



The Centre for the Analysis of  
South African Social Policy



## **Guidance on using the Census variables imputed by CASASP**

### **General information**

This document sets out the information needed to use the data that has been imputed using Sequential Regression Multiple Imputation (with IVEware) by the Centre for the Analysis of South African Social Policy (CASASP) at the University of Oxford, UK.

There are ten datasets, one for each imputation (casasp\_imputation1, casasp\_imputation2 etc). These contain observations for every individual for each of the ten imputations separately.

In order to process and analyse the data it is necessary to use software that can handle large datasets (approximately 3.6 million cases in each dataset). SAS and Stata have both been used successfully. The datasets are provided in Stata format but can be transferred to other programs.

It is recommended that the paper ‘Multiple Imputation of Missing Data in the 2001 South African Census’ by Barnes, Gutierrez-Romero and Noble is read prior to undertaking any analysis using the imputed data.

### **The datasets**

The datasets contain only the variables used in the sequential regression multiple imputation plus serial number and person number variables, a variable showing the imputation number (this will be the same for all observations in the separate imputation files) and a variable uniquely identifying the individual for each of the 10 imputations. People in institutions were excluded from the analysis and therefore are not included in the datasets. The datasets can be merged onto the persons, geography and household files of the 10% Census using either the serial number and person number (for the persons file) or just the serial number (for the geography and household files).

Further details on the 2001 Census and the 10% sample can be obtained from the Statistics South Africa website: [www.statssa.gov.za](http://www.statssa.gov.za)

## Metadata

The same variables are included in all datasets. Full details of how some of the variables were created (i.e. how the original Census variables were recoded to give new categories) are given in the paper cited above. The name of the variable as it appears in the dataset is given in parentheses below. The imputed variables can be identified by the suffix *\_imp* in the variable name.

### **Serial number (*serial\_number*)**

100000424 - 820014292

### **Person number (*person\_number*)**

1 - 49

### **Age (*age\_imp*)<sup>1</sup>**

0 - 180 in single years

### **Sex (*sex\_imp*)**

0 = Male

1 = Female

### **Population group (*pop\_group\_imp*)**

1 = Black African

2 = Coloured

3 = Indian/Asian

4 = White

### **Employment status (*emp\_status\_imp*)**

0 = Unemployed

1 = Employed

### **Years of education (*edu\_years\_imp*)**

0 - 18 in single years

### **Income (*income\_imp*)**

1 = no income (no income per annum)

2 = R 1 - R 400 (R 1 - R 4 800 per annum)

3 = R 401 - R 800 (R 4 801 - R 9 600 per annum)

4 = R 801 - R 1 600 (R 9 601 - R 19 200 per annum)

5 = R 1 601 - R 3 200 (R 19 201 - R 38 400 per annum)

6 = R 3 201 - R 6 400 (R 38 401 - R 76 800 per annum)

7 = R 6 401 - R 12 800 (R 76 801 - R 153 600 per annum)

8 = R 12 801 - R 25 600 (R 153 601 - R 307 200 per annum)

9 = R 25 601 - R 51 200 (R 307 201 - R 614 400 per annum)

10 = R 51 201 - R 102 400 (R 614 401 - R 1 228 800 per annum)

11 = R 102 401 - R 204 800 (R 1 228 801 - R 2 457 600 per annum)

12 = R 204 801 or more (R 2 457 601 or more per annum)

---

<sup>1</sup> As described in the full paper, age was imputed as a continuous rather than categorical variable and so the values for the imputed cases have a number of decimal places. The age variable can either be rounded or used in its current unrounded form.

**Occupation (*occup\_imp*)**

- 1 = Legislators, senior officials, managers and professionals
- 2 = Elementary occupations
- 3 = Other occupations
- 9 = No occupation

**Province (*province*)**

- 1 = Western Cape
- 2 = Eastern Cape
- 3 = Northern Cape
- 4 = Free State
- 5 = KwaZulu-Natal
- 6 = North West
- 7 = Gauteng
- 8 = Mpumalanga
- 9 = Limpopo

**Urban/rural (*rural*)**

- 0 = Urban
- 1 = Rural

**Imputation number (*imp\_number*)**

- 1 - 10

**Person number 2 (*unique\_person\_number*)**

- 1 - 3599972

**Contact details**

Please do not hesitate to contact CASASP for additional information or queries:  
[casasp@socres.ox.ac.uk](mailto:casasp@socres.ox.ac.uk)

**Disclaimer**

The University of Oxford has taken care to ensure that the information in this document and the accompanying data and paper are correct. However, no warranty, express or implied, is given as to its accuracy and the University of Oxford does not accept any liability for error or omission. The University of Oxford is not responsible for how the information is used, how it is interpreted or what reliance is placed on it. We do not guarantee that the information in this document, the data files or report is fit for any particular purpose. We do not accept responsibility for any alteration or manipulation of our data once it has been released.