




Alternative approaches for creating a wealth index: the case of Mozambique

Kexin Xie,¹ Achla Marathe ,^{2,3} Xinwei Deng,¹ Paula Ruiz-Castillo,⁴ Saimado Imputiua,⁵ Eldo Elobolobo,⁵ Victor Mutepa,⁵ Mussa Sale,⁵ Patricia Nicolas,^{4,5} Julia Montana,^{4,5} Edgar Jamisse,⁵ Humberto Munguambe,⁵ Felisbela Materrula,⁵ Aina Casellas,⁴ Regina Rabinovich,^{4,6} Francisco Saute,⁵ Carlos J Chaccour ,^{4,7,8} Charfudin Sacoor,⁵ Cassidy Rist ⁹

To cite: Xie K, Marathe A, Deng X, *et al.* Alternative approaches for creating a wealth index: the case of Mozambique. *BMJ Glob Health* 2023;**8**:e012639. doi:10.1136/bmjgh-2023-012639

Handling editor Seye Abimbola

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjgh-2023-012639>).

Received 21 April 2023

Accepted 31 July 2023

ABSTRACT

Introduction The wealth index is widely used as a proxy for a household's socioeconomic position (SEP) and living standard. This work constructs a wealth index for the Mopeia district in Mozambique using data collected in year 2021 under the BOHEMIA (Broad One Health Endectocide-based Malaria Intervention in Africa) project.

Methods We evaluate the performance of three alternative approaches against the Demographic and Health Survey (DHS) method based wealth index: feature selection principal components analysis (PCA), sparse PCA and robust PCA. The internal coherence between four wealth indices is investigated through statistical testing. Validation and an evaluation of the stability of the wealth index are performed with additional household income data from the BOHEMIA Health Economics Survey and the 2018 Malaria Indicator Survey data in Mozambique.

Results The Spearman's rank correlation between wealth index percentiles from four methods is over 0.98, indicating a high consistency in results across methods. Wealth rankings and households' income show a strong concordance with the area under the curve value of ~0.7 in the receiver operating characteristic analysis. The agreement between the alternative wealth indices and the DHS wealth index demonstrates the stability in rankings from the alternative methods.

Conclusions This study creates a wealth index for Mopeia, Mozambique, and shows that DHS method based wealth index is an appropriate proxy for the SEP in low-income regions. However, this research recommends feature selection PCA over the DHS method since it uses fewer asset indicators and constructs a high-quality wealth index.

INTRODUCTION

The elimination of extreme poverty in all its forms everywhere by 2030 is one of the major goals of the United Nations' Sustainable Development Agenda. Despite consistent and widespread progress, poverty remains a major problem in Africa.¹ Socioeconomic position (SEP) of households is a key indicator of poverty and generally measured in terms of income and consumer spending.

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Wealth index, derived from the assets of the household, is widely used as a proxy indicator for socioeconomic position (SEP) and living standard.
- ⇒ Importance of principal components analysis (PCA), in constructing the wealth index, has been well accepted by researchers.

WHAT THIS STUDY ADDS

- ⇒ This research provides an alternative to Demographic and Health Survey (DHS) methodology for constructing a wealth index in data poor regions, and gives insights into the effectiveness of using alternative approaches for creating a wealth index, including feature selection PCA, sparse PCA and robust PCA.
- ⇒ The feature selection PCA method, when applied to the district of Mopeia in Mozambique, shows an improvement over the DHS methodology since it reduces the required number of input variables by 40% and yet constructs a high-quality wealth index.
- ⇒ The method works for the Mopeia region as well as for the entire country, Mozambique.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ The proposed method reduces the burden of data collection on investigators and simplifies the construction of the wealth index.
- ⇒ For other low-income and middle-income countries which are data sparse, it will be easier to build the wealth index, which is a key indicator of the SEP of households.



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Achla Marathe; achla@virginia.edu

In low-income and middle-income countries (LMICs) such as African countries, where most of the population lives in rural areas, this information is not readily available. Hence, the wealth index is developed as a proxy measure for SEP in LMICs.² The wealth index is constructed using asset-based information including ownership of durable assets, housing characteristics and access to basic services, which are easier to obtain and

are more reliable in terms of data collection than income and consumption.²

Demographic and Health Survey (DHS) datasets are widely used with the classical PCA (principal components analysis) approach for the wealth index construction, known as the DHS Wealth Index.³ When it comes to explaining variation in education, child mortality, nutrition, fertility and healthcare, classical PCA wealth indices based on the DHS dataset frequently outperform spending data.²⁻⁷

Despite its long success, questions have been raised about the issues in the construction and interpretation of wealth indices. For example, both Houweling *et al*⁸ and Howe *et al*⁹ stated that the wealth index may differ between urban and rural areas. This is because the PCA procedure assigns more weight to indicators of assets owned by more urban households (eg, television, communication tools, electricity), and less or even negative weights to indicators of assets owned by rural households (eg, livestock, agricultural land), leading to a gross underestimation of the wealth of rural households. Thus, a recent development proposed a polychoric PCA wealth index with two principal components, to reduce the urban bias in standard PCA with one component.¹⁰ PCA could also be challenging to interpret due to extremely small weights,¹¹ and proxy methods such as combining categorical and sparse PCA (SPCA)¹² have been described. Such approaches, however, still have the issues of redundancy caused by hundreds of categories and sensitivity to outliers.

Here, we focus on building a wealth index for the district of Mopeia in Mozambique. With a per capita GDP of US\$541.5 in 2022 reported by the World Bank, 64% of the population in Mozambique is still below the extreme poverty line (international poverty line is US\$2.15 (2017 PPP) per day per capita).¹³ For LMIC and data sparse regions, this study offers new insights into the effectiveness of using alternative PCA approaches for creating a wealth index.

