Published in final edited form as:
Econ Model. 2006 ; 23(5): 822–835. doi:10.1016/j.econmod.2005.10.009.

The sensitivity to key data imputations of recent estimates of income poverty and inequality in South Africa

Cally Ardington[a,b], **David Lam**[c], **Murray Leibbrandt**[b,d,*], and **Matthew Welch**[b,e]

a*Department of Statistical Sciences, University of Cape Town, South Africa*

b*Southern Africa Labour and Development Research Unit, University of Cape Town, South Africa*

c*Department of Economics and Population Studies Center University of Michigan, South Africa*

d*School of Economics, University of Cape Town, Private Bag, Rondebosch, 7701, South Africa*

e*DataFirst, University of Cape Town, South Africa*

Abstract

Existing literature using South African censuses reports an increase in both poverty and inequality over the 1996 to 2001 period. This paper assesses the robustness of these results to a number of weaknesses in the personal income variable. We use a sequential regression multiple imputation approach to impute missing values and to explicitly assess the influence of implausible income values and different rules used to convert income that is measured in bands into point incomes. Overall our results for 1996 and 2001 confirm the major findings from the existing literature while generating more reliable confidence intervals for the key parameters of interest than are available elsewhere.

Keywords

Inequality; Poverty; Missing data imputation

1. Introduction

Changes in inequality and poverty are key dimensions of the transformation of any society. Given the twentieth century history of South Africa, these two dimensions of economic well-being and, in particular, their changing racial profiles have been of special interest. An important empirical tradition in tracking longer-run South African inequality and poverty changes has made use of records of personal income collected in the national censuses of 1970, 1991, 1996 and 2001 (McGrath, 1983; Whiteford and McGrath, 1994; Whiteford and van Seventer, 2000; Leibbrandt et al., 2004; Simkins, 2005). In the apartheid era, such empirical work was central to highlighting the destructive impact of racially driven policies on South Africa's non-white groups. In the post-apartheid era, these empirical analyses have taken on additional importance.

Whiteford and van Seventer (2000) documented a high and constant national income inequality for the 1991 to 1996 period. Both Simkins (2005) and Leibbrandt et al. (2004) showed that this inequality remained high and even took a turn for the worse in the period 1996-2001. As regards racial inequality, between 1996 and 2001, inequality within each race group increased.

[*] Corresponding author. School of Economics, UCT, Private Bag, Rondebosch, 7701, South Africa. Tel.: +27 21 650 2726; fax: +27 21 650 2854. *E-mail address:* mleibbra@commerce.uct.ac.za (M. Leibbrandt).

Formal decompositions showed that this within-group contribution to aggregate inequality increased while the between-group component decreased. This represented a continuation of the trend that Whiteford and van Seventer (2000) had noted for the 1991-1996 period and, indeed, from as far back as 1975.

The poverty analysis of Simkins (2005) and Leibbrandt et al. (2004) revealed that national poverty worsened over the period, particularly for Africans. This too suggested a continuation of the longer-run poverty trend revealed by Whiteford and van Seventer (2000). However, for the 1996-2001 period, the extent to which poverty was measured as increasing was dependent on the choice of poverty line. At lower poverty lines, the increase in poverty is significantly more muted than at higher poverty lines.

These analyses are based on the income variable from the census. While the broad reach of the census data is their strength, the income data are far from ideal. Three particular weaknesses of the personal income variable are a high rate of missing data; an implausible number of zero values and the fact that the question on personal income requested an appropriate income band for each person rather than an income value. The broad task of this paper is to ascertain the sensitivity of key income-based measures of well-being to these weaknesses.[1]

Sixteen percent of individuals in the 2001 census 10% sample have missing income values. Who are these people and their households? If they were not missing, where would they have fallen in the distribution of income and what impact would they have had on measured poverty and inequality? In the next section of the paper, we deal with missing data by imputing income values for those with missing income data for 2001 using contemporary multiple imputation techniques. Statistics South Africa offers users of the 2001 data a single hotdeck imputation for the missing 2001 personal income data. In line with current international best practice, multiple imputation approaches are preferred to such single imputations. In this section we compare our imputation results with the hotdeck results of Statistics South Africa.

A large percentage of individuals, including some who are employed, is recorded as having zero incomes. On aggregating these personal incomes into household incomes, this translates into a quarter of households with zero total income values. Even allowing for South Africa’s low labour market participation rates and high unemployment rates, it is highly unlikely that all of these zero income households had no adult members receiving any income. In analysing poverty and inequality, previous practice has been to ignore the zeros or to change them to some arbitrarily small number. The former practice is an arbitrary decision to effectively remove a group of households who currently make up the bottom of the distribution. As such, this decision sharply decreases measured poverty levels and also narrows inequality. The latter practice effectively accepts all recorded zeros as genuine zeros, possibly leading to an overestimate of measured poverty and inequality.[2] In the third section of the paper we consider the impact of the implausibly high percentage of households with zero incomes. Our approach is to use a set of decision rules to reclassify potentially problematic zero incomes as missing and then to re-run the multiple imputation process on these augmented missing data. The process allows for the possibility that any values that are reclassified from zero to missing to be imputed back to zero once more if the census data support such an imputation.

