Multiple Imputation of Missing Data in the 2001 South African Census

Helen Barnes
Roxana Gutierrez-Romero
Michael Noble

Working Paper No 4

Centre for the Analysis of South African Social Policy
University of Oxford

Acknowledgments

We would like to thank Prof. Peter Solenberger and Dr John Van Hoewyk for kindly and promptly answering all enquires concerning the use of IVEware. We are very grateful to Cally Ardington, Prof. Murray Leibbrandt, Prof. David Lam, and Matthew Welch for their advice on imputation and for sharing the method and results of their imputation work.

The imputation results obtained in this paper were estimated using the Unix computer of the Oxford Supercomputing Centre (OSC), at the University of Oxford. We would like to thank Dr Jon Lockley and his team at the OSC for all the technical assistance provided.

Multiple Imputation of Missing Data in the 2001 South African Census

Table of Contents

List of Figures

List of Tables

Multiple Imputation of Missing Data in the 2001 South African Census

Abstract

This article describes the imputation of missing data in key variables in the ten percent sample of the 2001 South African Census. Using the sequential multiple regression technique income, education, age, gender, population group, occupation and employment status were imputed. The main focus of the work was to impute income where it was missing or recorded as zero. The imputed results are similar to previous imputation work on the 2001 South African Census, including the single 'hot-deck' imputation carried out by Statistics South Africa.

1. Introduction

One of the main difficulties in producing indicators of well-being from the 2001 South African Census is the large proportion of missing values on a number of key variables such as age, education and income. In the 2001 Census, data on personal income were obtained by asking each person in the household "What is the income category that best describes the gross income of (this person) before tax?" The answer was recorded in income bands for each person. The use of the income data from the Census is problematic as there is a large proportion of missing income data (16% of individuals in the ten percent sample) and a large proportion of incomes reported to be zero (50.1% of individuals aged 18 or over). It is difficult to determine *a priori* whether reported zero incomes are actual observed levels or whether it is due to the reluctance of people to reveal their true income. Non-response or invalid income values can bias well-being indicators if data are not missing completely at random.[1] Imputation is one of a number of recognized methods for dealing with the problem of missing or invalid data.

Before releasing any Census products, Stats SA adjusted for non-response using a logical imputation method and a single 'hot-deck' imputation. Logical imputation replaces missing data using information from other variables available in the dataset. Single hot-deck imputation involves matching as closely as possible individuals with missing data on some variables to individuals who have complete records, and using the information from the latter to replace the missing values in the former. This procedure is particularly suitable when data are missing at random (MAR)[2] and when the number of outliers is small. When these conditions are not met, individuals with missing data may be

[1] Missing completely at random means that the probability of an item being missing is unrelated to any observed or unobserved characteristic for that unit.

[2] Missing at random means that the probability of an item being missing depends only on other items that have been observed for that unit and no additional information as to the probability of being missing would be obtained from the unobserved values of the missing items.

inappropriately matched to individuals with complete records who are outliers or for whom the information recorded is incorrect.

In the 2001 Census there is strong evidence which suggests that the missing data fulfils the MAR assumption. For instance, White people are more likely to have missing income data than Black Africans. People living in urban areas are more likely to have missing income data than those living in rural areas (see Appendix). Although the pattern of missing data fulfils the MAR assumption required by the hot-deck imputation method (and in general by other imputation methods), there is no way to assess a priori the reliability of the data imputed; that is whether there is a large proportion of outliers in the data that could potentially bias the imputed values. Furthermore, single hot-deck imputations do not provide a measure of the variance introduced with the imputation process.

The total variance introduced by imputing values can be estimated by repeating the imputation process a number of times, and thus the possible bias caused by the outliers in the data can be minimised. This technique is known as multiple imputation. To assess the reliability of the reported zero income, zero values can be set to missing and the values imputed, thereby checking whether the original value of zero was accurate.