In addition to the classical PCA method, we apply three alternative approaches for building the wealth index: (1) Feature selection PCA approach, which only uses a subset of asset indicators for estimating wealth. (2) SPCA approach, which uses the well-known SPCA¹⁴ method using LASSO (elastic net) to provide an easily interpretable modified wealth index. (3) Robust PCA approach, which uses the popular robust PCA method, ROBPCA,¹⁵ to overcome the sensitivity of classical PCA to outliers.

METHODS

Data

Data used for this study are drawn from the demographic survey of the population in Mopeia in 2021 conducted under the Broad One Health Endectocide-based Malaria Intervention in Africa (BOHEMIA) study,^{16 17} including 25 550 households and 131 818 people in Mopeia district, Zambezia province, in Mozambique. Several authors of

this study were engaged in the original data collection phase of the BOHEMIA survey. The present research, however, uses data extracted retrospectively from the public dataset provided by the BOHEMIA demographic study.¹⁷ This dataset offers an extensive set of 72 indicators that capture detailed information about each participating household. Out of these, we focus on 16 variables (table 1) as the socioeconomic indicators. These particular variables were selected in accordance with relevant literature obtained from the DHS website.³ This method ensures a thorough and well-founded socioeconomic analysis, grounding our research in established methodologies.

The wealth index is validated using the household income, which was collected by the BOHEMIA team through the 2022 Health Economics Survey in Mopeia. The Health Economics survey gathered income information from 537 households for six consecutive months, including labour income from each member of the household, households' non-farm business income, and households' agricultural income. The estimation of total household monthly income in our analysis is derived by combining the individual monthly incomes of all members and all types of household monthly income. Incomes were initially retrieved in 2022 Mozambican meticals and later converted to 2022 US dollars (US\$) under the 2022 exchange rate of 63.85 metical/US\$.¹⁸

Classical PCA wealth index

We followed the steps used in constructing the DHS Wealth Index¹⁹ to calculate our classical PCA wealth index. The very first step is to convert each category of the 16 asset ownership variables into 71 dummy variables to form an assets' binary dataset. Then PCA²⁰ is carried out on the correlation matrix of the standardised assets' binary data. The wealth index for each household is a linear combination of all assets with the PCA weights as corresponding coefficients according to the formula described in online supplemental appendix section 1.1.

Feature selection PCA wealth index

The feature selection PCA selects a much smaller number of asset indicators to build the wealth index. In the classical PCA wealth index, several asset variables have extremely small absolute weights indicating that they are only weakly correlated to the first component. By ignoring variables with small-magnitude loadings, important features can be retained without losing much information.²¹ Here an absolute weight threshold of 0.01, which is approximately equal to the median of all PCA weights, is applied to filter out the negligible asset variables. A sensitivity analysis of choice of threshold was conducted and is described in online supplemental appendix section 2. Results suggest that researchers should carefully consider the threshold to ensure enough indicators are retained, thereby maintaining the robustness of the wealth quintiles.

Table 1 Percentage (or mean number) of households owning each asset indicator across wealth quintiles, and the regression results between each asset indicator and classical PCA (DHS) rank

	Mean (SE) or % N=25 550	Wealth quintiles					Regression*
		Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	β
No of constructions (++++†)	1.565 (0.890)	1.320 (0.636)	1.404 (0.738)	1.559 (0.874)	1.716 (1.009)	1.827 (1.026)	0.133***
No of members per sleeping room (++++)	4.030 (2.013)	3.348 (1.712)	3.776 (1.805)	4.115 (1.929)	4.365 (2.088)	4.586 (2.243)	0.306***
No of bed nets (++++)	2.149 (1.288)	1.301 (0.954)	1.823 (1.056)	2.201 (1.124)	2.580 (1.272)	2.843 (1.374)	0.384***
Ownership of cattle or pigs (---‡)	7.86%	12.00%	7.51%	7.42%	7.01%	5.36%	0.193***
No of cows (> 1 year of age)							
1–4 (++)	0.20%	0.04%	0.04%	0.35%	0.29%	0.27%	0.386***
5–9 (++)	0.06%	0.00%	0.02%	0.06%	0.10%	0.12%	0.617**
10 or more (++)	0.03%	0.00%	0.00%	0.02%	0.02%	0.10%	1.183*
No of cows (< 1 year of age)							
1–4 (++)	0.27%	0.14%	0.29%	0.27%	0.33%	0.29%	0.134
5–9 (+)	0.07%	0.10%	0.04%	0.08%	0.10%	0.06%	0.026
10 or more (++)	0.02%	0.00%	0.00%	0.00%	0.02%	0.06%	1.604
No of pigs (>6 weeks of age)							
1–4 (---)	4.90%	8.29%	4.86%	4.27%	4.17%	2.90%	0.253***
5–9 (---)	0.97%	1.27%	0.90%	1.00%	0.88%	0.82%	0.095*
10 or more (+)	0.28%	0.25%	0.24%	0.29%	0.29%	0.31%	0.064
No of pigs (<6 weeks of age)							
1–4 (---)	2.77%	4.61%	2.63%	2.54%	2.19%	1.88%	0.224***
5–9 (---)	0.92%	1.29%	0.88%	0.82%	0.78%	0.80%	0.119*
10 or more (+)	0.20%	0.27%	0.20%	0.14%	0.20%	0.22%	0.058
Main housing building type							
Apartment (++)	0.19%	0.00%	0.02%	0.08%	0.41%	0.43%	0.860***
Conventional house (++++)	13.17%	0.00%	0.00%	0.00%	0.68%	65.15%	5.610***
Flat (+)	0.09%	0.08%	0.12%	0.08%	0.10%	0.08%	0.022
Hut (----)	29.01%	40.69%	38.53%	33.46%	26.42%	5.93%	0.417***
Precarious (---)	19.53%	13.48%	19.66%	23.60%	28.18%	12.74%	0.045***
Traditional mud house (----)	36.91%	45.57%	41.15%	41.82%	42.13%	13.86%	0.274***
Other (++)	1.11%	0.18%	0.53%	0.96%	2.07%	1.82%	0.482***
Wall material in the main house							
Adobe block (----)	54.55%	68.42%	67.63%	65.62%	53.87%	17.22%	0.501***
Bamboo (----)	26.02%	51.73%	31.73%	23.86%	18.77%	3.97%	0.632***
Bark (----)	4.17%	11.14%	4.43%	2.90%	1.76%	0.61%	0.708***
Brick block (++++)	15.67%	0.00%	0.04%	0.94%	13.29%	64.09%	2.524***
Cardboard (++)	0.09%	0.04%	0.08%	0.10%	0.08%	0.16%	0.269.
Cement blocks (++++)	3.55%	0.00%	0.00%	0.00%	0.25%	17.48%	4.443***
Other (----)	26.05%	40.14%	33.41%	28.16%	22.15%	6.36%	0.433***
Palm tree (----)	9.38%	19.82%	9.88%	8.32%	6.71%	2.17%	0.493***
Paper (++)	0.12%	0.08%	0.16%	0.14%	0.10%	0.14%	0.049
Plastic bags (---)	1.62%	5.28%	1.33%	0.57%	0.55%	0.39%	0.839***
Reed (---)	3.80%	8.13%	3.92%	3.42%	2.97%	0.53%	0.483***

