# Developing Poverty Assessment Tools based on Principal Component Analysis: Results from Bangladesh, Kazakhstan, Uganda, and Peru

Manfred Zeller*, Nazaire Houssou*, Gabriela V. Alcaraz*, Stefan Schwarze**, Julia Johannsen**

* Institute of Agricultural and Socioeconomics in the Tropics and Subtropics, University of Hohenheim, Germany

** Institute of Rural Development, University of Goettingen, Germany

Contributed paper prepared for presentation at the International Association of Agricultural Economists Conference, Gold Coast, Australia, August 12-18, 2006



Corresponding author:

Prof. Dr. Manfred Zeller
Institute of Agricultural and Socioeconomics in the Tropics and Subtropics, University of Hohenheim
Schloß, Osthof (490a)
70599 Stuttgart
GERMANY
Tel. +49-711-4592175 (Office), +49-711-6337161 (Residence)
Fax: +49-711-4592794
Email: manfred.zeller@uni-hohenheim.de

**Abstract**

Developing accurate, yet operational poverty assessment tools to target the poorest households remains a challenge for applied policy research. This paper aims to develop poverty assessment tools for four countries: Bangladesh, Peru, Uganda, and Kazakhstan. The research applies the Principal Component Analysis (PCA) to seek the best set of variables that predict the household poverty status using easily measurable socio-economic indicators. Out-of sample validations tests are performed to assess the prediction power of a tool. Finally, the PCA results are compared with those obtained from regressions models.

In-sample estimation results suggest that the Quantile regression technique is the first best method in all four countries, except Kazakhstan. The PCA method is the second best technique for two of the countries. In comparison with regression techniques, PCA models accurately predict a large percentage of households.

With regard to out-of sample validations, there is no clear trend; neither the PCA method nor the Quantile regression consistently yields the most robust results. The results highlight the need to assess the out-of-sample performance and thereby the robustness of a poverty assessment tool in estimating the poverty status of a new sample. We conclude that measures of relative poverty estimated with PCA method can yield fairly accurate, but not so robust predictions of absolute poverty as compared to more complex regression models.

*JEL: H5, Q14, I3*

Keywords: *poverty assessment, targeting, principal component analysis, Bangladesh, Peru, Kazakhstan, Uganda*

## 1. Introduction

Most of the world's poor live in rural areas and are directly or indirectly dependent on agriculture. A wide range of rural development policies and projects, for example in the area of agricultural extension, rural finance, and safety nets, seeks to target the poor in the provision of information, capital and services. However, the identification of those with incomes below the poverty line in an accurate, yet low-cost manner remains a challenge. This study aims at developing and testing newly designed poverty assessment tools. The paper uses primary, nationally representative data from four countries[i]: Bangladesh, Kazakhstan, Peru, and Uganda.

In contrast to previous research that employed multivariate regression to identify and weigh poverty indicators for the prediction of daily per-capita-expenditures (see, for example, Ahmed and Bouis, 2003), this paper is the first to our knowledge that applies Principal Component Analysis (PCA) to identify a set of variables that predict whether a household is below or above the poverty line. Confidence intervals for the accuracy ratios are estimated using the bootstrap technique and out-of sample validations tests are implemented to evaluate the models prediction power over a new set of observations derived from the same population. Furthermore, the PCA results are compared with those obtained by OLS, LPM, Probit, and Quantile regressions applied to the same data. Each of the four data sets contains variables that are usually enumerated in Living Standards Measurement Surveys (LSMS). Thus,

---

[i] The data stem from the project Developing Poverty Assessment Tools, which is carried out by the IRIS Center, University of Maryland. The project is funded by the United States Agency for International Development (USAID) under the Accelerated Microenterprise Advancement Project (AMAP) (Contract No. GEG-I-02-02-00029-009). The cleaning and aggregation of the data were carried out at the Institute of Rural Development, University of Göttingen. We gratefully acknowledge the source of the data. We are grateful for comments by Walter Zucchini regarding the design of out-of-sample tests.

indicators cover demography, education, food security, and especially ownership of consumption and production assets as well as financial capital of the household. The set of poverty indicators and their derived weights can be viewed as a tool to target ex-ante the poor, or to assess ex-post the poverty outreach of any poverty-targeted development policy or project.

Section 2 discusses the data, the PCA estimation procedure, including the construction of the confidence intervals, and briefly presents the regression methods. Section 3 presents the PCA results for four countries, whereas section 4 makes a within and cross-country comparison of accuracy performance. Section 5 concludes the paper.

## 2. Data and Methodology

### 2.1 Data Collection

In each country, the IRIS center of the University of Maryland worked with survey firms that carried out nationally representative household surveys and double entry of data (Table 1).

Table 1: Country survey

| Countries | Survey Firms | Sample Size (households) | Interview dates 2004 | Data entry software |
|---|---|---|---|---|
| Bangladesh | DATA | 800 | March-April | SPSS |
| Kazakhstan | Sange Research Center | 840 | September-October | SPSS |
| Peru | Instituto Cuánto | 800 | June-August | ISSA |
| Uganda | NIDA | 800 | August-October | SPSS |

Source: Country reports by Zeller et al. (2005) available for downloading at www.povertytools.org.
ISSA denotes Integrated System for Survey Analysis; SPSS is a Statistical Package for Social Sciences.

Two types of questionnaires were employed. The composite questionnaire enumerated indicators from many poverty dimensions. In order to measure absolute poverty, an LSMS-type household expenditure questionnaire was administered exactly 14 days after the interview with the composite questionnaire. The questionnaires were adapted to the country-specific context and can be downloaded at www.povertytools.org.

