DataFirst Technical Papers

October Household Survey 1994

by
Martin Wittenberg

Technical Paper Series
Number 9

About the Author(s) and Acknowledgments

Recommended citation

Wittenberg, M. (2008). October Household Survey 1994. A DataFirst Technical Paper Number 9. Cape Town: DataFirst, University of Cape Town

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

October Household Survey 1994

From DataFirst

Article prepared by: Martin Wittenberg (School of Economics and SALDRU, University of Cape Town January 2008)

Contents

Overview

The 1994 October Household Survey, like the other October Household Surveys, provides information on services supplied to the household, some socio-demographic information (including births and deaths) and a fair amount of labour force information. What distinguishes it, is that it is the first nationally representative household survey to be released by Statistics South Africa. The 1993 OHS did not cover the previous homelands. Unlike the later OHSs, however, the 1994 October Household Survey has a much more disproportional sampling scheme. The racial breakdowns of the 1994 and 1995 OHS are shown in Table 1. Because of this it is really important to use the weights released with the 1994 OHS or some alternative weights.

Table 1 Racial breakdown of the 1994 and 1995 OHS samples

	1994 OHS		**1995 OHS**	
	Number	**Percentage**	**Number**	**Percentage**
Africans	77,073	58.18	92,883	71.02
Coloured	24,412	18.43	17,456	13.35
Indian	10,404	7.85	4,485	3.43

White	20,580	15.54	15,963	12.21
Total	132,469	100.00	130,787	100.00

The 1994 OHS also shows some other differences with the later OHSs. The sampling scheme called for 1000 clusters with around 30 households per cluster. This is much greater level of clustering than in the 1995 OHS where there were 3000 clusters with around 10 households per cluster. Although the sample sizes are similar, the standard errors of the 1994 OHS are likely to be larger. Furthermore, it turns out that it is not straightforward to derive indicators for the cluster variables (see below). Consequently the 1994 OHS is a bit more difficult to work with.

In fact the 1994 OHS has not been used all that much in academic analyses. Indeed most analyses of post-apartheid socio-economic trends choose to base themselves on the 1995 OHS. The reason for this is not only that the 1994 OHS differs in design from the later ones, but that the 1995 OHS was linked to the 1995 Income and Expenditure Survey, which means that there are many more interesting analyses that can be run on that survey. Nevertheless the relative neglect of the 1994 OHS seems somewhat strange, given that it provides the first snapshot of the post-apartheid era. To our knowledge there are no good reasons not to work with this survey, provided that due attention is paid to the known problems listed below.

Data releases, files and documentation

The 1994 OHS has been released in at least two different ways and the two releases differ in some non-trivial respects.

Files and documentation released directly by Statistics South Africa

The files that were originally released by Statistics South Africa include the data in ASCII format:

HOUSE.RSA date stamped 1 July 1996, includes information from Section 1

PERSON.RSA date stamped 25 October 1995, information from Section 2

WORK.RSA date stamped 25 October 1995, information from Section 3

DEATHS.RSA date stamped 25 October 1995, information from Section 4

BIRTHS.RSA date stamped 25 October 1995, information from Section 5

Besides these ASCII data files there are another two files that describe the data:

CODELIST. date stamped 1 July 1996 that has codes for occupations, districts, countries of birth, provinces, cause of death and industries.

LAYOUT. date stamped 1 July 1996, which provides the key to which variables are located in which columns of the ASCII files.

Unlike with later Statistics South Africa releases, there is no “Metadata”, i.e. no direct information on the sampling scheme and the calculation of weights. Furthermore Statistics South Africa did not release an electronic copy of the questionnaire with the data either. A hard copy was supplied.

Files and documentation released by SADA

Between 2000 and December 2006 the main distributor of the 1994 OHS was the South African Data Archive (SADA). The data was distributed as SADA study number 61. The data was withdrawn (and has as of December 2007 not been made available again) after it was pointed out (Wittenberg 2006) that the electronic copy of the questionnaire that was distributed with the data set did not correspond to the hard copy.