Continued

Table 1 Continued

	Mean (SE) or % N=25 550	Wealth quintiles					Regression*
		Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	β
Tin (++)	4.74%	1.17%	2.61%	5.38%	10.33%	4.19%	0.317***
Tinned wood (++)	4.63%	1.09%	2.53%	5.36%	10.20%	3.97%	0.316***
Wood (++++)	9.77%	3.01%	6.13%	10.70%	16.58%	12.43%	0.347***
Zinc (++++)	12.13%	0.00%	0.04%	0.20%	7.36%	53.03%	2.745***
Main water sources for consumption							
Fountain (++++)	10.74%	4.05%	6.94%	10.51%	17.01%	15.19%	0.502***
Hole protected with hand pump outside backyard (++++)	50.58%	34.47%	48.21%	54.27%	56.58%	59.37%	0.373***
Hole with manual pump inside house (++)	0.49%	0.10%	0.31%	0.57%	0.68%	0.76%	0.508***
Mineral bottled water (++)	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	15.185
Other (-)	0.05%	0.02%	0.10%	0.02%	0.10%	0.00%	0.083
Piped water in neighbour house (++++)	1.39%	0.00%	0.04%	0.18%	1.37%	5.36%	1.589***
Piped water inside house (++++)	0.38%	0.00%	0.00%	0.00%	0.02%	1.90%	4.614***
Piped water within compound (++++)	1.48%	0.00%	0.00%	0.00%	0.08%	7.32%	4.633***
Protected inside backyard (++)	0.46%	0.14%	0.16%	0.43%	0.84%	0.72%	0.353***
Protected outside backyard (---)	4.81%	6.92%	5.82%	5.11%	4.19%	2.00%	0.441***
Rainwater (--)	0.08%	0.12%	0.14%	0.10%	0.04%	0.00%	0.463*
Surface (river, lake, Lagoon) (----)	11.04%	20.85%	14.21%	10.96%	7.61%	1.55%	0.236***
Unprotected inside household (--)	1.78%	2.25%	2.31%	2.11%	1.72%	0.53%	0.258***
Unprotected outside household (----)	16.56%	31.05%	21.71%	15.62%	9.55%	4.87%	0.235***
Water from tank truck (-)	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.000
Time for taking main water sources							
Under 10 min (++++)	31.84%	6.02%	18.85%	37.14%	47.01%	50.20%	0.592***
Between 10–30 min (----)	45.63%	61.50%	53.67%	41.29%	35.32%	36.34%	0.283***
Between 30–60 min (----)	17.75%	24.96%	21.67%	16.91%	14.48%	10.70%	0.250***
More than 1 hour (---)	4.76%	7.52%	5.78%	4.66%	3.17%	2.66%	0.281***
Main energy source for lighting							
Batteries (----)	68.79%	84.70%	81.13%	79.22%	62.74%	36.18%	0.593***
Candles (-)	0.78%	0.06%	0.43%	1.23%	1.88%	0.29%	0.255***
Electricity (++++)	11.65%	0.00%	0.00%	0.00%	7.14%	51.10%	2.774***
Firewood (----)	11.96%	14.75%	16.28%	15.52%	11.33%	1.94%	0.300***
Gas (-)	0.01%	0.00%	0.02%	0.04%	0.00%	0.00%	0.169
Generator (++)	0.04%	0.00%	0.00%	0.00%	0.00%	0.18%	17.383
Oil (+)	0.15%	0.00%	0.08%	0.06%	0.51%	0.10%	0.459***
Other (-)	1.55%	0.39%	1.49%	1.82%	3.15%	0.92%	0.180***
Solar panel (++++)	4.66%	0.10%	0.43%	1.86%	12.02%	8.90%	0.820***
Ownership of radio (++++)	24.66%	7.93%	14.13%	22.80%	37.34%	41.12%	0.523***
Ownership of television (++++)	7.47%	0.02%	0.00%	0.04%	0.45%	36.85%	4.577***
Ownership of cell phone (++++)	38.47%	8.46%	21.28%	38.12%	57.63%	66.87%	0.746***

*The β's are the slope coefficients of the logistic regression for all binary indicators, or the slope coefficients of the linear regression for number of constructions, number of members per sleeping room, number of nets; p<0.1, *p<0.05, **p<0.01, ***p<0.001.
 †++++Positive PCA weight larger than 0.01; +++Positive PCA weight smaller than 0.01 and larger than 0.005; ++Positive PCA weight smaller than 0.005 and larger than 0.001; +Positive PCA weight smaller than 0.001.
 ‡----Negative PCA weight less than -0.01; ---Negative PCA weight larger than -0.01 and smaller than -0.005; --Negative PCA weight larger than -0.005 and smaller than -0.001; -Negative PCA weight larger than -0.001.
 DHS, Demographic and Health Survey; PCA, principal components analysis.