Two types of poverty lines were used, as outlined by the *Amendment to the Microenterprise for Self-Reliance and International Anti-Corruption Act of 2000* by US congress (USAID, 2005). According to that legislation, a household is classified as "very poor" if either (a) the household is "living on less than the equivalent of a dollar a day" ($1.08

per day at 1993 Purchasing Power Parity) — the definition of “extreme poverty” under the Millennium Development Goals; or (b) the household is among the poorest 50 percent of households below the country’s own national poverty line. Table 2 provides the overall headcount index for the “very poor” in the four countries.

Table 2: Headcount index for the “very poor”, by country

| Countries | Poverty headcount (%) | Poverty line used |
|---|---|---|
| Bangladesh | 31.40 | International |
| Kazakhstan | 4.53 | National |
| Peru | 26.88 | National |
| Uganda | 32.36 | International |

Source: Own calculations described in Zeller et al. (2005a, 2005b, 2005c, 2005d)

In Bangladesh and Uganda, the international 1 dollar a day poverty line yields a higher headcount index of “very poor” whereas for the wealthier countries - Peru and Kazakhstan -, the alternative definition of the bottom 50 percent population below the national poverty line yields a higher headcount index.

## 2.2 Principal Component Analysis (PCA)

### *2.2.1 Theoretical Considerations*

Because the relative strengths of different indicators in capturing poverty are very likely to vary across countries, a method is called for that allows adjusting weights for each situation based on the country-specific poverty context existing therein. For example, in the case of nutritional indicators, Habicht and Pelletier (1990) show that the socio-economic context matters in the choice of appropriate nutrition-related indicators. Zeller et al. (2006) show that the relative poverty of households in very poor countries is better captured by several indicators for food security whereas the number and type of consumer assets matter more for explaining relative poverty in wealthier countries.

The method of principal component analysis (PCA) addresses, when used as an aggregation procedure, the concerns raised above in an objective and rigorous way. Earlier applications of PCA for the measurement of relative poverty or wealth include Filmer and Pritchett (1998), Sahn and Stifel (2000), and Henry et al. (2003). PCA assists in statistically identifying and weighing the most important indicators in order to calculate an aggregate index of relative poverty for a specific sample household.

Basically, the principal component technique slices information contained in a set of indicators into several components. Each component is constructed as a unique index based on the values of all the indicators. The main idea is to formulate a new variable, $z_1$, which is the linear combination of the original indicators so that it accounts for the maximum of the total variance in the original indicators (Basilevsky, 1994).

In other words, once data on k indicators are arranged in k columns to form a n x k matrix X, the method of principal components can be used to extract a small number of variables that accounts for most or all variations in X. This is done by obtaining a linear combination of the columns of X that provides the best fit to all columns of X as in

$$z_1 = Xw \qquad (1)$$

The first principal component is then described by the index variable $z_1$, as defined in equation 1. This index aggregates the information contained in the poverty indicators. Having identified the first principal component as the 'poverty' component, one can compute for each household denoted by the subscript j its poverty index $z_j$ with the following equation:

$$z_j = f_1 * ((X_{j1} - X_1) / S_1) + \ldots + f_N * ((X_{jN} - X_N) / S_N) \qquad (2)$$

where $f_1$ is the weight for the first of the N poverty indicator variables identified as significant in the PCA model, $X_{j1}$ is the jth household's value for the first variable, and $X_1$ and $S_1$ are the mean and standard deviation of the first variable over all households (Zeller et al., 2006).

In each of the countries presented here, the first component was always the one that was identified as the multidimensional index of relative poverty based on a number of criteria. This is because the poverty component and its significant underlying indicators can be identified by analyzing the signs and size of the indicators in relation to the new component variable (Henry et al., 2003; Zeller et al., 2006).

For example, according to theory, higher education should contribute positively – not negatively – to wealth, whereas more dependents such as children in a household are associated with lower wealth.

The PCA method, hence, can be used to compute weights that mark each indicator's relative contribution to the overall poverty component. Using these weights, a household-specific poverty index can be computed based on each household's indicator values as shown in equation 2 above. This poverty index is a measure of relative poverty. Having a negative value for the poverty index identifies a household as being poorer than the population mean, whereas positive values indicate an above-average wealth.

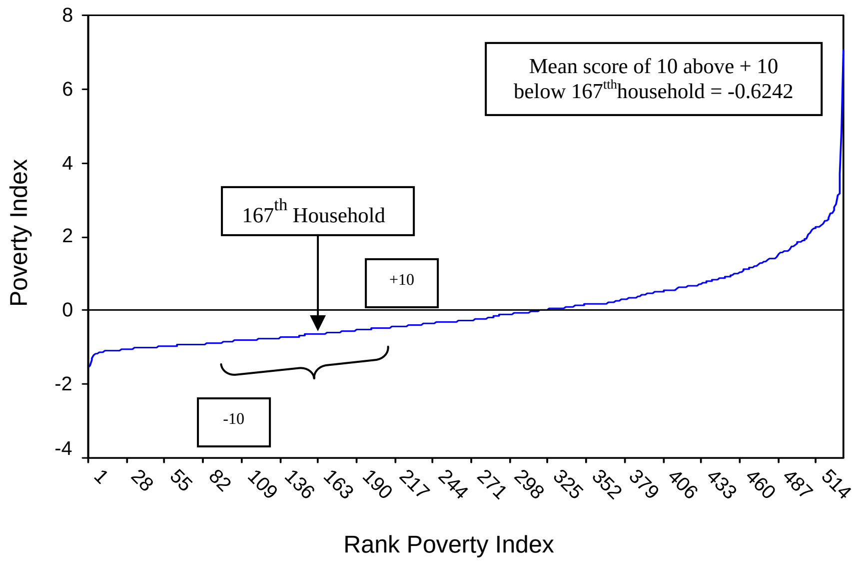
### *2.2.2 Methodological steps taken in estimating the poverty index using PCA*

In order to perform out-of samples tests, the samples were first split into two sub-samples in ratio 67:33 in all the methods considered, including the regressions. The larger samples were employed to identify the best set of variables and their weights, and the smaller samples were used to test out-sample the prediction accuracy of the constructed tools. In the out-sample test, we therefore applied the set of identified indicators and their derived weights to predict per-capita daily expenditures.

