the analysis.

### 5.1 PCA and cluster analysis

In order for the PCA to produce meaningful results, the selected asset variables should be correlated to certain extent and the distribution of variables should vary across households (Vyas and Kumaranayake, 2006). Asset variables which have zero variance (for example, none of the households own a motorcycle/car) will have no role in differentiating between poverty levels and hence, should be eliminated. Similarly, assets that are owned/ not owned by a major proportion of households (for example, 70% of households do not have a bicycle/rickshaw/van or 79% do not have fishing nets) can lead to clumping and hence, should be avoided. Some categorical variables such as occupation, religion and gender are not suitable for PCA as the categories are translated to a quantitative scale which has no meaning. Moreover, it should be noted that variables which show a high degree of variance across households will generate the highest weights or factor loadings in PCA. Thus, the key is to include variables that capture inequality in households. To avoid these issues, it is best to examine the descriptive statistics of each variable prior to their inclusion in the PCA.

In most studies, the first component scores for each household are treated as indicators of socio-economic status. A higher score refers to a greater level of well-being. Unlike other studies (Filmer and Pritchett, 2001), which disaggregated households into socio-economic status based on quintiles, in this study a k-means cluster analysis is used to differentiate between four different levels of poverty. In this case, the factor scores from $PC_1$ are used as inputs for the cluster analysis. The mean asset ownership of each cluster is calculated and the results show internal coherence, meaning that mean asset ownership varies among socio-economic groups. For example, if the ownership of almost all assets increases with socio-economic status, it indicates that the cluster analysis produced conceptually meaningful results.

The factor loadings generated are highly sensitive to the variables included, the number of factors extracted as well the numerical codes used for ordinal variables. In this study, different PCA and cluster analysis have been carried out using different sets of variables to find out which set of variables result in the most coherent and reliable categorisation of households by socio-economic groups. For simplicity, the results of only two of these PCAs have been included in this paper (Table 2). In the first PCA (PCA1), 22 variables under eight dimensions are included; these exact variables have also been used in FST (see section 5.2). The second PCA (PCA2) includes only those variables that produced the best results among all the iterations.

*Table 2. Indicators and number of factors extracted for each PCA*

| Dimensions | Variables | PCA 1 (22 variables; same as those included in FST) | PCA 2 (Only retaining variables that produce the best results) |
|---|---|---|---|
| Education | Highest education of any member | X | X |
| Durable assets | Quantity of television | X | |
| | Quality of TV (B&W/Coloured) | | X |
| | Mobile phones currently in use | X | X |
| | Number of tools | X | X |
| | Number of furniture items | X | X |
| Land | Homestead land (decimals) | X | X |
| | Arable land (decimals) | X | X |
| Livestock or poultry | Total livestock unit (1 cow = 1 LSU, 3 goats = 1 LSU) | X | X |
| | Poultry | X | |
| Housing materials | Wall material | X | X |
| | Roof material | X | X |
| | Floor material | X | X |
| Energy sources | Electricity connection | X | |
| | Cooking fuel | X | |

| | | | |
|---|---|---|---|
| | Lighting fuel | X | |
| Water and sanitation | Water for domestic purposes | X | |
| | Water for drinking | X | |
| | Latrine type | X | |
| Food consumption | Big fish consumption per week (kg) | | X |
| | Small fish consumption per week (kg) | | X |
| | Total fish consumption per week (kg) | X | |
| | Total meat consumption per week (beef/chicken) (kg) | X | X |
| | No. of eggs consumed per week | X | X |
| | Milk consumed per week (kg) | X | X |
| | Parboiled rice consumption per week (kg) | | X |
| | Non-parboiled rice consumption per week (kg) | | X |
| | Fine rice consumption per week (kg) | | X |

| Number of factors extracted | All factors with eigenvalue>1 (6 factors) | 4 factors | All factors with eigenvalue>1 (6 factors) |
|---|---|---|---|
| KMO statistic | 0.818 | 0.731 | 0.731 |
| Total variance accounted for | 22.6% + 8.6% + 7.2% + 5.5% + 5.1% + 4.7% = 53.8% | 20.1% + 9.1% + 7.7% + 6.1% = 42.9% | 20.1% + 9.1% + 7.7% + 6.1% + 6.0% + 5.3% = 54.3% |
| Percentage of households in different clusters (Poorest to richest group) | 23.1%<br>29.7%<br>22.9%<br>24.2% | 30.9%<br>44.1%<br>18.2%<br>6.7% | 50.6%<br>34.4%<br>13.0%<br>1.7% |
| Remarks | Cluster analysis shows similar percentages of households in four categories. | Cluster analysis shows reasonable percentage of households in four categories. | Cluster analysis categorises most households into the bottom two categories. |
| | $PC_1$ shows very poor correlations with other forms of classification (0.133 with consumption expenditures, 0.199 with deprived households as per CL method, and 0.135 with deprived households as per Ragin's method). | $PC_1$ shows relatively good correlations with other forms of classification (0.519 with consumption expenditures, 0.713 with deprived households as per CL method, and 0.653 with deprived households as per Ragin's method). | $PC_1$ shows poor correlations with other forms of classification (0.232 with consumption expenditures, 0.408 with deprived households as per CL method, and 0.509 with deprived households as per Ragin's method). |
| | Mean assets do not show internal coherence for most variables. | Mean assets show internal coherence, except for no. of goats. | Mean assets show internal coherence, except for housing, TV and cooking/lighting fuel. |

Based on the results shown in Table 2, PCA2 involving the four factors is chosen for the purpose of this study. Firstly, this PCA produce a reasonable categorisation of households into the four clusters and the mean ownership of assets by cluster shows higher coherence compared to the other PCAs. Secondly, in PCA2 all relevant asset variables are included; the exclusion of utilities is justified as access to water, sanitation, cooking and lighting fuel, and electricity do not show wide variation among households (see Table 4), leading to possible clumping. Thirdly, when the total non-food consumption expenditures (Figure 1) (not included in the PCAs) are correlated to the first principal component scores ($PC_1$), PCA2 shows the highest level of correlation (Pearson's two tailed correlation coefficient of 0.514, which is significant at p=0.01) compared to the others. Table 3 shows how the mean ownership of assets changes across socio-economic status.

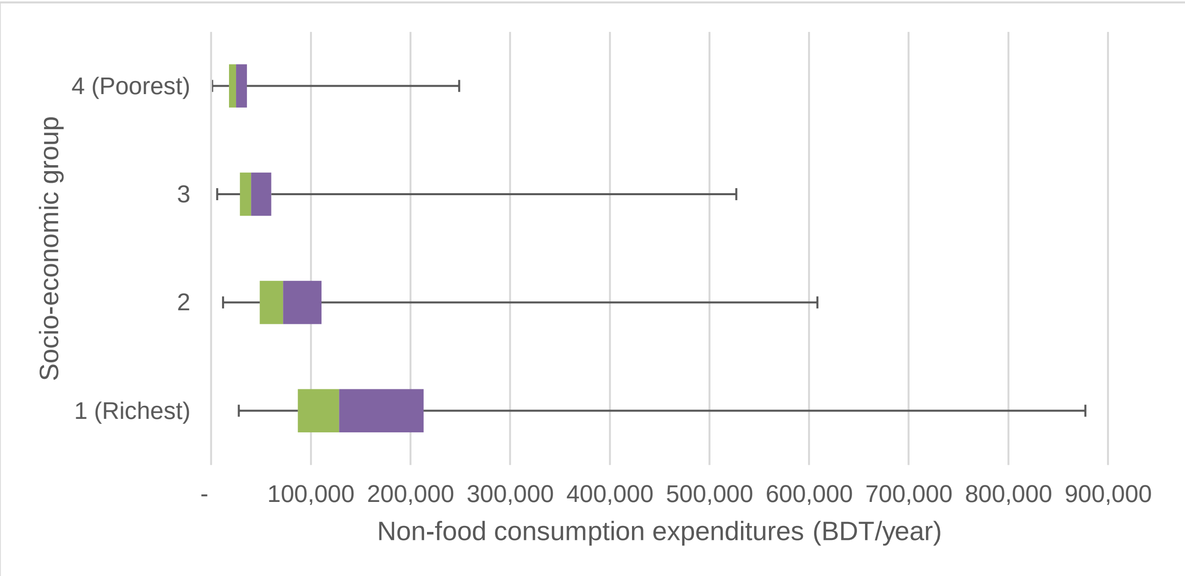


*Figure 1. Box and whisker plots showing the annual non-food consumption expenditures by socio-economic clusters (USD 1 = BDT 80)*

