Yoon, Jisu; Pasha, Atika

Working Paper

An alternative strategy to identify deprivations in multidimensional poverty: A partial least squares approach

Discussion Papers, No. 271

Provided in Cooperation with:
Courant Research Centre 'Poverty, Equity and Growth in Developing and Transition Countries', University of Göttingen

Suggested Citation: Yoon, Jisu; Pasha, Atika (2020) : An alternative strategy to identify deprivations in multidimensional poverty: A partial least squares approach, Discussion Papers, No. 271, Georg-August-Universität Göttingen, Courant Research Centre - Poverty, Equity and Growth (CRC-PEG), Göttingen

This Version is available at:
https://hdl.handle.net/10419/219035

Courant Research Centre

'Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis'

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 271

An alternative strategy to identify deprivations in multidimensional poverty: a partial least squares approach

Jisu Yoon, Atika Pasha
May 2020

Platz der Göttinger Sieben 5 · 37073 Goettingen · Germany
Phone: +49-(0)551-3921660 · Fax: +49-(0)551-3914059

Email: crc-peg@uni-goettingen.de Web: http://www.uni-goettingen.de/crc-peg

An alternative strategy to identify deprivations in multidimensional poverty: a partial least squares approach

Jisu Yoon* Atika Pasha †

May 31, 2020

Abstract

This study determines data driven weights for the indicators in the multidimensional poverty index (MPI), based on partial least squares (PLS), using income as the outcome variable. Consequently, the resulting MPI is particularly useful to income related policy and research questions. An innovative data driven procedure is proposed to determine the first cut-offs of the MPI inside of the PLS algorithm, which provides an alternative to the first cut-offs based on researchers' judgement. Another adjustment to the PLS procedure enables the weights to respect the existing practice in the MPI literature, that health, education and living standard dimension are equally important. The new MPI can consider heterogeneous observations by means of interaction terms in the weighting structure. Using this approach, a new MPI is created considering the additional deprivation of the black population in South Africa, compared to other racial groups. This MPI shows different weights and first cut-offs than the old MPI. It suggests that the first cut-offs for the year of education indicator needs to be 12 years instead of 5 years to have a practical relevance for the South African context. Additionally, the weight of assets is important and electricity less so. The black population shows higher deprivation for all considered deprivation indicators within the MPI using interaction terms.

*Courant Research Centre "Poverty, equity, and growth", Georg-August-Universität Göttingen, Germany, E-mail: jisu.yoon@zentr.uni-goettingen.de; This paper was conceptualized during Jisu Yoon's employment at this institution.

†Department of Econometrics, University of Mannheim, L7 3-5, 68167 Mannheim, Germany

1 Introduction

Increasing amounts of scientific literature establishes the needs for a multidimensional measure of poverty, departing from a unidimensional measurement, based on income or consumption for instance. The seminal concept of the capability approach, famously discussed by Sen (1985; 1999), has been influential in this research. Following his suggested approach of a "multidimensional" measurement of poverty, several indices (such as the Human Development Index (HDI) or the Gender Related Development Index (GDI); UNDP, 2017a,b) and dashboard approaches (for instance the Millennium Development Goals (MDG)) were implemented. These measures account for a broad-based measurement of wellbeing and development, and do not rely on a single indicator (say only income or consumption) to determine the poverty status.

The Multidimensional Poverty Index (MPI) is a typical example of such comprehensive measures. The MPI was proposed by the Oxford Human Development Initiative (OPHI) and the United Nations Development Programme (UNDP). This index operationalizes the capability approach to measure poverty based on three dimensions, which are health, education and living standards. Each dimension is comprised of multiple indicators, which measures deprivations in that particular dimension. Indicators are aggregated using weights, which determine the relative importance of the indicators in the MPI as well as the relative importance between the dimensions. Traditionally, each dimension is considered to be equally important (1/3 weights for health, education and living standards), and indicators in each dimension are treated as equally important. For example, 2 indicators (nutrition and child mortality) in the health dimension get 1/6 weight each (1/3 weight for health * 1/2 for each indicator).

The dual cut-off approach (Alkire and Foster, 2011a) is a widely accepted approach to construct the MPI. This approach involves two cut-offs, as its name suggests. The first cut-off transforms observed variables to dichotomous indicators, which identifies whether

a household[1] is considered to be deprived or not (taking values 1 and 0 respectively). This transformation usually relies on *a priori* selected thresholds, which are the first "cut-offs" determining poverty in each indicator. For example, consider a variable "drinking water", which is an ordinal variable showing various sources of drinking water. If accessing water through a public tap is the threshold, the categories ranking below this threshold are considered to be deprived, which in the case of our data are accessing water through a flowing water/stream, spring and dam/pool/stagnant water. Depending on the reported drinking water source, each household will be determined to be deprived (or not) in this indicator. The dichotomized indicators are then aggregated to build a weighted score for each household, where a higher score means more poverty. If the score is above a certain threshold, i.e., the second cut-off, the household is considered to be poor.

To generate the weighted index, the indicators are traditionally aggregated by normatively determined weights (e.g., indicators in each dimension being equally important), which are not free from the arbitrary judgment of a researcher. Moreover, it could be challenging to adjust the weights to a local context. For instance, cooking fuel might be more important compared to having access to drinking water in rural communities with abundant fresh water. Alternatively, if electricity is available to almost no household in a region, it may provide only little information on poverty compared to toilet type, for instance.

Typically, data driven weighting schemes such as principal component analysis (PCA; Hotelling, 1933) emphasize the largest variations in indicators to derive weights, which are not necessarily informative in practice. Therefore, this paper uses partial least squares (PLS; Wold, 1966) to determine the weights in the MPI. In our PLS application the covariance between per capita income of a household and multidimensional poverty will be maximized. In other words, the weights are based on the assumption that low per capita income is related to more poverty, and the resulting MPI will capture the part of poverty particularly relevant to income. With increasing income, a poor household

[1]It could also be an individual, which is out of scope of this paper.

will spend resources to escape from the poverty in the relevant indicator depending on its priority. Hence, the covariance between income and poverty can be beneficial in identifying the weights.

Apart from generating a weighting structure, this paper also utilizes an innovative data based approach to derive the first cut-offs in the MPI. So far in the literature, the thresholds for dichotomizing the indicators have been determined based on empirical understanding and practical considerations (Alkire and Santos, 2014). However, it is not easy to determine the thresholds suitable for local settings. For example, the deprivation in education can be defined in terms of the lack of tertiary, secondary or elementary education, and it is not clear which is most relevant for the population under consideration. Therefore, our work provides an adjustment to the PLS method, such that it determines the thresholds for each indicator. Analogous to the weights on indicators, the thresholds are identified based on the assumption that low income is related to more poverty. This approach provides new insights on the deprivation level for each indicator of the MPI and depends less on the judgement of researchers, which may be influenced by subjective beliefs.

An additional adjustment to the PLS algorithm enables us to reflect existing knowledge onto the data driven weighting scheme. The new PLS algorithm respects the existing tradition in the multidimensional poverty literature to treat the three dimensions (health, education and living standards) as equally important (Atkinson et al., 2002). The proposed method adjusts the optimization in the PLS algorithm, but includes certain constraints to accord equal weights to all three dimensions, i.e., one third each.

The normative weights to the three dimensions contrast to the data driven weights for the indicators. In other words, this study agrees with the existing procedure that the these three dimensions are equally important, but challenges that indicators in each dimension are equally important. The three dimensions are regarded as equally important "values",

but indicators just "measurements" for relevant dimensions. Of course, this procedure is based on our judgement and is arguable. But the proposed method can be adjusted to other kind of judgements and is useful, when researchers want to mix their judgements to data driven weighting structure.

To empirically test our approach, we use the National Income and Dynamics Study (NIDS; Southern Africa Labour and Development Research Unit, 2016) data from South Africa, using three waves from 2010 to 2014. The NIDS is a rich panel dataset, that allows one to track not only the spatial differences but also the dynamic changes within multidimensional poverty of a household. It contains all indicators to build a typical MPI, except the information on flooring. Therefore, the MPI from this dataset contains only nine indicators instead of ten. In addition, this dataset provides detailed information on the income earned by the household, upon which the thresholds and weights are based.

The presence of high income inequality in South Africa provides a ripe background to study these new sets of indices. The literature shows that the largest incidence of income and multidimensional poverty is found within the Black households, in comparison to any of the other races in South Africa (Alkire and Santos, 2014; Finn and Leibbrandt, 2013). With the help of the NIDS data, we explore the racial divide in poverty between the Black households, compared to the White, Indian/Asian and Coloured households in South Africa.

The PLS approach has an advantage that it is very easy to model the additional deprivation of the Black households by means of interaction terms. Therefore, this study presents a MPI with different weights for the Black population compared to the other races, in addition to the typical MPI with common weights for all races. The MPI with the different weights takes account with the average household income difference of poor Black households and others.

To our best knowledge, this paper is the first to use PLS to identify weights for the MPI.

Previous works in the literature focused on either normative weights or data driven weights emphasizing the largest variance. There have been no published articles that presented a data driven methodology to determine the thresholds for poverty measurement in a multidimensional index. There is a working paper from Dotter and Klasen (2019), which used median to determine the first cut-off. However, median has a limitation in identifying indicator poverty. For example, median education could be elementary school education. However, household welfare may depend largely on college or university education, which brings large difference in job opportunities. The restrictions on the weights (i.e., 1/3 weight on each dimension) and the interaction term approach are new to the literature as well.

