Incorporating uncertainty in poverty dynamics: how can we assess the economic impact of AIDS mortality in the presence of measurement error and missing data?

Alessandra Garbero

London School of Hygiene and Tropical Medicine

Summary

Disentangling the complex relationships between HIV and AIDS and poverty has proven to be methodologically difficult and many studies to date have tried to quantify their respective contribution to households' vulnerability, in Sub-Saharan Africa (Gillespie 2006; Carter, May et al. 2007; Gillespie, Greener et al. 2007; Gillespie, Kadiyala et al. 2007). Is AIDS exacerbating poverty or is the latter contributing to the spread of the epidemic, in Sub-Saharan Africa?

The literature has been mostly concerned with two econometric issues, inherent to the estimation of the impact of HIV and AIDS mortality on household welfare: the potential endogeneity of adult death and household unobserved heterogeneity. Adult mortality can be endogenous when is a function of unobservable characteristics of the households and the deceased individual; unobserved heterogeneity occurs when unobserved characteristics of the household and the individual affect the outcome variable of interest, i.e. the welfare or poverty indicator of interest, and also the likelihood of having a death in the household (Chapoto and Jayne 2006).

Some recent work by Murtin and Marzo (2008) has also addressed the endogeneity of AIDS mortality to welfare outcomes methodologically using Bayesian estimation.

While the relevant literature has mostly been concerned with such issues, there is no evidence to date on the implications of poor welfare proxies while estimating the impact of HIV and AIDS mortality on household consumption, or broadly its welfare. This paper complements this literature in this respect and addresses the problem of measurement error and missing data in the consumption module of the ACDIS data, the Africa Centre Demographic Surveillance System.

The paper is structured as follows. After evaluating the quality of the consumption module in the ACDIS data taking into account the standard practice as described in Deaton and Zaidi's recommendations for constructing consumption aggregates (Deaton and Zaidi 2002), different imputation methodologies to partially address these problems will be compared and tested.

Three scenarios are produced and compared: 1) consumption-based poverty indicators based on a scenario with no imputation (naïve scenario); 2) a crude imputation scenario where missing items are imputed with cluster medians (i.e. missing prices or unit values are replaced by the median of "similar" households in the neighborhood or geographical area, Deaton and Zaidi 2002) and 3) multiple imputation (MI) (Paulin and Ferraro 1994; Little and Rubin 2002),

specifically MI by chained equation - MICE (Van Buuren and Oudshoorn 1999; Royston 2004; Royston 2005). The robustness of results is evaluated and, the validity of the use of multiple imputation is assessed in this paper.

The emphasis in this paper is thus put on issues of data quality and reliability of estimates; specifically to what extent we can estimate and say something meaningful about the impact of AIDS mortality on household welfare when the consumption-based welfare indicator of interest (poverty) is measured with error? In the presence of a negative scenario, are asset indices clearly a second and necessary best?

The final aim of this work would be to measure the impact of AIDS mortality on consumption poverty, longitudinally, in the presence/absence of imputation and compare the results. By attaching a measure of uncertainty to the consumption-based poverty indicator, I will quantify to what extent the latter exercise can bias an assessment of the impact of AIDS mortality on household poverty.

Introduction

Any measurement of poverty, broadly defined as "not having enough today in some dimensions of well-being", starts with defining the welfare indicator of interest (Ravallion 1992).
Household welfare can be measured based on a money-metric approach, i.e. by aggregating all the income or consumption components at household level (Deaton 2003). Other approaches exist in the literature, "non-money metric approaches" i.e based on other dimensions of well-being, such as health, nutrition and assets (Filmer and Pritchett 1998).
This is not the place to deal with the theoretical issues underlying the choice between income and consumption as appropriate indicators of wellbeing. Generally measurement of household wellbeing has proven to be difficult given the reluctance of respondents to openly disclose their welfare in most settings.
There is a vast literature and particularly the seminal publication by Deaton and Zaidi (2002) which covers the issues just mentioned and summarizes strengths and weaknesses in detail. An accurate quantification of income components (production surplus, wages, remittances, pensions, borrowing), non-cash incomes, and households proxy flows (i.e. rent of own house) has proven to be challenging in socio-economic surveys that aim at measuring income poverty. Conceptually, income is a flow, subject to seasonality effects (drought shocks in agricultural settings), and volatility (unemployment and illness shocks that could all cause a temporary interruption in such a flow). As a consequence, measures of poverty based on income often fail to identify households' vulnerability to poverty.
On the other side, there is some consensus in the literature in favour of the measurement of household welfare based on consumption (expenditures), as this seems to provide a better representation of household permanent income. However the latter is also prone to difficulties given households complex expenditure patters (which can have varying timeframes, i.e. measured on a monthly or yearly basis for instance). Moreover households can forget the exact amount of quantities purchased, and their relative prices (recall loss).

Data

The Africa Centre Demographic Information System (ACDIS) has been collecting data on more than 11,000 households and 85,000 individuals in part of the Umkhanyakude district of KwaZulu-Natal since the beginning of 2000 (Tanser, Hosegood et al. 2007) Hosegood and Timaeus 2005).
Households visited every 6 months to obtain reports of births, deaths, migration and changes in household membership (rounds). Detailed socioeconomic data (HSE), including household

expenditure data, are collected on alternate visits. Verbal autopsies are used to determine causes of death. This analysis specifically employ three socio-economic waves, specifically HSE 2 – which has been collected during 2003- 2004; HSE 3 (2005) and HSE 4 (2006).
The ACDIS datasets present the methodological problem of missing data and measurement error (presence of zero values) in the consumption module of each socio-economic wave. Missing data arise from the failure to obtain a complete response from all individuals included in a survey sample. They may occur because individuals refuse to return their questionnaire (unit non-response) or do not provide an answer for some of the questions (item non-response), and may depend on both respondents' attitudes and survey procedures. Statisticians have acknowledged the impact of these types of non-sampling errors on poverty estimates (Ardington, Lam et al. 2006).

Objective of the paper

The objective of this paper is twofold: it aims at evaluating the quality of the consumption module in the ACDIS data, addressing measurement error and missing information. Estimates of household consumption are in fact sensitive to both issues (Ardington, Lam et al. 2006).
Secondly, it aims at assessing the extent to which contaminated data can affect poverty measures. This paper adopts a comprehensive approach to deal with defective data when measuring household welfare using consumption aggregates. The final aim is to produce poverty (or welfare) indicators for each socio-economic wave under various imputation scenarios.
This research is instrumental to study the changes in consumption-poverty induced by AIDS mortality in Kwazulu-Natal.

Methodological review: reconciling economists perspectives and biomedical literature

This paper is concerned with a measurement of welfare based on consumption. Following the recommendations for constructing consumption aggregates (Deaton & Zaidi, 2002), (Carletto, Covarrubias et al. 2007) such authors offer useful guidance on how to deal with missing and miscoded information while constructing such aggregates. The standard practice in Living Standard Measurement surveys[1] (Deaton 2003) is to implement ad-hoc procedures on the field that could minimize non-sampling errors. After data collection, data is checked for outliers and also miscoded information. Misunderstanding of units for quantities can cause errors in unit values identified (Carletto, Covarrubias et al. 2007).