To compute the poverty index, the PCA procedure involves a number of steps following Henry et al. (2003) that are illustrated using the example of Bangladesh. First of all, bivariate correlation analyses of *the per capita daily expenditures* (benchmark indicator) were run with the initial variable list of 117 variables. Sixty variables with highly significant coefficients (alpha < 0.001) and a theoretically consistent sign for the correlation coefficient were retained from the initial data set. Second, before applying the PCA, following Henry et al. (2003), we grouped these sixty variables into several dimensions of poverty. Within each dimension, we dropped variables that were redundant, i.e. they exhibited a high correlation with other variables contained in the same dimension. When dropping similar variables, we preferred to drop variables that appeared to be more difficult to ask in household interviews.

For example, if the value of land was highly correlated with the area of land, we dropped the former variable. Thus, closely related variables that effectively measure the same phenomenon were screened out. After this second step, a set of 20 variables was retained. Third, the PCA was then carried out with SPSS. Here, the maximum number of iterations was set at 25. The Eigen value was limited to 1. Since PCA does not provide an easy way to generate a best fit for a poverty index, a trial and error process using the final 20 variables was used to determine which combination yielded the best accuracy performance. After obtaining the first PCA results, an intermediate step consisted in checking the component matrix and removing variables with coefficients lower than 0.3, in accordance with Henry et al. (2003). Likewise, variables displaying theoretically unexpected signs were removed from the list. Positive coefficients indicate a positive correlation with relative wealth of the household and vice versa. Following Henry et al. (2003), variables with communalities coefficients lower than 0.1 were removed from the list. Applying these screening procedures leads to increases in the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO). The larger the KMO index, the higher is the fraction of variance explained by the model. In the case of Bangladesh, the final number of variables after the last PCA run was 13. This number was further reduced to the best 10 variables based on the coefficient size in the component matrix. As stated by Henry et al. (2003), the higher the coefficient size, the stronger the relation with the derived poverty index. Using this final model of best 10 variables, the poverty index was computed for each of the households. The result is illustrated in Figure 1.

Figure 1 Poverty Index Distribution in Bangladesh.




Source: Own calculations

The graph shows the distribution of the poverty index over the nationally representative sub-sample of 533 households in Bangladesh. A cut-off poverty index is needed in order to predict the status of a household with respect to absolute poverty. Therefore, the poverty index generated by the PCA was ranked first. Since 31.4% of households have incomes below 1 US-$ at PPP rates, the sample household with a rank poverty index of 167 (167 divided by 533 yields approximately 31.4 %) was identified. This corresponds to the 167th household on the graph. Hence, all households that have a lower rank than this household are considered very poor and all above belong to not very poor group. This is based on the assumption that the distribution of relative poverty as measured with PCA generates the same ranking of households as those based on absolute poverty as measured by per-capita daily expenditures.

However, in order not to base the calibration on the poverty index of one single household, the mean poverty index of the ten above and ten below the anchor household with rank 167 was taken as the cut-off poverty index. This somewhat arbitrarily chosen range of ten households below and above yielded the best accuracy results when compared with those generated from alternative ranges. We apply the same range for the other three countries.

### 2.2.3 Accuracy Ratios

Seven ratios have been proposed by IRIS (2005) to assess the accuracy of a poverty assessment tool (Table 3).

Table 3: Definitions of accuracy ratios

| Accuracy Ratios | Definitions |
|---|---|
| Total Accuracy | Percentage of the total sample households whose poverty status is correctly predicted by the estimation model |
| Poverty Accuracy | Households correctly predicted as very-poor, expressed as a percentage of the total very-poor |
| Non-Poverty Accuracy | Households correctly predicted as not very-poor, expressed as percentage of the total number of not very-poor |
| Undercoverage | Error of predicting very-poor households as being not very-poor, expressed as a percentage of the total number of very-poor |
| Leakage | Error of predicting not very-poor households as very-poor, expressed as a percentage of the total number of very-poor |
| Poverty Incidence Error (PIE) | Difference between the predicted and the actual (observed) poverty incidence, measured in percentage points |
| Balanced Poverty Accuracy Criterion (BPAC) | Poverty Accuracy minus the absolute difference between undercoverage and leakage, each expressed as a percentage of the total number of very-poor |

Source: IRIS (2005)

The first five measures are self-explanatory. Undercoverage and leakage are extensively used to assess the targeting efficiency of policies (Valdivia, 2005; Ahmed et al., 2004; Weiss, 2004). The performance measure PIE indicates the precision of a model in correctly predicting the observed poverty rate. Positive PIE values indicate an overestimation of the poverty incidence, whereas negative values show the opposite. It is an important accuracy criterion for assessing ex-post the poverty outreach of a given policy. The balanced poverty assessment criterion BPAC considers three accuracy measures that are especially relevant for poverty targeting: poverty accuracy, leakage, and undercoverage. These three measures exhibit trade-offs.

For example, minimizing leakage leads to higher undercoverage and lower Poverty Accuracy. Higher positive values for BPAC indicate higher Poverty Accuracy, adjusted by the absolute difference between leakage and undercoverage. In the following, BPAC is used as the overall criterion to judge the model’s accuracy performance.

Confidence intervals for the ratios were estimated using the technique of bootstrapping. Efron (1987) introduced the estimation of confidence intervals based on bootstrap computations. Bootstrap is a statistical procedure which models sampling from a population by the process of resampling from the sample (Hall, 1994).