The results show that the adjusted PLS identifies the first cut-offs differently to those from the traditional approach of the OPHI. Notable differences are found in the thresholds for categorizing households as poor in the indicators of education, water source, cooking fuel, phone and assets. In terms of the weighting structure, certain indicators receive very different weights compared to the traditional MPI weights. For example, in the living standard dimension, assets are found to be more than twice important than electricity. Finally, in the MPI using the interaction terms, different weights are assigned to the Black population, which can be interpreted as the additional deprivation of the black race compared to others.

This paper will be organized as follows. Section 2 briefly reviews the literature on multidimensional measures of poverty and then discuss the literature on the MPI. Section 3 recapitulates the PLS algorithm and describes the adjustments. Section 4 explains the data and provides a few descriptive statistics. Section 5 discusses the results from the analyses and the last section will conclude.

2 Literature Review

Money-metric measures of poverty are considered to be lacking complexity in defining "real" poverty, which not only implies a lack of income, food or housing, but also implies inaccessibility to other functionings such as being of sound health, having access to the desired level of education, being employed, living in a proper house, etc. (Sen, 1985, 1999; UNDP, 2015). Therefore, to capture a broader definition of deprivation, and its dynamics over time, recent literature has shifted focus to encompass the notion of *wellbeing* itself (Alkire and Foster, 2011b,a; Alkire and Santos, 2014). There are several studies that discussed the merits of multidimensional measures of poverty over unidimensional, or more precisely, income based measures (Alkire and Foster, 2011b; Klasen, 2000; Nussbaum, 2003; Sen, 1999). The changing focus, away from income and towards the real freedoms that people possess, based on their capability to undertake activities that are directly linked to the expression of this freedom (for instance, being able to read or being healthy due to affordable medical systems) was first clearly outlined by Sen in the Capability Approach (CA). This approach of linking the capabilities to the available functioning was then extended by several other philosophers and economists (Sen, 1985, 1999; Nussbaum, 2008; Nussbaum et al., 1993).

A more recent application of the CA is the multidimensional poverty index (MPI). In 2011, Alkire and Foster (2011b) proposed a multidimensional poverty measurement methodology, based on a dual cut-off approach. They provided directions on how to aggregate various dimensions of deprivation into a single composite index, and thereby construct a unified, measurable definition of wellbeing. The MPI is an application of the Alkire Foster (AF) method, using three dimensions of wellbeing- health, education and living standard- at the household level, thereby incorporating wider indicators of wellbeing than only income or expenditure.

There has been a large literature that describes the challenges of using a multidimensional

measure of poverty (as within the AF method). These are related to the selection of an appropriate weighting scheme within the chosen dimensions (Decancq and Lugo, 2013; Ravallion, 2011, 2012; Sen, 1999; Pasha, 2017), or the issues in measuring the inequality within the dimensions and populations (Chakravarty and D'Ambrosio, 2006; Jayaraj and Subramanian, 2010; Rippin, 2012a,b; Silber, 2011). Other studies have dealt with the difficulty in deriving the correct choice of indicators to represent these dimensions, or the need to adjust the dimensions in line with average wellbeing, such that the weakly relative nature of wellbeing and income correspond across a wide range of countries (Dotter and Klasen, 2014; Ravallion and Chen, 2011). [2]

Another aspect that has received little attention in the literature is the setting of the thresholds, which categorizes a household as poor or non-poor within a given indicator. So far these thresholds have been determined mostly arbitrarily by each study, i.e., relying on the opinion of a variety of experts. However, expert opinion may not always reflect a good judgement on the context of the indicator, i.e., geography, climate, religion, culture or any other socio-economic and demographic factors. Moreover, it may not always be possible to gather expert opinion rigorously in a typical research setting. Alkire and Santos 2014 didn't provide a straightforward justification for their thresholds either and referred to the Millennium Development Goals (MDG) as the basis for their thresholds. However, the MDG is an abstract concept and does not provide a clear cut solution on the threshold determination, e.g, the MDG does not say whether a household with "Water-Carrier/tanker" as the drinking water source can be considered to be poor quality drinking water in South Africa or not.

Given the current stand in the literature, as well as the typical normative approaches to multidimensional poverty measurement, this study employs a data driven approach to remedy these gaps.

[2] "the weakly relative nature of wellbeing" states that if all dimensions that form the wellbeing indicator are improving, the value of the indicator itself should also improve. In this case, setting an absolute cut-off might not work in countries that have vastly different wellbeing levels.

3 Methodology

This paper proposes two adjustments to the PLS algorithm. The adjustments determine the first cut-off of the indicators in the MPI (i.e., the thresholds to consider the household to be deprived), while respecting the tradition that considers the three dimensions (health, education and living standard) as equally important. After recapitulating the PLS algorithm briefly, these adjustments will be discussed.

PLS identifies the score (poverty score in our application) by maximizing the empirical covariance between the outcome variable and covariates. Consider a simple linear regression model, $y = X\beta + \varepsilon$. $y \in \mathbb{R}^N$ and $X \in \mathbb{R}^{N\times K}$ are centered, where $\mathbb{E}(\epsilon|X) = 0$, $cov(\epsilon|X) = \sigma^2 I_n$, $K \leq N$. The first score is a weighted average of covariates:

$$t_1 = Xw_1,\ t_1 \in \mathbb{R}^N. \tag{1}$$

Weights are defined as $w_1 = \underset{\|\omega\|=1}{\operatorname{argmax}} (\omega' X' y)^2 \in \mathbb{R}^K$. Since the MPI is a single vector, the parts of the PLS algorithm to calculate the second and later scores are not discussed. For a more detailed discussion on the PLS algorithm, see Höskuldsson (1988).

This paper proposes maximizing the empirical covariance in the PLS algorithm in terms of the first cut-offs within the MPI, in addition to weights for each indicator. Consequently, the first cut-offs can be defined by the PLS algorithm, which is not possible in the existing PLS method.

The MPI typically transforms each covariate to a Boolean variable (indicator) before building a weighted average, which can be expressed as follows:

$$f(x_{ij}, \theta_j) = \begin{cases} 1 & \text{if } x_{ij} < \theta_j \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where x_{ij} and θ_j are i-th observation of the j-th covariate and the j-th cut-off, respectively. $x_{ij} \in \{0, 1, ..., m_j\}$ is an ordinal variable and $\theta_j \in \{1, 2, ..., m_j\}$. For example, x_{ij} can be water source, where the responses are "well", "flowing water/stream", "spring",..., "piped

(tap) water in dwelling", listing the range of water sources possible to a household, starting from the worst to the best. If a household's water source being a spring is considered to be the cut-off, all households using well and flowing water/stream as their water source will be considered to be deprived in this indicator (water source).

The next step is applying this function to the columns of the covariate matrix and centralize $X^F = F(X, \Theta) = C(f(x_1, \theta_1), f(x_2, \theta_2), ..., f(x_K, \theta_K))$, where $C()$ denotes a function to centralize the columns of the matrix and $x_j = (x_{1j}, x_{2j}, ..., x_{Nj})'$.

Weight is defined as $w_1 = \underset{\|\omega\|=1,\Theta}{\operatorname{argmax}} (\omega' X^{F\prime} y)^2 = \underset{\|\omega\|=1,\Theta}{\operatorname{argmax}} (\omega' F(X, \Theta)' y)^2 \in \mathbb{R}^K$. The optimization of this target function proceeds by considering all possible combinations of Θ, while w_1 is calculated using the usual PLS approach (1). For example, drinking water has 10 candidates for thresholds and sanitation has 5 candidates for thresholds. There are then 10*5=50 combinations of potential thresholds from these two indicators. Note that the number of combinations increases quickly with increasing number of covariates. For instance, given the 9 indicators and various candidates, the number of thresholds in our application was 83160.

The next adjustment incorporates a normative judgement of applying equal weights to the three dimensions in the MPI (education, health and living standard), where each is considered equally important. This is achieved by putting a restriction on the scaling of indicators in the PLS algorithm. Indicators are usually scaled before applying PLS, but the proposed method determines the scaling inside of the PLS algorithm, which can be interpreted as setting a restriction on weights.

Consider the following adjustment to (1), which explicitly considers scaling: $t_1 = X^F S_D w_1$, where $S_D = diag(s_1, s_2, ..., s_K)$ is a diagonal matrix containing scaling for each indicator. An alternative notation is $t_1 = X^F S \odot w_1$, $S = (s_1, s_2, ..., s_K)'$, with $\odot$ being elementwise vector multiplication.

A linear restriction is imposed on scaling: $RS \odot w_1 = r$, where

$$R = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } r = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}. \tag{3}$$

The restriction on scaling can be interpreted as a restriction on weights. $Rw_1^S = r \leftrightarrow RS \odot w_1 = r$, where $S \odot w_1 = w_1^S$ are the weights in terms of unscaled indicators. The first row of the restrictions means that the weights for the two indicators in the health dimensions sum up to 1/3, while the next two rows have analogous meaning for the education (with 2 indicators) and the living standard (with 5 indicators) dimensions.