[1] Simkins (2005) makes a promising start down this road. For both 1996 and 2001, a set of decision rules is applied to allocate positive incomes to some adults with missing incomes and to adults with zero incomes that are in households with zero income. These decision rules are overt and replicable. However, they are not anchored in the imputation literature and there is no testing for the sensitivity of results to plausible rule changes.

[2] An example of the impact of these assumptions on measured poverty and inequality is contained in Appendix A of Leibbrandt et al. (2004).

Cronje and Budlender (2004) highlight another weakness of the census income variable; namely, that in both 1996 and 2001, the question on personal income asked for an appropriate income band for each person rather than a specific income value. The import of this is the fact that in order to estimate measures of poverty and inequality it is necessary to adopt some rule to translate these bands into point estimates. The problem of income bands is exacerbated in comparative work as the bands are not defined to be consistent real income categories across the two years, with particularly large differences at the top end.[3] The general practice in South Africa has been to attribute the band midpoints to individuals. This is only one of a number of possibilities and in Section 4, we examine the sensitivity of key results to the mid-point rule by imputing point estimates from intra-band empirical distributions of personal incomes that are derived from national sample surveys of incomes and expenditures.

In order to keep the discussion of Sections 2-4 manageable, we discuss the techniques and illustrate their impact using the 2001 census data. However, all exercises were replicated on the 1996 data. The final section of the paper briefly returns to the issue of comparing 1996 and 2001 poverty and inequality in the light of our imputation work.

2. Dealing with missing data

The potential bias in estimates caused by missing data is a pervasive problem in empirical work. Unless the data are missing completely at random (MCAR), estimates that exclude individuals with missing data from the analysis will be biased.[4] Missing data are particularly problematic in calculating measures of poverty and inequality. If those with missing data disproportionately fall into the bottom of the distribution, then levels of poverty will be underestimated. Alternatively, if non-response is higher among the wealthy, measures of inequality are likely to be biased downwards. Sixteen percent of individuals in the 2001 census 10% sample have missing income data. The missing data problem is exacerbated with analyses at the household level as more than a quarter of individuals belong to households where all or some of the household members have missing income data. If missing data are ignored, all these individuals are excluded from any household level analyses such as the calculation of per capita poverty and inequality measures.

There are significant differences in the response rates across a number of variables. Whites are much more likely to have missing income data (24%) than Black Africans (14%). Response rates are higher in rural areas than urban areas. There is a large variation in response rates across provinces with more than 23% of individuals in the Western Cape missing income data and less than 7% of individuals in the North West. It is evident that the income data are not MCAR. Rather, as Whites and the Western Cape province are, respectively, relatively better off racial and spatial sections of the South African population, these response rates suggest that higher incomes are more likely to be missing.

If the data are missing at random (MAR), then we can adjust for non-response in order to reduce bias in our estimates.[5] There are a range of methods for handling such missing data including

[3] The highest band for personal income in 1996 was R30000 or more. This is lower than the real income equivalent of the top three bands in 2001. This incompatibility of income bands in real terms needs to be dealt with in order to compare the data across time. Leibbrandt et al. (2004) compressed the top end of the 2001 distribution of personal incomes into the real income equivalent of the top band in 1996. As all of these bands are way above any plausible poverty line, this has no impact on the analysis of poverty. However, as this decision effectively compresses the top end of the 2001 income distribution, this decision impacts on the inequality analysis. See Table A.3 in Appendix A of Leibbrandt et al. (2004) for a detailed set of results. There is no particular subtlety to the decisions that analysts make in this regard and the most that can be asked for is that the decisions are spelled out explicitly and that there is some assessment of the sensitivity of any analysis to alternatives.

[4] Suppose that y_i is a response of interest (income in this case), x_i is a vector of information (province, rural/urban, race, age, sex, education, employment status, occupation) known about person i in the sample. If the probability that person i will respond does not depend on x_i, y_i or the survey design, the data are MCAR. If data are MCAR, the respondents are representative of the selected sample. The MCAR mechanism is implicitly adopted when non-response is ignored (Lohr, 1999: 264).

weighting, imputation and non-parametric techniques (Lohr, 1999;Little and Rubin, 2000). Indeed, the 2001 10% sample from Statistics South Africa offers a set of imputations for all the missing data. In this case, a single-imputation hotdeck method was used to impute missing income values for individuals.[6] This means that missing values "are replaced by values from similar responding sampling units. The hotdeck literally refers to the deck of matching computer cards for the donors available for a non-respondent".

