DataFirst Technical Papers

Sampling methodology and field work changes in the october household surveys and labour force surveys

by
Andrew Kerr and Martin Wittenberg

Technical Paper Series
Number 21

About the Author(s) and Acknowledgments

Andrew Kerr - Senior Research Officer, DataFirst, University of Cape Town. andrew.kerr@uct.ac.za
Martin Wittenberg - Director, DataFirst and Professor, School of Economics, University of Cape Town

We thank Cally Ardington for her suggestion to explore the number of households per EA in different surveys, which helped to lead to some of the findings reported in this paper, and Peter Buwembo for explaining the multiple household dwelling point sampling methods and for sourcing the OHS 1996 fieldworker manual.

This is a joint DataFirst and SALDRU paper.

Recommended citation

Kerr, A., Wittenberg, M. (2013). Sampling methodology and field work changes in the october household surveys and labour force surveys. A DataFirst Technical Paper Number 21 Cape Town: DataFirst, University of Cape Town

Corresponding Author: andrew.kerr@uct.ac.za

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

Sampling methodology and field work changes in the october household surveys and labour force surveys

Andrew Kerr and Martin Wittenberg

DataFirst Technical Paper 21
University of Cape Town
July 2013

Introduction

The 1999 October Household Survey was the first time that Statistics South Africa (Stats SA) introduced a master sample of Enumeration Areas (Stats SA, 2000a). There were several important changes in sampling and field worker practice that accompanied the introduction of the master sample of EAs, which have not been systematically documented[1] [2], and which make comparability of the surveys undertaken before and after this time difficult. We document these changes in this research note and provide evidence that these changes were partly responsible for the odd trends in the total number of single person households estimated from the October Household Surveys (OHSs) and Labour Force Surveys (LFSs), noted in Wittenberg and Collinson (2007) and Pirouz (2005), as well as rapid increases in employment, in the late 1990s.

Changes in field work practice and sampling methodology

The master sample was a database of Enumeration Areas from the 1996 census, and was used by Stats SA to sample EAs for the 1999 OHS and the subsequent Labour Force Surveys and General Household Surveys. Prior to OHS 1999 fieldworkers were responsible for listing

[1] Hodge (2009) comments that the introduction of the master sample may have affected the estimates of total employment in the OHSs, but did not investigate this any further.
[2] Buwembo (2010) explains how multi-household dwelling points were sampled differently before the introduction of the master sample but does not document the other changes we describe here.

all dwellings in a particular EA selected for the survey and then for drawing a random sample of 10 households themselves (Statistics South Africa, 1996). With the introduction of the master sample in OHS 1999, however, Stats SA office staff kept up to date listings of each EA selected as part of the master sample, which included numbering each dwelling unit in these EAs. Several samples of 10 dwelling units from the same EA were then selected by statisticians in Stats SA's methodology unit for use in a number of surveys from OHS 1999 onwards (until a new master sample was selected) and fieldworkers were instructed to visit only the 10 dwelling units selected for each survey (Buwembo, 2012, Buwembo, 2010), rather than do the listing and sampling themselves in a new EA every survey, as had occurred pre-OHS 1999 (Statistics South Africa, 1996)[3].

A major change to the sampling methodology in 1999 was that OHS 1999 was the first year in which dwelling units at which there was more than one household present had all households at that dwelling unit sampled (Buwembo, 2012, Buwembo 2010). Prior to this only one household per dwelling unit was sampled, meaning that not all households in an EA had an equal probability of selection. The 1996 fieldworker manual states that fieldworkers should randomise which household would be selected at dwelling units with multiple households present (Statistics South Africa, 1996). Buwembo (2010) describes this method as selection using probability proportional to size, because the method of randomisation given in the fieldworker manual implies that larger households were much more likely to be enumerated than small households (Statistics South Africa, 1996: 7-8). For example, the random number table from the OHS 1996 fieldwork manual suggests that if two households were present at one dwelling unit, one of size 1 and one of size 4 then there was only a 20% chance of selecting the smaller household.

A second major change was that substitution was disallowed from OHS 1999 onwards. The early OHS questionnaires allowed space for the enumerator to write in a substitute household number. The OHS 1996 fieldworker manual confirms this, noting that each household selected should be visited 3 times before an enumerator asks permission from their supervisor to substitute a household (Statistics South Africa 1996:7). The introduction of the master sample in OHS 1999 meant that substitution of households was no longer permitted. This was because Stats SA would not have wanted fieldworkers interviewing additional households from the same EA as substitutes, when these households might have been selected for a future sample in another survey using the same master sample.