SPCA wealth index

Another popular variable selection technique, which develops accurate and yet sparse models is LASSO (elastic net). Zou and Hastie¹⁴ proposed SPCA) by imposing the LASSO (elastic net) constraint on the regression optimisation problem such that the modified PCA produces sparse loadings (explained in online supplemental appendix section 1.2). This efficient approach is integrated into wealth index construction in our paper, producing SPCA weights and a more interpretable wealth index.

Robust PCA wealth index

Classical PCA, feature selection PCA and SPCA methods are sensitive to anomalous observations. This is because the sample covariance or correlation matrix is very sensitive to outliers. Robust PCA is an effective way of obtaining principal components with little impact from outliers.

A well-known robust PCA method, called ROBPCA,¹⁵ determines a robust subspace by obtaining an outlier-free subset. The data are then projected onto this subspace to robustly estimate the eigenvectors and eigenvalues. However, ROBPCA is typically suited for roughly symmetric distributed data, which is not common in assets' binary data, especially in LMICs. Hubert *et al*²² proposed an improved ROBPCA algorithm, skewness-adjusted ROBPCA, to address the issue of imbalanced data. In this study, the construction of robust PCA wealth index is performed using skewness-adjusted ROBPCA, due to the imbalance in the BOHEMIA data.

Statistical analysis of wealth indices

Per DHS wealth index methodology, missing data in our analysis are replaced by the average value of the respective variables, and all the asset variables are standardised before applying the PCA algorithm. As for the parameter setting, the number of principal components is set to one in all four approaches as suggested by the DHS method.¹⁹ The robustness parameter is set as 0.5 to yield maximal robustness in the robust PCA process.

Because the wealth index presents only a relative ranking of households, it is difficult to interpret and compare the values of scores. To address these issues, one of the popular approaches is to transform the wealth index into wealth quintiles. Wealth quintiles are calculated by dividing all households into equal quintiles (20%) based on the wealth index. Households are categorised into five ranks from 'rank 1' to 'rank 5' with the wealth index scores from the first quintile to the last quintile.

We examine the reliability of the wealth index from two perspectives: the internal coherence and its consistency with other wealth indices. The internal coherence is examined using the summary statistics (percentage of assets ownership or average number of asset indicator) of how the assets' ownership varies across five quintiles. An intuitive heatmap is used to visualise the agreement between the four indices, where misclassification between quintiles can be observed directly. The association

between different PCA wealth indices is determined by the Spearman's rank correlation coefficients, a typical non-parametric measure of rank correlation.

We validate the wealth indices through Spearman's rank correlation coefficient to measure the association between wealth ventiles (calculated by dividing all households into equal 5% quantiles based on the wealth index scores) and logarithmic household income. The validation of wealth indices is also evaluated by the receiver operating characteristic (ROC) analysis which is commonly used to calculate predictive capacity of a classification model. An ROC curve is obtained by plotting sensitivity against 1-specificity for all possible cut-off points of wealth indices. Thus, the area under the ROC curve (AUC) can be used as an informative measure of the discriminating capacity of wealth indices, and the closer the AUC is to one, the better is the performance.²³

In terms of the stability of the wealth index on the other dataset, the three alternative PCA approaches are applied to the 2018 Malaria Indicator Survey (MIS) data in Mozambique and are compared with the original DHS wealth index reported on the DHS website.²⁴

Data analysis in this paper is performed by using software R V.4.1.2. The classical PCA algorithm is achieved through 'principal' function in 'psych' package.²⁵ To accomplish the alternative PCA techniques, we use 'elasticnet' package²⁶ and 'robpcap' package²⁷ to carry out the SPCA and skewness-adjusted ROBPCA algorithm, respectively.

Patient and public involvement statement

It was not appropriate or possible to involve patients or the public in the design, or conduct, or reporting, or dissemination plans of our research.

Reflexivity statement

A structured reflexivity statement is provided in online supplemental appendix.

RESULTS

Importance of different asset indicators across methods

Figure 1 reports the PCA weights for each of the asset indicators for each of the four PCA approaches using all 69 variables from the BOHEMIA demographic survey data. The weights signify the relative contribution different assets make to the wealth index.

In the classical PCA result (see the first plot in figure 1), almost all wealth index coefficients have expected signs. Variables indicating wealth (eg, lighting by electricity, wall material made of zinc) have positive weights while those representing poverty (eg, living in a hut, lighting by firewood) have negative weights. However, there are also some unexpected results. For instance, the 'number of members per sleeping room' variable carries a positive weight, indicating that a wealthier household has more people per sleeping room.

More simplified PCA approaches, that is, feature selection PCA, SPCA and robust PCA—resulted in a



Figure 1 Coefficients of asset indicators from classical PCA, robust PCA, feature selection PCA and sparse PCA in Mopeia, Mozambique. PCA, principal components analysis.

succinct list of relevant household assets, trimming off variables with negligible weights. Interestingly, none of these methods assigned any weight to ‘cows’ ownership’ variable. Moreover, some variables that were used by a minority of households (eg, paper for wall material, natural gas for lighting) were also ignored by three alternative PCA methods due to sparseness. It is worth to noting that while the threshold choice in feature selection PCA may seem subjective, its outcomes align with those of SPCA. This agreement justifies setting 0.01 as the threshold value in the filtering criteria in feature selection PCA.