The SADA release consisted of the following:

ASCII format copies of the raw Stats SA data:

DA0036M.P1 date stamped 13 June 1995, includes information from Section 1

DA0036M.P2 date stamped 14 June 1995, information from Section 2

DA0036M.P3 date stamped 13 June 1995, information from Section 3

DA0036M.P4 date stamped 14 June 1995, information from Section 4

DA0036M.P5 date stamped 13 June 1995, information from Section 5

SPSS versions of the Stats SA data, all date stamped 20 April 1999

HOUSE94.SAV contains information from Section 1

PERSON94.SAV contains information from Section 2

WORK94.SAV contains information from Section 3

DEATH94.SAV contains information from Section 4

BIRTH94.SAV contains information from Section 5

Stata versions of the Stats SA data, all date stamped 15 March 2004 and presumably available only after that date:

house94.dta contains information from Section 1

person94.dta contains information from Section 2

work94.dta contains information from Section 3

death94.dta contains information from Section 4

birth94.dta contains information from Section 5

Several documentation files were released with the SADA data set:

README61.DOC date stamped 1 June 2000.

These are the original instructions for unzipping the SADA data sets from the files supplied on diskettes. Interestingly enough this file suggests that the ASCII files should have been named DA0061M.P* instead of the names actually contained in the data set.

RL0061M.DOC date stamped 1 June 2000

This file contains the record layout and codelist files as supplied by Statistics South Africa.

S0061 .DOC date stamped 8 November 2000 (presumably the release data)

S0061.PDF These files contain a brief study description and an electronic copy of the questionnaire.

SD0061.DOC date stamped 8 November 2000

This file contains the study description (also released in S0061.doc). This provides the only information on the sample design that is publicly available.

READ_94.SPS This is an SPSS syntax file. It reads in the raw ASCII data and coverts it to SPSS format. It reads the appropriate columns into the relevant variables and labels both the variable and in a number of cases creates value labels for the categories.

Differences between the Statistics SA and SADA raw data

The “raw” ASCII data supplied directly by Statistics South Africa and those supplied via SADA agree on substantial matters (i.e. the content of the data variables) but differ in a number of other, interesting ways:

- The StatsSA data sets have a number of additional variables not included in the SADA data sets. These always occur right at the end of the relevant files:
-
    - In the “House” data set, there are variables labelled “Province”, “Stratum”, “Race”, “Type of enumeration area” and “Population of magisterial district”.
    - In the “Person” , “Work”, “Death” and “Birth” data sets it is the variable “district”
- The information contained in the variables “Unique Number”, “Province”, “Urban”, “Population Group” at the beginning in the HOUSE.RSA file differs from that in the DA0036M.P1 file. In fact:
-
    - The “Unique Number” variable in the HOUSE.RSA file is not unique. It looks like a spatial identifier. The first three digits of it give the magisterial district.
    - The information in the “Province”, “Urban” and “Population Group” variables in the HOUSE.RSA variables does not look in the slightest as one might expect it. For instance there are ten values to the “province” variable, five for the “urban” variable and ten for the “Population Group” one. By contrast the additional “Province”, “Stratum” and “Race” variables at the end of the HOUSE.RSA file do correspond to the province, urban and “population group” variables of the DA0036M.P1 file.
    - The combination of the first two variables in the HOUSE.RSA does, however, seem to provide the unique PSU identifiers.
- The “created” variables “wk_old15”, “expanded”, “strict”, “not_econ” of the WORK.RSA file differ from the corresponding ones in the DA0036M.P3 file. In the latter the variables are true indicator variables. In the former, the variable is set to “missing” where the latter has a zero. This obviously does not affect the substantive information content of the variable.

In short the HOUSE.RSA file (perhaps unintentionally) gives more information on the geographic areas and clusters of the sample than the corresponding SADA data set.