There are an infinite number of scalings satisfying the aforementioned restrictions, because there are only 3 restrictions for the 9 scaling variables. It is necessary to make additional assumptions to the scaling structure. This study proposes applying autoscaling to the indicators before applying PLS, which can be interpreted as an additional assumption, which is, variables in each dimension (health, education and linving standard) are considered to be equally important. Autoscaling is a typical approach in the literature (e.g., Wold et al., 2001), which prevents any variable from having an advantage over another variable only for having large variance, which isn't always informative. Autoscaling will determine 9 scaling variables. These scaling variables in each dimension will be scaled up or down such that each dimension is equally important. This is implemented by allowing two parts in the scaling, $S = S_A \odot S_B$. The first term, $S_A = (Var(X_1^F)^{-\frac{1}{2}}, Var(X_2^F)^{-\frac{1}{2}}, ..., Var(X_9^F)^{-\frac{1}{2}})$, sets each indicator to be equally important (autoscaling). S_B scales the indicator in each dimension up or down, such that education, health and living standard are equally important.

$$RS \odot w_1 = RS_A \odot S_B \odot w_1 = \begin{bmatrix} s_{A1}s_{B1}w_{1,1} + s_{A2}s_{B2}w_{1,2} \\ s_{A3}s_{B3}w_{1,3} + s_{A4}s_{B4}w_{1,4} \\ s_{A5}s_{B5}w_{1,5} + ... + s_{A9}s_{B9}w_{1,9} \end{bmatrix} = r.$$

s_{Aj}, s_{Bj} and w_{1j} are the j-th element of S_A, S_B and w_1. Indicators in each dimension are scaled up or down in the same magnitude, i.e., $s_{B1} = s_{B2}$, $s_{B3} = s_{B4}$ and $s_{B5} = s_{B6} = ... = s_{B9}$. It means that indicators within each dimension are considered to be equally

important before applying PLS, since the relative size of S_A within each dimension will not change. In other words, autoscaling is maintained within each dimension.

$$\begin{bmatrix} s_{B1}(s_{A1}w_{1,1} + s_{A2}w_{1,2}) \\ s_{B3}(s_{A3}w_{1,3} + s_{A4}w_{1,4}) \\ s_{B5}(s_{A5}w_{1,5} + ... + s_{A9}w_{1,9}) \end{bmatrix} = r$$

The terms in the parenthesis are obviously non-zero.

$$\begin{bmatrix} s_{B1} \\ s_{B3} \\ s_{B5} \end{bmatrix} = \begin{bmatrix} r_1(s_{A1}w_{1,1} + s_{A2}w_{1,2})^{-1} \\ r_2(s_{A3}w_{1,3} + s_{A4}w_{1,4})^{-1} \\ r_3(s_{A5}w_{1,5} + ... + s_{A9}w_{1,9})^{-1} \end{bmatrix}$$

Therefore, $S_B = R'((R(S_A \odot w_1))^{-1} \odot r)$, where $(R(S_A \odot w_1))^{-1}$ is the reciprocal of vector $RS_A \odot w_1$. An additional restriction $S_B \geq 0$ is introduced in this equation, by means of taking the absolute value of w_1. Negative scaling, i.e., $S_B \leq 0$, will change the interpretation of the score (low value meaning more poverty) without adding any merit. Without this restriction, S_B may change its sign in every iteration in the PLS algorithm, preventing the convergence during optimization.

$$S_B = R'((R(S_A \odot abs(w_1)))^{-1} \odot r).$$

The adjustments in this section can be implemented via Algorithm 1.

4 Data

This paper uses the South African NIDS panel data (Southern Africa Labour and Development Research Unit, 2016), considering 3 waves of individual and household information, collected over a period of 2010, 2012 and 2014.[3] Of the typical 10 deprivation indicators in the original MPI, information on flooring is missing in the NIDS data. Therefore, the deprivation score for each household was calculated using the remaining 9 indicators.

For the PLS algorithm, it is important to choose an outcome variable, which has high covariance with the deprivation indicators. Household per capita income was chosen as the

[3]The NIDS data 2008 wave was not considered due to the missing information on the date of death of the child, which was essential for calculating the indicator "mortality".

Algorithm 1: PLS algorithm with the first cut-off estimation and partly normative weights

input : X, Y, Θ, R, r
output: w_1, c_1, t_1, u_1, S_A, S_B
for $l \leftarrow 1$ **to** L **do**
  $S_A = diag(Var(F(X, \Theta_l)))^{-\frac{1}{2}}$
  initialize u_1 and S_B
  repeat
    $w_1 = (F(X, \Theta_l)S_{DA}S_{DB})'u_1 / \|(F(X, \Theta_l)S_{DA}S_{DB})'u_1\|$
    $t_1 = F(X, \Theta_l)S_{DA}S_{DB}w_1$
    $c_1 = Y't_1 / \|Y't_1\|$
    $u_1 = Yc_1$
    $S_B = R'((RS_{DA}abs(w_1))^{-1} \odot r)$
  until *convergence of* $S_{DA}S_{DB}w_1$
  if *l=1* **then**
    Store the outputs
  else
    Replace the outputs if $t_1'u_1$ is larger than the stored $t_1'u_1$.
  end
end
Note: $S_{DA} = diag(S_A)$, $S_{DB} = diag(S_B)$, $\Theta = (\Theta_1, ..., \Theta_L)$

outcome variable because a strong covariance between household income and deprivation is expected. The per capita value was used instead of household aggregate income, which is likely to be more relevant to the capability of household members, and to account for the size of the household. For example, with an increasing number of children and a fixed household income, the amount of resources to be allocated towards the education of each child decreases. Moreover, all indicators building the MPI score are observed at the household level, which likely have a lower covariance with individual level income than household level income. Logarithm was taken on household per capita income, to take account of the typical skewness of income variable. Per capita household income has advantages over per capita expenditure as the outcome variable. It is well known that rich households spend a lower proportion of income compared to poor households. It suggests that expenditure has less variance, and likely a lower covariance, to poverty.

Table 1 shows summary statistics of the data, which contains in total over 98000 obser-

Table 1: Summary statistics of the data, per year

Variable	Total	2014	2012	2010
Individuals	98,417	37,787	32,822	27,808
Households	23,878	9,574	7,982	6,322
Household expenditure	4,477.81	5,060.09	4,043.60	4,199.08
Household income	6,477.77	7,372.16	6,107.53	5,699.41
Per capita household income	1,487.98	1,748.85	1,390.67	1,248.34
Number of children	2.9	2.9	2.9	3.1
Household Size	6.0	5.9	6.0	6.3
Black	83.4%	84.0%	82.6%	83.7%
Coloured	13.3%	13.3%	13.8%	13.0%
Asian/Indian	0.9%	0.8%	1.0%	1.0%
White	2.3%	1.9%	2.6%	2.4%
Years of Education	7.8	7.9	7.7	7.6
Age	39.2	38.8	39.3	39.4

Monetary variables are in the nominal Rand scale.
Variable names in italic for individual level averages.

vations, which comprise about 24000 households. It can be seen that monthly per capita household income, on average, is around 1488 Rand, which comes to around 100$, while household income itself is around 4 times the per capita income. Also, noticeably, the household income average in the data is always higher the household expenditure average. The sample has on average around 3 children and 6 individuals in a household. The largest share of the data consists of Blacks, (around 84%), while the colored, Asian/Indian and white South Africans make up the remaining with 13%, 1% and nearly 2% respectively. Finally, the average person in the data has around 8 years of education, and is around 40 years of age. Our descriptives here were calculated on individual level data and include individual level averages, to provide a better representation of the entire population.

Table 2: Traditional MPI

Dimensions	Indicators	Deprived if	Weights
Health	Child mortality	Any dead child in the household	1/6
	Nutrition	Any one of the two condition is met: 1. adult malnutrition in terms of BMI is identified (BMI $\leq$ 18.5) 2. child malnutrition in terms of z-score is identified in the household (z-score $\leq$ -2)	1/6

Education	Years of education	No household member with at least one 5 year education	1/6
	School attendance	Any school-aged child (age 7 to 15) not attending school	1/6
Living standard	Electricity	No electricity	1/15
	Drinking water	No Piped (tap) water in dwelling, on site or in yard	1/15
	Sanitation	No flush toilet	1/15
	Cooking fuel	Cooking fuel other than electricity, gas and paraffin	1/15
	Assets	Any one of the two condition is met: 1. Less than 2 small assets (radio, tv, phone, bicycle, fridge) 2. not having car	1/15

Note: A household is considered to be without a phone if it has neither landline nor cell phone.

Table 2 shows the traditional MPI as defined in Fintel and Zoch (2015): The data were prepared following Fintel and Zoch (2015). There were 16 covariates, from which 9 indicators were built. The table depicts the relationship between observed covariates and deprivation indicators. Note that in the traditional MPI, the weights do not vary within a dimension.

The proposed PLS method requires ordinal variables. The majority of the covariates had a clear order, but some did not. Drinking water, sanitation and cooking fuel had many categories. The NIDS provides an order on those covariates, but the ordering can be contested. For example, public tab water is considered to be a better water source compared to borehole on site. However, public tab water may involve several difficulties (e.g., long queue and efforts to get water, rationing, available only certain time on a day) and could be an inferior category. And public tab water may be used by poor households in the city, with less capacity compared to well-off households in the rural area using borehole on site as the water source.