Evaluation of wealth indices

All households are ranked into five levels: rank 1 (poorest), rank 2, rank 3, rank 4, rank 5 (richest). Each

group represents 20% of total households. In this section, we investigate the internal coherence and cross-method agreement of the wealth indices.

Internal coherence of wealth indices

Table 1 compares the average asset ownership across the households in five wealth quintiles based on the classical PCA approach. Since the other three techniques have similar results as the classical one, their data are not shown here (see online supplemental appendix section 3). The percentage of households, owning assets with positive weights, generally rises from rank 1 to rank 5. Conversely, the opposite trend is observed for assets with negative weights. For example, only 8.4% of households in rank 1 own cell phones, but this percentage increases for higher ranks, reaching 66.9% in rank 5.

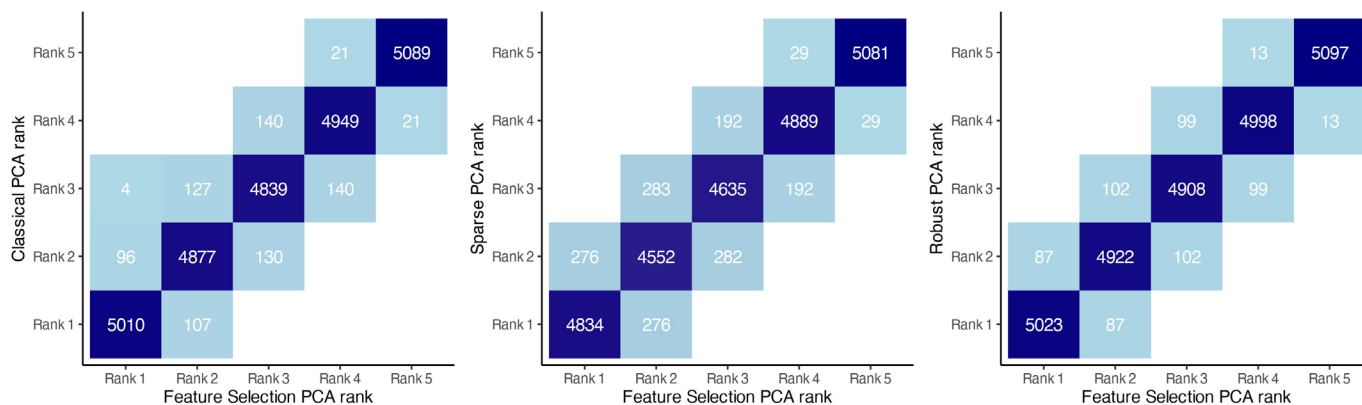


Figure 2 Numbers of households under different wealth quintiles created using classical PCA, feature selection PCA, robust PCA and sparse PCA. PCA, principal components analysis.

Table 1 also presents the regression analysis between each asset indicator and classical PCA (DHS) rank. The linear regression and the logistic regression are used to examine the association between wealth ranks with assets for numerical indicators and binary indicators, respectively. The signs and the magnitude of regression coefficients are consistent with those of the PCA weights, particularly when the absolute PCA weights are more than 0.001. Conversely, indicators with tiny absolute weights (<0.001) exhibit limited discriminatory capacity between wealth groups, as indicated by the statistically insignificant results of the regression analysis.

Cross-method agreement of wealth indices

Figure 2 demonstrates the wealth ranks according to the four PCA techniques. The value in each cell represents the number of households under the corresponding combination of wealth quintiles from three approaches. For instance, among all households that have rank 1 in classical PCA, 99.75% of households are classified in rank 1, with remaining 0.25% of households classified into rank 2 in robust PCA. Almost all households are classified in the same group in all four indices. Approximately, only 10% of the households appear in a different group, which is an adjacent group with a difference of only one rank. Additionally, the strong Spearman's correlation coefficients (all >0.98 with $p < 2.2 \times 10^{-16}$) across all methods confirm this agreement (the table of spearman's rank correlation is shown in online supplemental appendix table S4). Consequently, the wealth quintiles derived from four different PCA approaches are robust to insignificant asset indicators and outliers.

External validation of wealth indices

To further validate our indices, we cross-checked against household income data and the 2018 DHS wealth index.

Consistency with household income classification

Income information for 537 households in the Mopeia district is available from the Health Economics Survey data collected by BOHEMIA project in 2022. However, nearly all households reported zero personal income (93.67%, N=503), and a majority of households reported

zero household income (77.47%, N=416). Here, we only illustrate the association between the wealth indices and 138 households with non-zero total income. The analysis for households with zero total income is demonstrated in online supplemental appendix section 6.

The feature selection PCA wealth ventiles showed a moderate association (Spearman's rank correlation=0.26) yet significant positive linear relationship ($p < 0.02$) with the average monthly income on a log scale of 138 households with non-zero total income. All wealth ventiles exhibit similar results, which are included in online supplemental appendix table S5.

Classifying these households as 'rich' or 'poor' according to the 2022 international poverty line (US\$2.15 (2017 PPP) per day per capita) from World Bank Report,¹³ we used ROC curves to compare income classification with wealth indices. Figure 3 shows the ROC curves, with the feature selection PCA wealth index demonstrating a high AUC value of 0.76 for the classification results, suggesting its robust discriminating capacity. Notably, all other wealth indices also exhibit strong performance, with AUC values exceeding 0.75 (see online supplemental figure S3). Both the Spearman's correlation and ROC analysis show significant coherence between wealth indices and households' average monthly income.