The method described in this article utilised the sequential regression multiple imputation (SRMI) technique developed by Raghunathan et al. (2001) to impute the missing values. The software used in the imputation was IVEware, which was developed exclusively to perform SRMI imputations. The imputation work was carried out to test whether Stats SA's hot deck imputations, particularly for the income variable, are reliable enough to be used for a low income measure in the Income and Material Deprivation Domain of the Provincial Indices of Multiple Deprivation for South Africa 2001 (PIMD 2001) (Noble et al., 2006, forthcoming).[3]

The rest of the paper is organised as follows. In Section 2 statistics for the missing variables and implausible values are presented. In Section 3 the imputation method used is discussed. In Section 4 the imputation results are presented and compared with those produced by Ardington et al. (2005) and Stats SA. Ardington et al. (2005) also employed the SRMI imputation method, but used different types of regression in the sequence of imputation and different software than in this study. It was found that the imputation results from this work are similar to both sets of imputations. Finally, in Section 5 some conclusions are presented.

[3] The PIMD 2001 is a relative measure of multiple deprivation created at ward level for each province in South Africa using the 2001 Census. The conceptual model is based on the idea of distinct domains of deprivation which can be recognised and measured separately. These are experienced by individuals living in an area. People may be counted as deprived in one or more of the domains, depending on the number of types of deprivation that they experience. The overall PIMD is conceptualised as a weighted area level aggregation of these specific domains of deprivation. One of the domains of the PIMD 2001 is an Income and Material Deprivation Domain which includes a measure of the number of people living in a household that has a household equivalent income below 40% of the mean equivalent household income. This indicator makes use of the income variable in the 2001 Census.

2. Extent of missing data and implausible values

The imputation analysis on the income variable in the Census for the PIMD 2001 was performed on both original missing incomes and ‘implausible’ zero incomes. The implausible zero incomes were defined according to rules devised by Ardington et al. (2005):

1. 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.
2. For those younger than 15 with recorded income greater than R6 400 per month, income was set to missing.
3. For those recorded as being employed but with zero income, income was set to missing.

Table 1 below, describes the frequency of the implausible cases that were set to missing as well as the proportion of individuals with missing income. In total, 27.1% of cases in the dataset were either missing originally, or were set to missing for the imputation.

Table 1. Missing income and implausible income values set to missing[4]

Case	Frequency (%)
Original missing income	15.59[5]
Individual older than 15 in household with zero income	12.45[6]
Individual younger than 15 with income greater than R6 400 per month	0.04
Individual who is employed with zero income	0.35

Implausible values set to missing

3. Imputation method

There are various ways of dealing with missing data. Davern et al. (2001) describe four of the main methods used:

1. Analyse complete cases only;
2. Only use cases with reported data on the problematic item;
3. Weight complete cases to make up for missing cases;
4. Impute missing data values.

[4] 7.65% of cases were younger than 15 years and in a household with zero income. According to the rules, hese cases were set to zero income rather than missing.

[5] 1.35% of cases were changed from missing to zero by rule 1.

[6] 0.53% of these cases were already missing.

Imputation is generally preferred when (a) there is a substantial proportion of non-response or missing data (more than 10%); (b) imputation can correct for potential distributional differences between respondents with missing data and those with reported data; and (c) it is possible to maintain relationships among associated variables. Given that the proportion of missing data is large in the 2001 Census, the preferred solution to the problem is to impute the missing values. There are various methods for imputing data and the best imputation method depends on the type of missing data. The main imputation methods are (Steen Larsen and Madsen, 2000):