Therefore, these covariates were coded as ordinal variables based on the mean per capita household income. The rationale behind this approach is that with increasing income, a

Table 3: Coding of drinking water

mean per capita household income	category	coding
705.10	Well	0
833.00	Flowing water/stream	1
860.59	Spring	2
895.24	Dam/pool/stagnant water	3
1250.92	Public tap	4
1303.49	Other	5
1365.99	Rain-water tank on site	6
1379.33	Water-Carrier/tanker	7
1930.90	piped (tap) water on site or in yard	8
3672.08	Borehole on site	9
3978.56	Borehole off site/communal	10
3994.66	piped (tap) water in dwelling	11

household will shift to the water source category providing higher welfare. Table 3 shows the mean household income for each category of drinking water. The category with the lowest mean household income value (Well) was coded as 0, the lowest ordinal value. Other categories were coded in integer value in ascending order with higher household income value. The coding for other covariates will be explained when we discuss the analysis results in Table 4 in Section 5.

The proposed PLS method considers various combinations of first-cutoffs, which is a major difference to the traditional MPI. For ordinal variables, the candidates of first-cutoffs and their combinations are straightforward. If there are 3 ordinal variables with 5 categories, there are $(5-1)^3 = 64$ combinations of first-cutoffs. However, the data contains three numerical variables, child and adult malnutrition and year of education. Child and adult malnutrition were measured in z-scores and BMI, respectively, and education was measured in years of completed education. These variables were transformed to ordinal variables using several potential thresholds. Child malnutrition is defined as having one or more malnourished child in the household. Having at least one severely undernourished (z-score<-3) child in the household is the lowest category, at least one moderately

undernourished (z-score<-2) child is the medium category and no undernourished child is the highest. Having at least one severely underweight (BMI<16) adult in the household is the lowest category of adult malnutrition. At least one underweight (BMI<15) adult in the household is the next category, followed by no underweight adult category (ranked highest).[4] Having no educated persons in a household is the lowest category in education followed by at least one person with at least 5 years, 12 years and 14 years of education. The 5 year cut-off was motivated by the traditional approach (Fintel and Zoch, 2015), while the 12 year and 14 year cut-offs were chosen due to the natural tendency of individuals to graduate from the education system, as these are the years required to get high school diploma and a two year college diploma, respectively.[5]

The aforementioned transformations of numerical variables to ordinal variables (on z-score, BMI and education in year) are not the only solutions. For example, 16 year of education could have been considered to account for 4 year university bachelor's degree and 18 year for master's degree. However, it is not possible to consider all potential thresholds, because the computational costs increase multiplicatively with additional candidates. Therefore, a limited number of thresholds were selected based on literature and practical consideration.

The data from 2014, from which we built the MPI and determined the proposed thresholds, had 9574 household level observations after keeping only the complete observations of the outcome variable (household log per capita income), the covariates and the black dummy. The black dummy is defined as more than 50% of household member being of African origin. The data from 2010 and 2012 are used only for descriptive statistics, which will be elaborated further in Section 5.3.

[4]See the WHO websites for the classification based on z-score and BMI (WHO, 2017a,b).

[5]Additionally, the mean household income showed large increases for these thresholds, establishing empirical evidence for our assumptions as well.

5 Empirical analysis

5.1 Empirical models

This study used the following empirical model to determine the association between household per capita income and the MPI indicators.[6]

$$y = \beta_0 + X_1^F\beta_1 + X_2^F\beta_2 + ... + X_9^F\beta_9 + \varepsilon \tag{4}$$

Another empirical model was used to investigate the interaction between the black population and deprivations.

$$y = \beta_0 + X_1^F\beta_1 + X_2^F\beta_2 + ... + X_9^F\beta_9 + X_1^F * B\beta_{10} + X_2^F * B\beta_{11} + ... + X_9^F * B\beta_{18} + \varepsilon \tag{5}$$

The black dummy is denoted by B. The interaction terms model the differences between the black and the non-black households in terms of the association between the deprivation in an indicator and per capita household income. For example, if black households deprived in water source have lower per capita household income compared to the others with the same deprivation, the corresponding coefficient will be negative. Note also that the black dummy appears in the interaction terms, but does not appear as a level term (i.e., no $B\beta_{19}$ in Eq. (5)). It means that the MPI with the interaction term will consider the different meaning of deprivation indicator for black households, but being a black household itself does not contribute to poverty.

The PLS scores were extracted from the aforementioned two models (Eq. 4 and Eq. 5), using Algorithm 1[7].

[6]The household per capita income, y, was centralized when performing PLS, but when linear regressions were performed, it was not centralized to identify the intercept, β_0.

[7]Note that there was a small deviation to Eq. (2), to transform the observed ordinal variables to the required deprivation indicators. As discussed in Section 4, some indicators were built from several variables involving special transformations (e.g., assets). These transformations were easy to be incorporated to the algorithm, simply by making $f(x_{ij}, \theta_j)$ depend on several variables and incorporate those special transformations in the program.

5.2 Results

Table 4 shows the first cut-offs identified by the proposed method. Each row shows the coding of each variable, with 0 being the lowest category and better categories taking higher integer values. The columns "Deprived" and "Not deprived" show the deprivation identified by our methodology. The deprivation following the traditional approach (Fintel and Zoch, 2015) is shown in red and italicized. The column "Covariate" shows names of the 16 covariates, which are used in the construction of the 9 indicators in the MPI. In case several covariates build up one indicator, the name of the indicator is shown in the parenthesis (for nutrition and assets). The 17th row, "# small assets" is an intermediate product to build the indicator "assets" (i.e., the number of assets possessed by the household among radio, television, phone, bicycle and fridge).

The health dimension is reported in rows 1 to 3. All italicized categories in red (traditional first cut-offs) are in the column "deprived" (first cut-offs from our approach), meaning that our approach identified the same first cut-offs as the traditional approach. Therefore, being non-deprived in the *child mortality* indicator is identified as having no dead children in the household, and similarly, if no adult in the household has a BMI of less than 15 and no child has a z-score less than -2, then the household is not deprived in the *nutrition* indicator.

The education dimension is reported in rows 4 and 5. As per the PLS approach, to be categorized as non-deprived in the *years of education* indicator, a household needs to have at least one person with at least 12 year education in our approach. This differs from the traditional approach, where having one person with only 5 years of education suffices to be non-deprived. Since elementary education in South Africa is mandatory, 5 years of education captures the majority of the population (see the statistics from World Bank, 2017) and are likely less informative compared to categorization based on 12 years of education. Deprivation in *school attendance* is identified as having even a single school

aged child in the household that is not attending school- same as the traditional approach. The living standard dimension is reported from rows 6 to 17. For Boolean variables, our approach and the traditional one is always same, since there was no room for any alternative cut-offs. This is the indicator *electricity* and the covariates *radio*, *tv*, *cell phone*, *bicycle*, *fridge* and *car*, which build the indicator *assets*. Another deviation in our approach is that it considers only piped (tap) water in the dwelling as non-deprived in the *drinking water* indicator. The traditional approach considers both, piped (tap) water on site or in yard and piped (tap) water in dwelling, as non-deprived. Since the category piped (tap) water on site or in yard is associated with lower per capita household income, compared to borehole on site and borehole off site/communal, the traditional approach may create a distortion in the overall poverty figure (at least in terms of income poverty). Water from borehole can have better quality than piped (tap) water depending on the environment and region. Additionally, piped (tap) water on site or in yard may signal that several households sharing water, which can be associated with several limitations (waiting in queue, rationing).

Table 4: The first cut-offs

	Covariate	Deprived	Not deprived
1	Child mortality	*0. One or more dead children*	1. No dead child
2	Adult malnutrition (nutrition)	*0. One or more malnourished adults, BMI threshold 16*, *1. One or more malnourished adults, BMI threshold 18.5*	2. No malnourished adult
3	Child malnutrition (nutrition)	*0. One or more malnourished children, z-score threshold -3*, *1. One or more malnourished children, z-score threshold -2*	2. No malnourished children
4	Years of education	*0. No educated person*, 1. At least one 5 year educated person	2. At least one 12 year educated person, 3. At least one 14 year educated person
5	School attendance	*0. One or more than one not-enrolled children*	1. No not-enrolled child
6	Electricity	*0. No*	1. Yes

7	Drinking water	*0. Well*, *1. Flowing water/stream*, *2. Spring*, *3. Dam/pool/stagnant water*, *4. Public tap*, *5. Other*, *6. Rain-water tank on site*, *7. Water-Carrier/tanker*, 8. piped (tap) water on site or in yard, *9. Borehole on site*, *10. Borehole off site/communal*	11. piped (tap) water in dwelling
8	Sanitation	*0. None*, *1. Pit latrine with ventilation pipe*, *2. Bucket toilet*, *3. Pit latrine without ventilation pipe*, *4. Chemical toilet*, *5. Other*	6. Flush toilet with offsite disposal, 7. Flush toilet with onsite disposal
9	Cooking fuel	*0. Animal Dung*, *1. Other*, *2. Wood*, *3. Coal*, *4. Solar energy*, 5. Paraffin	6. Electricity from mains, *7. None*, 8. Electricity from generator, 9. Gas
10	Radio (assets)	*0. No*	1. Yes
11	TV (assets)	*0. No*	1. Yes
12	Landline (assets)	*0. No*	*1. Yes-Currently not in working condition*, 2. Yes-Currently in working condition
13	Cell phone (assets)	*0. No*	1. Yes
14	Bicycle (assets)	*0. No*	1. Yes
15	Fridge (assets)	*0. No*	1. Yes
16	Car (assets)	*0. No*	1. Yes
17	# small assets	*0*	*1*, 2, 3, 4, 5

A household is considered non-deprived in the indicator *sanitation* if it has a flush toilet with off-site disposal or a flush toilet with onsite disposal in both approaches. *Cooking fuel* is considered to be non-deprived if a household use electricity from mains, none, electricity from generator and gas in our approach. The traditional approach identifies non-deprivations as using paraffin, electricity from mains, electricity from generator and gas. Our approach differs from the traditional approach that paraffin is considered to be deprived, while none is considered to be non-deprived. We didn't expect the category none to be highly associated with income, but this category covers only 16 households out of 9574 in total. These small number households could be regarded as exceptions, potentially

urban households with high income but without any cooking needs. It is also reassuring that the data based coding approach takes account with unexpected phenomenon in the data. In our approach, having a land line suffices to be considered as having access to phone, but in the traditional approach a household need to have a land line in a working condition. Rows 10 to 15 are small assets. A household is non-deprived in *assets*, if it has more than certain number of small assets and a car. In the traditional approach, the number of small assets to be considered not deprived is 2 or more, but in our approach it is enough to have 1 small asset. Overall, our approach has led to a different first-cutoffs of in the *years of education*, *drinking water*, *cooking fuel* and *assets* indicators, compared to the traditional approach.