Consistency with DHS wealth index based on 2018 MIS data in Mozambique

The DHS household wealth index for Mozambique has been estimated and reported on the DHS website based on the 2018 MIS data.²⁴ Applying all three alternatives to the 2018 MIS data, we investigate the consistency between the alternative PCA wealth index with the original DHS wealth index, and therefore, justify the stability of the alternative methods over different data sets. Figure 4 demonstrates a strong correlation (Spearman's rank correlation=0.99, $p < 2.2 \times 10^{-16}$) between the DHS wealth index and the feature selection PCA wealth index as well as between the wealth quintiles (Spearman's rank correlation=0.95, $p < 2.2 \times 10^{-16}$). Using a threshold of 0.02 (approximately equal to the median of all PCA weights), feature selection PCA filtered out 63 of 107 asset indicators in the

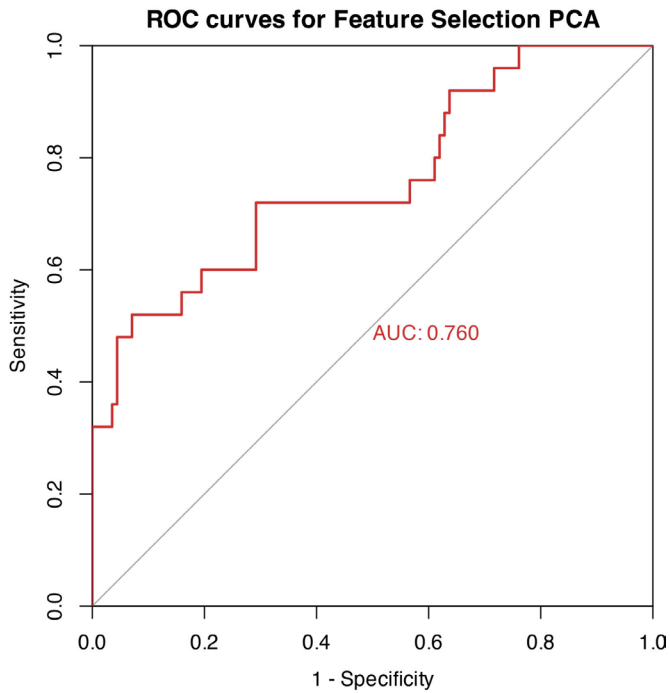


Figure 3 ROC curves and AUC values for classification results of feature selection PCA wealth index according to the income-based poverty status (N=138). AUC, area under the curve; PCA, principal components analysis.

2018 MIS data from Mozambique, without significantly affecting the quality of the wealth index.

DISCUSSION

Wealth indices have been widely used as a proxy measure of SEP for households in LMICs. One popular way of creating these indices within DHS datasets is the PCA

approach, despite criticisms regarding its reliability and the data burden it creates.^{9 28} This work offers a new way of calculating the wealth index for the Mopeia district in Mozambique. This alternative methodology removes less significant features and handles outliers more efficiently.

The analysis shows that wealth indices produced by four PCA techniques exhibit strong internal coherence and offer a viable indicator of the SEP of households. The three alternative indices are highly consistent with the classical PCA wealth index, a benchmark for the standard wealth index. Each wealth index exhibits a significant correlation with household income and hence an ability to discriminate between households' levels of poverty.

The alternative methods, feature selection PCA, SPCA and robust PCA, use fewer asset indicators (about 40% less), while still being able to construct a reliable wealth index. A concise set of indicators and a simpler model could make it easier for researchers to identify and comprehend the major contributors. These results are in line with other studies^{29 30} which support the idea of designing a simpler questionnaire with fewer number of questions to assess SEP. A shorter questionnaire has a positive effect on response rate and response quality, which further improves the accuracy of the follow-up studies or strategies.³¹

Both SPCA and robust PCA, while guaranteeing the high quality of the wealth index, may compromise computational simplicity. Therefore, the feature selection PCA wealth index is recommended over the classic DHS wealth index. Its efficiency is demonstrated by its ability to generate comparable results while requiring 40% less information—a significant reduction in the burden of data collection and computational load. However, one