In this paper, we use a multivariate regression technique to multiply impute missing values. This technique has a number of advantages over the single hotdeck imputation approach adopted by Statistics South Africa. Firstly, a multivariate multiple imputation approach is more robust than a single hotdeck imputation. While all imputation based approaches rely on the observed data to impute values for missing items, the hotdeck technique is particularly sensitive to the problems of badly measured variables. If the data on some of your respondents are not good then there is a chance of drawing a bad respondent as a donor to replace your non-respondent. For example in the 2001 data, most fifteen-year olds are not earning any income. However, of the few that report that they are employed and are earning, one in two are earning implausibly high levels of income. With these values in the data set, a fifteen-year old with a missing income value might draw one of these cards from the hotdeck and be given an implausibly high income. In multivariate regression imputation, the impact of these outliers in generating imputed values is lessened. In addition, final estimates are obtained by averaging over multiple imputations, further reducing the problems of badly measured variables.

Secondly, any single imputation technique does not distinguish between observed and imputed values in the resultant data set and as such the variance of any estimates is understated. Multiple imputations generate a distribution of imputed values and a distribution of parameters of interest. This allows for the uncertainty due to imputation to be reflected in the standard errors of the estimates. Given such advantages, the imputation literature has a strong preference for running multiple imputations using a suitable multivariate technique.

2.1. The technique for multiply imputing missing values using a sequence of regression models

The multiple imputation approach adopted in this paper follows the sequential regression multiple imputation (SRMI) approach of Raghunathan et al. (2001). Most directly, our task is to impute an appropriate personal income band for each individual with missing income data. In the 2001 census, there are 12 income bands with 0 being the lowest "band" and R204 801 a month (R2 457 601 a year) being the lower bound of the top band. Thus, this task is to specify an ordered logit model that uses the best set of variables that are available in the census to allocate missing data into these income bands.[7] The explanatory variables used in the ordered logit can be broken into two matrices **X** and **Y**. The **X** matrix contains the set of predictor variables with no missing values. The **Y** matrix contains the variables with missing values. Let the k variables Y_1, Y_2,..., Y_k represent these variables ordered by the amount of missing values from least to most.

[5] If the probability of response depends on x_i, but not on y_i, the data are MAR as the non-response depends only on observed variables. We can successfully model the non-response, since we know that values of x_i for all sample units. If the probability of non-response depends on the value of y_i and cannot be completely explained by values of the xs, then the non-response is non-ignorable as we are unable to model it (Lohr, 1999: 265).

[6] Statistics South Africa used a combination of two kinds of imputations with the 2001 data. The first were "logical" imputations and the second were "hotdeck" imputations. For the logical imputations, "a consistent value is calculated or deduced from other information relating to the individual or household. For example, a married person with invalid sex would be assigned to the opposite sex of his or her spouse" (Statistics South Africa, 2001a: 3). If a logical imputation was not possible then the "hotdeck" procedure was used. While the conceptual distinction between logical imputations and other imputations is clear, tabulations of the logical imputations for some of our variables throw up some results that are not immediately obvious. For example, it is clearly "logical" to code two year olds as having no education but it is not clear how a number of adults were "logically" imputed to have no education. Therefore in this paper, we do not distinguish between Statistics South Africa's "logical" and "hotdeck" imputations.

With the census data we have complete data for each person on province of residence, whether they resided in an urban or rural area and race. These variables therefore make up the **X** matrix. The set of **Y**s ordered from least to most missing values consisted of age (a count variable), a gender dummy variable, an employment dummy variable, a four category occupation variable, years of education (a count variable), and income (an ordered categorical variable of 12 income bands).[8] The inclusion of income in the **Y** matrix goes to the heart of the SRMI approach to imputation. In this approach, income is imputed as part of a process of imputing missing values for all of the variables in the **Y** matrix. More formally, all missing values are imputed as part of a process to estimate the joint conditional density of Y_1, Y_2,..., Y_k given **X**. This density can be factored as:

$$\begin{aligned} f \quad & (Y_1, Y_2, \ldots, Y_k | X, \beta_1, \beta_2, \ldots, \beta_k) \\ & = f_1 (Y_1 | X, \beta_1)\, f_2 (Y_2, | X, Y_1, \beta_2) \ldots f_k (Y_k | X, Y_1, Y_2, \ldots, Y_{k-1}, \beta_k) \end{aligned} \tag{1}$$

where f_i represent the conditional density functions and $\boldsymbol{\beta}_i$ is a vector of parameters in the conditional distribution. In all cases the $\boldsymbol{\beta}_i$ vectors are estimated coefficients as well as estimates of the disturbance term. As mentioned above, our **Y** matrix contains count variables (age and education), binary categorical or dummy variables (gender, employment), a multiple category variable (occupation) and an ordered categorical variable (income). If Y_i is a count variable, then a poisson distribution is used to estimate f_i. If Y_i is binary, f_i is estimated using a logistic distribution. If Y_i is categorical, a multinomial logistic regression model is estimated and if Y_i is ordinal, an ordinal logistic regression model is estimated.