*Table 3. Mean asset ownership by socio-economic cluster**

| **Socio-economic Status** | | **Poorest** | → | | **Richest** |
|---|---|---|---|---|---|
| **Cluster centres based on $PC_1$** | | **-1.0003** | **-0.0653** | **0.9702** | **2.3970** |
| **Percentage (Number) of households** | | **30.9% (2011)** | **44.1% (2871)** | **18.2% (1182)** | **6.7% (439)** |
| Education | Highest education of any member (0 = never attended school, 1 = in/passed primary school, 2 = in/passed secondary school, 3 = in/passed SSC exams, 4 = in/passed HSC exams, 5 = Doing/ completed bachelor's degree, 6 = Doing/ completed professional or master's degree) | 1.05 | 1.92 | 2.87 | 3.54 |
| Employment | Common occupation of Household Head | Agri-labour (23%); Working in own farm (13%); share cropper (11%); | Working in own farm (25%); share cropper (12%); agri-labour (10%) | Working in own farm (28%); Small/ medium trader (14%) | Working in own farm (24%); Medium/ large trader (19%) |
| Durable assets | Number of TV | 0.06 | 0.22 | 0.50 | 0.83 |
| | Mobile phones currently in use | 0.40 | 0.95 | 1.52 | 2.13 |
| | No. of tools | 2.31 | 2.86 | 3.57 | 4.29 |
| | No. of furniture items | 2.83 | 5.58 | 9.37 | 14.71 |
| Land and water bodies | Homestead land (decimals) | 7.3 | 9.6 | 14.3 | 19.8 |
| | Arable land (decimals) | 37.3 | 67.3 | 103.2 | 167.3 |
| Livestock or poultry | No. of bullock/milk cow/buffalo | 0.96 | 1.07 | 1.01 | 1.14 |
| | No. of goat/sheep | 0.55 | 0.47 | 0.40 | 0.39 |
| | No. of hen/duck/other birds | 4.00 | 6.16 | 8.57 | 11.54 |
| Housing materials | Condition of dwelling (1 = Extremely damaged, 2 = Very damaged, 3 = Somewhat damaged, 4 = Slightly damaged, 5 = No damage) | 2.59 | 3.16 | 3.67 | 4.22 |
| | Wall material (1 = Golpata/ palm leaves/ grass/ straw/ cardboard/ plastic, 2 = Jute sticks, 3 = Bamboo, 4 = Mud or unfired brick, 5 = Wood, 6 = Tin/ corrugated iron, 7 = Concrete/ brick) | 3.73 | 5.43 | 6.10 | 6.71 |
| | Roof material (1 = Golpata/ palm leaves/ grass/ straw/ cardboard/wood, 2 = Plastic sheets, 3 = Tin/ corrugated iron, 4 = Concrete/ brick) | 2.68 | 2.99 | 3.05 | 3.30 |

| | | | | | |
|---|---|---|---|---|---|
| | Floor material<br>(1 = Mud or unfired brick, 2 = Wood, 3 = Concrete/ brick) | 1.00 | 1.02 | 1.54 | 2.69 |
| Energy sources | Electricity connection | | | | |
| | Cooking fuel<br>(1 = saw dust/ leaves/ others, 2 = Cow dung/ coal, 3 = Firewood, 4 = Electricity/ supply gas/ lpg/ kerosene) | 1.92 | 2.13 | 2.42 | 2.66 |
| | Lighting fuel<br>(1 = Kerosene/ candles, 2 = Solar energy, 3 = Electricity/ private generators) | 1.47 | 2.01 | 2.48 | 2.75 |
| Water and sanitation | Water for domestic purposes<br>(1 = Pond or river water, 2 = Community TW/ rainwater, 3 = Own TW, 4 = Supply water) | 1.92 | 2.13 | 2.29 | 2.61 |
| | Water for drinking<br>(1 = Pond or river water, 2 = Community TW/ rainwater, 3 = Own TW, 4 = Supply water) | 2.13 | 2.38 | 2.57 | 2.79 |
| | Latrine type<br>(1 = Open defecation, 2 = Kacha, 3 = Community latrine/other, 4 = Pakka, 5 = Sanitary without flush, 6 = Sanitary with flush) | 3.36 | 3.76 | 4.22 | 4.67 |
| Food consumption | Big fish consumption per week (kg) | 0.37 | 0.75 | 1.30 | 1.96 |
| | Small fish consumption per week (kg) | 0.47 | 0.56 | 0.69 | 0.97 |
| | Meat consumption per week (kg) | 0.22 | 0.52 | 1.17 | 1.79 |
| | No. of eggs consumed per week | 1.47 | 3.19 | 5.63 | 9.87 |
| | Milk consumed per week (kg) | 0.35 | 0.64 | 1.07 | 1.64 |
| | Parboiled rice consumption per week (kg) | 9.03 | 10.74 | 10.38 | 8.94 |
| | Non-parboiled rice consumption per week (kg) | 2.24 | 1.44 | 1.96 | 3.54 |
| | Fine rice consumption per week (kg) | 0.71 | 1.27 | 1.97 | 3.25 |

*Note: This table shows more variables than those included in the PCAs, to illustrate how asset ownership varies by socio-economic cluster.

## 5.2 Fuzzy set theory

Similar to PCA1, a total of 22 indicators grouped into 8 main dimensions have been included in the FST (Table 4). The calibration of indicator values into membership scores has been performed separately using two different methods: 1) Ragin's direct method and 2) Cheli and Lemmi's frequency based method [5] and for each method, different weighing systems have been used for aggregating the variables. The purpose of using two methods of calibration and aggregation is to demonstrate the extent to which the results differ according to the method used.

### Ragin's method

For calibration using Ragin's method, the different categories for each indicator have been arranged in increasing order of socio-economic status and the percentage of households belonging to each category has been noted (Table 4). Information about the frequency distribution of each indicator is used to decide on the three qualitative anchors of Ragin's direct method. For each indicator, the category in which the lowest 5% of the households belonged to is used as the 0.95 breakpoint, the category in which the lowest 50% of the households belonged to is assigned as the 0.5 cross-over point and the category in which the highest 5% of the households belonged to is selected as the 0.05 non-membership threshold value. For example, as 14.6% of households own between 0-2 decimals of homestead land, choosing 2 decimals as the anchor for full membership gives a value of 0.95 to those owning 2 decimals, 0.97 to those having 1 decimal and 0.99 to those with no homestead land. This method of selecting qualitative anchors is a data-driven method that avoids arbitrariness in selecting cut-offs. Moreover, unlike PCA, irrelevant variation is not given much importance. For instance, although the highest amount of agricultural land ownership is 2380 decimals, those with land over 271 decimals (that is, 5.3% of all households) have been assigned fuzzy values of 0.05 or less. The software fs/QCA Version 2.5 (Ragin and Davey, 2012) is used for calibration and aggregation using Ragin's method. While calibrating using this software, the variables should be coded such that increasing number indicates less deprivation.