[1] Living Standard Measurement Surveys are routine socio-economic surveys have been implemented by the World Bank in order to measure poverty since the ?date!

Missing data are not uncommon in any survey and especially in large socio-economic studies. According to Schafer et al. (1997) missing data falls in the category of "coarsened data" which are defined as a combination of point, interval and missing responses. Coarsened data include also censoring, heaping and rounding issues. These methodological problems are very common in surveys that aim at measuring welfare through income, assets and earnings (Heitjan and Rubin 1991; Philip K. Hopke 2001; Vermaak 2008).

In the statistical jargon the missing data mechanism, defined as the relationship between "missingness" or the underlying cause for missing and the data, can be defined as missing at random (MAR), missing completely at random (MCAR) and non-missing at random (MNAR) (Rubin 1976; Schafer and Graham 2002). MCAR substantially means that there is no relationship between the missingness process and the observed data and/or unobserved data: this assumption is quite unrealistic and almost never met in practice. MAR occurs instead when the probability of missing depends only on observed characteristics. However data are rarely missing at random and NMAR is the most complex case (i.e. the probability of a datum missing depends on the unobserved data or missing values). There are methods designed to deal with this assumption, however they are quite complex and may aggravate the biases in the data if the model is not correctly specified (Diggle and Kenward 1994).

There are lenghty reviews on missing data methods and this is not the scope of this article as this paper will specifically focus on multiple imputation (Little and Rubin 2002). However a brief overview of traditional missing data methods is provided. Multiple imputation or MI falls within the realm of the modern methods and will be described later.

Traditional missing data methods include a complete case analysis (or listwise deletion), and single imputation methods. A frequent approach common to any scientific discipline is to restrict the analysis to the completers, or perform a complete-case analysis. However such an analysis entails excluding all the missing values and it results in a loss of precision and also produces substantive bias unless the data are MCAR.

Single imputation methods include arithmetic mean substitution (where the missing value is replaced by the arithmetic mean of the complete cases); regression imputation (where the missing value is replaced by predicted scores from a linear regression equation); stochastic regression imputation (where the problem of exclusion of residual variation of regression imputation is solved by including a residual component to each imputed value sampled randomly from a normal distribution); and last observation carried forward (the latter method is specific to longitudinal studies were the missing value is replaced by its last observation) (Enders 2006).

Multiple imputation was suggested by Rubin in the 70s (Rubin 1976; 1987; 1996), and has been gaining ground as an alternative to likelihood based methods for addressing the issue of missing data. According to van Buuren, Rubin's original publication did not deal with multivariate data imputation (van Buuren and Oudshoorn 1999). As such several problems occur when imputing

multivariate data 1) the need to select a reasonable number of predictor variables to be used in the imputation of large data sets; 2) the fact that missing data can occur also within predictor variables 3) variables' different levels of measurement (nominal, ordinal and continuous) and 4) dependency while imputing variables (for example: Y1 is imputed given Y2 and Y2 given Y1).

Recent literature has suggested methods to impute multivariate data (Rubin and Schafer 1990). The substance of such methods is that they are "Bayesian simulation algorithms that draw imputations from the posterior predictive distribution of the missing data given the observed data" (van Buuren and Oudshoorn 1999). The latter author underlines both the limitations of the Rubin-Schafer method (assumption of a MAR missingness mechanism and data following a multivariate normal distribution) and Schafer's (1997) variation of algorithms that mostly assume normality in distributional assumptions.

Other approaches exist in the literature that do not assume that the data can be modeled by a multivariate probability distribution (for more info see van Buuren and Oudshoorn 1999).

This article uses the Van Buuren et al's MICE or multiple imputation by chained equation method. The approach allows the user to specify a conditional distribution for the incomplete variables given the other variables. Different models can be employed depending on the nature of the variables, i.e. logistic regression for incomplete binary variables, ordinal/multinomial regression for categorical data, and linear regression for continuous variables. MICE uses an algorithm which is based on GIBBS sampling, essentially imputing variable by variable iteratively [see (Oudshoorn, van Buuren et al. 1999)-but also the manual on S plus, for more details on the algorithm]. MICE has been implemented in STATA by Royston (Carlin, Li et al. 2003; Royston 2004; Royston 2005). The attractiveness of the technique is that it is a stochastically, Bayesian-driven procedure that takes into account the uncertainty in the imputation procedure.

Model-based imputation in income and expenditure surveys has been already used in the literature (Paulin and Ferraro 1994; Fisher 2006). Specifically, to our knowledge, multiple imputation has only been used to impute income data (Ardington, Lam et al. 2006) and earnings (Vermaak 2008) in South Africa. The reason for this is that standard practice recommended by World Bank guidelines (Deaton and Zaidi 2002) is to impute missing values (the missing price or quantity) using market reference (i.e. looking at the average quantity or average price reported by other households with similar characteristics, such as place of residence, size, etc.). Essentially, the imputation of missing prices or unit values is conducted by using as a proxy the median expenditure of "similar" households in the neighborhood or geographical area (cluster) (Deaton and Zaidi 2002). Some authors like Vermaak (2008) have also highlighted the methodological costs associated with multiple imputation which makes the technique rather time consuming for the researcher.

Rationale for ad-hoc approach

The ACDIS dataset contains the following methodological problems: attrition, typical of longitudinal data, but relevant to the calculation of a consumption-based welfare indicator, measurement error in the welfare proxy (consumption). Measurement error essentially means a large number of zeros and missing values in each expenditure item.

The ACDIS survey instruments aimed at measuring consumption, present some (complex) design issues: there are no prices and no quantities specified (just the amount spent per month), and also the consumption module questionnaire is not fully standardized across socio-economic waves (fewer and aggregate items in HSE2 and key items missing in HSE 3 and 4- such home production and clothing expenditures).

The rationale for devising an ad-hoc approach is that using the raw data as they are, would introduce a substantive overestimation of poverty in the area if the data are not missing completely at random and also coarsened completely at random, and won't help discriminating the poor. It would possibly bias any analysis aiming at differentiating poor vs. very poor targeting measures.