Settling on a set of imputed values for the missing **Y**s is analogous to settling on a satisfactory estimate of the joint conditional density of **Y** given **X**. The model is settled over a number of rounds. As reflected in Eq. (1) above, the first round starts with obtaining an estimate of the vector $\boldsymbol{\beta}_1$ in a regression of Y_1 on **X**. The missing values in Y_1 are then replaced by random draws from the posterior predictive distribution. That is, by first drawing a vector β_1^* from the posterior distribution of $\boldsymbol{\beta}$1 and then using β_1^* to generate a set of predicted values to replace the missing Y_1 values. This is followed by an estimate of Y_2 given **X** and the newly derived Y_1, including imputed values, on the right hand side. The first round finishes when the ordered logit is estimated to derive initial imputed values for missing Y_k (income bands) conditional on **X** and Y_1 to $Y_{k\text{-}1}$. Once this model has been estimated, the first complete data set with no missing values is available.

At the start of the second round, Y_1 is re-estimated including all first round **Y** imputations on the right hand side. The first round missing value imputations for Y_1 are replaced by a new set of imputations derived from this re-estimation. All in all, the essence of the Raghunathan et al. (2001: 88) method is that “the new imputed values for a variable are conditional on the previously imputed values of the other variables and the newly imputed values of variables that preceded the currently imputed variable”. These authors state that although it is theoretically possible that such a process does not converge to a stationary distribution, they have never encountered this in an empirical setting. As recommended by Raghunathan et al. (2001), we ran five rounds or iterations before settling on final imputed values.

This provides the sequential regression equivalent to the single vector of hotdeck imputations that were derived by Statistics South Africa. For multiple imputations we proceed to generate

[7] Van Buuren et al. (1999: 5) summarise a literature showing that “including as many predictor variables as possible tends to make the MAR assumption more plausible”. Given that most data sets are very large, computationally, it is not really possible to include every possible variable. However it is also not necessary as the increase in explained variance is often minimal once the best set of variables have been included. Therefore, at the very least, one wants to include the best set of variables that are available in the census for predicting income bands for individuals. We include 9 variables which is below the maximum of 15 to 25 variables that Van Buuren et al. (1999) suggest as a rule of thumb.

[8] More detail on all of these X and Y variables is provided in Ardington et al (2005).

a set of such imputations to give us a distribution of imputed values and a distribution of parameters of interest. Each imputation starts in the same way as described above by estimating a starting value for Y_1. Clearly the first estimated regression coefficients will always be the same but as the missing values are imputed after each regression using a random draw from the posterior distribution of the regression coefficients; i.e., including a randomly drawn error component, a different set of imputed values is generated from the outset. Thus, from the very first imputation, this round generates different imputed values. At the end of this process, *m* complete data sets have been derived incorporating *m* equally plausible sets of imputations for income bands for all individuals who had missing values on the income variable.

Parameters of interest can be derived from each data set. In the context of this paper, three parameters of interest are mean household per capita income, an index of poverty (as measured by the head count index) and an index of inequality (as measured by the Gini coefficient). The multiple imputation variance formula suggests that, in each of these three cases, the best estimate of each multiply imputed parameter is the mean of the *m* estimates of that parameter. The variability associated with this estimate consists of two parts; namely, the average within-imputation variance (V_w) and the between-imputation variance (V_b). A within sample variance is calculated for each parameter of interest each of the *m* times the parameter is calculated. The V_w is the mean of these *m* variances. The V_b is calculated as the variance of the *m* parameter estimates. The total variance V_t equals $V_w+((m+1)/m)V_b$ where *m* is the number of imputations and $((m+1)/m)$ is an adjustment for the fact that the V_b is being calculated off a finite number of parameter estimates. The square root of this total variance is the standard error associated with the best estimate of each of our three parameters. Interval estimates are then based on the *t*-distribution with $(m\text{-}1)(1+(1/(m+1))V_w/V_b)^2$ degrees of freedom Little and Rubin (2000: 87). In any single imputation model, one would be offered a single parameter estimate and a single variance associated with this estimate with no sense of how this parameter estimate might vary across equally plausible sets of missing data imputations.

The number of imputations required for a desired level of efficiency depends on the rate of missing information for the quantity being estimated. The rate of missing information, as distinct from the rate of missing data, is a measure of the relative increase in the variance due to non-response. According to Rubin's (1987) formula we calculated that fifteen imputations would be needed to achieve an efficiency of 95% for estimates of mean income.[9]

2.2. Estimates of mean per capita income, poverty and inequality

The technique outlined above was used to derive a set of imputed values and ascertain their influence on measures of poverty and inequality. Table 1 presents point estimates and 95% confidence intervals for a range of poverty and inequality measures for 2001.[10] Mean per capita income and Gini coefficients are presented in Panel A and poverty headcounts for both a R124 and R340 per capita poverty line are shown in Panel B. The lower line is a $2 per day poverty line, which is widely used for international poverty comparisons. The upper line is arbitrarily chosen to represent a generous poverty line in 2001 terms. The first row presents estimates that were calculated ignoring all missing income data. The estimates in the second row were calculated using Statistics South Africa's hotdeck imputations. The third row presents the combined estimates from our multiple imputations. The estimates in the fourth and fifth rows will be discussed in Sections 3 and 4, respectively.