The reason for using this methodology is that the above ratios are highly aggregated. Unlike traditional confidence intervals estimation, bootstrap does not require the assumption of a normal distribution. The original dataset is used to create 1000 new randomly selected samples with replacement. Then, the above seven accuracy ratios are computed for each sample. This yields a set of 1000 observations for each of the ratios. The percentile method is applied to derive the confidence intervals. The $2.5^{th}$ and $97.5^{th}$ percentiles are calculated for a 95% confidence level.

### 2.3 Overview of regressions methods

In the country reports by Zeller et al. (2005), four different single-step regressions methods were used to identify and test the accuracy of alternative poverty assessment tools. These include: the Ordinary Least Square method (OLS), the Linear Probability Model (LPM), the Probit, and Quantile regressions.

The present study applies the above-mentioned methods to the data being used. These methods seek to identify the best set of ten regressors for predicting the household poverty status. For the OLS and LPM models, the MAXR routine of SAS was used to identify a set of the best ten regressors that maximizes the model's explained variance. It is not feasible to identify the set of best ten for Probit and Quantile regressions using the MAXR routine of SAS. Therefore, the ten regressors from the LPM and OLS models were then used in the Probit and Quantile models, respectively.

Obviously, the models do not seek to identify the causal determinants of poverty, but identify variables that can best indicate about the current poverty status of a household. For purposes of comparisons, we also allow only ten indicators in the PCA analysis.

## 3. Results from Principal Component Analysis (PCA)

### 3.1 Empirical Results from Bangladesh

The above-mentioned measures of model performance are illustrated here using the results of the PCA for Bangladesh. This model uses only 10 indicators to allow for comparison with regression models (Table 4).

Table 4: Summary of PCA results for Bangladesh

| **Variables (10)** | **Component Loadings 1** |
|---|---|
| **Kaiser-Meyer-Olkin measure of sampling adequacy:** 0.790 | |
| Black and white TV ownership | 0.478 |
| Any household member has a checking account | 0.558 |
| Number of adult household members who can read and write | 0.707 |
| Poultry number | 0.444 |
| Room size in square feet | 0.610 |
| Log value of kantha (a digging tool used in farming) | 0.434 |
| Public grid with legal socket in house | 0.592 |
| Household has improved toilet | 0.520 |
| Number of *saris* (woman's clothing) owned by household | 0.781 |
| Amount of remittances received divided by remittances sent | 0.518 |

Source: Own calculations

The ten indicators are fairly easy to measure in household surveys, and capture different dimensions of poverty. Some indicators are directly observable through a visit to the household's homestead. All the components loadings are far higher than 0.3 and display theoretically expected signs which indicate a good variable screening. Likewise, the Kaiser-Meyer-Olkin measure of sampling adequacy is relatively high. Results from the PCA models for the other three countries are shown in the annex.

The model for Bangladesh yields the following prediction matrix when calibrated to the absolute poverty line as described above using Figure 1.

Table 5: Observed and predicted household poverty status for Bangladesh

| Observed poverty status | Predicted poverty status | | |
|---|---|---|---|
| | Not very-poor | Very-poor | Total |
| Not very-poor | 297 | 67 | 364 |
| Very-poor | 71 | 98 | 169 |
| Total | 368 | 165 | 533 |

Source: Own calculations

From Table 5, one can calculate the seven measures of accuracy performance (Table 6). The bootstrapped confidence intervals are presented in Table 7.

Table 6: Measures of accuracy performance of PCA model for Bangladesh

| **Bangladesh** | Total Accur. | Pov. Accur. | Under-coverage | Leakage | PIE | BPAC |
|---|---|---|---|---|---|---|
| **Principal Component Analysis** | | | | | | |
| Random 2/3 sample (N=533) | 74.11 | 57.99 | 42.01 | 39.65 | -0.75 | 55.26 |
| Predictions for remaining 1/3 sample (N=266) | 71.05 | 50.00 | 50.00 | 43.90 | -1.88 | 43.90 |

Source: Own calculations

Table 7: Confidence intervals for the accuracy performances

| **Bangladesh** | **95% Bootstrap confidence intervals for 2/3 sample (1000 replications)** | |
|---|---|---|
| **Accuracy ratios** | Upper limit | Lower limit |
| **Principal Component Analysis** | | |
| Total Accuracy: | 77.67 | 70.17 |
| Poverty Accuracy: | 64.42 | 52.05 |
| Non-Poverty Accuracy: | 84.32 | 78.39 |
| Undercoverage: | 47.95 | 35.58 |
| Leakage: | 52.70 | 30.06 |
| Predicted Poverty Incidence: | 31.89 | 30.39 |
| PIE: | 3.56 | -4.13 |
| BPAC: | 61.58 | 41.72 |

Source: Own calculations

As concerns Tables 5 and 6, the results were obtained at a cutoff score for the poverty index of -0.6242. This value is equivalent to the mean of the poverty index of the ten above and ten below the 167$^{th}$ household that has a rank equivalent to the poverty rate. Households with a value lower than or equal to -0.6242 are considered ‘very poor.’ About 74% of households were correctly predicted by the calibrated PCA model. Yet, among poor households, this accuracy is lower. The same trend applies to the results yielded by the out-of sample validations. Compared to in-sample results, the out-sample BPAC drops by about 12 percentage points, whereas the poverty and the total accuracy drop by 7% and 3% respectively. These results indicate that the identified tool is capable of achieving fairly comparable results with some moderate drops in performances when applied to a different set of households drawn from the same population.

Table 7 provides the bootstrap confidence intervals for in-sample ratios, based on 1000 replicated samples. Strikingly, the results suggest that all the ratios are different from zero, except the PIE. As indicated in the formula, the PIE could be estimated at zero. However, the constructed intervals are fairly large for most of the ratios considered.