1. **Deductive**: This method is used when there is only one possible response to the question, for example, when all the values are given but the total or subtotal is missing.
2. **Substitution**: This method relies on the availability of comparable data. Imputed data are extracted from the respondent's record from a previous wave of the survey.
3. **Group mean imputation**: This method uses information from answers to other questions. A missing value is replaced with the average value from the responding units with the same set of predetermined characteristics.
4. **Predicted mean imputation**: This method uses information from answers to other questions. Imputed values are predicted using an ordinary least-squares (OLS) multiple regression. The OLS regression is used when the variables to be imputed are continuous or ordinal.
5. **Last value carried forward**: This method uses the last observed value of a longitudinal variable and is only used in longitudinal surveys.
6. **Cold deck**: This method replaces missing data with a fixed set of values from historical data.
7. **Hot deck**: This method uses data from other records (donors). First, a list of possible donors with the same set of predetermined characteristics is created. Then one of the donors is randomly selected. The donor's response replaces the missing data. Donors can be found with matching criteria or a method called the nearest neighbour imputation. A problem with hot deck occurs if many auxiliary variables are available because it is impossible to find a donor which matches with all the variables.
8. **Predictive model based method (multiple imputation)**: This method is similar to the predicted mean imputation. However, each parameter is randomly drawn from the posterior distribution. This method is useful when there are many auxiliary variables.
9. **Propensity score method (multiple imputation)**: The propensity score is the estimated probability that a particular element of data is missing. The missing data are filled in by sampling from the cases that have a similar propensity score. This is done with bootstrap procedures.
10. **Sequential regression multiple imputation (multiple imputation)**: This method computes the maximum likelihood of the covariance matrix and mean vector when data are missing and imputes the data as requested. This method imputes missing data (one or more variables) conditional on a set of predictor variables with no missing values.

The imputation method used in this study is the sequential regression multiple imputation. There are three main reasons why SRMI is preferred to other multiple imputation methods. First, SRMI is a multiple imputation technique which allows estimation of the variance introduced in the imputations. Second, SRMI can handle very complex data structures (e.g. count, binary, continuous and categorical variables) that other imputation methods find problematic. Third, given that SRMI imputes values through a sequence of multiple regressions, covariates include all other variables observed and imputed from previous rounds. This sequence of imputing missing values builds interdependence among imputed values and exploits the correlation structure among covariates (Raghunathan et al., 2001).

The aim of this study was not only to impute missing income, but also to impute the variables that could be used as explanatory variables of income such as age, gender, years of education, population group, province, etc. In this case the SRMI method divides the dataset into two matrices, one composed of all the variables that contain missing data, denoted by Y and another composed of variables that have no missing data, denoted by X.

All variables included in the matrix Y are ordered from least to most missing values, i.e. Y_1, Y_2, Y_3, Y_4, Y_n and all Y variables are considered to be dependent variables to which missing values are imputed by maximizing the joint conditional density of matrix Y given matrix X. Each of the variables included in matrix Y can have a different type of distribution, such as dichotomous (e.g. gender), categorical (e.g. banded income) or continuous (e.g. age). These dependent variables are estimated using the type of regression model that most suits them. For instance, dichotomous variables are estimated using logit regression models, continuous variables are estimated using OLS regression models and so on.

In this method the imputation is performed in a series of steps and rounds. The first step is to estimate the missing values of Y_1 given the variables in the X matrix. This step consists of up to 250 maximum likelihood iterations which are needed for maximizing the joint conditional density of Y_1 given matrix X. Then, this is followed by an estimate of Y_2 given X and the newly derived $\hat{Y}_1$ which contains both observed and imputed values (in other words $\hat{Y}_1$ is included in the X matrix. The first round of imputations, $\hat{Y}^{R=1}$, is completed once each of the variables included in the Y matrix is estimated, as in the steps explained above. This first round of imputation contains no missing values and is equivalent to the single vector of hot deck imputations derived by Stats SA.

In the second round Y_1 is re-estimated including all first round $\hat{Y}^{R=1}$ 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. The new imputed values are conditional on the previously imputed values of the preceding imputation round. Figure 1 shows the variables that were imputed and the regression types that were used.