[9] The efficiency of an estimate based on m imputations is approximately $(1+\gamma/\mathrm{m})^{-1}$ where γ is the rate of missing information. The rate of missing information is given by $\gamma=(1+1/m)V_b/V_t$. Little and Rubin (2000).

[10] A continuous measure of personal income was generated by allocating each individual the midpoint income of the band within which they are found. The highest (unbounded) band was assigned the lower bound value. Furthermore, because we are interested in per capita income, we summed all positive personal income for each household and then divided by household size to obtain a monthly per capita measure of income. For comparability between the two censuses and to avoid problems in calculating household size, we excluded all data on people living in institutions, and all results were weighted using the weights supplied by Statistics South Africa.

The imputations suggest that ignoring the missing values results in downwardly biased estimates of mean income and inequality and upwardly biased estimates of poverty at both poverty lines. When we adjust for non-response, mean per capita income increases, the percentage of households in poverty decreases and inequality increases. These results are consistent with our previous observation that response rates were lower in urban areas and amongst Whites, suggesting that individuals with imputed incomes have on average higher incomes than individuals with non-missing income.

The parameter estimates for Statistics South Africa's hotdeck imputation are closer to the combined SRMI estimates than to the no imputation results. However, the confidence intervals for the hotdeck estimates are noticeably tighter than those of the combined multiple imputation estimates. This is because the variances for the hotdeck results are akin to the within-imputation variances of any of the fifteen imputations that are derived as part of the SRMI multiple imputation process. The between-imputation variance is ignored by the single imputation hotdeck. This variance is clearly seen by comparing the estimates across each of the multiple imputations. The estimates of mean income and the poverty headcount ratio for each of the fifteen imputed data sets are shown in Table 2. As shown, the single imputation technique that does not take into account uncertainty due to imputation overstates the precision of the estimates.

3. Assessing the importance of implausible values

Given our warnings above about imputing off a data set that contains outliers and implausible values, as a next step we investigate the sensitivity of our results to such values in the data set. This is an especially important stage in the analysis of income as it allows us to acknowledge and deal with the implausibly high proportion of zero income households that are recorded. These households clearly have an impact on estimates calculated from the observed data values. In this section, we present a fresh set of imputations that begins by taking problematic zero values and recoding them as missing before re-running the imputations on missing values. In this way, through the imputation process zero income households are screened for plausibility and then either are assigned a positive income amount or affirmed as a zero income household.

Clearly there are some households that genuinely receive zero income even though that may appear to be implausible. By setting all individuals in such households to missing, we remove these valid observations from our observed data, thus affecting our imputations. These observations only come back into the imputation process at the end of the first round of regression estimations and there is some chance that they are imputed to have positive incomes at this point. In other words, our screen for plausibility has some biases and cannot be seen as deriving an unambiguously superior set of estimates. Rather it should be seen as investigating the sensitivity of our results to outliers and implausible values in the observed data.

We began by recoding problematic values to missing using the following rules:

- If household income was zero, income was set to missing for household members aged 15 and older and to zero for those younger than 15.
- For those younger than 15 with recorded income greater than R6400 per month, income was set to missing.
- For those recorded as being employed but with zero income, we set income to missing.

This gave us a new base data set of individuals in which income was coded as missing or as one of 12 income bands. We then undertook the same multiple imputation process as before on this new base data set in order to transform all missing values into one of the 12 income bands.

Table 3 presents the percentage of households and individuals whose income values warranted closer inspection. The table shows the percentage and number of households reporting zero income, the percentage of employed people earning zero income and the percentage of people aged 15 and under earning over R6400 per month. In all cases, these are national figures derived using sampling weights. These percentages were calculated under four different data assumptions. The first row presents estimates based on a complete case analysis where the missing data were ignored. The second set of estimates use Statistics South Africa's hotdeck imputations. The third row presents the combined estimate of our fifteen imputed data sets of the previous section of the paper. The final row presents the combined estimate for our new set of imputations where implausible values were reset to missing as described above.

For both the hotdeck imputation and our multiple imputation of the previous section, the percentage in each category is very similar to the data set where missing values were ignored. The dependence of the imputed values on the observed data is clear. In the second multiple imputation implausible values were set to zero and the proportion of households or individuals in each category is significantly reduced. While the imputation process allows for some households to be reclassified as earning zero, many of the households previously classified as earning zero income are now imputed to earn some positive income. As stated above, this should be viewed as a lower bound and we would expect the true percentage of zero income household to lie somewhere between 13% and 23%. The results for employed people earning zero income and high income children are similar.

The analysis of high income children shows interesting differences across each category. With no imputations, 0.14% of the population aged 15 and under is captured as such high earners. Earlier in the paper, it was mentioned that one in two fifteen-year olds that reported positive incomes were earning implausibly high amounts. It was mentioned that this situation resulted in the hotdeck imputation having a non-negligible probability of drawing high earning fifteen-year olds in its imputation of missing values. Given this possibility, it is interesting to note that the percentage of high earning children increases to 0.17% (about 9000 children) as a result of the hotdeck imputation. In the first multiple imputation process, about 1000 additional children are imputed to be high-earning. However, this group now represents a smaller percentage of the population aged 15 and under (0.12%). In the second multiple imputation, these high-earning children are set to missing at the start. Given this new base data set, the multiple imputation process gives high earnings to a mere 0.01% of children. This is a lower bound value rather than a clearly more correct value. Nonetheless, the well-being of children is a key issue in South Africa and this demonstration of the sensitivity of estimates of children's income to a high-earning group in the data set is an important cautionary note.