## 3.2 Comparison of PCA and Regression Results

### 3.2.1 Within Country Comparison of Accuracy Results

Table 8 compares the accuracy performances of PCA with those of single-step regression techniques for four countries. Like the PCA, each regression model uses 10 indicators.

Table 8: Comparison of PCA and regression results for Bangladesh

| | Model 9 | Adj. $R^2$ | Total Accur. (%) | Poverty Accur. (%) | Under-coverage (%) | Leakage (%) | PIE (% point) | BPAC (% point) |
|---|---|---|---|---|---|---|---|---|
| Bangladesh | Overall poverty rate: 31.41% | | | | | | | |
| | **OLS** | **59.44** | | | | | | |
| | In-sample | | 81.43 | 56.71 | 43.29 | 17.07 | -8.07 | 30.46 |
| | Out-sample | | 78.20 | 52.87 | 47.13 | 19.54 | -9.02 | 25.29 |
| | **LPM** | **38.14** | | | | | | |
| | In-sample | | 83.68 | 61.59 | 38.42 | 14.63 | -7.32 | 37.81 |
| | Out-sample | | 79.32 | 60.92 | 39.08 | 24.14 | -4.89 | 45.98 |
| | **Probit** | | | | | | | |
| | In-sample | | 83.87 | 66.46 | 33.54 | 18.90 | -4.50 | 51.83 |
| | Out-sample | | 78.95 | 68.97 | 31.03 | 33.33 | 0.75 | 66.67 |
| | **Quantile P=42nd** | | | | | | | |
| | In-sample | | 82.36 | 71.34 | 28.66 | 28.66 | 0 | 71.34 |
| | Out-sample | | 80.45 | 72.41 | 27.59 | 32.18 | 1.50 | 67.82 |
| | **PCA** | | | | | | | |
| | In-sample | | 74.11 | 57.99 | 42.01 | 39.65 | -0.75 | 55.26 |
| | Out-sample | | 71.05 | 50.00 | 50.00 | 43.90 | -1.88 | 43.90 |

Source: Own calculations based on IRIS survey data. P = Percentage point of estimation used in quantile model.

The results regarding Bangladesh show that the best estimation technique which maximizes the BPAC is the Quantile regression technique. Through an iterative procedure involving a series of regressions with the given set of the best ten regressors as identified by the MAXR routine of SAS in the OLS model, alternative percentile points of estimation for the Quantile model are tested in order to maximize BPAC.

With an optimal point of estimation identified at the 43rd percentile, the Quantile regression achieves a PIE of 0 percentage points. Moreover, the Poverty Accuracy amounts to about 70%, and the BPAC is estimated at 71.34 percentage points. In terms of BPAC as our overall criterion, the PCA model is the second best method with a value of 55.26 percentage points. The PCA also achieves a PIE of -0.75, which implies a good prediction of the observed poverty rate in the sample. However, the achieved Poverty Accuracy is lower compared to Probit, LPM, and OLS methods

Likewise, the out-of sample validations results suggest the Quantile regression identifies the set of indicators that yields the most stable (and equally most accurate) results, since in and out-samples ratios, especially for the Poverty Accuracy and BPAC, are very comparable. The latter drops by about 4 percentage points, whereas the former increases by about 1%. The PCA is one of the most inferior methods, with a drop of about 8% in Poverty Accuracy and a drop of about 11 percentage points in BPAC.

Table 9: Comparison of PCA and regression results for Kazakhstan

| | **Model 9** | **Adj. $R^2$** | **Total Accur. (%)** | **Poverty Accur. (%)** | **Under-coverage (%)** | **Leakage (%)** | **PIE (% point)** | **BPAC (% point)** |
|---|---|---|---|---|---|---|---|---|
| **Kazakhstan** | Overall poverty rate: 4.52% | | | | | | | |
| | **OLS** | **53.60** | | | | | | |
| | In-sample | | 95.05 | 10.71 | 89.29 | 7.14 | -4.22 | -71.43 |
| | Out-sample | | 96.70 | 22.22 | 77.78 | 22.22 | -1.84 | -33.33 |
| | **LPM** | **20.69** | | | | | | |
| | In-sample | | 95.23 | 7.14 | 92.86 | 0 | -4.77 | -85.71 |
| | Out-sample | | 96.32 | 0 | 100 | 11.11 | -2.94 | -88.89 |
| | **Probit** | | | | | | | |
| | In-sample | | 96.15 | 32.14 | 67.86 | 7.14 | -3.12 | -28.57 |
| | Out-sample | | 95.96 | 22.22 | 77.78 | 44.44 | -1.10 | -11.11 |
| | **Quantile P=23rd** | | | | | | | |
| | In-sample | | 92.84 | 32.14 | 67.86 | 71.43 | 0.18 | 28.57 |
| | Out-sample | | 93.38 | 55.56 | 44.44 | 155.56 | 3.68 | -55.56 |
| | **PCA** | | | | | | | |
| | In-sample | | 95.41 | 45.00 | 55.00 | 70.00 | 0.55 | 30.00 |
| | Out-sample | | 92.65 | 11.76 | 88.24 | 29.41 | -3.68 | -47.06 |

Source: Own calculations based on IRIS survey data. P = Percentage point of estimation used in quantile model.

As concerns Kazakhstan, in-sample results described in Table 9 suggest that the PCA is the best method followed by Quantile regression which yields a BPAC of 28.57 percentage points and a PIE of 0.18 percentage points. The latter implies an almost perfect prediction of the poverty rate compared with the PCA, which overestimates the rate. Nonetheless, the Poverty Accuracy of the PCA is much higher.

With regard to out-of sample tests, the results exhibit no clear trend with regard to accuracy performance. On the one hand, the BPAC drops significantly in the case of the PCA and Quantile regression, but only slightly for the LPM. One the other hand, this ratio increases for the Probit and more substantially for the OLS method. Likewise, the Poverty Accuracy drops substantially for the PCA, moderately for the Probit, and estimates at zero for the LPM, whereas it increases moderately and substantially for OLS and Quantile regressions respectively.