Figure 1. Variables with missing data to be imputed

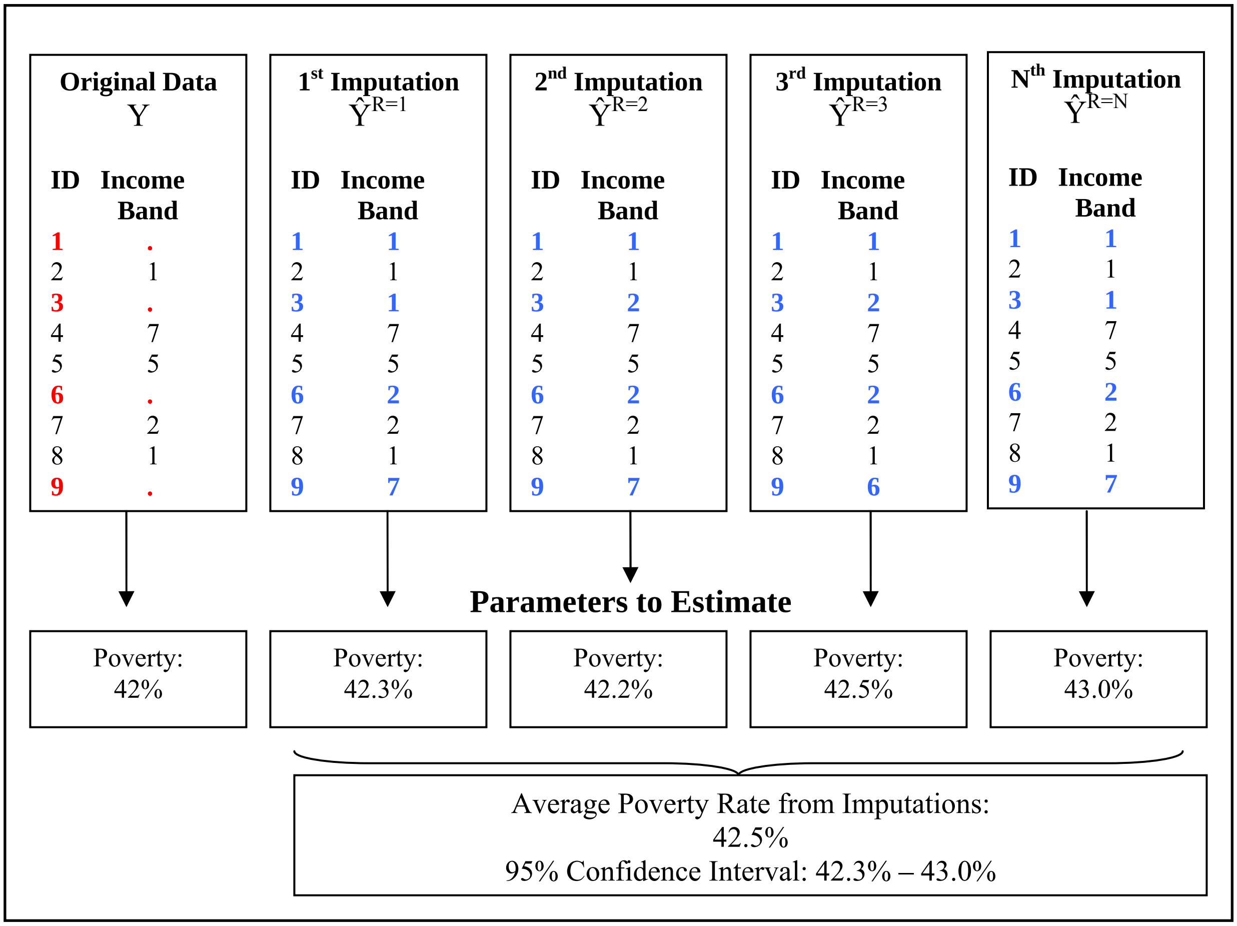
Y Matrix*	Regression Type	X Matrix
Age	OLS	Province
Gender	Logit	Location
Population group	Logit	
Employment status	Logit	
Occupation	Logit	
Education	OLS	
Income	Logit	

* Variables ordered from least to most number of missing cases.

3.1. Combining imputed datasets

Each round of imputations produces individual values of matrix $\hat{Y}$, which contain no missing values and from which it is then possible to carry out further analysis such as measuring poverty or inequality. The imputed values for each variable may differ from round to round, and hence the estimated level of poverty, for example, may also differ from one round to another. The uncertainty about the actual level of poverty, say, is overcome by obtaining an estimate of poverty from every round, and then averaging over these estimates. The standard error of the resulting estimate of (average) poverty is obtained using the multiple imputation rule developed by Little and Rubin (2000). This rule applies to the derived measures only, such as poverty and inequality, and not to the imputed values that were obtained (see Figure 2).