In the following, the simple and interaction models are reported, which estimated the association between the MPI from our approach and per capita household income.

Table 5: Coefficients

	Coefficients in simple model	Coefficients in interaction model
intercept	7.20***	7.20***
MPI	-1.54***	-1.67***
r^2	0.161	0.168

Note: *** p<0.01, ** p<0.05, * p<0.1. N=9574

Table 5 shows the regression coefficient between log per capita household income and the MPI score (before the second cut-offs) by our approach. The conditional correlation between the MPI and the outcome variable is negative and statistically significant at 1% for the simple and interaction model. The inference was based on the Jack-knife confidence interval from Martens and Martens (2000). This confidence interval is constructed by resampling the data, while keeping the first cut-offs as fixed.

The coefficients in Table 5 can be transformed to the coefficients in terms of the indicators

Table 6: Weights and coefficients in terms of the variables in the MPI

	Weights in simple model	Coefficients in simple model	Weights in interaction model	Coefficients in interaction model
Child mortality	0.174	-0.27***	0.083	-0.14***
Nutrition	0.159	-0.24***	0.076	-0.13***
Years of education	0.173	-0.27***	0.085	-0.14***
School Attendance	0.160	-0.25***	0.079	-0.13***
Electricity	0.043	-0.07***	0.022	-0.04***
Drinking water	0.062	-0.10***	0.032	-0.05***
Sanitation	0.071	-0.11***	0.037	-0.06***
Cooking fuel	0.068	-0.10***	0.035	-0.06***
Assets	0.090	-0.14***	0.047	-0.08***
Child mortality*black			0.091	-0.15***
Nutrition*black			0.083	-0.14***
Years of education*black			0.084	-0.14***
School attendance*black			0.085	-0.14***
Electricity*black			0.024	-0.04***
Drinking water*black			0.030	-0.05***
Sanitation*black			0.036	-0.06***
Cooking fuel*black			0.035	-0.06***
Assets*black			0.034	-0.06***

Note: *** p<0.01, ** p<0.05, * p<0.1. N=9574

in the MPI using a simple transformation.

$$\begin{aligned} y &= \hat{\beta}_0 + MPI\hat{\gamma}_1 + \hat{\varepsilon} \\ &= \hat{\beta}_0 + X^F(S_A * S_B * w_1)\hat{\gamma}_1 + \hat{\varepsilon} \\ &= \hat{\beta}_0 + X^F\hat{\beta} + \hat{\varepsilon} \end{aligned} \tag{6}$$

This can be interpreted as a regularized regression of Model (4) and Model (5). These coefficients are reported in Table 6 (weights $S_A * S_B * w_1$ and coefficients $\hat{\beta}$).

The column "Weights in simple model" shows the weights of the indicators in a model without interaction terms (Model (4)). In the health dimension, child mortality has a little larger weight than nutrition, meaning that mortality covaries with the outcome variable stronger than nutrition variable[8]. The years of education and school attendance

[8]The covariance between log per capita household income and deprivation indicators was always negative in our data. The PLS weights were negative at the beginning, but they were rotated 180 degree

indicators in the education dimension also have similar weights, while the years of education indicator has slightly stronger covariance to the outcome variable, compared to child school enrolment. Based on the weights, assets is the most important indicator in the living standard dimension, followed by sanitation, cooking fuel, drinking water and electricity. The difference in weights is substantial in this dimension. For example, assets are more than twice important than electricity.

Note that the weights in each dimension sum up to 0.333 (1/3) in analogy to the traditional approach. Therefore, health, education and living standards are treated equally important in this MPI.

The column "Coefficients in simple model" shows the coefficients of the model without interaction. All indicators in the MPI score predict low per capita household income. The size of the coefficients are proportional to the size of the weights and all of them are statistically significant. The coefficients can be interpreted as the predicted percentage change in the outcome variable if a household is deprived in the respective deprivation indicator. For example, if a household is deprived in the years of education indicator in the education dimension (row 3), the household is predicted to have approximately 27% lower per capita household income than the non-deprived households.[9]

The column "Weights in interaction model" shows the weights of the indicators in the model with the interaction terms (Model (5)). The weights are composed of level terms, from the first to the ninth row, and interaction terms, from the tenth to the eighteenth row. The weights of the level terms are roughly half the size of the weights of the simple model and the relative importance of indicators within each dimension is similar to the simple model. The interpretation of the level terms is therefore analogous to the weights of the simple model.

(i.e., multiplied by -1). Consequently, a high value of MPI score implies a higher level of poverty in this MPI, analogous to the traditional MPIs.

[9]Of course, one may avoid the log approximation, which leads to $e^{-0.27} - 1 \approx -0.237$. It means 23.7% lower predicted per capita household income.

All weights of the interaction terms are positive, meaning that the black households with respective deprivation have lower per capita household income than other households. This can be interpreted as the additional burden of deprivation that the black households face. The interactions in the health dimension shows that mortality has a slightly higher weight than nutrition, similar to the weights of level term. However, the interaction terms in the education and living standard dimensions are not similar to the weights of the level terms. The interaction term of school attendance is larger than the years of education indicator in the education dimension, while it is opposite in the level terms. This illustrates that school attendance matters more for black households compared to other households. In the living standard dimension, among the interaction terms, sanitation has the largest weight, followed by cooking fuel, assets, water source and electricity. In other words, the interaction term approach not only identifies the additional deprivation of black households, but also the different association between income and deprivation indicators for black households compared to others.

Note that the sum of the level and interaction terms show the total weights for the black households. For example, in the living standard dimension, the sum of the level and the interaction terms is 0.046, 0.062, 0.073, 0.070 and 0.079 for electricity, water source, toilet, cooking fuel and assets. Additionally, the property of equal weighting among the three dimensions is still maintained, even after including interaction terms. The sum of the weights of the level and interaction terms of each dimension is always 1/3.For example, in the health dimension, the sum of weights are 0.083+0.076+0.091+0.083=0.333.

The column "Coefficients in interaction model" shows that both level and interaction terms predict lower log per capita household income statistically significantly. The size of the coefficients is proportional to the weights and the interpretation is similar to the simple model coefficients. The interaction terms show that black households have lower log per capita household income than other households. For example, a household deprived in

years of education is predicted to have 14% lower per capita household income than the non-deprived. But if a black household has the same deprivation, it's predicted to have 14+14=28% lower per capita household income than the non-deprived.

The interaction model identifies the additional deprivation of the black households compared to others. The empirical model predicted that if a black household has a deprivation in an indicator, it suffers additionally from lower income compared to other households with the same deprivation. We found this to be an interesting empirical exercise for a composite index taking account of the heterogeneity among observations. However, if this index is used at the policy level with resource allocation, it may invite a sensitive ethical question, as it will discriminate people according to ethnic origin. For example, a white household may not be qualified for a poverty alleviation program, while a black household with exactly the same poverty indicators qualifies. Such practice could harm social cohesion. Another limitation of the interaction model approach is that the regression coefficients are consistent (in the statistical sense) if multidimensional poverty can indeed be summarized in a single vector. Although taking a single vector is common in the literature, there is no guarantee that it reflects the reality. Considering several PLS scores will however beat the purpose of having a single easy-to-interpret summary measure.

Table 7: Discrepancies in weights between new and traditional method

Indicator	Traditional	New	Percentage change
Child mortality	0.167	0.174	4.4
Nutrition	0.167	0.159	−4.6
Years of education	0.167	0.173	3.8
School Attendance	0.167	0.160	−4.0
Electricity	0.067	0.043	−35.5
Drinking water	0.067	0.062	−7.0
Sanitation	0.067	0.071	6.5
Cooking fuel	0.067	0.068	2.0
Assets	0.067	0.090	35.0

Table 7 shows the difference in weights between the new MPI (with level terms) and the traditional MPI. The traditional approach employs equal weighting in each dimension (in column "Traditional"), which contrasts to the new approach (in column "New"). Both electricity and assets show large changes, where the weight decreased by around 35% for the electricity indicator, and similarly increased for the assets indicator, compared to the traditional MPI. Other indicators show moderate differences.

5.3 Descriptive statistics

In the following, comparisons between the traditional and the new MPI will be made. These exercises aim to provide insights on how the two MPIs assess multidimensional poverty in South Africa. These comparisons consider 3 data waves in the NIDS data, 2010, 2012 and 2014. Data for the waves from the years 2010 and 2012 were prepared in the same manner as the data from year 2014, as described in Section 4. The number of household (individual) level observations, after keeping complete observations (those with no missings for any of the indicators) are 6322 (27808) and 7982 (32822), respectively.