After calibration, the next step is aggregating the individual indicators under each of the eight dimensions. The variables that have been calibrated using Ragin's method are aggregated using different operators and weights as mentioned in Table 4. Indicators such as livestock, durable assets, housing materials and protein consumption are aggregated using strong intersection, so that the minimum of the fuzzy scores of the variables are selected. In other words, the household is assigned the membership score of the variable for which it is least deprived. For example, as one type of protein can be compensated by other types, as long as a household is consuming good quantity of one of the protein foods, it is considered as less deprived. In case of other dimensions such as land, energy source and water sanitation, where one variable cannot totally compensate for another, a weighted average has been calculated. The weights were calculated using CL formula [9].

Table 6 shows the percentage of households deprived (fuzzy scores ≥0.5) and extremely deprived (fuzzy scores ≥0.95) in each dimension. The fuzzy scores of the eight dimensions are not aggregated further; rather the number of variables for which each household is deprived (Figure 3) and extremely deprived (Figure 4) is found. Based on the number of deprived variables, each household is assigned a pre-determined poverty membership score.

*Table 4. Indicators, membership scores assignment criteria, and aggregating operators used for FST using Ragin's method of calibration*

| Dimension (aggregating operators) | Indicators (qualitative anchors - 0.95,0.5,0.05) | Categories [Codes] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Increasing socio-economic status → | | | | | | | |
| Education<br><br>(none; only one indicator) | Education (0,2,4) | Never attended school<br><br>[0] | Completed primary school (upto grade 5) [1] | Completed secondary school (upto grade 8)<br><br>[2] | Completed SSC (upto grade 10)<br><br>[3] | Completed HSC (upto grade 12)/ vocational training<br>[4] | Doing/ completed bachelor's degree<br><br>[5] | Doing/ completed professional or master's degree<br><br>[6] | |
| | | 9.2% | 34.0% | 24.7% | 22.7% | 6.4% | 1.9% | 0.9% | |
| Land<br><br>(weighted average) | Homestead land (2,7,30) | 0-2 | 3-5 | 6-8 | 9-12 | 13-19 | 20-30 | >30 | |
| | | 14.6% | 28.2% | 17.7% | 15.7% | 10.6% | 8.2% | 5.1% | |
| | Agricultural land (0,30,271) | 0 | 1-20 | 21-40 | 41-60 | 61-90 | 91-140 | 141-270 | >271 |
| | | 37.6% | 7.8% | 10.3% | 8.0% | 9.7% | 10.4% | 10.8% | 5.3% |
| Livestock/ poultry<br><br>(strong intersection) | Livestock (0,1,5) | 0 | | 1 | 2 | 3-4 | | 5 and above | |
| | | 49.3% | | 15.8% | 15.1% | 14.6% | | 5.2% | |
| | Poultry (0,2,20) | 0 | | 1-3 | 4-8 | 9-19 | | 20 and above | |
| | | 31.1% | | 21.5% | 24.2% | 18.0% | | 5.2% | |
| Durables<br><br>(strong intersection) | Quantity of tools (0,3,5) | 0 | 1 | 2 | | 3 | 4 | 5 and above | |
| | | 6.7% | 18.6% | 22.4% | | 18.5% | 17.5% | 7.5% | |
| | Number of furniture items (1,5,13) | 0-1 | 2-4 | 5-7 | | 8-12 | | 13 and above | |
| | | 12.8% | 29.8% | 28.1% | | 21.3% | | 8.0% | |
| | Cellular phone (0,1,2) | 0 | | 1 | | 2 | | 3 and above | |
| | | 27.4% | | 53.6% | | 15.2% | | 3.7% | |
| | Television (0,1,2) | 0 | | 1 | | 2 and above | | | |
| | | 74.7% | | 24.3% | | 1% | | | |
| Housing | Wall material | Grass/ | Jute stick [2] | Bamboo | Mud/ unfired brick | Wood [5] | Tin/ | Concrete/ brick [7] | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (strong intersection) | (2,6,7) | straw/ leaves [1] | | [3] | wall [4] | | Corrugated iron sheets [6] | |
| | | 3.3% | 5.4% | 10.0% | 16.5% | 2.6% | 46.5% | 15.5% |
| | Roof material (1,2,4)* | Grass/ straw/ leaves [1] | | Plastic sheets [2] | | Tin/ Corrugated iron sheets [3] | | Concrete/ brick [4] |
| | | 5.1% | | 0.3% | | 91.4% | | 3.2% |
| | Floor material (1,2,3) | Mud/ unfired brick [1] | | | | Wood [2] | | Concrete/ brick [3] |
| | | 88.8% | | | | 0.1% | | 11.1% |
| Energy/ Fuel (weighted average) | Electricity connection (1,1.5,2) | No [1] | | | | Yes [2] | | |
| | | 53.2% | | | | 46.8% | | |
| | Cooking fuel (1,2,3) | Saw dust/ dried leaves [1] | | Dried cow dung/Coal [2] | | Firewood [3] | | Kerosene/Supply gas/ LPG/ Electricity [4] |
| | | 33.7% | | 20.7% | | 42.2% | | 3.3% |
| | Lighting fuel (1,2,3) | Kerosene/ candles [1] | | Solar energy [2] | | | Electricity/ private generator [3] | |
| | | 48.7% | | 4.8% | | | 46.5% | |
| Water and sanitation | Toilet (2,4,5) | Open field [1] | Kacha [2] | Community shared latrine [3] | | Pakka (unsealed) [4] | Sanitary (without flush) [5] | Sanitary (with flush) [6] |
| (strong intersection for aggregating the two water indicators, followed by weighted average for toilet and both water) | | 3.0% | 19.4% | 2.3% | | 47.6% | 27.2% | 0.5% |
| | Drinking water (1,2,3) | Pond/river water/ others [1] | | Community tubewell/ Rain water [2] | | Own tubewell [3] | | Supply water [4] |
| | | 14.1% | | 36.9% | | 46.9% | | 1.9% |
| | HH water (1,2,3) | Pond/river water/ others [1] | | Community tubewell/ Rain water [2] | | Own tubewell [3] | | Supply water [4] |
| | | 33.7% | | 21.2% | | 43.4% | | 1.6% |
| Protein consumption | Meat (0,0.5,2.8) | 0 | | 0.1-1.1 | | 1.2-2.8 | | Above 2.8 |
| | | 62.0% | | 18.0% | | 12.7% | | 5.3% |
| (strong | Eggs (0,2,11) | 0 | | 1-3 | 4-6 | 7-11 | | Above 11 |
| | | 44.6% | | 14% | 20.5% | 13.6% | | 7.3% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| intersection) | Milk (0,0.5,3.2) | 0 | 0.1-2.0 | | 2.1-3.2 | | Above 3.2 |
| | | 69.7% | 19.3% | | 2.9% | | 8.1% |
| | Fish (0,1,4) | 0 | 0.01-0.49 | 0.50-0.99 | 1.00-1.99 | 2.00-3.99 | Above 4.00 |
| | | 12.2% | 10.7% | 19.9% | 30.4% | 20.6% | 6.2% |

*As tin/corrugated iron falls both within the lowest 50% and the highest 5% marks, the qualitative anchors have been modified for roofing material. Similar issues have been encountered in case of floor materials and TV.

### CL method

For calibration using CL frequency based method, the formulae given in [4a and 4b] have been used. In contrast to Ragin's method, in this case, the variables should be coded such that increasing number indicates more deprivation.

For those that have been calibrated using CL method, the frequency based weighting system [9] has been used to aggregate the variables into the 8 dimension (greater weight is assigned to those indicators in which lower numbers of households are deprived). For example, as most households (88.8%) have a mud/unfired brick floor (that is, the lowest category of flooring material), it carries the smallest weight compared to all other indicators in Table 5.