Figure 2. Example: Combining imputed datasets



Original Data Y		1st Imputation $\hat{Y}^{R=1}$		2nd Imputation $\hat{Y}^{R=2}$		3rd Imputation $\hat{Y}^{R=3}$		Nth Imputation $\hat{Y}^{R=N}$	
ID	**Income Band**	**ID**	**Income Band**	**ID**	**Income Band**	**ID**	**Income Band**	**ID**	**Income Band**
1	**.**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
2	1	2	1	2	1	2	1	2	1
3	**.**	**3**	**1**	**3**	**2**	**3**	**2**	**3**	**1**
4	7	4	7	4	7	4	7	4	7
5	5	5	5	5	5	5	5	5	5
6	**.**	**6**	**2**	**6**	**2**	**6**	**2**	**6**	**2**
7	2	7	2	7	2	7	2	7	2
8	1	8	1	8	1	8	1	8	1
9	**.**	**9**	**7**	**9**	**7**	**9**	**6**	**9**	**7**

Parameters to Estimate

Poverty: 42%	Poverty: 42.3%	Poverty: 42.2%	Poverty: 42.5%	Poverty: 43.0%

Average Poverty Rate from Imputations:
42.5%
95% Confidence Interval: 42.3% – 43.0%

Note: The data and the poverty rates shown are only for explanatory purposes.

4. Imputation results

The imputation results produced in this work are very similar to those of Stats SA and of Ardington et al. (2005). Ardington et al. also employed the SRMI imputation method, but used different types of regression in the sequence of imputation and also different software.

The tables below compare the proportion of cases in each income band obtained by the different imputation methods (and the last column in the first table shows the proportion in each income band when no imputation or recoding of implausible zero incomes had taken place). The first table shows the distribution of income for all cases (i.e. imputed and not imputed cases), while the second table shows the distribution for the imputed cases only (i.e. the missing and implausible zero cases).

Each method assigns roughly the same proportion of people overall to the first three income bands (84 to 85%). For the imputed cases only the proportion of people in the first three income bands ranges from 90 to 98%. In terms of the zero incomes, for each method between 65 and 69% of people overall either reported having no income or were placed in the zero income band by the imputation process. For the imputed cases only the proportion of people assigned to the zero income band is very similar for all three methods (79 to 81%).

The similarity between the results of Ardington et al. and this work is reassuring, but more surprising is the fact that the imputation results are also similar to those obtained using the single hot-deck imputation method. This is likely to reflect the fact that the number of outliers among the observations on the variables used for imputation was not large. Therefore, the large number of zero incomes reported are indeed reasonably accurate and do not represent outliers. Given the similarity in the results it was felt it would be acceptable to produce the Income and Material Deprivation Domain using Stats SA's hot deck imputations, rather than run the sequential regression multiple imputation on the full Census.

Table 2. Observed and imputed cases

	Sequential Regression Multiple Imputation				**Hot deck imputation**		**No imputation**	
	Ardington et al. imputation 15		**CASASP imputation 10**		**Stats SA imputation 1**			
Income band	**%**	**Cum %**	**%**	**Cum %**	**%**	**Cum %**	**%**	**Cum %**
1	65.42	65.42	66.66	66.66	69.34	69.34	67.20	67.20
2	7.28	72.70	6.70	73.36	5.33	74.68	5.57	72.78
3	11.22	83.92	11.94	85.30	9.67	84.35	10.42	83.19
4	5.46	89.38	5.04	90.34	5.14	89.49	5.60	88.80
5	4.52	93.90	4.18	94.52	4.43	93.93	4.77	93.57
6	3.31	97.20	3.01	97.53	3.28	97.21	3.49	97.06
7	1.74	98.94	1.55	99.08	1.73	98.94	1.83	98.89
8	0.66	99.60	0.57	99.65	0.65	99.59	0.68	99.57
9	0.23	99.83	0.20	99.85	0.22	99.81	0.24	99.81
10	0.09	99.92	0.08	99.93	0.09	99.90	0.09	99.90
11	0.06	99.97	0.06	99.99	0.07	99.97	0.07	99.97
12	0.03	100.00	0.02	100.00	0.03	100.00	0.03	100.00