The fourth row in Table 1 presents the combined estimates from fifteen imputations where implausible values were set to missing at the start of the imputation process. The impact of recoding implausible values to missing is seen in an increase in mean per capita income and a decrease in both the Gini coefficient and the percentage of households in poverty. It is interesting to note the increases in mean income given our earlier discussion about the greatly reduced number of high-earning children in the second imputation. Statistically speaking, the estimates from our sensitivity analysis in Table 1 are significantly different. However, perhaps the key point to note is that the implausible values sensitivity analysis does not change the substantive interpretation of our estimates of interest.

4. From bands to point incomes

In both the 1996 and 2001 censuses personal income was collected in bands. In order to estimate measures of poverty and inequality, a continuous measure of income is required. Therefore, a further “imputation” step is required in order to translate the bands into point estimates. In

Sections 2 and 3 of this paper we generated a continuous measure of personal income by allocating to each individual the midpoint income of their income band. The lowest band is zero income and is therefore already a point income. The highest (unbounded) band was assigned the lower bound value for this band.

All in all these rules represent the most common approach to these bands. However, individuals can be assigned incomes within their bands according to other equally plausible intra-band distributional rules. An alternative intra-band allocation rule requires the generation of point incomes by randomly sampling from a specified distribution within each band. While the uniform, normal or lognormal distributions are used, there is generally no dominant theoretical basis for the choice of distribution or the value of its parameters. Indeed there is no theoretical basis for the conventionally adopted mid-point approach. Ideally one would want the distribution to be as close as possible to the true distribution of personal incomes. We therefore decided to use an empirical distribution based on another appropriate data set; the 2000 Income and Expenditure Survey (IES).[11]

Personal incomes in the 2000 IES were first adjusted to 2001 equivalents using a single price inflator.[12] Then, an empirical cumulative distribution was generated for each income band. Random probabilities were generated for each individual in the Census 2001 data and individuals were assigned an income such that the cumulative probability of observing such a value from the empirical distribution was greater than or equal to the generated probability. It is important to note that this does not replicate the full distribution of personal incomes within the IES but rather replicates the intra-band IES distribution. This approach would seem to be particularly useful in imputing point incomes for those in the top band. While decisions about this top band have no implications for poverty, measured inequality is likely to be sensitive to changes at the top of the income distribution.

Theoretically, we cannot predict whether assigning all imputed incomes to the midpoints of bands will generate more or less inequality than distributing the incomes across the bands. While the effect at the top band is clear - assigning everyone to the bottom cutoff of the band must generate less inequality than distributing individuals across the band - the effect in lower bands is indeterminate. While individuals within a band get compressed when midpoints are used, many individuals in adjacent bands will be spread further apart when midpoints are used. The net effect is theoretically ambiguous. Similarly, we cannot predict a priori which of the two approaches will generate the higher poverty headcount. This will depend on the position of the poverty line relative to the midpoint of a band, the movement across that threshold caused by adjustments for household size, and other complex interactions between the poverty line and the income imputations. While the effects of using midpoints versus full distribution across the band would be relatively easy to analyse in the case of individual incomes and an individual poverty line, it is considerably more complicated in the case where we are using aggregate household income adjusted for household size.

To assess the influence of these allocation rules, we use the fifteen data sets that were generated in multiply imputing all missing income values in Section 2 of the paper. The fifth row of Table 1 presents the resultant parameter estimates obtained using the empirical distribution of personal income in the IES within each band on each of the fifteen multiply imputed data sets. The combined parameter estimates and confidence intervals are derived from these fifteen estimates in the same way that they have been for all of the multiple imputations in this paper.

[11] Total regular personal income was used to generate an empirical distribution (Statistics South Africa, 2000 Section 24.1: 44-47). While this excluded household level income, it is not clear that household level income would have been well captured by the 2001 Census.

[12] The percentage chance in CPI between October 2000 and October 2001 was 4% (Statistics South Africa, 2001b: 1).

Comparing the estimates based on mid-points of the income bands with those obtained using the IES empirical distribution it can be seen that the point estimate of mean per capita income is lower using the IES distribution. The falling mean suggests that on average within bands, the distribution of income is skewed to the right with the mean below the mid-point of the band. Estimates of poverty at both poverty lines are slightly higher and the estimated Gini coefficient rises; suggesting an increase in inequality. As noted above, this will in part be attributed to the fact that the IES rule stretches out personal incomes at the top end, with effects in other bands either reinforcing or offsetting the increase in inequality at the top.

Thus, in sum, it does not appear that our results are very sensitive to the assumptions underlying the “imputation” of a continuous variable from the bands. However, as the distribution of income within bands is more likely to follow the empirical distribution of personal income in the IES, we prefer this technique to the mid-point approach.