Difference between the SADA raw data and the SADA Stata and SPSS data sets

In the process of reading in the ASCII data into the original SPSS files some variables in the "Work" file became corrupted. Only the first column of the variable was read in. The variables affected are:

nat_food, nat_oth, inc_dedc, occ_empl, ind_empl, emp_code, expenses, tot_unpd, unpd_u15, tot_paid, paid_u15, exp_sal, famwork, absence, usualjob, occ_unem, inc_exp, prev_occ, whynowk7, age, level_ed, occ_main and ind_main.

Difference between the SADA questionnaire and the hard copy of the questionnaire

The SADA electronic questionnaire differs from the original hard copy on the crucial "race" variable. This is shown in Table 2 below. The reason for this divergence has not been established. The codes given in the electronic copy of the questionnaire are identical to those given on the electronic copy of the 1993 questionnaire. The rest of the questionnaire is, however, not identical to the earlier one. It is clear that at some stage the 1994 questionnaire was retyped to accompany the data release (with at least the one typing error), since the "error codes" have been typed into the electronic copy, whereas they were not in the original questionnaire.

Table 2 Codes in the electronic and hard copy of the OHS 94 questionnaire

Codes	Electronic copy of the questionnaire	Hard copy of the questionnaire	Sample proportions (person file)	Population estimates (weighted sample)
1	Asian	Asian	10,404	1,038,851
2	Black	Coloured	24,412	3,472,178
3	Coloured	White	20,580	5,192,498
4	White	Black	77,073	30,613,467
Source: taken from Wittenberg (2006, Table 1)				

More baffling is the fact that the SADA "value labels" for the race variable (where these have been used) agree with the hard copy and not the SADA electronic version. These labels have, however, not been used in the "Work" file.

Sampling design

The only description of the sample design is contained in the "Sampling" section of the "Study description" given by SADA. It reads:

"The 1991 Population Census (the latest available then) served as the basis of the sample framework for the OHS. A stratified multi-stage cluster sample was used. The size and geographical distribution of the population were taken into account in stratifying according to (the predominant) population group and province (urban and non-urban). Depending on the size of each stratum, between 20 and 70 enumerator areas (EAs) were drawn at random for inclusion in the sample. Within each EA a cluster of 30 households was drawn at random. Altogether 30300 households in 1010 EAs (Blacks 530 + Coloureds 160 + Asians 90 + Whites 230) were included in the sample." (page 4 of S0061.doc)

Identifying the primary sampling units

The primary sampling units are not identified as such in any of the releases of the data sets. As mentioned in Section 2.2, however, the combination of the first two variables in the HOUSE.RSA file seem to give the primary sampling units. The evidence for this claim is as follows:

- There are 1016 unique combinations of these two variables. This agrees quite closely with the 1010 clusters that the sampling scheme calls for. If five “clusters” that have only one household in each are ignored the agreement would be even closer.
- In the vast majority of cases (see Table 3) the “PSU”s defined in this manner have 30 households in them, as required by the sampling scheme.
- The “PSU”s created in this manner are nested within province and within “type of enumerator area” as they should be.
- Indeed they are nested within magisterial districts, since the first three digits of the first variable correspond to the district identifier (as given in the “PERSON.RSA” and “WORK.RSA” files).

Table 3 Number of households per "PSU"

Number of households within “PSU”	Frequency	Percent
1	5	0.49
10	1	0.1
15	1	0.1
17	1	0.1
22	1	0.1
24	1	0.1
25	2	0.2
26	1	0.1
27	1	0.1
28	7	0.69
29	21	2.07
30	951	93.6
31	14	1.38
32	4	0.39
33	1	0.1
35	1	0.1
38	1	0.1

40	2	0.2
Total	1,016	100
Note: The “PSU” is defined by the first two variables in the HOUSE.RSA file. It is likely that this corresponds to the primary sampling units used in the OHS 1994		

Identifying the strata

The paragraph from the SADA study description suggests that enumerator areas were picked by race and province. Presumably the “race” classification was based on the historical “Group Area” designation. We do not have this information available. We do, however, have the empirical “race” classification of the heads of the household in the cluster. Since in most cases the complexion of group areas did not change all that rapidly, we assume that the majority group within the cluster reflects the designation of the enumerator area.