Descriptive analysis

The following table presents a descriptive analysis of the prevalence of zero and missing values in each socio-economic wave of the ACDIS longitudinal dataset. It indicates the number of households with missing or zero expenditure information on **all** expenditure items (excluding education). It is important to note that the expenditure on education aggregate is derived from the HSE individual module, and, as such, is almost never missing and/or zero. The ACDIS household level module include food expenditures (total shopping per month) and non-food expenditures for HSE2 (water; electricity; fuel; telephone; payments for goods bought by hire purchase or lay-bye; health; transportation; religious expenses; funeral expenses for households members; life insurance, burial societies and funeral policy; large expenses; other usual expenses –the latter also includes other items which are unique to HSE 2 i.e. rent and bond payment, cell phone, clothing and shoes, expenses on outsiders, funeral expenses on outsiders). In HSE 3 food expenditures were aggregated and include the following: mealiemeal, rice, beans, samp, flour, cooking oil, sugar and salt, tea and coffee, milk, vegetables, meat, bread, tinned goods, soap. Non-food expenditures included water; electricity; fuel; telephone; payments for goods bought by hire purchase or lay-bye; health; transportation; religious expenses; funeral expenses for households members; life insurance, burial societies and funeral policy; large expenses; other usual expenses. In HSE 4, the questions on food expenditures were similar to those in HSE 3 but

include three further items (snacks, fruit and eggs) while non-food expenditure are identical to HSE 3.

The total number of households per wave is included in Table 1.

At first glance the amount of missing information globally increases across waves. There are 246 households that have missing information on all household-level expenditure aggregates in HSE4, although they do have a record in the corresponding HSE individual level file.

Table 1: Number of households presenting missing or zero information on all households level expenditure items, ACDIS data.

	HSE2		**HSE3**		**HSE4**	
Missing or zero information on all hh level exp items	Freq.	Percent	Freq.	Percent	Freq.	Percent
Non-missing	10,711	98.98	9,649	98.77	9,137	97.38
Missing	110	1.02	120	1.23	246	2.62
Non-zeros	10,806	99.86	9,726	99.56	9,347	99.62
Zeros	15	0.14	43	0.44	36	0.38
Both						
Non-missing or non-zeros	10,696	98.84	9,606	98.33	9,101	96.99
missing or zeros	125	1.16	163	1.67	282	3.01
Total sample	10,821	100	9,769	100	9,383	100
*excluding education						

The following two tables present more disaggregated information on the percentage of missing and zero information in each HSE wave, looking at food and non food items in turn.

Table 2: Percent number of zeros and missing values for food expenditure items.

	HSE2		HSE3		HSE4	
Food expenditure items	**% zeros**	**% missing values**	**% zeros**	**% missing values**	**% zeros**	**% missing values**
Total shopping	**1**	**12**				
Mealiemeal			2	3	2	7
Rice			6	3	6	6
Beans			8	4	8	6
Samp			51	3	48	7
Flour			35	4	31	6
Cooking oil			3	3	4	6
Sugar and Salt			3	5	3	9
Tea and Coffee			7	5	6	13
Milk			36	6	19	27
Vegetables			5	18	4	46
Meat			4	26	3	49
Bread			17	27	14	53
Tinned goods			37	15	35	33
Soap			7	7	6	23
Snacks					25	44
Fruit					46	33
Eggs					30	14
TOTAL HH	**100**	**100**	**100**	**100**	**100**	**100**

Table 3: Percent number of zeros and missing values for non-food expenditure items.

	HSE2		HSE3		HSE4	
Non-food items	**% zeros**	**% missing values**	**% zeros**	**% missing values**	**% zeros**	**% missing values**
Education	23	0	19	0	20	0
Water	76	4	82	4	81	8
Payment for goods bought by hire-purchase or lay-buy	71	12	74	9	67	19
Health	52	26	59	23	55	35
Transportation	22	37	13	31	18	62
Religious Expenses	43	30	45	25	40	43
Life insurance, Burial societes, Funeral policies	53	17	46	19	38	35
Electricity	48	5	43	4	39	10
Other usual expenses	49	12	58	18	57	24
Fuel	31	14	30	8	34	24
Large expenses	86	5	73	25	81	18
TOTAL HH	**100**	**100**	**100**	**100**	**100**	**100**

Given the high percentage of missing and zero values, it is therefore necessary to explore and understand the missingness mechanism, describe the missing data pattern and whether the data are missing at random. If the latter assumption is verified, we could then apply multiple imputation.

Missing data patterns: Descriptives

In order to use MICE (Multiple imputation by Chained Equation), we will test the data for missing at random assumption (MAR). As the NMAR is untestable, future work will perform departure from MAR assumption via sensitivity analysis (Carpenter, Kenward et al. 2007).
Zeros are assumed to be structural, meaning real quantities and they will be neglected for now.

Patterns of missing data patterns are explored by using the Missing Value Analysis module (MVA) in SPSS. The latter provides the separate-variance t tests table that can help identify the variables that could influence the quantitative variables of interest (in my case the expenditure items list, presented in table 2 and 3). The test is computed by creating a binary variable for each

missing/non missing variable (for an individual case). Separate means are reported for the two groups.

The same analysis was performed for HSE 2, 3, 4, selecting the following household level variables:

1. *quantitative (scale variables):* number of deaths per household, total AIDS deaths, no of AIDS and NOAIDS deaths in various age groups, proportion of females, household size, proportion of individuals in various age groups, number of unemployed, all food expenditure items and all non food items, number of assets, number of old-age pensions.
2. *Categorical variables*: asset index (proxy for wealth), maximum level of education in the household, the Isigodi area to which the household belong, whether the households resides in a rural/urban or periurban area, whether the household has electricity, whether the household has a toilet or whether the household has piped water.

The Little's Chi-square statistics for testing whether the data are MCAR was also performed. If the p-value is less than 0.05 level, the data are not MCAR. Such test is reported while running the EM (Expectation-Maximization) algorithm in SPSS. The EM method assumes that there is a distribution for the missing data and makes inferences on the likelihood under that distribution (SPSS, missing value analysis reference book).

The Little's Chi-square statistics for HSE 2 is 17125, df= 16446, Sig=.000.

From the descriptive analysis we can conclude that data are not MCAR. The location variables are quite interesting because the number of missing items seems to be related to geographical location (Isigodi area and rural/urban location, Tables available upon request). However this difference might be also due to chance.

Methodology

In order to address such a complex methodological setting, a comprehensive approach that combine methods to deal with censored observations (tobit models) and multiple imputation techniques (MICE) was implemented to take into account missing data and generate standard errors that reflect the imputation process uncertainty. From now on, such an approach will be named as "ad-hoc approach". Substantially the latter imputation process is divided into three steps: the treatment of positive outliers (only non-zero values), the treatment of zeros, and the missing data treatment.

Different imputation methodologies are implemented and three scenarios are produced:

1. a "naïve scenario" with no imputation (listwise deletion)
2. a "crude imputation" scenario; it entails a two-step procedure:
   - the imputation of extreme outliers through medians computed by logical variables. An outlier is defined as a value higher or lower than 3 Standard Deviations (SD) from the median (Carletto, Covarrubias et al. 2007)
   - imputation of missing values based on cluster median expenditures (cluster defined as *Isigodi areas*) –World Bank approach (Deaton and Zaidi 2002)
   - an "*Ad-hoc"* approach which consist of a three step procedure:
     a. Imputation of Outliers (as in crude imputation)
     b. Tobit regressions
     c. Multiple Imputation by Chained Equation-MICE (Van Buuren-Royston)

Results are presented in terms of a comparison of poverty lines. A poverty line is defined as a subjective judgment in terms of a socially acceptable minimum standard of living (Ravallion 1992).