Table 10: Comparison of PCA and regression results for Peru

| | **Model 9** | **Adj. $R^2$** | **Total Accur. (%)** | **Poverty Accur. (%)** | **Under-coverage (%)** | **Leakage (%)** | **PIE (% point)** | **BPAC (% point)** |
|---|---|---|---|---|---|---|---|---|
| **Peru** | Overall poverty rate: 26.88% | | | | | | | |
| | **OLS** | **77.87** | | | | | | |
| | In-sample | | 85.74 | 67.14 | 32.86 | 21.43 | -3.00 | 55.71 |
| | Out-sample | | 84.27 | 60.00 | 40.00 | 16.00 | -6.74 | 36.00 |
| | **LPM** | **40.54** | | | | | | |
| | In-sample | | 84.43 | 55.71 | 44.29 | 15.00 | -7.69 | 26.43 |
| | Out-sample | | 81.65 | 48.00 | 52.00 | 13.33 | -10.86 | 9.33 |
| | **Probit** | | | | | | | |
| | In-sample | | 84.80 | 60.71 | 39.29 | 18.57 | -5.44 | 40.00 |
| | Out-sample | | 81.65 | 57.33 | 42.67 | 22.67 | -5.62 | 37.33 |
| | **Quantile P=43rd** | | | | | | | |
| | In-sample | | 85.56 | 72.14 | 27.86 | 27.14 | -0.19 | 71.43 |
| | Out-sample | | 85.02 | 65.33 | 34.67 | 18.67 | -4.49 | 49.33 |
| | **PCA** | | 72.05 | 47.10 | 52.90 | 55.07 | 0.56 | 44.93 |
| | In-sample<br>Out-sample | | 70.41 | 48.05 | 51.95 | 50.65 | -0.37 | 46.75 |

Source: Own calculations based on IRIS survey data. P = Percentage point of estimation used in quantile model.

As concerns Peru, Table 10 indicates that in-sample, the best regression technique in terms of BPAC is the Quantile model. This technique achieves a BPAC of 71.43 percentage points and a PIE of -0.19 percentage points. The second best method is OLS with a BPAC of 55.71

percentage points and a PIE of -3.00. The estimated Poverty Accuracy in both cases amounts about 70% which indicates that a considerable proportion of poor households have been correctly predicted by the methods. The PCA is the third best method with a BPAC of about 45 percentage points and a Poverty Accuracy of almost 47%.

Considering the similarity between in and out-of sample results, a different trend applies. The PCA yields the most similar performances in terms of both BPAC and Poverty Accuracy. These ratios increase slightly regarding out-of sample predictions. This indicates that the PCA method identifies the set of indicators that yields the most stable, but one of the less accurate for Peru. The Probit method yields the second most stable set with moderate performances. LPM and OLS regressions follow the Probit with a relatively high drop in BPAC, but moderate reduction in Poverty Accuracy.

Table 11: Comparison of PCA and regression results for Uganda

| | **Model 9** | **Adj. $R^2$** | **Total Accur. (%)** | **Poverty Accur. (%)** | **Under-coverage (%)** | **Leakage (%)** | **PIE (% point)** | **BPAC (% point)** |
|---|---|---|---|---|---|---|---|---|
| **Uganda** | Overall poverty rate: 32.36% | | | | | | | |
| | **OLS** | **54.77** | | | | | | |
| | In-sample | | 77.14 | 59.41 | 40.58 | 30.00 | 3.43 | 48.82 |
| | Out-sample | | 69.20 | 45.88 | 54.12 | 41.18 | 4.18 | 32.94 |
| | **LPM** | **30.05** | | | | | | |
| | In-sample | | 80.19 | 62.35 | 37.65 | 23.53 | 4.57 | 48.24 |
| | Out-sample | | 69.58 | 56.47 | 43.53 | 50.59 | -2.28 | 49.41 |
| | **Probit** | | | | | | | |
| | In-sample | | 80.38 | 60.58 | 39.41 | 21.18 | 5.90 | 42.35 |
| | Out-sample | | 69.20 | 54.12 | 45.88 | 49.41 | -1.14 | 50.59 |
| | **Quantile P=46th** | | | | | | | |
| | In-sample | | 78.10 | 65.29 | 34.71 | 32.94 | 0.57 | 63.53 |
| | Out-sample | | 69.20 | 54.12 | 45.88 | 49.41 | -1.14 | 50.59 |
| | **PCA** | | | | | | | |
| | In-sample | | 69.14 | 51.98 | 48.02 | 43.50 | -1.52 | 47.46 |
| | Out-sample | | 64.64 | 53.85 | 46.15 | 73.08 | 7.98 | 26.92 |

Source: Own calculations based on IRIS survey data. P = Percentage point of estimation used in quantile model.

In the case of Uganda (Table 11), the best method is again the Quantile regression, followed by the OLS method which yields a BPAC of 48.82 and a PIE of 3.43 percentage points. Nonetheless, the BPAC achieved by the OLS, LPM, and PCA methods are comparable.

Considering the Poverty Accuracy, the Quantile regression is still the first, followed by the LPM and Probit methods respectively. The PCA is the worst method.

With respect to out-sample predictions, the LPM appears to yield the most robust results in terms of the BPAC, followed by the Probit regression. The Quantile regression is the third, whereas the PCA is the last method. The latter yields, however the most comparable results considering the Poverty Accuracy ratio, followed by the LPM and Probit methods. These results seem to suggest that neither of the methods has a clear advantage with respect to in-sample accuracy and out-sample robustness of predictions. Moreover, a method that yields the most comparable results in terms of BPAC does not necessarily generate the most similar results in terms of Poverty Accuracy and vice-versa. This is explained by the relationship between both ratios which is not linear.