*Table 5. Weights for each of the 22 variables used in CL method*

| Dimensions | Variables | Weights |
|---|---|---|
| Education | Education | 0.874 |
| Land | Homestead land | 0.756 |
| | Agricultural land | 0.377 |
| Livestock/ poultry | Livestock | 0.375 |
| | Poultry | 0.460 |
| Durable assets | Tools | 0.806 |
| | Furniture | 0.744 |
| | Cellular phone | 0.863 |
| | Television | 0.279 |
| Housing materials | Wall | 0.992 |
| | Floor | 0.118 |
| | Roof | 2.501 |
| Energy source/ fuel | Electricity | 0.760 |
| | Cooking fuel | 0.691 |
| | Lighting fuel | 0.632 |
| Water and sanitation | Toilet | 1.067 |
| | Drinking water | 0.711 |
| | HH water | 1.014 |
| Protein consumption | Meat | 0.249 |
| | Eggs | 0.370 |
| | Milk | 0.184 |
| | Fish | 0.622 |

For final poverty analysis, the same procedure applied in Ragin's method (that is, aggregating score by dimensions and then categorising households into 'deprived' and 'extremely deprived' based on the number of dimensions for which scores are ≥0.50 and ≥0.95 respectively) has been used to enable comparison, as shown in Figure 3 and Figure 4.

### Comparing fuzzy results from Ragin's and CL methods

In terms of calibration of variables, the two methods do not show much difference. However, if fuzzy scores of ≥0.50 are taken as deprivation cut-off lines, Ragin's method turns out to be more stringent (that is, categorises more households as deprived) than CL method in case of education and durable asset variables while the latter seems more stringent in case of livestock/poultry and protein consumption variables (Table 6).

*Table 6. Percentage of households 'deprived (≥0.50)' and 'extremely deprived (≥0.95)' in each dimension*

| Dimensions | Ragin's method | | CL method | |
|---|---|---|---|---|
| | **Deprived** | **Extremely deprived** | **Deprived** | **Extremely deprived** |
| Education | 67.9 | 9.2 | 43.2 | 9.2 |
| Land ownership | 61.5 | 9.1 | 56.7 | 3.0 |
| Livestock/poultry | 34.4 | 20.0 | 75.0 | 18.0 |
| Housing materials | 4.9 | 1.1 | 5.5 | 2.4 |
| Durable assets | 39.2 | 1.2 | 43.0 | 1.2 |
| Water and sanitation | 42.4 | 2.3 | 35.4 | 0.3 |
| Protein consumption | 24.9 | 4.7 | 75.3 | 10.4 |
| Energy sources | 42.1 | 0.0 | 54.3 | 0.7 |

For example,

Figure 2 shows the fuzzy membership scores for number of mobile phones based on calibration by Ragin's and CL methods. Ragin's method assigns a score of ≥0.50 to households owning ≤1 phone, hence, categorising 81% of households as deprived. The CL method assigns a score of ≥0.50 only for 0 phone ownership, resulting in only 27% of households being identified as deprived. Moreover, when the scores for mobile phone and TV are aggregated in the CL method, greater weight is placed on mobile phone, resulting in a further decrease in the number of households being categorised as deprived.

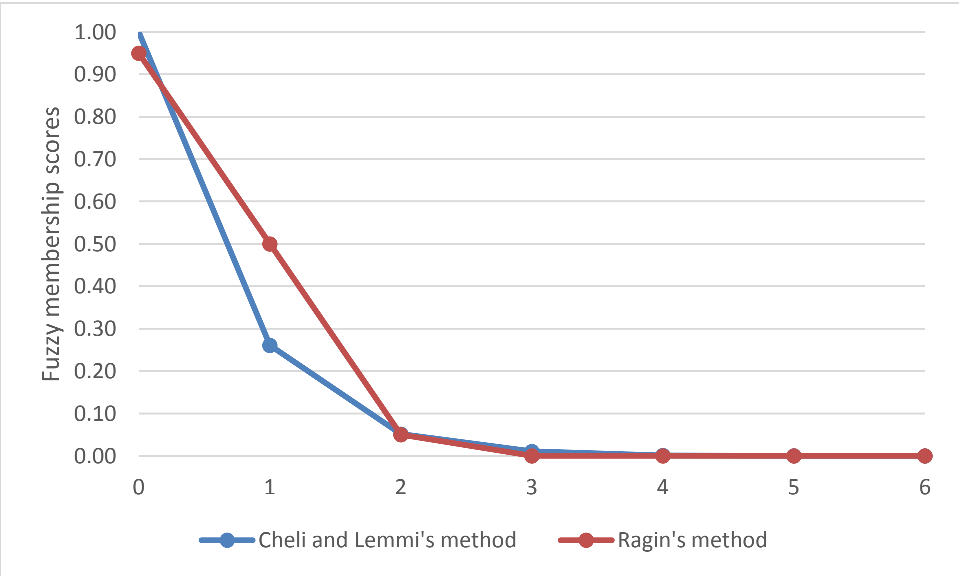


*Figure 2. Fuzzy set membership scores for number of mobile phones by Ragin's and CL method*

In case of the livestock dimension (comprising of cattle/goat and poultry as variables), Ragin's method gives a score of ≥0.50 to households owning ≤2 hens/ducks, while in CL method assigns the same scores for owning ≤5 hens/ducks. For the cattle/goat variable, both methods assign scores of ≥0.50 to households owning ≤1 animal. Moreover, during aggregation CL method gives greater weight to poultry than cattle/goats, leading to more households being classified as deprived. In comparison, while aggregating using Ragin's method, the minimum score is selected, resulting in less deprivation score. Similar issues are observed in case of the protein consumption dimension.

Figure 3 and Figure 4 show fuzzy membership scores for 'deprived' and 'extremely deprived' households based on the number of dimensions on which the membership scores are ≥0.50 and ≥0.95 respectively. In this case, membership scores have been pre-determined. For example, in Figure 3 a score of 0.00 indicates households that are not deprived on any of the dimensions, while 1.00 indicates deprivation on all eight dimensions. The final membership scores are relatively similar regardless of the method of calibration and aggregation. However, in order to investigate whether the same households have received the same scores, the correlation co-efficient between the 'deprived' fuzzy scores of the two methods is obtained. The Pearson's two tailed correlation coefficient is 0.822, which is significant at p=0.01. Moreover, the correlation coefficients between the fuzzy scores of the two methods and the factor score of PCA2 have been found. The results show similar correlations of 0.653 between Ragin's fuzzy scores and PCA2 scores, and 0.713 between CL's fuzzy scores and PCA2 scores.

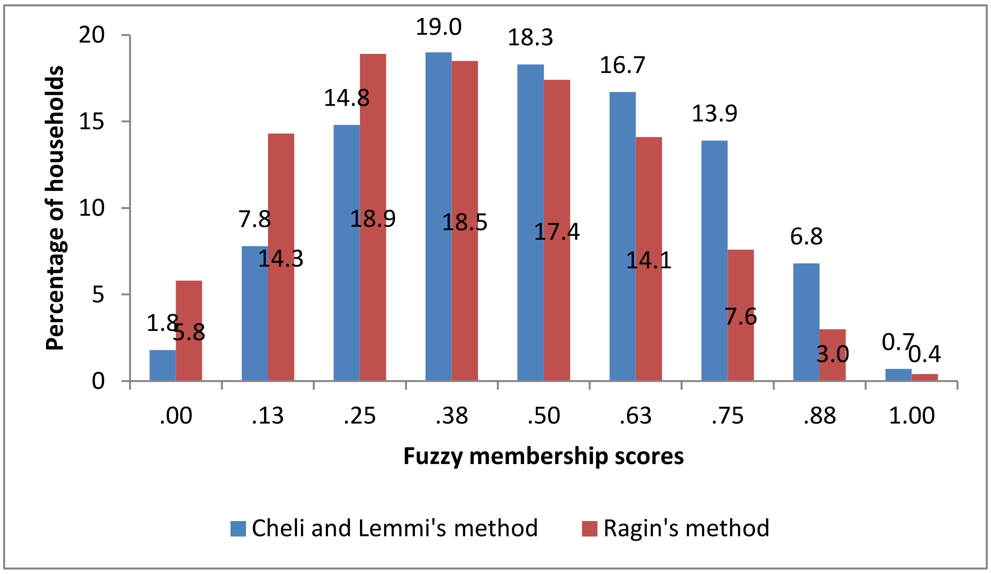


*Figure 3. Fuzzy membership scores for 'deprived' households based on the number of dimensions on which the membership score is ≥0.50 (0.00 = deprived on none of the dimensions; 1.00 = deprived on all 8 dimension)*