Poverty lines (PL) can be absolute, i.e. fixed to a set threshold in terms of living standards, or relative, which vary with the average standard of living in a country or region/group. In this paper, an **absolute** PL is set at 2$ a day PPP and equivalent to 240 Rands per capita a month at 2003 prices.

A relative PL is computed in terms of 50% of "median" consumption (Ravallion 1992).

"Ad hoc approach"

The ad-hoc imputation procedure consists of three steps. The **first step** is the trimming of outliers, or implausible positive values different from zero, where an outlier is defined as a value higher or lower than 3 standard deviation from the median expenditure after its log transformation (Carletto, Covarrubias et al. 2007). The trimming of outliers was performed on each food and non food expenditure item on all HSE waves, and is in common with the crude imputation procedure.

.

The **second step** consisted of performing tobit regressions in order to deal with the large number of zeros in expenditure data vector. Heeringa (2002) defines this feature of the data "semicontinuous-distributions".

In our case, the untransformed distribution is highly skewed to the left. Skewness and non-normality here is addressed via logarithmic transformation of non-zero amounts.

There is a vast literature that takes care of this methodological problem. Whether considering the zeros as missing values, or truncated distribution, is widely debated (Little and Su 1987; Heeringa, Little et al. 1997; Heeringa, Little et al. 2002).

Here models that deal with distributions that are restricted to non-negative values, such as Tobit regression were used. Tobin (1958) was the first to suggest a solution to the censoring problem; it became known as "Tobin's probit" or tobit model.

Specifically we implement a tobit model on the untrasformed distributions of food expenditures to address the number of zeros in HSE 3 and 4. The assumption underlying this decision is that while it is possible to spend nothing on non-food items, households should be spending a minimal amount on food. Given the large number of zeros in the food expenditure items list, this might hide the fact that some of them might be missing values instead. As such, the "zero problem" was tackled as a censoring problem. Long (1997) also provides an exhaustive treatment of the topic.

The tobit regression model can be summarized as follows (Wooldridge and Wooldridge 2006). The observed response or expenditure item, y, is expressed in terms of a latent variable y*:

$$y^* = \beta_0 + x\beta + u, u \mid x \sim Normal(0, \sigma^2) \quad (1)$$

$$y = \max(0, y^*) \quad (2)$$

The variable y is equal to y* when the following is true :

$$y = \{ \begin{matrix} y^* \text{if} \;\; y^* > 0 \\ 0 \;\; \text{if} \;\; y^* \leq 0 \end{matrix} \}$$

The predicted probabilities are estimated based on each regression model. Each food expenditure item is regressed on other food expenditures items following a decreasing order of presence of zeros and other predictors (number of assets, maximum level of education in the household, Isigodi area, rural/urban location and household size).

After performing the tobit regressions, the "false" zeros are identified according to the prediction model in the following way. The tobit regression gives a probability - *x* – of a food expenditure item being non-zero. A random number *u* was then drawn from a uniform on [0,1]. If u occurred to be less than x (i.e. in the long run with probability *x*), their imputed value that time was non-zero, and was set to missing (i.e. will be drawn using MICE). The latter statement is equivalent to saying that the "false zero" is set to missing when the positive predicted probabilities are larger than the random number. If *u* was found to be larger than *x* (i.e. in the long run with probability *y* less than *x*), their imputed value that time was zero. The food expenditure quantities are updated at each round.

In summary, the added value of this procedure is that instead of replacing the value with the fitted value, the value is attributed via the MICE imputation model (third step in the methodology).
The aim of this is to increase the variability between the imputed data sets.

Ad-hoc Imputation procedure

The third stage of the ad-hoc imputation procedure consisted of employing "MICE"- multiple imputation by chained equations (Van Buuren and Oudshoorn 1999; Royston 2004; Royston 2004; Royston 2005; Royston 2005) which is implemented in STATA via the ICE module.
The general idea underlying the method in STATA is a regression model where missing values are replaced with "plausible" substitutes (based on distribution of given data). The assumption behind the method is that data are assumed to be missing at random as opposed to MCAR (i.e. it assumes that the probability that a datum is missing does not depend on unobserved information). The procedure entails producing *m* imputed datasets, where the rule of thumb has been *m*=5 imputations (Schafer 1999). Royston suggested increasing the number of imputations up to 10 or greater than 20 as he proposes in his empirical studies (Royston 2004) according to a coefficient of variation of the confidence coefficient $tv\sqrt{T}$ less than 5%. T is the total variance or adjusted sum of within and between imputation variance, of Q^2 bar (-averaged Q-the population quantity to be estimated). There is no acknowledged consensus on the ideal number for m.
The method averages estimates of the parameters of interest (Q) and standard errors and confidence intervals are calculated according to Rubin's rules (combining information on within and between imputation variation-the latter is extremely important to reflect the variability due to imputation uncertainty). The algorithm type is the Gibbs sampler, where the distribution of missing values of a covariate, is sampled conditional on the distribution of the remaining covariates.
The choice of the imputation model and the appropriate number of predictors require careful thought as a correct imputation model should include all the predictors and interactions that will be in the researcher final analysis model. The latter issue is not trivial for analysis purposes.

A combined imputation procedure of logged food and non-food expenditures was run, with m=50. Food expenditures were passively[3] imputed with ICE i.e. the updated food expenditure

[2] Q bar is taken to be an average of its repeated estimates across the m imputed datasets.

[3] The passive option in Ice (STATA command) allows the use of "passive imputation variables" that depend on other variables, some of which are imputed. Basically the imputed food aggregate was updated with the new values that depend on each food expenditure item imputed values.

aggregate was then used as input to the imputation of non-food expenditures. The match option was also used. The final step was to aggregate over all imputed food and non-food items to obtain total imputed expenditure or imputed total consumption. The 1st imputed dataset was selected for exploratory analysis.

Crude imputation scenario

This scenario entailed basically two steps, first the treatment of outliers as already described, and second, the imputation of missing values. The latter consisted of replacing the missing item with the median expenditure of the Isigodi area where the household is located. Isigodi are areas in the demographic surveillance area for which single indunas are responsible (chiefs) (ACDIS reference insert). The latter was found in the descriptive analysis to be a key predictor of missingness. Isigodi is a proxy for location and one of the few geographical indicators which are present in the ACDIS dataset.

Results: ***Poverty lines & FGT measures***
How (& how much) do contaminated data affect poverty measures?