These descriptive statistics were calculated for individuals. Household level data was preferred for applying PLS because it does not need to consider the clustering within household. However, individual level data was preferred for depicting the descriptive statistics, because it is likely to provide more policy relevant information. Poor households are likely to be larger than non-poor households, and showing descriptive statistics at the household level may under-emphasize the deprivation level, due to under-counting of poor individuals.

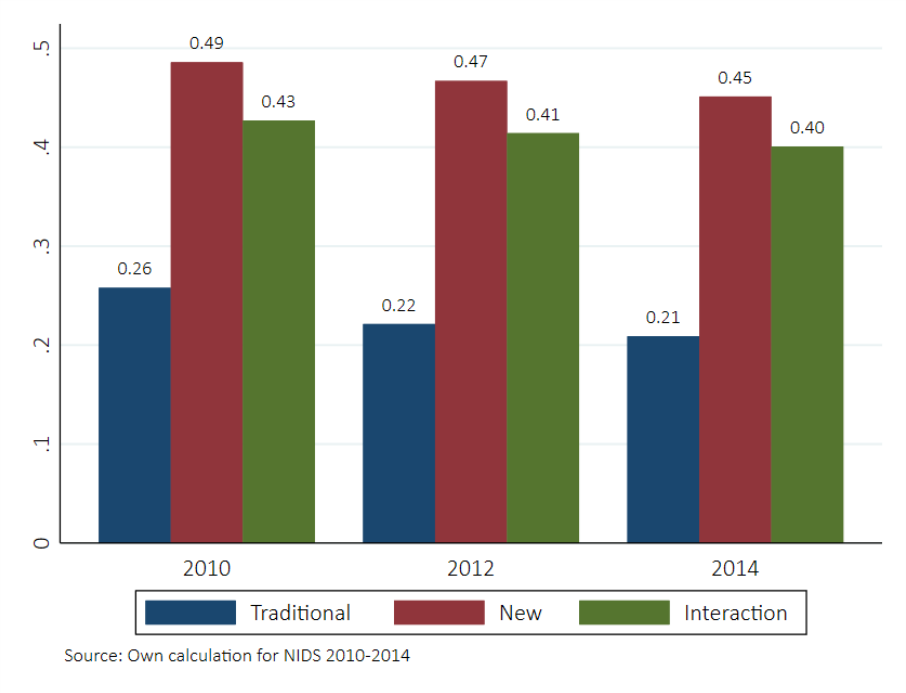
The MPIs for these old waves were calculated using the same weights and thresholds calculated using the data from 2014. The new MPI applies autoscaling to indicators (i.e., centralization and unit variance) before applying PLS weights. The traditional MPI applies equal weights on indicators in each dimension without autoscaling. Due to the different scaling of indicators before the aggregation, the range of resulting MPI differs.

The traditional MPI ranges from 0 to 1, while the new MPI has no limited range, including negative values. For the sake of comparison, the new MPI in this subsection is transformed using the following function, so that the range of the old and new MPI is equal ([0,1]).

$$MPI^{*}_{new} = \frac{MPI_{new} - min(MPI_{new})}{max(MPI_{new}) - min(MPI_{new})}$$

Traditionally, a household with the MPI score less than 0.33 is considered to be multidimentionally poor, and this threshold (0.33) is called the second cut-off. By transforming the new MPI to have the same range as the old MPI, the second cut-off can be applied to both MPIs.

Figure 1: Headcount ratio of traditional, level and interaction methods of poverty calculation



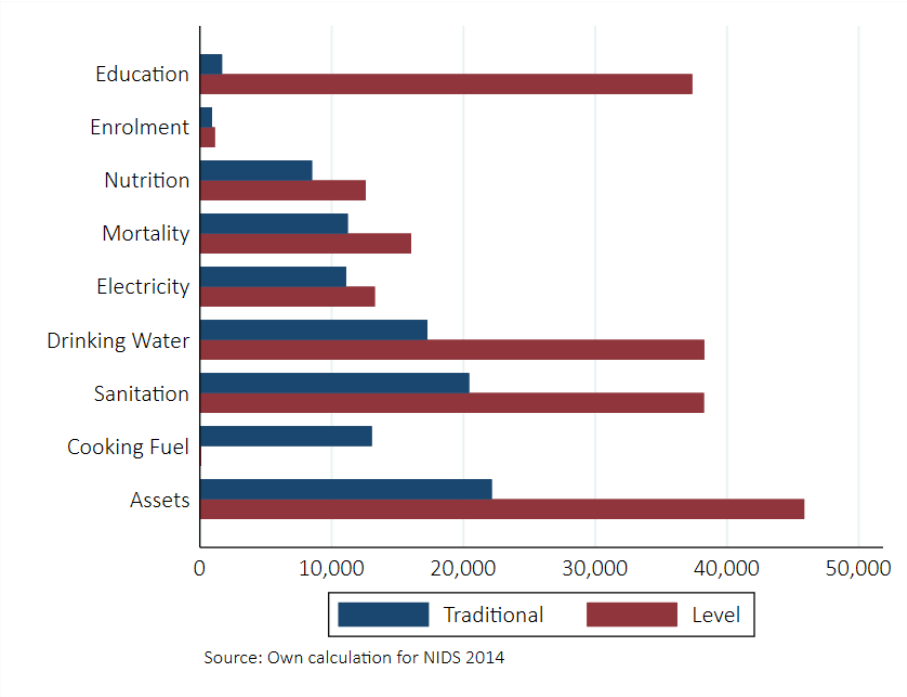
Headcount ratio

Figure 1 depicts the headcount ratio using the traditional method, and the two new methods in this paper (level and interaction). The headcount ratio shows the proportion of individuals belonging to multidimensionally poor households, after applying the second cut-off (poor individuals hereafter). The proportion of poor individuals nearly doubles

using the new methods compared to the traditional one. The difference occurs because of the different first cut-offs as well as the different weights. The level MPI shows a slightly higher proportion of poor individuals (between 5 and 6 % points) than the interaction MPI. The interaction MPI assigns higher weights to black households than the level MPI, which implies that the number of poor black individuals increases, while the number of poor non-black individuals decreases. Hence, the smaller poverty headcount of the interaction MPI indicates that the decrease in the number of poor non-black individuals is larger than the increase in the number of poor black individuals.

Figure 1 shows the falling trend in multidimensional poverty level in South Africa, over the period of 2010 to 2014, for all the three methods. The magnitude of the decline in poverty seems to be similar across the three MPIs, where the largest drop is found using the traditional method (5% points) and the smallest using the interaction method (3% points).

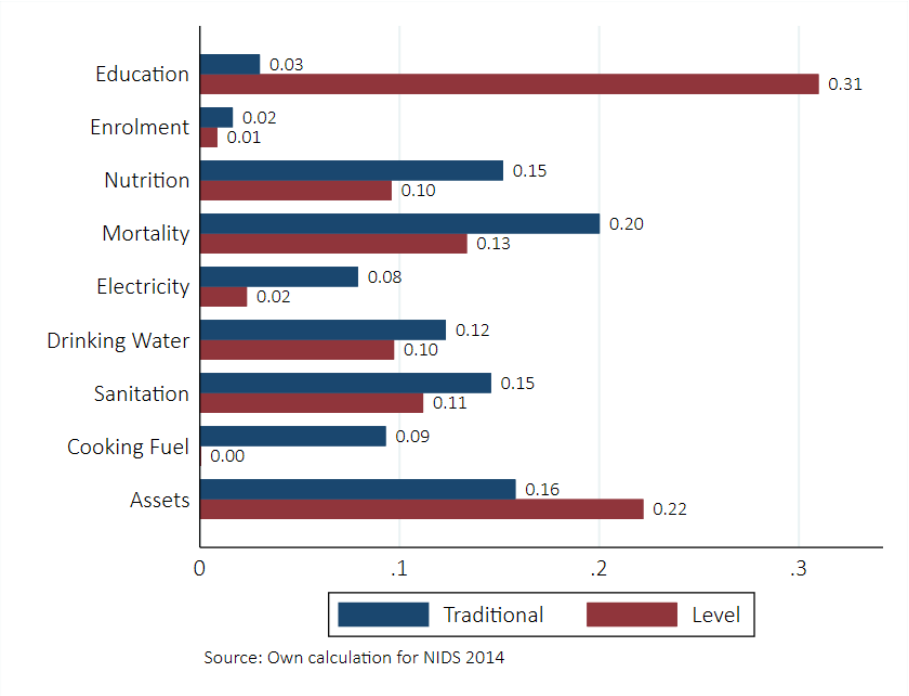
Figure 2: Absolute number of deprived individuals of each indicator, using traditional and new method of multidimensional poverty



Absolute number of deprived individuals of each indicator

Figure 2 shows the absolute numbers of poverty (in individuals), comparing the new (level) and traditional methods of multidimensional poverty calculation. The interaction model was not considered, since the traditional method and the interaction model included different set of indicators, which makes this kind of comparison misleading. In this figure, the difference is caused by the different definition of first cut-offs, but not by weights. A comparison between the blue versus the red bars (traditional versus new, respectively) shows that the poverty levels are higher using the new method for all indicators except cooking fuels. The difference is most salient in the years of schooling, where the new MPI considered only households having at least one household member of $\geq$12 years of education as above the first cut-off, while the traditional MPI considered household already when one member had $\geq$5 years of education. The new MPI seems to offer a more practical measurement of deprivation, because South Africa has free primary schooling, and as a result there are very few individuals in households without at least one household member having completed five years of schooling. Indeed, inequalities in schooling in South Africa start much later than primary education, where only 52% of the eligible students remain enrolled in school in grade 12 (Department of Basic Education, Republic of South Africa, 2015). The new method, with the threshold of 12 years of schooling, accounts for this specific feature in South Africa. Another interesting difference is that the new MPI shows a much higher number of poor individuals deprived in sanitation, drinking water and assets compared to the traditional MPI. This finding is consistent with existing literature, that in South Africa, there is a large inequality in wealth (Wittenberg and Leibbrandt, 2017), which is visible in the access to drinking water and sanitation facilities, and the accumulation of assets (Frayne and McCordic, 2015). The only exception, by which the new MPI shows smaller number of poor individuals than the traditional MPI, is cooking fuel, where the traditional method identifies around 13000 poor individuals and the new method almost none.