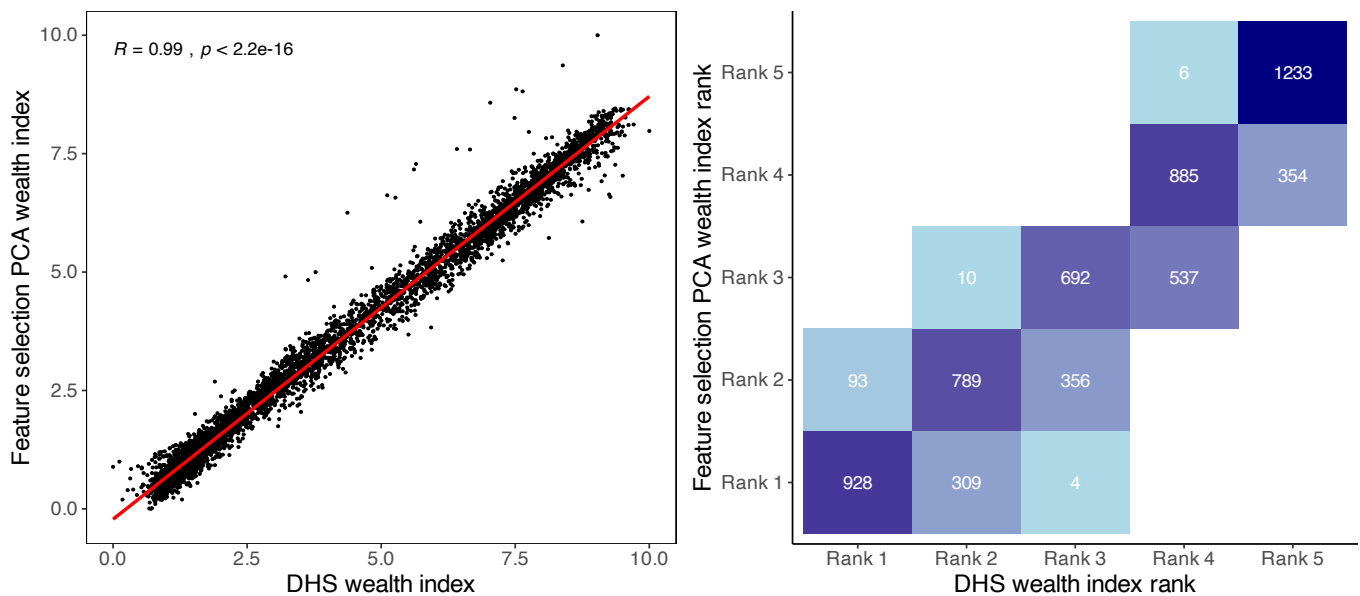


Figure 4 Correlation between DHS household wealth index and the feature selection PCA wealth index, and numbers of households under different wealth quintiles created using DHS wealth rank and feature selection PCA wealth rank. DHS, Demographic and Health Survey; PCA, principal components analysis.

aspect to consider is that the selection of the optimal threshold may not be universally applicable across regions. If a wealth index in other regions is needed, the DHS data from previous years can be analysed, if available, to identify valid thresholds and to eliminate insignificant asset indicators before the questionnaire-development stage, reducing the burden on investigators and survey respondents.

Limitations

There could be some limitations in the proposed method and analysis. First, the first principal component from all PCA methods only explains a fraction of variation in the data. Some studies consider multiple principal components to capture more wealth effect in the data.^{10,32} However, there is an inevitable trade-off between the higher explained variance and a clear interpretation of the contribution of each asset indicator. Moreover, the total proportion of variance explained may not be considerably higher since the successive higher-order components always explain smaller proportions than the first component.²⁸

Another drawback is possibly the lack of generalisability due to the high level of rurality in Mozambique. Pursuant to the DHS method, we initially created a composite wealth index by combining different urban and rural wealth indices.¹⁹ However, the analysis revealed that it was indistinguishable from a unified wealth index. Hence, we developed only one wealth index for Mozambique, without differentiating between urban and rural areas. Although our method works well for Mozambique (as shown in figure 4), its performance in more urbanised LMICs is unclear. Further analysis using diverse datasets from various socioeconomic settings is warranted to evaluate the generalisability of our method.

CONCLUSIONS

This research presents a new approach for calculating a wealth index for Mopeia, Mozambique. The commonly used DHS method based wealth index is effective, but we find the feature selection PCA approach achieves comparable performance while using 40% less variables. We identify variables that make minimal contribution in calculating the wealth index, omit them and show that their elimination does not affect the quality of the wealth index. This simplifies the data collection process and reduces the cost of data collection while improving the quality of the survey results. Despite using fewer asset indicators, feature selection PCA delivers a stable and robust wealth index, and shows consistency in performance with other methods, including the DHS method. Thus, we recommend the feature selection PCA approach as a practical alternative for wealth index calculations in similar LMIC regions.

Author affiliations

¹Department of Statistics, Virginia Tech, Blacksburg, Virginia, USA

²Network Systems Science and Advanced Computing Division, Biocomplexity Institute, University of Virginia, Charlottesville, Virginia, USA

³Department of Public Health Sciences, University of Virginia, Charlottesville, Virginia, USA

⁴Barcelona Institute for Global Health (ISGlobal), Hospital Clínic-Universitat de Barcelona, Barcelona, Spain

⁵Centro de Investigação em Saúde de Manhiça, Maputo, Mozambique

⁶Harvard University T H Chan School of Public Health, Boston, Massachusetts, USA

⁷Centro de Investigación Biomédica en Red de Enfermedades Infecciosas, Madrid, Spain

⁸Facultad de Medicina, Universidad de Navarra, Pamplona, Spain

⁹Department of Population Health Sciences, Virginia-Maryland College of Veterinary Medicine, Virginia Tech, Blacksburg, Virginia, USA

Acknowledgements We acknowledge support from the Spanish Ministry of Science, Innovation and Universities through the 'Centro de Excelencia Severo Ochoa 2019-2023' Program (CEX2018-000806-S), and support from the Generalitat de Catalunya through the CERCA Program. CISM is supported by the Government of Mozambique and the Spanish Agency for International Development (AECID). We thank the residents and authorities of Mopeia for their support. We acknowledge the work and dedication of over 360 local staff during implementation and data collection.

Contributors KX, AM, XD, CJC and CR conceived the idea presented. The methodology was developed by KX, AM, XD and CS. The project implementation and social science study are supported by SI, VM, MS, PN, JM, EJ, HM, FM, AC and RR. The investigation process and the project administration were conducted by AM, XD, PR-C, SI, EE, VM, MS, PN, FS and CS. EE coordinated the data curation and software. Writing the original draft and preparing the figures and tables was performed by KX and all authors contributed to the review and editing of the final draft. All authors contributed cognitively to various areas of this work, participated in later changes, and read and approved the final publication. All authors had access to all study data and held the ultimate responsibility for the publication decision. AM is responsible for the overall content as guarantor of this paper.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The study protocol for the BOHEMIA study was approved by the Internal Scientific Committee and Institutional Review board from the Centro de Investigação em Saúde de Manhiça (Ref: CIBS-CISM/004/2021), Hospital Clínic of Barcelona Clinical Research Ethics Committee (Ref: HCB/2019/0938) and The Ethics Research Committee of the WHO (Protocol ID: ERC.0003265).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. The BOHEMIA consortium has agreed to make the data underlying each manuscript openly available on publication. The data supporting this paper is generating from BOHEMIA project, which can be found in the ISGlobal dataverse repository (<https://doi.org/10.34810/data682>).