Results are presented in terms of comparisons of poverty lines (Ravallion and Bidani 1994). Poverty lines are based on per capita expenditure (derived as total household expenditure divided household size) deflated to 2003 prices in order to adjust for inflation. In this analysis there is no adjustment for economies of scale and size (Lanjouw and Ravallion 1995).
It is important to underline that this paper also aims at assessing the extent to which contaminated data can affect poverty measures. Oversimplifying the subject matter, poverty measurement entails defining a poverty line and calculating poverty indices. The literature is vast and technical and I will present my results calculating the Foster-Greer-Thorbecke poverty indices (Foster, Greer et al. 1984). This class of poverty indices comprises the poverty headcount (FGT0) which reflects the extent or incidence of poverty and indicates the proportion poor; the poverty gap index (FGT1) also named the depth of poverty which measures the distance separating the poor to the poverty line; and the squared poverty gap index (FGT2) named the severity of poverty and attributes more weight to the poorest among the poor. The latter is also harder to interpret and it is not essential to the scope of my analysis.
The absolute poverty line is defined as 240 Rands per capita per months and equivalent to a 2 $ a day PPP.
The relative poverty lines are calculated based on 50% of median consumption. I produced relative poverty lines based on each scenario and for each HSE wave. In summary:

1. Relative Poverty line (PL) based on naive scenario (no imputation)

2. RPL based on crude imputation scenario
3. RPL based on ad-hoc imputation (*ICE)*

So does imputation methodology matter? And, if so, does it outweigh its "methodological" costs?

Figure 1 presents the distribution of log per capita expenditure per resident member using Kernel density estimates. The dotted line represents log per capita expenditure without imputation, the uninterrupted line the "ad-hoc imputation" and the dashed line the "crude imputation". Looking at the graphs, we can see how the two imputations modify the distribution of log per capita expenditure for HSE 2 and 4 to a lesser extent for HSE 3.

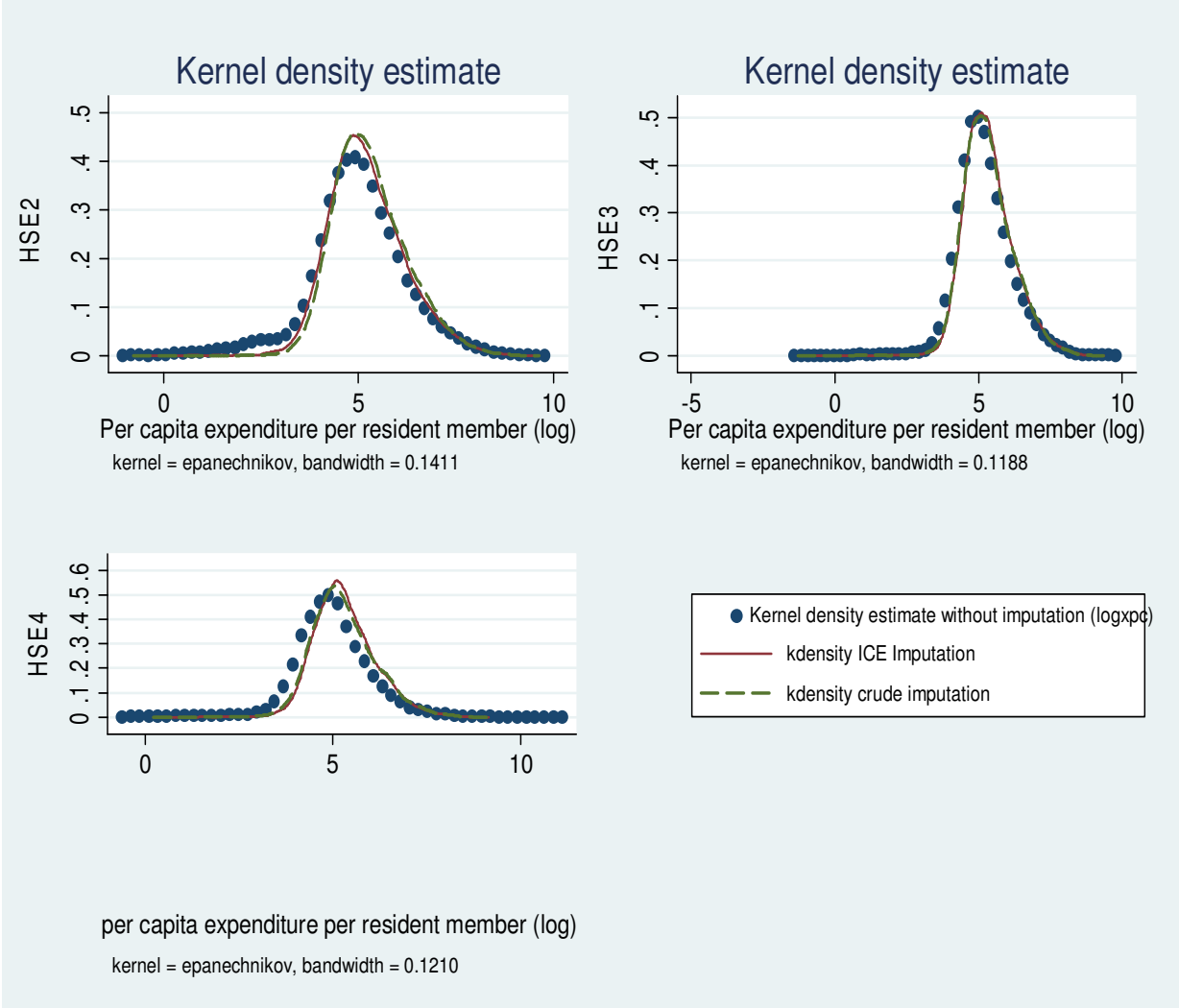


Table 5 presents poverty indices calculated using the above mentioned absolute poverty line.

Using a fixed threshold, we can assess to what extent the imputation procedures have an effect on the poverty indices. Looking at the naïve imputation results, we would tend to conclude that the poverty headcount (FGT0) remained the same and then increased in HSE 4. The latter finding is in line with most of the South African literature on poverty trends (Hoogeveen and Özler 2005) which essentially stress, notwithstanding the divergence in magnitudes, that post-apartheid poverty trends declined and then rose recently.

However, while results for the two imputations methodologies present the same trend, the magnitudes are lower when compared to the naïve scenario. The naïve imputation tends to overestimate the proportion poor, by construction, as there is no treatment for the large number of zeros and missing values.

The proportion poor in the ad-hoc and crude imputation scenario tend to be quite similar.

Note that in table 5 and 6, Ice minus is a consumption aggregate that excludes the three extra items contained in HSE 4 and as such has an identical consumption aggregate to the ad-hoc scenario (ICE) in HSE 3.

Table 4: Foster-Greer-Thorbecke poverty indices, FGT(a), Absolute poverty line.