5. Discussion and conclusion

We began this paper with reference to the literature assessing medium-run changes in poverty and inequality in South Africa using census data. According to this literature, over the 1996 to 2001 period both poverty and inequality increased. We pointed out that this literature has paid considerable attention to issues of comparability over time but has made little attempt to deal with the large percentage of individuals and households for whom income data were missing. We went on to use the sequential regression multiple imputation approach to impute missing values for the 2001 census data. The results suggested that, at the national level, the imputation increased the estimates of mean income and the Gini coefficient measure of inequality and decreased measured poverty. This was true even accounting for the wider confidence intervals that arise from the uncertainty that the imputations bring into the estimation process. There are a number of implausible personal income values in the data set. In the next section, we assessed the influence of these values by setting them to missing and re-doing the multiple imputation process with this augmented set of missing values as a sensitivity analysis. This resulted in an increase in mean income and a decrease in both the poverty headcount and Gini coefficient. While the differences were statistically significant the magnitude of the differences would not alter our substantive interpretation of poverty and inequality in South Africa.

Up to this point in the paper all parameters of interest were calculated by taking personal incomes recorded in bands and attributing the appropriate band midpoint to each individual. The final imputation exercise assessed the sensitivity of results to this practice by using information on intra-band empirical distributions of personal incomes that was available through a national income and expenditure survey and imputing intra-band point incomes in a manner that would replicate these empirical distributions. Assigning midpoints leads to lower estimates of inequality than this alternative approach, although the differences are relatively small.

In order to keep the discussion of the imputation technique manageable, in the main body of the paper we limited all empirical work to the 2001 data. However, the same analyses recorded in this paper for 2001 were undertaken on the 1996 census. Given the starting point of this paper it is appropriate to finish it with a brief comparison of multiply imputed poverty and inequality results for 1996 and 2001. The comparison across the years uses point incomes for each person that are fitted from the intra-band empirical distributions of personal incomes found in the 1995 and 2000 national income and expenditure surveys.

Table 4 presents the resultant comparisons of poverty and inequality in 1996 and 2001. There has been an increase in poverty with around 5% more individuals finding themselves in poverty in 2001 than in 1996 at both poverty lines. Turning to the poverty gap ratios, we also see an increase in the poverty gap ratio indicating that the depth of poverty has also increased over the period. The Gini coefficients in 1996 and 2001 suggest a marked increase in inequality over the period. In all cases these differences are statistically significant. Encouragingly the comparison between 1996 and 2001 is robust to any choice of estimate from Table 1.

Thus, at the end of a lot of careful imputation work on the 10% micro samples of the 1996 and 2001 censuses, our results confirm the major findings from the existing literature in that we find small increases in poverty for the poorest of the poor between 1996 and 2001, more marked increases when a higher poverty line is used and unambiguous increases in inequality. For both 1996 and 2001, our estimates of the poverty and inequality parameters are combined estimates embodying fifteen sequential regression imputations for missing values. As a consequence of this they have higher standard errors, serving as an appropriate signal of the additional uncertainties that imputations bring to parameter estimation.

Acknowledgements

We acknowledge the useful comments that we have received on this paper from a referee of this journal, Servaas van der Berg, Debbie Budlender and the participants in a seminar held in the School of Economics at the University of Cape Town. Support for this research was provided by the U.S. National Institute for Child Health and Human Development and Andrew W. Mellon Foundation. The usual disclaimer applies.

References

Ardington, C.; Lam, D.; Leibbrandt, M.; Welch, M. CSSR Working Paper 05/106. Feb. 2005 The Sensitivity of Estimates of Post-Apartheid Changes in South African Poverty and Inequality to Key Data Imputations.

Cronje M, Budlender D. Comparing census 1996 and census 2001: an operational perspective. South African Journal of Demography 2004;9(1):67–89.

Leibbrandt, M.; Poswell, L.; Naidoo, P.; Welch, M.; Woolard, I.; Centre for Social Science Research. CSSR Working Paper. 84. University of Cape Town; 2004. Measuring recent changes in South African inequality and poverty using 1996 and 2001 census data.

Little, RJ.; Rubin, DB. Statistical Analysis with Missing Data. Wiley; New York: 2000.

Lohr, SL. Sampling: Design and Analysis. Brooks/Cole Publishing Company; Pacific Grove: 1999.

McGrath, MD. University of Natal Durban; 1983. Inequality in the Size Distribution of Personal Income in South Africa in Selected Years Over the Period from 1945 to 1980. Unpublished PhD. Dissertation

Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 2001;27(1):85–95.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. Wiley; New York: 1987.

Simkins, C. University of the Witwatersrand; 2005. What Happened To The Distribution Of Income In South Africa Between 1995 And 2001?. Unpublished paper

Statistics South Africa. Income and Expenditure Survey 2000. Statistics South Africa; 2000.

Statistics South Africa. Census 2001 10% Sample. Statistics South Africa; Pretoria: 2001a.