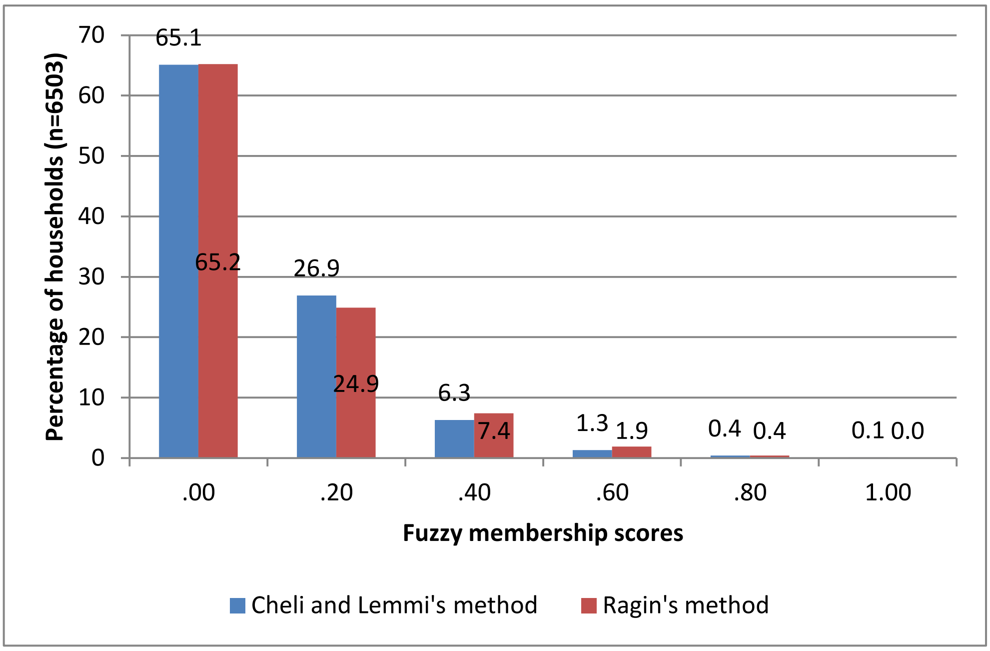


*Figure 4. Fuzzy membership scores for 'extremely deprived' households based on the number of dimensions on which the membership score is ≥0.95 (0.00 = extremely deprived on none of the dimensions; 1.00 = extremely deprived on at least 5 dimensions)*

## 6. Discussion

In order to enhance comparison of results from PCA and FST, the initial aim was to include 22 variables grouped under eight dimensions. However, PCA1, which used these variables, produced in poor results when cluster analysis was used to classify households. This is because none of the variables, other than quantity of TV, electricity connection and lighting fuel, were loaded onto the first principal component ($PC_1$) (see the rotated component matrix of PCA1 in Table 9 of appendix). Variables under the education and durables dimensions, the land and livestock dimensions, the food consumption and the housing materials dimension were loaded on $PC_2$, $PC_3$, $PC_4$ and $PC_6$ respectively. This highlights that using the $PC_1$ scores for poverty assessment may not always lead to reliable results as important variables may have lower loadings on $PC_1$. Authors who have had good results using $PC_1$ scores in previous studies cited in section 3.2, have restricted the number of variables to more homogenous ones. Alternatively, if the scores of all the six principal components are included in cluster analysis, the households are categorised into just two groups.

In case of PCA2 with four factors, most of the important variables were loaded onto $PC_1$, and those that were loaded onto other components at least produced second highest loadings on $PC_1$ (see the rotated component matrix of PCA2 in Table 11 of appendix). Hence, PCA2 produced the best results among all possible trials, including those that have not been shown in this paper. Even in PCA2, extraction of 6 factors leads to lower loadings on $PC_1$ for most important variables (see the rotated component matrix of PCA1 in Table 13 of appendix).

Since the variables in this study are not highly correlated to each other, inclusion of more variables requires more factors to be extracted in order for the PCA to account for greater variance in the data. Consequently, less number of variables are loaded onto $PC_1$, making the results of cluster analysis based on $PC_1$ less reliable. Thus, a trade-off has to be made between the number of variables to be included, the number of factors to be extracted and the amount of variance accounted for. It is also important to understand how the inclusion or exclusion of one additional variable can lead to significantly different results in PCA. For instance, excluding the number of tools and furniture items in PCA1, or just replacing the quality of TV (B&W or colour) with the quantity of TV, results in entirely different outcomes. However, PCA is useful when trying to investigate the number of relevant dimensions in the data-set and each factor can be considered as representative of a given dimension.

Contrarily, FST does not suffer from many of the limitations of PCA described above. In FST, inclusion/exclusion of one variable does not affect the calibration results of another variable, but affects the aggregated results to some extent. Among the two FST methods applied in this study, Ragin's method provides more flexibility in terms of allowing the researcher to assign cut-off points based on theoretical or contextual understanding and to choose suitable aggregating operators. The CL method, based on frequency distribution, is more rigid, and temporal comparisons may not be made using longitudinal datasets, as each data-set will have different frequency distribution. However, the process of generating weights under the CL method proves to be useful when aggregating variables.

The methodological advantages and disadvantages of PCA and FST do not necessarily suggest the superiority of one method over another. Depending on the data-set and the research objective, both methods can prove to be useful once the technical considerations are taken into account. Moreover, the methods can also be integrated; for example, PCA can be used to generate weights or identify the dimensions for FST, while weights obtained from CL method can be used to aggregate variables calibrated by Ragin's method.

## 7. Conclusion

This paper aimed to investigate the extent to which the results of asset-based multi-dimensional poverty assessment depend on the method of data analysis. In doing so, two distinct methods, principal component analysis and fuzzy set theory, have been applied to household survey data from rural Bangladesh. Moreover, within each method, the procedures are varied to capture their effect on the final outcomes. Two different PCAs have been conducted using different sets of variables and changing the number of factors extracted. In FST, Ragin's and Cheli and Lemmi's methods of calibration and aggregation have been used. The findings from the multi-dimensional poverty assessment in this study, based on different forms of PCA and FST, have revealed the following issues.

1. Both PCA and FST can lead to reliable results in terms of poverty analysis; however, as changes in procedures (such as the variables used, the number of

factors extracted and the type of cluster analysis applied, in case of PCA or the calibration and aggregation methods used in case of FST) can lead to different results, the outputs need to be checked with some form of external data to ensure validity.