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Achla Marathe <http://orcid.org/0000-0002-0258-1588>

Carlos J Chaccour <http://orcid.org/0000-0001-9812-050X>

Cassidy Rist <http://orcid.org/0000-0002-7558-8094>

REFERENCES

- 1 United Nations. Ending poverty. n.d. Available: <https://www.un.org/en/global-issues/ending-poverty>
- 2 Filmer D, Pritchett LH. Estimating wealth effects without expenditure data--or tears: an application to educational enrollments in States of India. *Demography* 2001;38:115–32.
- 3 Rutstein SO, Johnson K. The DHS wealth index. 2004. Available: <https://dhsprogram.com/publications/publication-cr6-comparative-reports.cfm>
- 4 Yaya S, Bishwajit G, Shah V. Wealth, education and urban–rural inequality and maternal Healthcare service usage in Malawi. *BMJ Glob Health* 2016;1:e000085.
- 5 Booysen F, van der Berg S, Burger R, et al. Using an asset index to assess trends in poverty in seven sub-Saharan African countries. *World Development* 2008;36:1113–30.
- 6 Boah M, Azupogo F, Amporfro DA, et al. The epidemiology of Undernutrition and its determinants in children under five years in Ghana. *PLOS ONE* 2019;14:e0219665.
- 7 Vuković D, Bjeđović V, Vuković G. Prevalence of chronic diseases according to socioeconomic status measured by wealth index: health survey in Serbia. *Croat Med J* 2008;49:832–41.
- 8 Houweling TA, Kunst AE, Mackenbach JP. Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter *Int J Equity Health* 2003;2:8.
- 9 Howe LD, Hargreaves JR, Gabrysch S, et al. Is the wealth index a proxy for consumption expenditure? A systematic review. *Journal of Epidemiology & Community Health* 2009;63:871–7.
- 10 Martel P, Mbofana F, Cousens S. The Polychoric dual-component wealth index as an alternative to the DHS index: addressing the urban bias. *J Glob Health* 2021;11:04003.
- 11 Jolliffe IT, Trendafilov NT, Uddin M. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics* 2003;12:531–47.
- 12 Merola GM, Baulch B. Using sparse categorical principal components to estimate asset indices: new methods with an application to rural Southeast Asia. *Rev Dev Econ* 2019;23:640–62. 10.1111/rode.12568 Available: <https://onlinelibrary.wiley.com/toc/14679361/23/2>
- 13 Global_Poveq_Moz. 2023. Available: https://databankfiles.worldbank.org/public/ddpext_download/poverty/987B9C90-CB9F-4D93-AE8C-750588BF00QA/current/Global_POVEQ_MOZ.pdf
- 14 Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *J Computat Graphical Stat* 2006;15:265–86.
- 15 Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: A new approach to robust principal component analysis. *Technometrics* 2005;47:64–79.
- 16 Chaccour C, Casellas A, Hammann F, et al. BOHEMIA: broad one health Endectocide-based malaria intervention in Africa—a phase III cluster-randomized, open-label, clinical trial to study the safety and efficacy of Ivermectin mass drug administration to reduce malaria transmission in two African settings. *Trials* 2023;24:128.
- 17 Ruiz-Castillo P, Imputiua S, Xie K, et al. BOHEMIA a cluster randomized trial to assess the impact of an Endectocide-based one health approach to malaria in Mozambique: baseline demographics and key malaria indicators. *Malar J* 2023;22:172.
- 18 Banco de Moçambique. *Taxas de Câmbios de Referência do Mercado Interbancário (Cotações)*, Personal communication. 2023.
- 19 Rutstein SO. *Steps to constructing the new DHS Wealth Index*. Rockville, MD: ICF International, 2015.
- 20 Pearson K. On lines and planes of closest fit to systems of points in space. *London, Edinburgh, and Dublin Philosophical Magazine J Sci* 1901;2:559–72.
- 21 Cadima J, Jolliffe IT. Loading and correlations in the interpretation of principle Components. *J Appl Stat* 1995;22:203–14.
- 22 Hubert M, Rousseeuw P, Verdonck T. Robust PCA for SKEWED data and its Outlier map. *Computat Stat Data Analy* 2009;53:2264–74.
- 23 Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters* 2006;27:861–74.
- 24 National Institute of Health/INS, National Institute of Statistics/INE, National Malaria Control Program/PNCM, ICF. In: *National Survey on Malaria Indicators in Mozambique 2018 [Internet]*. Maputo, Mozambique: INS/Mozambique, INE, PNCM, ICF, 2019. Available: <http://dhsprogram.com/pubs/pdf/MIS33/MIS33.pdf>
- 25 Revelle W. Psych: procedures for psychological, Psychometric, and personality research. 2022. Available: <https://CRAN.R-project.org/package=psych>
- 26 Hastie HZ, elasticnet T. Elastic-net for sparse estimation and sparse PCA. 2020. Available: <https://CRAN.R-project.org/package=elasticnet>
- 27 Reynkens T, Hubert M, Schmitt E, et al. Rospca: robust sparse PCA using the ROSPCA algorithm. n.d. Available: <https://CRAN.R-project.org/package=rospca>
- 28 Howe LD, Hargreaves JR, Huttly SRA. Issues in the construction of wealth indices for the measurement of socio-economic position in low-income countries. *Emerg Themes Epidemiol* 2008;5:3.
- 29 Chakraborty NM, Fry K, Behl R, et al. Simplified asset indices to measure wealth and equity in health programs: A Reliability and validity analysis using survey data from 16 countries. *Glob Health Sci Pract* 2016;4:141–54.
- 30 Garenne M, Hohmann-Garenne S. A wealth index to screen high-risk families: application to Morocco. *J Health Popul Nutr* 2003;21:235–42.
- 31 Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value Health* 2011;14:1101–8.
- 32 Poirier MJP, Grépin KA, Grignon M. Approaches and alternatives to the wealth index to measure socioeconomic status using survey data: A critical interpretive synthesis. *Soc Indic Res* 2020;148:1–46.