This works fairly well on the whole. Table 4 shows that roughly 92% of the heads of household in areas that we designated as “Indian” were, in fact, Indian. The corresponding figures for Coloured, White and African areas were 88.4%, 93.2% and 96.4% respectively. This suggests that the “PSU”s that we identified were, on the whole, still reasonably homogeneous in terms of racial classification. Nevertheless the figures also indicate that there had already been some “mixing”. Indeed in some of the enumerator areas there was no single majority group. As “tie-breaking” mechanism, we assumed that Africans would move more readily into “White”, “Indian” and “Coloured” areas than *vice versa*, that Indians would more likely be living in “Coloured” or “White” areas, than *vice versa* (due to the size of the group) and that “White” areas would have received a much larger influx than outflow.

Table 4 Racial classification of heads of household by classification of area

Area designated as:	Percent Indian	Percent Coloured	Percent White	Percent African
Indian	91.9	3.3	0.9	3.9
Coloured	1.0	88.4	4.7	6.0
White	1.1	1.9	93.2	3.8
African	0.7	1.4	1.6	96.4

Once we classify our enumerator areas in this way and cross-tabulate this variable against province we get the “sampling scheme” given in Table 5. Except in the case of Indians and Coloureds, the number of PSUs per “stratum” is close to or larger than the twenty that the SADA description calls for. Indeed if one counts up the number of “Indian” PSUs in the Western Cape, Eastern Cape, North West, Mpumalanga and Limpopo one gets 18, while there are twenty “Coloured” PSUs in the Free State, KwaZulu Natal, North West, Mpumalanga and Limpopo. It looks highly likely that the sampling scheme was not one with thirty-six strata (four race groups by nine provinces) but twenty-six: three strata for the “Indian” group (KZ, GT and the rest), five for the “Coloured” group (WC, EC, NC, GT and the rest) and nine each for “Whites” and “Africans”.

Table 5 Breakdown of “PSU”s by race and province

	Group Area				

Province	Indian	Coloured	White	African	Total
WC	7	81	28	23	139
EC	5	36	19	95	155
NC	0	22	20	40	82
FS	0	6	19	41	66
KZ	42	9	41	97	189
NW	2	3	18	53	76
GT	20	20	47	65	152
MP	3	1	17	43	64
LP	1	1	16	75	93
Total	80	179	225	532	1,016
Note: PSUs and their racial classification are defined by the algorithm described in the text.					

Note also that the distribution of “PSU”s across the four group areas is in the same ballpark as that given by the SADA description of the sampling scheme. We seem to have picked up too many “Coloured” areas and not enough “Indian” ones, but the numbers agree remarkably well with the scheme.

Nevertheless given that the agreement is not perfect, it would be wise to check that any results are not sensitive to this set of strata, i.e. run analyses without setting strata and then with the stratification variables set.

Weights

Although there is no discussion of weights in any of the documentation, it is clear that the weights have been poststratified, presumably to achieve the age-gender-race distribution of the demographic model employed by Statistics South Africa at the time. The weights are certainly not common within household.

The weights in the “House” file are not household weights, but the person weights of the head of the household. This is probably not ideal. Slightly better might have been to assign the average weight of the members of the household. Then, at least, a population count done using the “household size” variable and the “household weights” would give the same total as the “person weights” do.

Known bugs/issues and fixes

The data set as a whole

- Primary sampling unit and Stratum information is missing

Fix: see the discussion in section 3 above.

- The race codes are different in the data set from those listed in the electronic copy of the questionnaire released by SADA. (Note that the race codes are also different from those in later OHSs.)

For more details, see section 2 above.

- Weights: In the raw ASCII data as well as in the SADA Stata release, the weights have to be divided by 100 000 in order to make them valid. This is noted in the “LAYOUT” file, but this may be missed by the typical end user.

Section 1 (“House” data set)

- The “Unique Number” variable in the original HOUSE.RSA file is not unique.

Fix: Fortunately the sort order of this file is identical to the sort order of the “person” file. The correct household identifier can be generated as

Observation number * 10 +5

- The “Province”, “Urban” and “Population Group” variables in the HOUSE.RSA file do not actually give the province, urban area indicator or population group of the head of household.