Table 3. Imputed cases only

Income band	Ardington et al. imputation 15		CASASP imputation 10		Stats SA imputation 1	
	%	Cum %	%	Cum %	%	Cum %
1	80.12	80.12	79.00	79.00	80.95	80.95
2	7.40	87.51	7.35	86.35	4.03	84.98
3	6.96	94.47	11.61	97.96	5.62	90.61
4	2.11	96.57	1.15	99.11	2.67	93.27
5	1.40	97.97	0.57	99.68	2.59	95.86
6	1.02	98.99	0.22	99.90	2.12	97.98
7	0.59	99.58	0.05	99.95	1.21	99.19
8	0.25	99.83	0.01	99.96	0.47	99.66
9	0.09	99.92	0.00	99.96	0.15	99.81
10	0.04	99.96	0.00	99.96	0.07	99.88
11	0.03	99.99	0.03	99.99	0.09	99.98
12	0.01	100.00	0.00	100.00	0.02	100.00

4.1. Differences with previous work

Besides the type of software used to run the sequential regression multiple imputation, there are a few other differences between this work and the work of Ardington et al. (2005). In contrast to Ardington et al., a Poisson regression model was not used in this study to estimate years of education and age. The reason for not using a Poisson regression model is that it assumes a process in which a rare event occurs with a constant probability for each subject, independent of how many times the event has occurred to the subject in the past. Years of education and age probably do not fit this model, and thus we decided to impute these using OLS-type regression models.

A second difference is that Ardington et al. (2005) estimated banded income using an ordered logit regression model, whilst only a logit regression was used in this study, since IVEware cannot estimate an ordered logit regression model. Because age, gender, population group, employment status, occupation, education, province and location were controlled for, there are only small differences in the results obtained using an ordered logit regression compared to a logit regression. Therefore the difference in type of regression used is not considered to be an issue.

The final difference is that Ardington et al. chose to use Stats SA's imputations for the population group variable, rather than impute the variable as part of the process of imputing income. This was partly because the majority of the imputations were logical (in most instances individuals were assigned the population group of other people in the household), but mainly because they were limited by the computational capacity of Stata and their computers: "imputing missing values for race would require the fitting of 25 multinomial logit models, each with 26 independent variables… for each imputation. With a dataset as large as the 10 percent micro-sample, the computational requirements are very demanding" (Ardington *et al*., 2005: 9). This analysis was carried out using IVEware, which runs independently of statistical software packages, on a powerful UNIX system, which allowed population group to be imputed alongside the other variables.

5. Conclusion

In the analysis for this paper key missing variables and implausible income values in the ten percent sample of the 2001 South African Census were imputed. The sequential regression multiple imputation method was used to impute income, education, age, gender, population group, occupation and employment status

The imputation was performed on both original missing incomes and implausible zero incomes (together comprising 27.1% of individuals in the ten percent sample of the 2001 Census). The imputations results are very similar to those of Ardington et al. (2005) and Stats SA.

Although SRMI and the single hot deck imputation show similar results, the SRMI results also allow one to compute the standard errors of the resulting measures, by distinguishing observed data from imputed data and measuring the potential error introduced by estimating missing variables. This standard error cannot be obtained from single hot deck imputations. Furthermore, SRMI minimises any possible bias in the imputation results that might arise as a consequence of outliers in the data. SRMI produces more than a single set of imputation results, so it is possible to derive a distribution of imputed values and assess the variability of the imputed values across the rounds of imputation.

Although the imputation was not used in the final PIMD 2001, the imputed values have subsequently been used to estimate poverty and a series of inequality and polarisation measures (Gutierrez-Romero, Barnes and Noble, 2006, forthcoming), as well as to estimate eligibility for the Child Support Grant for the South African national government Department of Social Development (Noble et al., 2005a; Noble at al., 2005b).

References

Ardington, C., Lam, D., Leibbrandt, M. and Welch, M. (2005) ‘The Sensitivity of Estimates of Post-Apartheid Changes in South African Poverty and Inequality to Key Data Imputations’, Working Paper No. 106, Centre for Social Science Research, University of Cape Town.