2. In PCA, selection of variables is a very important step. While some variables, such as electricity, water and sanitation, are conceptually very important in poverty assessment, inclusion of these variables may lead to clumping, as these services are often not within the control of the household and are shared. Similarly, variables which are contextually more important, such as agricultural land in rural Bangladesh, may gain less weight than other factors.
3. Compared to PCA, FST provides more flexibility, in terms of involving conceptual and theoretical inputs, and also leads to outputs that are easier to visualise and interpret.
4. While both methods of FST have strong conceptual foundations, the CL method is totally data-driven, which can often give more importance to certain variables (such as ownership of poultry over housing materials) that may not be as important in poverty determination. However, the frequency-based weighing system if very useful if applied with care, as has been done in the aggregation of variables into land or energy dimensions in Ragin’s method.
5. PCA and CL method may not be appropriate when making temporal or spatial comparisons, because as these methods are data-driven, the same category/variable can have different scores in different times or sites. In this case, Ragin’s calibration method may be more useful, as the qualitative anchors can be kept same.

## 8. References


AHMED, A. 2013. Bangladesh Integrated Household Survey (BIHS) 2011-2012. Washington, D.C.: International Food Policy Research Institute.

ALKIRE, S. & SANTOS, M. E. 2010. Acute Multidimensional Poverty: A New Index for Developing Countries Oxford Poverty and Human Development Initiative (OPHI).

BERMAN, R. J., QUINN, C. H. & PAAVOLA, J. 2014. Identifying drivers of household coping strategies to multiple climatic hazards in Western Uganda: implications for adapting to future climate change. *Climate and Development*, 1-14.

CERIOLI, A. & ZANI, S. 1990. A fuzzy approach to the measurement of poverty. *Income and wealth distribution, inequality and poverty.* Springer.

CHELI, B. & LEMMI, A. 1995. Totally fuzzy and relative approach to the multidimensional analysis of poverty. *Economic Notes,* 24**,** 115–134.

CHIAPPERO MARTINETTI, E. 1994. A new approach to evaluation of well-being and poverty by fuzzy set theory. *Giornale degli economisti e annali di economia*, 367-388.

CHIAPPERO MARTINETTI, E. 2000. A multi-dimensional assessment of well-being based on Sen's functioning theory. *Revista Internazionale di Scienza Soziali,* 58**,** 207-239.

CHIAPPERO MARTINETTI, E. 2006. Capability approach and fuzzy set theory: description, aggregation and inference issues. *In:* LEMMI, A. & BETTI, G. (eds.) *Fuzzy set approach to multidimensional poverty measurement.* Berlin: Springer.

DECANCQ, K. & LUGO, M. A. 2012. Weights in Multidimensional Indices of Wellbeing: An Overview. *Econometric Reviews,* 32**,** 7-34.

DESAI, M. & SHAH, A. 1988. An econometric approach to the measurement of poverty. *Oxford Economic Papers,* 40**,** 505-522.

FILMER, D. & PRITCHETT, L. H. 2001. Estimating Wealth Effects Without Expenditure Data—Or Tears: An Application To Educational Enrollments In States Of India*. *Demography,* 38**,** 115-132.

HAQ, M. U. 1995. *Reflections on Human Development,* New York, Oxford University Press.

HOUWELING, T. A., KUNST, A. E. & MACKENBACH, J. P. 2003. Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter? *International Journal for Equity in Health,* 2**,** 8.

ISLAM, M. M. 2013. *Vulnerability and Adaptation of Fishing Communities to the Impacts of Climate Variability and Change: Insights from Coastal Bangladesh.* PhD in Earth and Environment, University of Leeds.

KLASEN, S. 2000. Measuring poverty and deprivation in South Africa. *Review of income and wealth,* 46**,** 33-58.

LELLI, S. 2001. *Factor Analysis vs. Fuzzy Sets theory: Assessing the influence of different techniques on sen's functioning approach*, Katholieke Universiteit Leuven.

MCKENZIE, D. J. 2005. Measuring Inequality with Asset Indicators. *Journal of Population Economics,* 18**,** 229-260.

MONTGOMERY, M. R., GRAGNOLATI, M., BURKE, K. A. & PAREDES, E. 2000. Measuring Living Standards with Proxy Variables. *Demography,* 37**,** 155-174.

MOOI, E. & SARSTEDT, M. 2011. *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics,* Heidelberg, Springer.

NARDO, M., SAISANA, M., SALTELLI, A., TARANTOLA, S., HOFFMAN, A. & GIOVANNINI, E. 2005. Handbook on constructing composite indicators: methodology and user guide. OECD publishing.

NEFF, D. 2013. Fuzzy set theoretic applications in poverty research. *Policy and Society,* In press.

NOORBAKHSH, F. 1998. The human development index: some technical issues and alternative indices. *Journal of International Development,* 10**,** 589-605.

NUSSBAUM, M. C. 2000. *Women and Human Development: A Study in Human Capabilities,* Cambridge, Cambridge University Press.

QIZILBASH, M. 2002. A note on the measurement of poverty and vulnerability in the South African context. *Journal of International Development,* 14**,** 757-772.

QIZILBASH, M. & CLARK, D. A. 2005. The capability approach and fuzzy poverty measures: an application to the South African context. *Social Indicators Research,* 74**,** 103-139.

RAGIN, C. & DAVEY, S. 2012. *fs/QCA [Computer Programme], Version 2.5* [Online]. Irvine, CA: University of California. Available: http://www.u.arizona.edu/~cragin/fsQCA/software.shtml [Accessed 02 February 2014].

RAGIN, C. C. 2000. *Fuzzy-set social science*, University of Chicago Press.

RAGIN, C. C. 2006. User's Guide to Fuzzy-Set/Qualitative Comparative Analysis 2.0. . Tucson, Arizona: Department of Sociology, University of Arizona.

RAGIN, C. C. 2008. *Redesigning social inquiry: Fuzzy sets and beyond*, Wiley Online Library.

ROCHE, J. M. 2008. Monitoring Inequality among Social Groups: A Methodology Combining Fuzzy Set Theory and Principal Component Analysis *Journal of Human Development,* 9**,** 427-452.

SALLU, S. M., TWYMAN, C. & STRINGER, L. C. 2010. Resilient or vulnerable livelihoods? Assessing livelihood dynamics and trajectories in rural Botswana. *Ecology and Society,* 15**,** 3.

SEN, A. 1985. *Commodities and Capabilities,* Amsterdam, North-Holland.

SEN, A. 1993. Capability and Well-being. *In:* NUSSBAUM, M. C. & SEN, A. K. (eds.) *The Quality of Life.* Oxford: Clarendon Press.

SEN, A. 1999. *Development As Freedom,* Oxford, Oxford University Press.

SEN, A. 2008. The economics of happiness and capability. *In:* BRUNI, L., COMIM, F. & PUGNO, M. (eds.) *Capability and Happiness.* New York: Oxford University Press.

SOMARRIBA, N. & PENA, B. 2009. Synthetic indicators of quality of life in Europe. *Social Indicators Research,* 94**,** 115-133.

UNDP 2013. The Rise of the South: Human Progress in a Diverse World. *Human Development Report.* New York: United Nations Development Programme (UNDP).

VYAS, S. & KUMARANAYAKE, L. 2006. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy and Planning,* 21**,** 459-468.

WOLFF, H., CHONG, H. & AUFFHAMMER, M. 2011. Classification, Detection and Consequences of Data Error: Evidence from the Human Development Index*. *The Economic Journal,* 121**,** 843-870.

ZADEH, L. A. 1965. Fuzzy sets. *Information and control,* 8**,** 338-353.