Figure 3: Percentage contribution of each indicator, using traditional and new method of multidimensional poverty

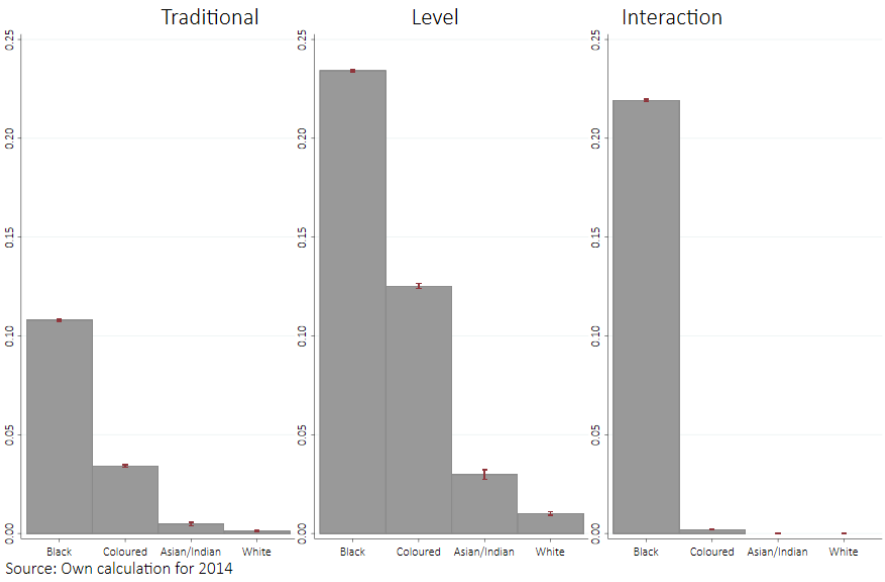


Percentage contribution of each indicator

The contribution to the overall deprivation levels can be observed in Figure 3. Again the interaction model was not considered because it contains different indicators than the traditional one. Contribution of each indicators is the censored headcount of that indicator (the proportion of individuals deprived in that indicator divided by the total number of multidimentionally poor) multiplied by the weight of that indicator (Oxford Poverty & Human Development Initiative, 2020). This can be interpreted as the relative influence of each indicator on the final MPI, where indicators with higher weights or with a large number of deprived individuals (or both) have a larger influence. A larger number of individuals are found to be deprived using the new method, compared to the traditional, in the years of education indicator. The weight of the indicator within the new method was also slightly larger. As a result of the higher weight and therefore, more demanding deprivation threshold of the education indicator, the contribution of this indicator to overall poverty figures is 31% using the new method, compared to 3% in the

case of the traditional method. Likewise, assets also have a much higher contribution in the new MPI compared to the traditional MPI, due to the large increase in the number of individuals that are now considered to be deprived in assets, as well as the larger weight.

Figure 4: Average deprivation level for each race using the traditional, level and interaction models of multidimensional poverty in South Africa

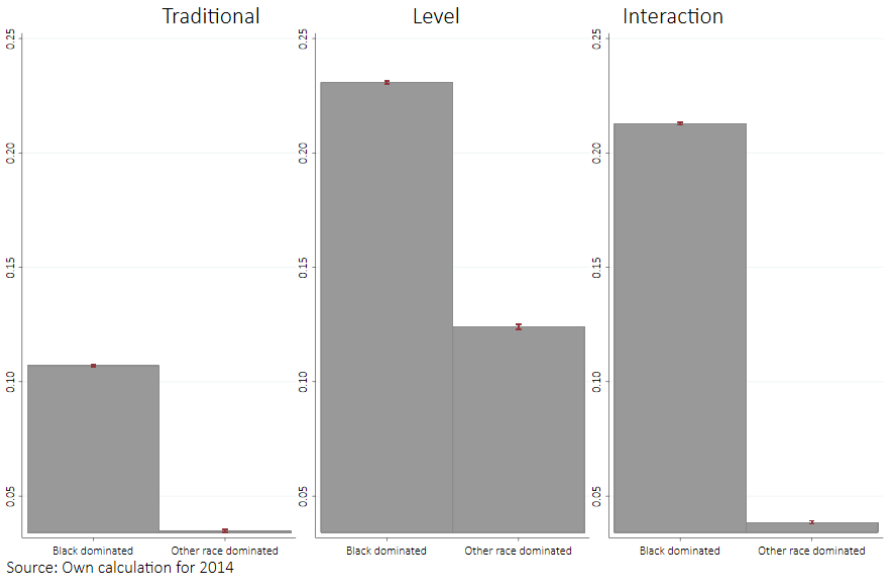


Proportion of deprived individuals by race

Figure 4 shows the proportion of deprived individuals using the three methods of multidimensional poverty calculation, disaggregated by the four races in the data (black, Coloured, Asian/Indian and white). The multidimensional poverty level of the black population is particularly interesting, since they are the majority, accounting for over 80% of the sample. In all three MPIs, Blacks have the highest percentage of poor individuals, compared to other racial groups. The traditional MPI shows the lowest proportion of poor Blacks among the three MPIs (around 11%), followed by the interaction (around 22%) and the level MPI (around 23%). Across all three methods, compared to the Blacks, the Coloured population shows a lower proportion of poor individuals, followed by Asian/Indians and Whites. There is a noticeable difference in the proportion of poor individuals using the interaction MPI, compared to the level MPI- only Blacks show

a large proportion of poor individuals, while other racial groups shows either very small proportion or a non-existing share of poor individuals. This is because the interaction model takes into account the extra deprivation of the black population in South Africa in terms of lower income.

Figure 5: Average deprivation level in black dominated provinces and other race dominated provinces in South Africa



Average deprivation level by black dominated provinces and others

Figure 5 shows the proportion of deprived individuals in black dominated and non-black dominated provinces. The provinces with more than 50% of black population are considered to be black dominated, while others are non-black dominated.[10] In the traditional MPI, black dominated provinces show around 11% point higher poverty than white dominated provinces. This pattern remains in the level MPI, but both black dominated and white dominated provinces show higher poverty compared to the traditional MPI. In the interaction model, the difference in poverty between black and non-black dominated provinces is more salient compared to the other MPIs (around a 20% point difference). It

[10]Western Cape (76.2%) and Northern cape (66.6%) are the two districts that are non-black dominated. These comprise only 17% of the data.

is again because the interaction MPI takes account of the extra income poverty of black population.

6 Conclusion

In this paper, a novel data driven approach to determine the first cut-offs in the dual cut-off based MPI is proposed, utilizing the intuitive association between low income and multidimensional poverty. This approach is helpful to reduce biases/errors, which can result from the arbitrary judgment of researchers, leading to a distorted figure of deprivation. Our review of the literature shows the scant work done in this area, highlighting the inventiveness of this approach.

A highly practical aspect of this approach is that if external knowledge or judgement is considered relevant in the particular context, the weighting structure can be restricted to reflect such judgements. For instance, our approach of creating the MPI respects the traditional value judgement in the poverty indexes literature that allocates equal importance to the dimensions of education, health and living standards to measure poverty.

This study generates a new weighting scheme for the MPI, based on PLS, again utilizing the association between low income and multidimensional poverty. Thereby, the weights for each indicator are derived using the covariance between household per capita income and multidimensional poverty. This is also relevant for policies targeting poverty reduction within a multidimensional framework, especially when the decision maker is interested in those aspects that are related to income.

Another novelty within this paper is the consideration given to the additional deprivation faced by the black population. Using PLS, it was possible to create an MPI that allows interaction terms between each of the deprivation indicators and a dummy reflecting whether a household is black, or of another race.

The results show that the new MPI identifies different first cut-offs compared to the

traditional MPI. Differences are found in the indicators of years of education, drinking water, cooking fuel, landline and the number of small assets (within assets). Education showed particularly large changes, given that the new minimum threshold for being non-deprived is when at least one individual in the household has 12 years of education, compared to the traditional method using a 5 year threshold.

From the new weights proposed by PLS, the indicators within the health and education dimension show modest differences compared to the traditional weights (equal weights for each dimension), while the differences are large in the living standard dimension. The new weights emphasize child mortality slightly more than nutrition in the health dimension and years of education slightly more than school attendance in the education dimension. In the living standard dimension, however, assets are much more important than electricity. The Black interactions show that the black population suffers additional deprivation of low income for the same indicator poverty compared to others. In most indicators, the weights for the black deprived households are nearly double that of the other households.