Davern, M., Blewett, L., Bershadsky, B. and Arnold, N. (2001) ‘Possible Bias in the Census Bureau’s State Income and Health Insurance Estimates’, Working Paper, University of Minnesota.

Gutierrez-Romero, R., Barnes, H. and Noble, M. (2006, forthcoming) ‘Poverty and Polarisation in South Africa in 2001’, Working Paper, Centre for the Analysis of South African Social Policy, University of Oxford.

Little, R.J. and Rubin, D.B. (2000) *Statistical Analysis with Missing Data*, New York: Wiley.

Noble, M., Babita, M., Barnes, H., Dibben, C., Magasela, W., Noble, S., Ntshongwana, P., Phillips, H., Rama, S., Roberts, B., Wright, G. and Zungu, S. (2006, forthcoming) *The Provincial Indices of Multiple Deprivation for South Africa 2001*, University of Oxford, UK.

Noble, M., Wright, G., Barnes, H., Noble, S., Ntshongwana, P., Gutierrez-Romero, R., McLennan, D. and Avenell, D. (2005a) *The Child Support Grant: A Sub-Provincial Analysis of Eligibility and Take Up in January 2004*, National Department of Social Development, South Africa.

Noble, M., Wright, G., Barnes, H., Noble, S., Ntshongwana, P., Gutierrez-Romero, R. and Avenell, D. (2005b) *The Child Support Grant: A Sub-Provincial Analysis of Eligibility and Take Up in January 2005*, National Department of Social Development, South Africa.

Raghunathan, T., Lepkowski, J., van Hoewyk, J. and Solenberger, P. (2001) ‘A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, 27(1), pp. 85-95.

Steen Larsen, B. and Madsen, B. (2000) ‘Evaluation of Solas 2.0 for Imputing Missing Values’, Working Paper No.22, Statistical Commission and Economic Commission for Europe.

Appendix: Rates of missing data in 2001 South African Census

Table (A.1) Rates of missing data in variables used in imputation (not weighted) before recoding of implausible zero incomes

	Number of individuals with missing data	**Percentage of sample with missing data**
Age	25976	0.72
Gender	45358	1.26
Population group	49719	1.38
Employment status	196918	5.47
Occupation	218855	6.08
Education	236578	6.57
Income	561095	15.59
Income (after recoding of implausible values)	975016	27.08

Source: Own estimates using ten percent sample of 2001 Census (people living in institutions not included in analysis).

Table (A.2) Rates of missing income data by variables used in imputation (not weighted) before recoding of implausible zero incomes