		All obs	a=0	a=1	a=2
		Naïve	0.71	0.41	0.28
		Crude	0.64	0.31	0.19
HSE2	PL (2$)	Ice	0.66	0.34	0.21
		Naïve	0.71	0.37	0.23
		Crude	0.66	0.31	0.18
HSE3	PL (2$)	Ice	0.65	0.31	0.18
		Naïve	0.79	0.45	0.30
		Crude	0.69	0.34	0.20
HSE4	PL (2$)	Ice-minus	0.69	0.33	0.19

*Ice-minus =excluding snacks/fruit/eggs in Food Expenditure (FE) in HSE 4

Poverty trends depend on the poverty line used and the extent of poverty and inequality changes with the definition of consumption (Lanjouw and Lanjouw 1997). Specifically poverty indicators such as FGT class measures and also inequality, change when different measures of consumptions are used. In empirical findings, while the headcount seems to appear fairly stable, FGT1 and FGT2 seem to take ambiguous directions while changing poverty lines)(Lanjouw and Lanjouw 1997). However as we can see from Table 5 the direction of chianges in FGT0, FGT1 and FGT2 is the same regardless the imputation procedure.

Table 6 presents estimates of poverty defined by a relative poverty line set at 50% of median consumption. Obviously setting the poverty line to 50% of median consumption is a severe criteria but this is not essential to the analysis. Different measures of consumption (proxied by per capita expenditure, in the naïve, crude, and ad-hoc scenario) define different relative poverty lines.

Based on the three measures (log per capita expenditure in the naïve scenario, log imputed per capita expenditure in the crude imputation scenario, log imputed per capita expenditure in the ad-hoc imputation scenario), a higher poverty line is obtained in the case of imputed measures. The

two imputations seem to have yielded similar results (PLs in crude vs. ICE are not so distant in magnitude) except than for HSE 2.

In conclusion, it is important to note that relative poverty lines are higher in the imputation scenarios, and as such reflect a lower poverty headcount, by construction.

The direction of changes in the poverty indices (FGT0-1-2) is consistent in the three scenarios (naïve, crude, Ice). In summary, poverty trends are consistent and robust under the different imputation scenarios, a finding that also has been acknowledged by Vermaak (2008).

Table 5: Foster-Greer-Thorbecke poverty indices, FGT(a), Relative PL=50% of median consumption.

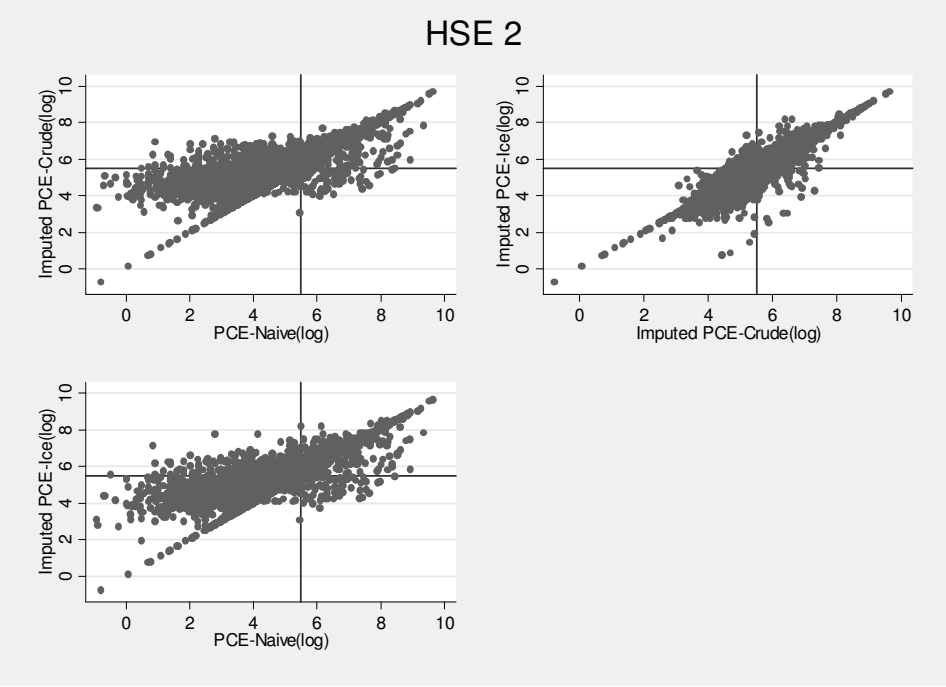
	RL	All obs	a=0	a=1	a=2
	68.15	Naïve	0.23	0.11	0.08
	86.56	Crude	0.19	0.06	0.02
HSE2	**80.24**	Ice	0.20	0.06	0.03
	73.12	Naïve	0.18	0.06	0.03
	84.44	Crude	0.17	0.05	0.02
HSE3	**85.93**	Ice	0.16	0.05	0.02
	58.71	Naïve	0.19	0.07	0.05
	77.48	Crude	0.16	0.05	0.02
HSE4	**79.77**	Ice-minus	0.15	0.04	0.02

*Ice-minus =excluding snacks/fruit/eggs in Food Expenditure (FE) in HSE 4

Finally, visual representation is given to the various per capita expenditure measures (log) created for the three waves. Figure 1, 2 and 3 crosstabulates log imputed per capita expenditure in the crude imputation scenario (crude) vs. the log per capita expenditure in the naïve imputation scenario (naïve); log imputed per capita expenditure in the ad-hoc imputation scenario (ICE) vs. log imputed per capita expenditure in the crude imputation scenario (crude), and log imputed per capita expenditure in the ad-hoc imputation scenario (ICE) vs. log imputed per capita expenditure in the naive scenario(naïve), for HSE 2, 3, and 4 respectively (Figure 2 and 3 available upon request). The solid lines in each figure represents the log of an absolute poverty line equal to 240 Rands or 2$ a day PPP. The log imputed per capita expenditure in the ad-hoc imputation scenario classifies fewer people as poor. Such an imputed measure is slightly more refined than its imputed counterpart (crude), as the log imputed per capita expenditure in the ad-hoc imputation scenario pushes those households classified as rich by the crude imputation below the PL.

Lastly figure 4 and 5 compare expenditure-based measures versus asset-based indicators.
Figure 4 plots a simple count index, based on the number of assets in the different waves, by the number of households that are classified as poor and non-poor under the various imputation scenarios for HSE 2. On average, non-poor households should have/own more assets compared with poor counterparts. However Figure 4 can convey two different messages: the first is that the number of assets that a household possess or share is a different indicator of wealth when compared with expenditure-based indicators. The second is that they don't seem to be correlated, even when acknowledging the naïve scenario as the gold standard or the "true" scenario under a missing completely at random assumption: the non-poor as categorized by the naïve scenario, own quite a diverse number of assets across the various groups.

Figure I[4]: Comparisons of log per capita expenditure per resident member, in the naïve, crude and ad-hoc imputation scenario, HSE2.



[4] Figure II and III were omitted due to document size constraints. Available upon request.

Figure IV: Number of assets vs. Poor/Non-poor households, in the various scenarios.