### 3.2.2 Cross-country Comparison of Accuracy Results

In Table 12, the performances across countries are compared.

Table 12: Accuracy performance by estimation method and country (BPAC in % points)

| Methods | Countries | Bangladesh | Kazakhstan | Peru | Uganda | Mean |
|---|---|---|---|---|---|---|
| PCA | In-sample | 56.21 | 30.00 | 44.93 | 47.46 | 44.65 |
| | Out-sample | 47.56 | -47.06 | 46.75 | 26.92 | 18.54 |
| OLS | In-sample | 30.46 | -71.43 | 55.71 | 48.82 | 15.89 |
| | Out-sample | 25.29 | -33.33 | 36.00 | 32.94 | 15.23 |
| LPM | In-sample | 37.81 | -85.71 | 26.43 | 48.24 | 6.69 |
| | Out-sample | 45.98 | -88.89 | 9.33 | 49.41 | 3.96 |
| Probit | In-sample | 51.83 | -28.57 | 40.00 | 42.35 | 26.40 |
| | Out-sample | 66.67 | -11.11 | 37.33 | 50.59 | 35.87 |
| Quantile | In-sample | 71.34 | 28.57 | 71.43 | 63.53 | 58.72 |
| | Out-sample | 67.82 | -55.56 | 49.33 | 50.59 | 28.05 |

Source: Own calculations based on IRIS survey data

Table 12 suggests that in-sample, the Quantile regression method yields on average the best results in terms of BPAC for the four countries, followed by the PCA. At individual country level however, some clarifications need to be made. The Quantile regression is still the best, except for Kazakhstan for which PCA yields a slightly higher BPAC. The PCA is the second best for Bangladesh, but the third best for Uganda, yielding a slightly lower BPAC compared to the OLS which is the second method. Likewise, the PCA is the third best method for Peru.

Considering out-of sample predictions, on average the most robust performances are achieved with the OLS. While its in-sample accuracy is on overage the lowest, the out-sample accuracy levels do not deviate much from the in-sample estimates. In terms of robustness, the LPM and Probit are the second and third best methods, whereas the PCA and Quantile yield the least stable results with a relatively high drop in BPAC. With respect to individual countries, however, the out-sample performance greatly varies across the different models.

**4. Concluding Remarks**

This paper focuses on the application of Principal Component Analysis (PCA) estimation method to identify the best indicators for predicting the poverty status. As poverty indicators, we use variables related to demography as well as human, physical, and financial assets that are usually contained in Living Standard Measurement Surveys. Our analyses cover four countries: Bangladesh, Kazakhstan, Peru, and Uganda.

The PCA models accurately predicted a large percentage of households. In all four countries, the Non-Poverty Accuracy (not reported) of the PCA model is higher than the Poverty Accuracy. The accuracy performance of PCA was further compared with poverty assessment tools identified by four different types of regression models. With respect to BPAC, the first best method in all the countries is the Quantile regression method, except for Kazakhstan.

The PCA method is the second best method for two of the countries, the third best for Uganda and one of the last methods for Peru.  With regard to out-of sample validations which seek to assess the robustness of a poverty assessment tool in terms of its accuracy in correctly predicting the poverty status of households, there is no clear trend. Neither the PCA method, nor the Quantile regression consistently yields the most robust results. Despite the large losses in out-sample accuracy for three of the four countries, the Quantile regression still achieves the highest BPAC.

The sets of indicators and their derived weights can be viewed as a potential means-tested poverty assessment tools which could be used to target the “very poor” households or to assess ex-post the poverty outreach performance of development policies and projects targeted to those living below the chosen poverty lines. The main conclusion drawn is that measures of relative poverty estimated with PCA can yield fairly accurate redictions of absolute poverty in nationally representative samples. However, the accuracy performance, especially the robustness of poverty assessment tools derived from regression models is generally higher.

We recommend that the comparisons of different regression techniques and the PCA be done for other LSMS-type data sets to either confirm or reject the findings of this paper. Our tentative conclusions – based on the test of five different methods for four countries- are as follows. In countries where recent nationally representative data sets with per-capita daily expenditures are available, the use of regression techniques, especially Quantile regression is more appropriate for the development of poverty assessment tools. In countries where nationally representative data on per-capita daily expenditures and suitable poverty indicators (such as from LSMS-type surveys) are not available, a second alternative consists of using data from the Demographic and Health Surveys (DHS) for the calibration of a nationally representative poverty assessment tool. Since DHS data do not contain expenditure variable, regression analysis is not feasible. DHS data contain few, but relatively simple poverty indicators related to demography, housing, food security, and nutrition as well as asset possession. DHS data has been used in the past to estimate the so-called wealth or poverty indices by means of the PCA (see, for example, Filmer and Pritchett, 1998). Our results now demonstrate that these wealth indices can be calibrated to predict absolute poverty status with relatively high accuracy. Thus, PCA is an alternative, second-best calibration technique for the calibration of means-tested poverty assessment tools.

**References**

Ahmed, A., Rashid, S., Sharma, M., Zohir, S., 2004. Food aid distribution in Bangladesh: leakage and operational performance, Disc. Pap. 173, International Food Policy Research Institute, Washington, D.C.

Ahmed, A., H. Bouis., 2003. Weighing what's practical: Proxy means test for targeting food subsidies in Egypt, Disc. Pap. 213, International Food Policy Research Institute, Washington, D.C.

Basilevsky, A., 1994. Statistical factor analysis and related methods, John Wiley and Sons, New York.

Efron, B., 1987. Better bootstrap confidence intervals, Journal of the American Statistical Association 82, 171-185.