Statistics South Africa. Consumer price index (CPI). Statistical Release P0141.1. Statistics South Africa; Pretoria: Oct. 2001b 2001

Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 1999;18:694–981.

Whiteford, AC.; McGrath, MD. Distribution of Income in South Africa. Human Sciences Research Council; Pretoria: 1994.

Whiteford AC, van Seventer DE. South Africa's changing income distribution in the 1990s. Journal of Studies in Economics and Econometrics 2000;24(3):7–30.

Table 1

Comparison of poverty and inequality measures 2001

Panel A. Mean per capita income and Gini coefficients

	Mean per capita income			Gini coefficient		
	Estimate	95% C.I.		Estimate	95% C.I.	
No imputed values	910.457	909.041	911.873	0.773	0.772	0.774
Statistics South Africa’s hotdeck imputation	1033.606	1032.325	1034.886	0.819	0.818	0.819
SRMI multiple imputation (mid-points)	1023.630	1020.314	1026.946	0.818	0.818	0.819
SRMI multiple imputation (implausible values set to missing)	1032.742	1030.031	1035.453	0.798	0.797	0.799
SRMI multiple imputation (IES empirical distribution)	986.070	969.865	979.589	0.822	0.821	0.823

Panel B: Poverty headcount ratios

	Poverty headcount ratio (Poverty line at R124 per capita per month)			Poverty headcount ratio (Poverty line at R340 per capita per month)		
	Estimate	95% C.I.		Estimate	95% C.I.	
No imputed values	0.452	0.451	0.453	0.685	0.685	0.686
Statistics South Africa’s hotdeck imputation	0.422	0.421	0.422	0.656	0.656	0.657
SRMI multiple imputation (mid-points)	0.423	0.423	0.424	0.639	0.639	0.640
SRMI multiple imputation (implausible values set to missing)	0.378	0.378	0.379	0.624	0.623	0.625
SRMI multiple imputation (IES empirical distribution)	0.428	0.428	0.429	0.651	0.650	0.652

Source: Census 2001 (authors’ own calculations).

All results were weighted using weights supplied by Statistics South Africa.

Table 2

SRMI multiple imputation estimates of mean per capita income and poverty headcount ratio 2001

	Mean per capita income		
	Estimate	**95% C.I.**	
Combined	1023.630	1020.314	1026.946
Imputation 1	1024.028	1022.746	1025.310
Imputation 2	1023.139	1021.862	1024.416
Imputation 3	1022.912	1021.634	1024.190
Imputation 4	1023.000	1021.722	1024.277
Imputation 5	1025.808	1024.526	1027.090
Imputation 6	1023.556	1022.280	1024.833
Imputation 7	1021.253	1019.981	1022.525
Imputation 8	1023.666	1022.387	1024.945
Imputation 9	1023.822	1022.543	1025.102
Imputation 10	1023.712	1022.433	1024.992
Imputation 11	1021.140	1019.867	1022.412
Imputation 12	1024.350	1023.074	1025.627
Imputation 13	1025.176	1023.892	1026.460
Imputation 14	1023.123	1021.847	1024.400
Imputation 15	1025.764	1024.482	1027.046

Source: Census 2001 (authors' own calculations).

All results were weighted using weights supplied by Statistics South Africa.

Table 3

Households reporting zero income, employed people earning zero income and people aged 15 and under earning in excess of R6400 per month

	Households reporting zero income	Employed people earning zero income	People aged 15 and under with incomes in excess of R6400 per month
No imputed values	2168820 (25.27%)	157834 (1.90%)	16493 (0.14%)
Hotdeck imputation	2541034 (23.18%)	231560 (2.39%)	25510 (0.17%)
SRMI multiple imputation	2535924 (23.14%)	199462 (2.17%)	17445 (0.12%)
SRMI multiple imputation with implausible values set to missing	1476989 (13.48%)	43333 (0.47%)	2104 (0.01%)

Source: Census 2001 (authors' own calculations).

All results were weighted using weights supplied by Statistics South Africa.

Table 4

Comparison of poverty and inequality in 1996 and 2001

	SRMI multiple imputation 1996 (IES 1995 empirical distribution)			SRMI multiple imputation 2001 (IES 2000 empirical distribution)		
	Estimate	95% C.I.		Estimate	95% C.I.	
Poverty headcount (Poverty line: R124 in 2001, R91 in 1996)	0.380	0.379	0.380	0.428	0.428	0.429
Poverty headcount (Poverty line: R340 in 2001, R250 in 1996)	0.598	0.597	0.598	0.651	0.650	0.652
Poverty gap ratio (Poverty line: R124 in 2001, R91 in 1996)	0.269	0.269	0.270	0.306	0.305	0.306
Poverty gap ratio (Poverty line: R340 in 2001, R250 in 1996)	0.418	0.417	0.418	0.468	0.467	0.468
Gini coefficient	0.744	0.743	0.745	0.822	0.821	0.823

Source: Census 1996; Census 2001 (authors’ own calculations).

All results were weighted using weights supplied by Statistics South Africa.