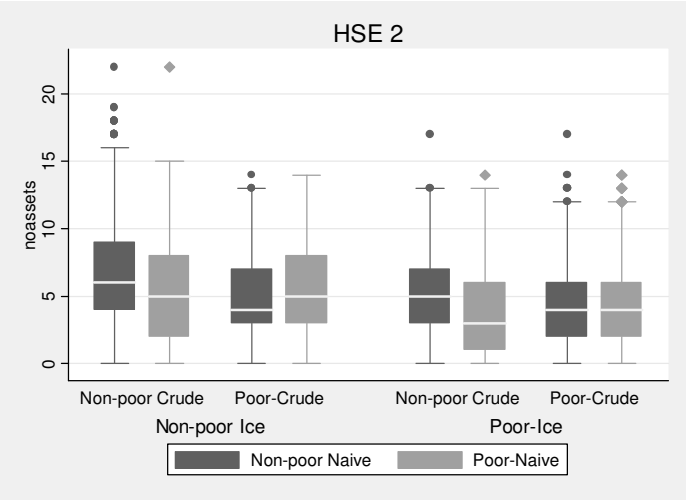


Figure V: Performance of asset index vs. imputed and non-imputed per capita expenditure (log)

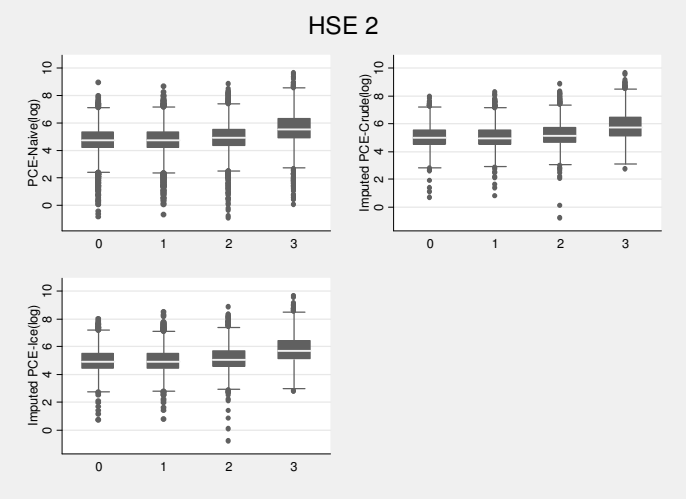


Figure 5 presents another visual assessment of the performance of asset-based measures (specifically here an asset index based on principal component analysis) versus imputed and non imputed log-per capita expenditure for HSE 2. This is not the place to assess whether an asset index (Figure 5) could perform better than a simple asset count (Figure 4). The aim of the figure is substantially twofold: the first is to highlight how such indices divided into quartiles, fare, in classifying poor vs. the non-poor vis a vis expenditure-based measures; the second is,

instrumental to this paper, to show whether different imputation methods could give different messages when compared with yet another indicator of wealth. The construction of such an asset index is based on principal component analysis as presented by Filmer and Pritchett in their seminal publication (Filmer and Pritchett 1998). The authors suggested using the statistical procedure of principal component to determine the weights for an index of the asset variables. In our analysis, the single asset variables have been used to create an index of assets that proxies for household wealth for each wave of the ACDIS data.

Figure 5 shows how the first 3 quartiles of the asset index have similar mean per capita expenditure (log). However, there is an upward trend in mean per capita expenditure across the asset index quartiles.

Figure 4 and 5 suggest the possibility that asset indices could be a poor proxy for household level wealth. Alternatively, asset-based wealth could just be poorly correlated with expenditure. The two measures may well identify two different dimensions of wealth: current consumption (expenditure) vs. permanent income (as proxied by asset ownership).

Conclusions

The objective of this paper has been to evaluate the quality of the consumption module for each socio-economic wave (HSE 2, 3, and 4) of the ACDIS longitudinal database. We addressed the presence of missing values and measurement error via the development of an ad-hoc approach that could address the problem in a comprehensive way.

The impact of different imputation methods (a naïve, crude and ad-hoc imputation scenario) on the Foster-Greer-Thoerbecke class of poverty indices was assessed. Lastly, it compares imputed and non-imputed per capita expenditure (log) versus other dimensions of wellbeing such as the number of assets and asset indices based on PCA.

Our findings highlight the fact that poverty indicators are quite robust to the choice of imputation method; however the crude imputation appears to be too generous, classifying households as wealthier than the ad-hoc approach or ICE method. The ad-hoc imputation scenario (ICE) performs better. It decreases the proportion poor by shifting the distribution up in each wave (Figure 1, 2 and 3), is statistically more accurate (standard errors are derived via the Rubin's rule), and confidence intervals are better estimated. Also additional predictors add more information and significance to the analysis. The added value of ICE compared with other imputation methods is that the former takes into account variability due to imputation uncertainty. The former will eventually lead to a better estimation of the final model of interest. This consideration strengthens our belief that the benefits of multiple imputation outweigh its

methodological costs and should be considered while addressing such methodological problems in socio-economic surveys.

When we derived poverty indices based on absolute and relative poverty lines based on 50% of median consumption (proxied by per capita expenditure, in the naïve, crude, and ad-hoc scenario), we found that estimated poverty trends are affected little by the different definitions of consumption employed and that even using such different consumption based welfare indicator measures, relative PL derived from imputed measures are similar in magnitude.
Specifically when comparing standard poverty indicators using the broader Foster-Greer-Thorbecke (FGT)[5] class of poverty measures (Foster, Greer et al. 1984), which give weight to the depth of poverty as well as to the number of the poor, we found that FGTs are quite consistent using an absolute poverty line of 2 dollars a day per capita, specifically that FGT0 and FGT1 (incidence and depth of poverty) are both lower and in the same direction in imputed measures compared to a non-imputation or naïve scenario.
The latter finding is of particular importance, as poverty and inequality changes with the definitions of consumption (Lanjouw and Lanjouw 1997). Specifically, poverty indicators such as FGT class measures and also inequality, change when different measures of consumptions are used. While the headcount seems to appear fairly stable, FGT1 and FGT2 seem to take ambiguous directions based on traditional and austere poverty lines (including food and non food items) (Lanjouw and Lanjouw 1997).

Obviously the imputation procedure has an impact on the poverty rate, in terms of magnitude, by construction, and also the choice of the poverty line matters for poverty analysis. These two issues are highlighted all along this paper. Notwithstanding the lack of a gold standard, or the "true" poverty indicator, what is really important here is to obtain a relative ranking of households.

The final remark (Figures 4 and 5), which deserves further investigation, is that consumption-based imputed measures (regardless of imputation procedure) seem to convey a different message when compared with asset indices based on either principal components analysis or simple assets count. Asset indices could be either a poor proxy for household level wealth or alternatively, asset-based wealth could just be poorly correlated with households spending

[5] The Foster-Greer-Thorbecke measures: the poverty headcount (FGT0) reflects the extent of poverty, the poverty gap index (FGT1) the depth of poverty, and the squared poverty gap index (FGT2) the severity of poverty.

patterns. The two measures could well identify two different dimensions of wealth: current consumption (expenditure) vs. permanent income (as proxied by asset ownership).