Filmer, D., Pritchett, L., 1998. Estimating wealth effects without expenditure data or tears: with and application to educational enrollments in states of India, Work. Pap. 1994, Poverty and Human Resources, Development Research Group, The World Bank, Washington, D.C.

Henry, C., Sharma, M., Lapenu, C., Zeller, M., 2003. Microfinance poverty assessment tool, Tech. T. S. 5, Consultative Group to Assist the Poor (CGAP) and The World Bank, Washington, D.C. (PDF-File at http://www.cgap.org/).

Grootaert, C., Braithwaite, J., 1998. Poverty correlates and indicator-based targeting in Eastern Europe and the Former Soviet Union, Poverty Reduction and Economic Management Network, Environmentally and Socially Sustainable Development Network, The World Bank, Washington D.C.

Habicht, J. P., Pelletier, D.L., 1990. The importance of context in choosing nutritional indicators, J. Nutr.120, 1519-1524.

Hall, P., 1994. Methodology and theory for the bootstrap, Mathematical Sciences Institute, Australian National University, Canberra, (PDF-File at http://wwwmaths.anu.edu.au/).

IRIS. 2005. Note on assessment and improvement of tool accuracy, Mimeograph, revised version from June 2, 2005. IRIS center, University of Maryland.

Sahn, D.E., Stifel, D. C., 2000. Poverty comparisons over time and across countries in Africa, World Development. 28, 2123-2155.

Sharma, S., 1996. Applied multivariate techniques, John Wiley and Sons, New York.

United States Agency for International Development (USAID), 2005. Poverty assessment tools, AMAP, (Available [online] at http://www.povertytools.org), Accessed November 10, 2005.

Valdivia, M., 2005. Is identifying the poor the main problem in targeting nutritional program? Disc. Pap. 7, The World Bank, Washington D.C.

Weiss, J., 2004. Reaching the poor with poverty projects: what is the evidence on social returns? Res. Pap. 61, Asian Development Bank Institute, Tokyo.

Zeller, M., Alcaraz, V. G., Johannsen, J., 2005a. Developing and testing poverty assessment tools: results from accuracy tests in Bangladesh, IRIS Center, University of Maryland, College Park (Available [online] at http://www.povertytools.org).

Zeller, M., Johannsen, J., Alcaraz V.G., 2005b. Developing and testing poverty assessment tools: results from accuracy tests in Peru, IRIS Center, University of Maryland, College Park (Available [online] at http://www.povertytools.org).

Zeller, M., Alcaraz, V. G., 2005c. Developing and testing poverty assessment tools: results from accuracy tests in Kazakhstan, IRIS Center, University of Maryland, College Park (Available [online] at http://www.povertytools.org).

Zeller, M., Alcaraz, V. G., 2005d. Developing and testing poverty assessment tools: results from accuracy tests in Uganda, IRIS Center, University of Maryland, College Park (Available [online] at http://www.povertytools.org).

Zeller, M., Sharma, M., Henry, C., Lapenu, C., 2006. An operational tool for assessing the poverty outreach performance of development policies and projects: results of case studies in Africa, Asia and Latin America, World Development 34, 446-464.

## Annex

Table 1 Summary of PCA results for Kazakhstan

| Variables (10)<br>Poverty rate: 4.52% | Component Loadings<br>1 |
|---|---|
| **Kaiser-Meyer-Olkin measure of sampling adequacy: 0.804** | |
| Household head completed superior education | 0.526 |
| Do you have a mobile cell phone in the house | 0.627 |
| Floor is linoleum, dutch tile, or parquet | 0.591 |
| Toilet: shared or own flush toilet | 0.581 |
| Ownership of a blanket | 0.587 |
| Log of total resale value of animals and other assets | 0.602 |
| Pipe water ownership | 0.601 |
| Log value of dishes | 0.529 |
| Log value of air conditioner | 0.566 |
| Log value of metal pots | 0.715 |

Source: Own calculations based on IRIS survey data

Table 2 Summary of PCA results for Peru

| Variables (10)<br>Poverty rate: 26.88% | Component Loadings<br>1 |
|---|---|
| **Kaiser-Meyer-Olkin measure of sampling adequacy: 0.871** | |
| Percentage of adult household members who read and write | 0.554 |
| Number of rooms in the dwelling have | 0.540 |
| Mobile cell phone in the house | 0.490 |
| Ownership of a color TV | 0.743 |
| Number of refrigerators | 0.725 |
| Cooking fuel is bamboo/wood/sawdust collected | -0.724 |
| Toilet: pit toilet | -0.525 |
| Dummy: untreated piped/river water | -0.577 |
| Household has electricity (autobattery, own generator included) | 0.792 |
| Dummy, if any household member has a passbook savings account | 0.320 |
| Log value of food processing assets | 0.736 |

Source: Own calculations based on IRIS survey data

Table 3 Summary of PCA results for Uganda

| Variables (10)<br>Poverty rate: 32.36% | Component Loadings<br>1 |
|---|---|
| **Kaiser-Meyer-Olkin measure of sampling adequacy: 0.821** | |
| Floor is brick/stone, cement, or cement with additional covering | 0.762 |
| Do you have mobile (cell phone) in the house? | 0.550 |
| Dummy: private borehole or piped water | 0.626 |
| Dummy: roof with banana leaves, fibre, grass, bamboo or wood | -0.508 |
| Toilet: shared or own ventilated, improved latrine or flush toilet | 0.490 |
| Number of black/white TVs | 0.464 |
| Lighting source: gas lamp or electricity (neighbor, public or own socket) | 0.740 |
| Cooking fuel is charcoal or paraffin | 0.768 |
| Dummy: if household head has any account | 0.489 |
| Log value of jewelry | 0.452 |

Source: Own calculations based on IRIS survey data

Note: For purposes of brevity, the regression results are not shown in the annex. They can be obtained from the authors upon request.