## APPENDIX

*Table 7. Descriptive statistics of the 22 variables under eight dimensions*

| Dimension | Variable* | Type | Minimum | Maximum | Mean | Std. Deviation | Skewness |
|---|---|---|---|---|---|---|---|
| Education | Highest education of any member | Ordinal | 0 | 7 | 1.93 | 1.24 | 0.70 |
| Durable assets | Cellular telephone in working condition | Scale | 0 | 6 | 0.96 | 0.79 | 0.88 |
| | Quantity of TV | Scale | 0 | 3 | 0.26 | 0.47 | 1.48 |
| | Quantity of tools | Scale | 0 | 13 | 2.86 | 1.86 | 0.92 |
| | Quantity of furniture | Scale | 0 | 39 | 6.04 | 4.51 | 1.50 |
| Land | Homestead land (decimals) | Scale | 0 | 187 | 10.41 | 12.52 | 4.45 |
| | Arable land (decimals) | Scale | 0 | 2380 | 71.30 | 119.99 | 4.95 |
| Livestock | Total livestock units | Scale | 0 | 25 | 1.27 | 1.73 | 2.06 |
| | Poultry | Scale | 0 | 1950 | 6.29 | 30.97 | 46.04 |
| Housing materials | Wall material | Ordinal | 1 | 7 | 5.11 | 1.63 | -0.87 |
| | Roof material | Ordinal | 1 | 4 | 2.93 | 0.49 | -2.90 |
| | Floor material | Ordinal | 1 | 3 | 1.22 | 0.63 | 2.47 |
| Energy sources | Electricity connection | Ordinal | 1 | 2 | 1.47 | 0.50 | 0.13 |
| | Cooking fuel | Ordinal | 1 | 4 | 2.15 | 0.93 | -0.06 |
| | Lighting fuel | Ordinal | 1 | 3 | 1.98 | 0.98 | 0.05 |
| Water and sanitation | Water for domestic purposes | Ordinal | 1 | 4 | 2.13 | 0.90 | -0.13 |
| | Water for drinking | Ordinal | 1 | 4 | 2.36 | 0.75 | -0.43 |
| | Latrine type | Ordinal | 1 | 6 | 3.78 | 1.15 | -0.80 |
| Protein consumption | Big fish consumption per week (kg) | Scale | 0 | 14 | 0.81 | 1.11 | 2.91 |
| | Meat consumption per week (kg) | Scale | 0 | 12 | 0.63 | 1.10 | 2.66 |
| | Eggs consumption per week | Scale | 0 | 100 | 3.55 | 5.18 | 4.02 |
| | Milk consumption per week (kg) | Scale | 0 | 26 | 0.69 | 1.50 | 3.69 |

*Codes for ordinal variables are provided in Table 3*

*Table 8. Correlation matrix of the 22 variables under eight dimensions*

| Variables | Education | No. of TV | Mobile phones | Tools | Furniture | Homestead land | Arable land | LSU | Poultry | Wall | Roof | Floor | Electricity | Cooking fuel | Lighting fuel | Domestic water | Drinking water | Latrine | Total fish | Meat | Eggs | Milk |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Education** | 1.00 | | | | | | | | | | | | | | | | | | | | | |
| **No. of TV** | 0.32 | 1.00 | | | | | | | | | | | | | | | | | | | | |
| **Mobile phones** | 0.46 | 0.38 | 1.00 | | | | | | | | | | | | | | | | | | | |
| **Tools** | 0.23 | 0.20 | 0.30 | 1.00 | | | | | | | | | | | | | | | | | | |
| **Furniture** | 0.48 | 0.43 | 0.52 | 0.40 | 1.00 | | | | | | | | | | | | | | | | | |
| **Homestead land** | 0.19 | 0.10 | 0.19 | 0.23 | 0.28 | 1.00 | | | | | | | | | | | | | | | | |
| **Arable land** | 0.24 | 0.16 | 0.22 | 0.29 | 0.34 | 0.29 | 1.00 | | | | | | | | | | | | | | | |
| **LSU** | 0.11 | 0.05 | 0.12 | 0.24 | 0.15 | 0.17 | 0.38 | 1.00 | | | | | | | | | | | | | | |
| **Poultry** | 0.06 | 0.05 | 0.09 | 0.06 | 0.09 | 0.06 | 0.07 | 0.05 | 1.00 | | | | | | | | | | | | | |
| **Wall** | 0.27 | 0.22 | 0.25 | 0.13 | 0.33 | 0.11 | 0.15 | 0.04 | 0.02 | 1.00 | | | | | | | | | | | | |
| **Roof** | 0.11 | 0.13 | 0.14 | 0.07 | 0.21 | 0.07 | 0.09 | 0.02 | 0.01 | 0.28 | 1.00 | | | | | | | | | | | |
| **Floor** | 0.27 | 0.31 | 0.28 | 0.13 | 0.41 | 0.13 | 0.14 | 0.00 | 0.02 | 0.35 | 0.21 | 1.00 | | | | | | | | | | |
| **Electricity** | 0.29 | 0.46 | 0.30 | 0.12 | 0.32 | 0.02 | 0.07 | -0.02 | 0.03 | 0.25 | 0.14 | 0.23 | 1.00 | | | | | | | | | |
| **Cooking fuel** | 0.18 | 0.15 | 0.16 | 0.08 | 0.19 | 0.02 | 0.01 | -0.05 | -0.02 | 0.15 | 0.04 | 0.16 | 0.17 | 1.00 | | | | | | | | |
| **Lighting fuel** | 0.30 | 0.48 | 0.33 | 0.13 | 0.35 | 0.04 | 0.10 | -0.01 | 0.03 | 0.28 | 0.16 | 0.25 | 0.95 | 0.18 | 1.00 | | | | | | | |
| **Domestic water** | 0.14 | 0.20 | 0.17 | 0.09 | 0.19 | 0.06 | 0.06 | 0.13 | 0.01 | 0.10 | 0.13 | 0.17 | 0.17 | 0.04 | 0.17 | 1.00 | | | | | | |
| **Drinking water** | 0.19 | 0.20 | 0.20 | 0.15 | 0.26 | 0.11 | 0.11 | 0.14 | 0.03 | 0.14 | 0.13 | 0.17 | 0.17 | 0.05 | 0.18 | 0.74 | 1.00 | | | | | |

| | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Latrine** | 0.28 | 0.22 | 0.25 | 0.15 | 0.33 | 0.10 | 0.14 | 0.05 | 0.04 | 0.20 | 0.10 | 0.23 | 0.20 | 0.16 | 0.21 | 0.06 | 0.09 | 1.00 | | | | |
| **Total fish** | 0.21 | 0.19 | 0.28 | 0.23 | 0.30 | 0.16 | 0.22 | 0.09 | 0.02 | 0.13 | 0.07 | 0.17 | 0.14 | 0.11 | 0.15 | 0.02 | 0.06 | 0.09 | 1.00 | | | |
| **Meat** | 0.20 | 0.18 | 0.25 | 0.14 | 0.25 | 0.12 | 0.14 | 0.04 | 0.01 | 0.13 | 0.08 | 0.16 | 0.16 | 0.13 | 0.17 | 0.11 | 0.11 | 0.13 | 0.24 | 1.00 | | |
| **Eggs** | 0.25 | 0.22 | 0.26 | 0.17 | 0.27 | 0.12 | 0.18 | 0.06 | 0.07 | 0.15 | 0.07 | 0.18 | 0.17 | 0.10 | 0.18 | 0.10 | 0.12 | 0.13 | 0.26 | 0.29 | 1.00 | |
| **Milk** | 0.17 | 0.17 | 0.20 | 0.16 | 0.23 | 0.12 | 0.18 | 0.27 | 0.03 | 0.14 | 0.08 | 0.14 | 0.10 | 0.05 | 0.12 | 0.10 | 0.12 | 0.10 | 0.17 | 0.17 | 0.19 | 1.00 |

| **Legend for correlation matrix** | |
|---|---|
| | Correlations between 0.20 and 0.30 |
| | Correlations between 0.30 and 0.50 |
| | Correlations above 0.50 |