Our empirical findings emphasize the need to address non-sampling errors while incorporating consumption modules in data collection effort that are not necessarily designed at measuring poverty. Demographic surveillance systems need to balance the trade off between detailed demographic data collection and meaningful (and sufficient) economic information.

The final aim of this work would be to measure the impact of AIDS mortality on consumption poverty, longitudinally, in the presence/absence of imputation, compare the results and assess whether such conclusions diverge. By attaching a measure of uncertainty to the consumption-based poverty indicator, I will quantify to what extent the latter exercise can bias an assessment of the impact of AIDS mortality on household poverty.
Also further research will examine departure from MAR assumption.

References

Ardington, C., D. Lam, et al. (2006). "The sensitivity to key data imputations of recent estimates of income poverty and inequality in South Africa." Economic Modelling **23**(5): 822-835.

Carletto, G., K. Covarrubias, et al. (2007). Rural Income Generating Activities Study: Methodological note on the construction of income aggregates. RIGA Publications. Rome, Italy, Food and Agriculture Organization.

Carlin, J. B., N. Li, et al. (2003). "Tools for analyzing multiple imputed datasets." The Stata Journal **3**: 226-244.

Carpenter, J. R., M. G. Kenward, et al. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. **16:** 259.

Carter, M. R. a., J. b. May, et al. (2007) "The economic impacts of premature adult mortality: panel data evidence from KwaZulu-Natal, South Africa. [Editorial]." **Volume**, DOI:

Chapoto, A. and T. S. Jayne (2006). "Socioeconomic Characteristics of Individuals Afflicted by AIDS-Related Prime-Age Mortality in Zambia." AIDS, Poverty, and Hunger: Challenges and Responses.

Deaton, A. (2003). "Household Surveys, Consumption, and the Measurement of Poverty." Economic Systems Research **15**(2): 135.

Deaton, A. and S. Zaidi (2002). Guidelines for constructing consumption aggregates for welfare analysis. Washington, DC, World Bank.

Diggle, P. and M. G. Kenward (1994). "Informative drop-out longitudinal data analysis." Applied statistics **43**(1): 49-93.

Enders, C. K. (2006). "A Primer on the Use of Modern Missing-Data Methods in Psychosomatic Medicine Research." Psychosom Med **68**(3): 427-436.

Filmer, D. and L. Pritchett (1998). "Estimating Wealth Effects without Expenditure Data or Tears: With an Application to Educational Enrollments in States of India." World.

Fisher, J. (2006). Income Imputation and the Analysis of Expenditure Data in the Consumer Expenditure Survey, U.S. Bureau of Labor Statistics.

Foster, J., J. Greer, et al. (1984). "A Class of Decomposable Poverty Measures." Econometrica **52**(3): 761-766.

Gillespie, S. (2006). AIDS, Poverty, and Hunger: Challenges and Responses, Int Food Policy Res Inst IFPRI.

Gillespie, S., R. Greener, et al. (2007). Investigating the empirical evidence for understanding vulnerability and the associations between poverty, HIV infection and AIDS impact. [Editorial], AIDS November 2007;21 Suppl 7:S1-S4.

Gillespie, S., S. Kadiyala, et al. (2007). Is poverty or wealth driving HIV transmission?. [Editorial], AIDS November 2007;21 Suppl 7:S5-S16.

Heeringa, S. G., R. J. A. Little, et al. (1997). Imputation of Multivariate Data on Household Net Worth. University of Michigan.

Heeringa, S. G., R. J. A. Little, et al. (2002). Multivariate Imputation of Coarsened Survey Data on Household Wealth**:** 357–371.

Heitjan, D. F. and D. B. Rubin (1991). "Ignorability and coarse data." Ann. Statist **19**(4): 2244-2253.

Hoogeveen, J. G. and B. Özler (2005). Not Separate, Not Equal: Poverty and Inequality in Post-Apartheid South Africa, William Davidson Institute at the University of Michigan Stephen M. Ross Business School.

Lanjouw, J. O. and P. Lanjouw (1997). "Poverty Comparisons with Noncompatible Data: Theory and Illustrations." World.

Lanjouw, P. and M. Ravallion (1995). "Poverty and Household Size." The Economic Journal **105**(433): 1415-1434.

Little, R. J. A. and D. B. Rubin (2002). Statistical analysis with missing data . Hoboken, NJ: Wiley.

Little, R. J. A. and H. L. Su (1987). Missing data adjustments for partially scaled variables**:** 644-649.

Murtin, F. and F. Marzo (2008). HIV/AIDS and Poverty in South Africa: a Bayesian Estimation.

Oudshoorn, K., S. van Buuren, et al. (1999). Flexible Multiple Imputation by Chained Equations of the AVO-95 Survey, TNO Prevention and Health.

Paulin, G. D. and D. L. Ferraro (1994). "Imputing Income in the Consumer Expenditure Survey." Monthly Labor Review **117**: 23.

Philip K. Hopke, C. L. D. B. R. (2001). "Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic." Biometrics **57**(1): 22-33.

Ravallion, M. (1992). "Poverty comparisons: A guide to concepts ad methods."

Ravallion, M. and B. Bidani (1994). How Robust Is a Poverty Profile?, World Bank. **8:** 75-102.

Royston, P. (2004). "Multiple imputation of missing data: an implementation of van Buuren's MICE, and more."

Royston, P. (2004). "Multiple imputation of missing values." STATA JOURNAL **4**: 227-241.

Royston, P. (2005). "MICE for multiple imputation of missing values."

Royston, P. (2005). "Multiple imputation of missing values: update." STATA JOURNAL **5**(2): 188.

Rubin, D. B. (1976). "Inference and missing data." Biometrika **63**(3): 581-592.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys.

Rubin, D. B. (1996). "Multiple imputation after 18+ years." Journal of the American Statistical Association **91**(434): 473-489.

Rubin, D. B. and J. L. Schafer (1990). Efficiently creating multiple imputations for incomplete multivariate normal data.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data, Chapman & Hall/CRC.

Schafer, J. L. (1999). Multiple imputation: a primer. **8:** 3.

Schafer, J. L. and J. W. Graham (2002). "Missing Data: Our View of the State of the Art." PSYCHOLOGICAL METHODS **7**(2): 147-177.

Tanser, F., V. Hosegood, et al. (2007). "Cohort Profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey." International Journal of Epidemiology.

Van Buuren, S. and C. G. M. Oudshoorn (1999). "Flexible multivariate imputation by MICE." Leiden, The Netherlands: TNO Prevention Center.

van Buuren, S. and K. Oudshoorn (1999). "Flexible multiple imputation by MICE." Leiden: TNO Prevention and Health, TNO-PG **99**.

Vermaak, C. (2008). The impact of multiple imputation of coarsened data on estimates of the. working poor in South Africa. Development Policy Research Unit Conference "The Regulatory Environment and Its Impact on the Nature and Level of Economic Growth In South Africa".

Wooldridge, J. M. and J. M. Wooldridge (2006). Introductory Econometrics: A Modern Approach, Thomson/South-Western.