Fix: There are variables at the end of the data set (“Province”, “Stratum” and “Race”) that do contain this information.

Section 2 (“Person” data set)

No known bugs.

Section 3 (“Work” data set)

There are problems with the Income variables:

- The “Net income of employee (Rand)” and “Net income of employee (Code)” variables derived from Question 3.13 include many imputed values. Some of the imputations are dubious.
- The “Gross income of employer (Rand)” and “Gross income of employer (Code)” variables (Q3.19) also include many imputed values.

Fix: It is possible to isolate many of the imputed values. See the detailed discussion in Wittenberg (2007).

The following are variables that are truncated in the SADA Stata data sets. Only the first digit is given.

- “Income in kind: food” and “Income in kind: other” (“Natura – food” and “Natura – Other”) variables derived from Q3.13
- “Income deductions” variable derived from Q3.14
- “Occupation of employer/own account worker” variable (Q3.16)

- “Industry of employer/own account worker” variable (Q3.16)
- “Gross income of employer (Code)” variable (Q3.19)
- “Expenses” variable (Q3.19)

- “Total unpaid employees”, “Unpaid employees – under 15 years”, “Total paid employees”, “Paid employees – under 15 years”, “Expenses – salaries” and “Family workers” variables (all from Q3.21)
- The variable “Why did [you] not work last week?” (Q 3.24)
- “What kind of job does […] usually do?” (Q 3.25)
- “Occupation of the unemployed” variable derived from Q 3.30.
- “Minimum wage prepared to work for” (Q3.31)
- The “Last occupation” variable (Q 3.33)
- “Reason for not working” variable (Q 3.35)
- The “Age”, “Highest level of education”, “Occupation – Main Groups” and “Industry – Main Groups” variables added to the data set

Fix: The ASCII data set contains the full information. The data needs to be read in again with a suitable dictionary file.

Special features

Unlike later OHSs, this data set asks (an admittedly crude) question about reservation wages (Q3.31).

References

Wittenberg, Martin (2006), Research note: errors in the October Household Survey 1994 available from the South African Data Archive, *South African Journal of Economics*, 74(4): 766-768.

Wittenberg, Martin (2007), Income in the October Household Survey 1994, mimeo.

Files to accompany this report

OHS94.DO [1] This file reads in the raw data from the HOUSE.RSA, PERSON.RSA and WORK.RSA files and creates basic House.dta, Person.dta and Work.dta files. It uses the following dictionary files:

House.dct [2] Dictionary file for HOUSE.RSA. It will also work on DA0036M.P1

Person.dct [3] Dictionary file for PERSON.RSA. It will also work on DA0036M.P2

Work.dct [4] Dictionary file for WORK.RSA. It will also work on DA0036M.P3.

The variables have been named identically to the variable names given in the SADA release.

OHS94Stratum.dta [5] This file provides PSU, Stratum and district variables for individuals that have access only to the SADA data release. It is created from the HOUSE.RSA file by the following do files:

OHS94Strat.do [6] This creates PSU and Stratum variables according to the algorithm outlined in Section 3.

Magist.do [7] This creates a value label for the district codes.

Retrieved from
"http://data1st.com.uct.ac.za/mediawiki/index.php/October_Household_Survey_1994"

Category: OHS 1994

- This page was last modified 10:34, 1 April 2008.

About DatatFirst

DataFirst is a research unit at the University of Cape Town engaged in promoting the long term preservation and reuse of data from African Socioeconomic surveys. This includes:

- the development and use of appropriate software for data curation to support the use of data for purposes beyond those of initial survey projects
- liaison with data producers - governments and research institutions - for the provision of data for reanalysis
- research to improve the quality of African survey data
- training of African data managers for better data curation on the continent
- training of data users to advance quantitative skills in the region.

The above strategies support a well-resourced research-policy interface in South Africa, where data reuse by policy analysts in academia serves to refine inputs to government planning.

DataFirst

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad •
SPES BONA