*Table 9. Rotated component matrix for PCA1**

| Variables | Principal components | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** |
| Highest education of any member | 0.22 | 0.58 | 0.15 | 0.23 | 0.09 | 0.11 |
| Cellular telephone in working condition | 0.25 | 0.54 | 0.18 | 0.30 | 0.12 | 0.09 |
| Quantity of TV | 0.55 | 0.33 | 0.08 | 0.19 | 0.13 | 0.11 |
| Quantity of tools | 0.07 | 0.36 | 0.48 | 0.16 | 0.05 | -0.04 |
| Number of furniture items | 0.22 | 0.60 | 0.29 | 0.26 | 0.13 | 0.27 |
| Homestead land (decimals) | -0.12 | 0.31 | 0.44 | 0.09 | 0.03 | 0.08 |
| Arable land (decimals) | 0.04 | 0.21 | 0.68 | 0.11 | -0.03 | 0.09 |
| Total livestock units | 0.02 | -0.08 | 0.76 | 0.01 | 0.12 | -0.03 |
| Poultry | 0.02 | 0.38 | 0.13 | -0.16 | 0.04 | -0.27 |
| Wall material | 0.16 | 0.24 | 0.07 | 0.07 | 0.01 | 0.68 |
| Roof material | 0.06 | -0.02 | 0.09 | -0.02 | 0.08 | 0.73 |
| Floor material | 0.11 | 0.37 | -0.02 | 0.16 | 0.12 | 0.54 |
| Electricity connection | 0.94 | 0.13 | -0.02 | 0.08 | 0.06 | 0.10 |
| Cooking fuel | 0.06 | 0.39 | -0.29 | 0.24 | 0.00 | 0.09 |
| Lighting fuel | 0.94 | 0.15 | 0.00 | 0.09 | 0.06 | 0.13 |
| Water for domestic purposes | 0.10 | 0.04 | 0.04 | 0.05 | 0.92 | 0.07 |
| Water for drinking | 0.09 | 0.11 | 0.10 | 0.05 | 0.90 | 0.09 |
| Latrine type | 0.11 | 0.56 | 0.02 | 0.00 | -0.02 | 0.18 |
| HH fish consumption per week (kg) | 0.07 | 0.15 | 0.19 | 0.60 | -0.08 | 0.04 |
| HH Meat consumption per week (kg) | 0.06 | 0.08 | -0.01 | 0.71 | 0.09 | 0.06 |
| HH eggs consumption per week | 0.09 | 0.16 | 0.07 | 0.65 | 0.06 | 0.01 |
| HH milk consumption per week (kg) | 0.13 | -0.13 | 0.42 | 0.40 | 0.07 | 0.17 |

**Highlighted cells show the highest factor loadings for each variable*

*Table 10. Results of k-means cluster analysis using the first principal component of PCA1*

| Cluster number | Final cluster centre | Number of households |
|---|---|---|
| 1 (Poorest) | -1.132 | 1502 |
| 2 | -0.710 | 1932 |
| 3 | 0.709 | 1493 |
| 4 (Richest) | 1.277 | 1576 |

*Table 11. Rotated component matrix for PCA2 (4 factors)*

| Variables | Principal components | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Highest education of any member | 0.63 | 0.13 | 0.11 | 0.01 |
| Quality of TV | -0.03 | 0.05 | -0.01 | -0.02 |
| Cellular telephone in working condition | 0.66 | 0.14 | 0.15 | 0.02 |
| Quantity of tools | 0.36 | 0.41 | 0.30 | -0.04 |
| Number of furniture items | 0.75 | 0.19 | 0.15 | 0.00 |
| Homestead land (decimals) | 0.29 | 0.36 | 0.19 | -0.13 |
| Arable land (decimals) | 0.30 | 0.60 | 0.16 | -0.03 |
| Total livestock units | 0.01 | 0.77 | -0.01 | 0.04 |
| Wall material | 0.62 | -0.10 | -0.27 | -0.03 |
| Roof material | 0.44 | -0.12 | -0.35 | -0.04 |
| Floor material | 0.64 | -0.19 | 0.01 | 0.08 |
| HH big fish consumption per week (kg) | 0.43 | 0.16 | 0.01 | 0.03 |
| HH small fish consumption per week (kg) | 0.14 | 0.16 | 0.46 | -0.05 |
| HH Meat consumption per week (kg) | 0.42 | 0.10 | 0.11 | 0.11 |
| HH eggs consumption per week | 0.45 | 0.17 | 0.04 | 0.11 |
| HH milk consumption per week (kg) | 0.25 | 0.47 | -0.15 | 0.31 |
| HH parboiled rice consumption per week (kg) | 0.04 | 0.42 | -0.38 | -0.70 |
| HH non-parboiled rice consumption per week (kg) | 0.04 | -0.19 | 0.83 | 0.08 |
| HH fine rice consumption per week (kg) | 0.15 | 0.08 | -0.10 | 0.86 |

*Table 12. Results of k-means cluster analysis using the first principal component of PCA2 (4 factors)*

| Cluster number | Final cluster centre | Number of households |
|---|---|---|
| 1 (Poorest) | -1.000 | 2011 |
| 2 | -0.064 | 2875 |
| 3 | 0.972 | 1178 |
| 4 (Richest) | 2.397 | 439 |

*Table 13. Rotated component matrix for PCA2 (all factors)*

| Variables | Principal components | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Highest education of any member | 0.26 | 0.45 | 0.36 | 0.14 | 0.01 | 0.01 |
| Quality of TV | -0.01 | -0.01 | 0.00 | 0.00 | 0.02 | 0.96 |
| Cellular telephone in working condition | 0.26 | 0.44 | 0.44 | 0.17 | 0.01 | -0.03 |
| Quantity of tools | 0.54 | 0.16 | 0.19 | 0.22 | -0.03 | -0.06 |
| Number of furniture items | 0.39 | 0.58 | 0.35 | 0.18 | 0.01 | -0.03 |
| Homestead land (decimals) | 0.49 | 0.19 | 0.05 | 0.13 | -0.10 | 0.07 |
| Arable land (decimals) | 0.71 | 0.14 | 0.08 | 0.04 | 0.00 | 0.02 |
| Total livestock units | 0.76 | -0.13 | -0.01 | -0.20 | 0.07 | -0.02 |
| Wall material | 0.07 | 0.71 | 0.07 | -0.14 | 0.02 | 0.05 |
| Roof material | 0.01 | 0.61 | -0.06 | -0.22 | 0.01 | -0.01 |
| Floor material | 0.02 | 0.66 | 0.15 | 0.15 | 0.11 | -0.04 |
| HH big fish consumption per week (kg) | 0.07 | 0.10 | 0.64 | -0.03 | -0.03 | -0.01 |
| HH small fish consumption per week (kg) | 0.28 | 0.01 | 0.08 | 0.42 | -0.04 | 0.21 |
| HH Meat consumption per week (kg) | 0.01 | 0.06 | 0.68 | 0.07 | 0.04 | 0.03 |
| HH eggs consumption per week | 0.08 | 0.09 | 0.68 | 0.00 | 0.04 | 0.01 |
| HH milk consumption per week (kg) | 0.39 | 0.01 | 0.36 | -0.26 | 0.28 | -0.08 |
| HH parboiled rice consumption per week (kg) | 0.29 | -0.03 | 0.12 | -0.50 | -0.71 | 0.06 |
| HH non-parboiled rice consumption per week (kg) | 0.02 | -0.08 | 0.03 | 0.85 | 0.06 | -0.10 |
| HH fine rice consumption per week (kg) | 0.08 | 0.11 | 0.10 | -0.09 | 0.87 | 0.05 |

*Table 14. Results of k-means cluster analysis using the first principal component of PCA2 (all factors)*

| Cluster number | Final cluster centre | Number of households |
|---|---|---|
| 1 (Poorest) | -0.719 | 3293 |
| 2 | 0.280 | 2251 |
| 3 | 1.545 | 847 |
| 4 (Richest) | 3.846 | 112 |