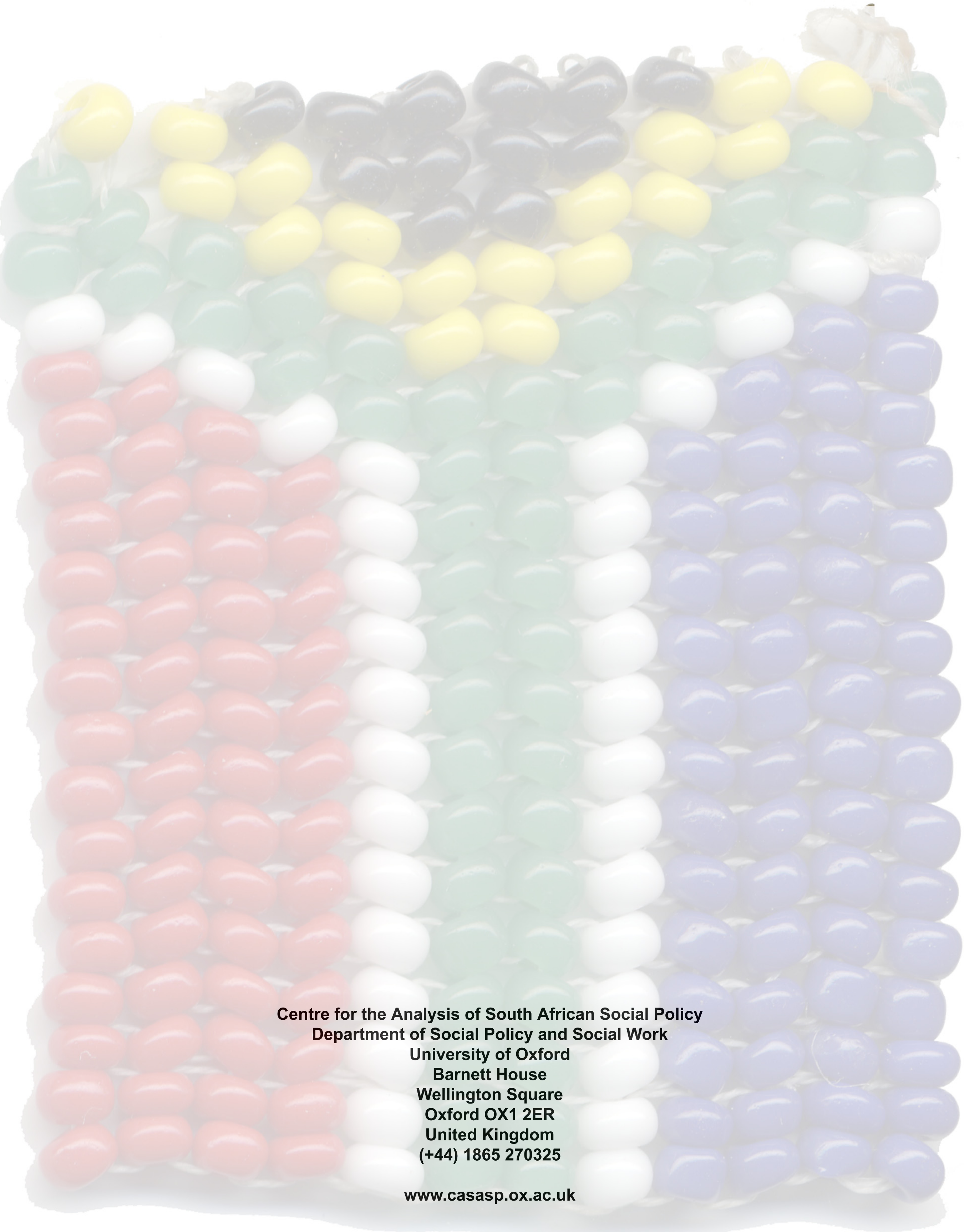
	Number of individuals with missing income data	**Percentage of category with missing income data**
Age		
Under 20	328306	20.71
20-29	88832	14.20
30-39	51773	10.45
40-49	37122	10.02
50-59	24762	10.81
60-69	13737	8.98
70-79	6711	8.22
80 or over	2777	8.21
Gender		
Male	252492	15.13
Female	296450	15.72
Population group		
Black African	390308	13.82
Coloured	66465	20.47
Indian/Asian	15295	16.90
White	72609	23.33
Employment status		
Not employed	438229	16.23
Employed	36959	5.26

Continuation of Table (A.2) Rates of missing income data variables used in imputation (not weighted) before recoding of implausible zero incomes

	Number of individuals with missing income data	**Percentage of category with missing income data**
Occupation*		
Legislators, senior officials, managers and professionals	6125	6.61
Elementary occupations	5471	2.91
Other occupations	19662	4.92
Education		
0 years	138255	16.95
1-6 years	136131	15.37
7-13 years	210884	13.73
14 or more years	12159	9.62
Location		
Urban	356030	17.47
Rural	205065	13.13
Province		
Western Cape	86006	23.38
Eastern Cape	99473	18.39
Northern Cape	8675	12.76
Free State	32968	15.11
KwaZulu-Natal	111707	15.61
North West	20761	6.94
Gauteng	131930	19.01
Mpumalanga	27492	10.77
Limpopo	42083	9.53

Source: Own estimates using ten percent sample of 2001 Census (people living in institutions not included in analysis).

*Employed people only.



Centre for the Analysis of South African Social Policy
Department of Social Policy and Social Work
University of Oxford
Barnett House
Wellington Square
Oxford OX1 2ER
United Kingdom
(+44) 1865 270325

www.casasp.ox.ac.uk