The new MPIs (level and interaction) show a much higher poverty headcount compared to the traditional MPI. With the new thresholds, more individuals are identified as poor in the education, water, sanitation and assets indicators, compared to the old thresholds, while the number of poor individuals in cooking fuel became smaller. In the new MPI (level), the contribution of education and assets were larger than the traditional MPI, while for other indicators, the contributions were smaller. The burden of poverty lies more severely on the black population across all MPIs, where the interaction MPI identified the strongest difference between the black and other racial groups, followed by the level MPI, and finally the traditional MPI. The same trend was identified when disaggregating the data between black and non-black dominated provinces, where the difference was the most salient in the interaction MPI, followed by the level MPI and the traditional MPI.

The proposed method's limitation is its computationally intensive nature, since it considers all possible combination for the first cut-offs. With an increasing number of covariates or number of categories, the computational costs will increase multiplicatively. Furthermore, only ordinal variables are admissible to the proposed algorithm. In our data, there were an infinite number of coding strategies available, to transform the numerical variables to ordinal variables. To reduce the computational costs to an acceptable level, the coding for numerical variables were decided by external data, and based on authors' judgment. Consequently, although it was possible to reduce the arbitrariness compared to traditional approach, the computationally intensive nature of this approach implies that some form of judgment was required nonetheless. Further, the estimation of standard error could not consider the uncertainty coming from the determination of the first cut-offs, due to the large computational costs. The interaction model approach assigns different deprivation scores based on racial origin, which will invite sensitive ethical questions if used at the policy level. The inclusion of the black interactions was authors' choice (i.e., these are not completely objective) and there is no guarantee that the black interaction coefficients are consistently estimated, which will make the ethical question even more challenging. Considering these limitations of the interaction approach, we prefer the MPI with only level terms.

There are many interesting future research avenues that can be availed from this method. Alone for the MPI, one could consider using a wellbeing or life-satisfaction index to calculate the weights of the deprivation score. Satisfaction is by itself an interesting outcome and some researchers may prefer not to have an MPI containing income information. This paper didn't pursue this approach for several reasons. Satisfaction is measured in an ordinal scale and an appropriate link function needs to be incorporated, which we believe to be possible, but more complicated. Another challenge is that satisfaction is measured in the individual level data. Satisfaction of individuals may show large correlation within a household. Large correlation between observations may deteriorate the performance

of the PLS algorithm, which may require additional adjustments to the PLS algorithm. Lastly, satisfaction may not be strongly correlated to poverty (i.e., miserable rich and contented beggar), which will make the use of the PLS algorithm less attractive.

One can also use our approach with PCA instead of PLS. PLS can be used to obtain PCA solution (by considering the PLS score as the outcome variable). This approach is attractive if a researcher seeks to make the composite index exogeneous to the outcome variable.

For composite indices other than the MPI, the method could have the following extensions. The method uses a special non-linear functions ($F(X, \Theta)$ in Algorithm 1) of the MPI in the optimization of the PLS algorithm. This function and the related optimization approach can be tailored to other composite index applications. For example, we may able to adjust this method to determine the association parameter of the Correlation Sensitive Poverty Index (CSPI; Rippin, 2012a). The restrictions on weights can be applied to other PCA and PLS applications with a few simple adjustments. For example, the KOF index of globalization (Dreher et al., 2008) also has three dimensions (economic, social and political globalization), where the approach in this paper can be used to consider the three dimensions to be equally important or potentially estimate the relative importance between the dimensions from data.

References

Alkire, S. and Foster, J. (2011a). Counting and multidimensional poverty measurement. *Journal of public economics*, 95(7):476–487.

Alkire, S. and Foster, J. (2011b). Understandings and misunderstandings of multidimensional poverty measurement. *The Journal of Economic Inequality*, 9(2):289–314.

Alkire, S. and Santos, M. E. (2014). Measuring Acute Poverty in the Developing World: Robustness and Scope of the Multidimensional Poverty Index. *World Development*, 59:251–274.

Atkinson, A., Cantillon, B., Marlier, E., and Nolan, B. (2002). Social Indicators: The EU and Social Inclusion. OUP Catalogue, Oxford University Press.

Chakravarty, S. R. and D'Ambrosio, C. (2006). The Measurement of Social Exclusion. *Review of Income and Wealth*, 52(3):377–398.

Decancq, K. and Lugo, M. A. (2013). Weights in Multidimensional Indices of Wellbeing: An Overview. *Econometric Reviews*, 32(1):7–34.

Department of Basic Education, Republic of South Africa (2015). Education Statistics in South Africa 2013. Technical report.

Dotter, C. and Klasen, S. (2014). The Multidimensional Poverty Index : achievements, conceptual and empirical issues. Technical report, United Nations Development Programme.

Dotter, C. and Klasen, S. (2019). An absolute multidimensional poverty measure in the functioning space (and relative measure in the resource space): An illustration using indian data. Technical report, Courant Research Centre: Poverty, Equity and Growth-Discussion Papers.

Dreher, A., Gaston, N., and Martens, P. (2008). *Measuring globalisation: Gauging its consequences*. Springer Science & Business Media.

Finn, A. and Leibbrandt, M. (2013). The dynamics of poverty in the first three waves of NIDS. SALDRU Working Paper 119, Southern Africa Labour and Development Research Unit, University of Cape Town.

Fintel, M. v. and Zoch, A. (2015). The dynamics of child poverty in South Africa between 2008 and 2012: An analysis using the National Income Dynamics Study. Working Paper 05/2015, Stellenbosch University, Department of Economics.

Frayne, B. and McCordic, C. (2015). Planning for food secure cities: Measuring the influence of infrastructure and income on household food security in southern african cities. *Geoforum*, 65:1–11.

Höskuldsson, A. (1988). Pls regression methods. *Journal of Chemometrics*, 2(3):211–228.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Jayaraj, D. and Subramanian, S. (2010). AChakravarty-D'Ambrosio View of Multidimensional Deprivation: Some Estimates for India. *Economic and Political Weekly*, 45(6):53–65.

Klasen, S. (2000). Measuring Poverty and Deprivation in South Africa. *Review of Income and Wealth*, 46(1):33–58.

Martens, H. and Martens, M. (2000). Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (plsr). *Food quality and preference*, 11(1):5–16.

Nussbaum, M. (2003). Capabilities as fundamental Entitlements: Sen and Social Justice. *Feminist Economics*, 9(2-3):33–59.

Nussbaum, M. C. (2008). *Women and human development: the capabilities approach*. Number 3 in The John Robert Seeley lectures. Cambridge Univ. Press, Cambridge, 13. print edition.

Nussbaum, M. C., Sen, A., and World Institute for Development Economics Research, editors (1993). *The Quality of life*. WIDER studies in development economics. Clarendon Press ; Oxford University Press, Oxford [England] : New York.

Oxford Poverty & Human Development Initiative (2020). The alkire-foster counting methodology. https://multidimensionalpoverty.org/chapter-5/ [Accessed: 23/04/2020].

Pasha, A. (2017). Regional Perspectives on the Multidimensional Poverty Index. *World Development*, 94(Supplement C):268–285.

Ravallion, M. (2011). Troubling Tradeoffs in the Human Development Index. *Journal of Development Economics*.

Ravallion, M. (2012). Mashup Indices of Development. *The World Bank Research Observer*, 27(1):1–32.

Ravallion, M. and Chen, S. (2011). Weakly Relative Poverty. *Review of Economics and Statistics*, 93(4):1251–1261.

Rippin, N. (2012a). Integrating inter-personal inequality in counting poverty indices: the correlation sensitive poverty index. In *32nd IARIW conference, Boston*.

Rippin, N. (2012b). Operationalising the Capability Approach: A German Correlation Sensitive Poverty Index. Courant Research Centre: Poverty, Equity and Growth - Discussion Paper 132, Courant Research Centre PEG.

Sen, A. (1985). *Commodities and capabilities*. Number v. 7 in Professor Dr. P. Hennipman lectures in economics. North-Holland ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co, Amsterdam ; New York : New York, N.Y., U.S.A.

Sen, A. (1999). *Development as freedom*. Knopf, New York, 1st. ed edition.

Silber, J. (2011). A comment on the MPI index. *Journal of Economic Inequality*, 9(3):479–481.

Southern Africa Labour and Development Research Unit (2016). National income dynamics study (NIDS). http://www.nids.uct.ac.za/nids-data/data-access.

UNDP (2015). Technical note 5. Multidimensional Poverty Index. Human Development Report 2015. United Nations Development Programme (UNDP).

UNDP (2017a). Gender development index (GDI). http://hdr.undp.org/en/content/gender-development-index-gdi.

UNDP (2017b). Human development index (HDI). http://hdr.undp.org/en/content/human-development-index-hdi.

WHO (2017a). BMI classification. http://apps.who.int/bmi/index.jsp?introPage=intro_3.html.

WHO (2017b). Global database on child growth and malnutrition. http://www.who.int/nutgrowthdb/about/introduction/en/index5.html.

Wittenberg, M. and Leibbrandt, M. (2017). Measuring inequality by asset indices: A general approach with application to south africa. *Review of Income and Wealth*, 63(4):706–730.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. In *Research papers in statistics*. Wiley, New York.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130.

World Bank (2017). Gross enrollment ratio, primary, both sexes (%) for south africa. https://data.worldbank.org/indicator/SE.PRM.ENRR?locations=ZA.