[3] Statistics South Africa (2005: xxiv) defines a dwelling unit as "any structure or part of a structure or group of structures occupied by one or more than one household; or which is vacant or under construction but could be lived in at the time of the survey. The dwelling unit is the major listing unit for this survey." In the early OHSs dwelling units were called visiting points.

A description of the effects of changes in fieldwork and sampling methodology

In all the OHS and LFS surveys from 1994-2007 Statistics South Africa used a two stage sample design. The country was divided up into enumeration areas (EAs) and in all of the surveys from OHS 1995-LFS 2007:2, except February 2000, Statistics South Africa targeted 10 households per EA for enumeration. Despite the same number of households per EA being targeted in all these surveys, Figure 1 shows a dramatic difference in the actual number of households enumerated per EA between the earlier OHSs on one hand and the 1999 OHS and the LFSs on the other. Figure 1 shows that 95% of EAs had 10 households enumerated in OHS 1995-1998 but that in the later surveys only around 35% of EAs had 10 households enumerated in them and there are many EAs with either a larger or smaller number of households per EA, as a result of interviewing several households at multiple household dwelling units or non-response respectively.

The sampling changes resulting from the introduction of the master sample imply that, whilst the percentage of households enumerated at multiple household dwelling units should be much lower in the early OHSs (because only one household was actually then enumerated at each of these dwelling units prior to OHS 99), the percentage of dwelling units containing multiple households should not be any different before or after the introduction of the master sample. But it turns out that this is not the case- fieldworkers reporting dwelling units containing multiple households is less common in the pre-master sample period, at least where we can investigate this.

In the pre-Master sample OHSs we have information on the number of households at a dwelling unit for OHS 95, from the "hspnt_no" variable, and in OHS 96, from the "hhnumber" variable[4]. In OHS 95 only 2% of the dwelling units were reported to contain multiple households and only 1.3% in OHS 96. There is no information for OHS 97 and OHS 98. There is information from OHS 99 onwards and the percentage varies from between 3.77% in LFS 2002:2 and 8.71% in LFS 2002:1. The variation is itself a potentially worrying phenomenon but it is always much higher than in the early OHSs, with an average of 5.64% of dwelling units having multiple households reported between OHS 99 and LFS 2007:2 inclusive, as the second to last column of Table 2 shows.

We noted above that there was a change in control over dwelling unit selection from fieldworkers to office staff after the introduction of the master samples. Given this change in control, that the fieldworkers were paid by questionnaire completed, one for each household (Statistics South African, 1996), and that multiple household dwelling units required extra work for only one questionnaire (including finding out the household size

[4] According to the OHS questionnaires this data was collected in every wave of the OHSs before 1999. Unfortunately this was not made available in the public release data for OHS 97 and 98.

from each household at the dwelling point selected (Statistics South Africa, 1996)), this is likely to have reduced incentives to report multiple households at dwelling units and may explain the low reported numbers of multiple household dwelling units reported before OHS 1999. However, without investigation of fieldworker incentives we cannot firmly pin down the cause of this increase.

The impact of PPS sampling of households at multiple household dwelling units

Ignoring the possible under reporting of dwelling units with multiple households, we now explore whether the correct weights were calculated for households enumerated at multiple household dwelling units in the early OHSs, given the "probability proportional to size" method of sampling when a multiple household dwelling unit was found. If the data was weighted correctly then we should find that households selected from dwelling units where there were multiple households have much higher weights.

The first column of Table 1 shows the results from a regression of the original Stats SA household weight in OHS 1995 on a multiple household dwelling unit dummy with cluster fixed effects. The results give a value for the constant of 305 but only 5.75 for the multiple household dummy coefficient with a standard error of 4.95. If the data were correctly weighted the dummy coefficient would have a value higher than the constant and be highly significant. That it is not much above zero implies that the sampling and weighting method produced an under count of all households at multiple household dwelling units, and an even larger undercount for small households. Column 2 shows that OHS 96 weights were not adjusted at all within clusters, again suggesting that no weighting correction was undertaken if more than one household was found at a visiting point.

We can also examine this issue in OHS 97-98, although only indirectly, by looking at whether households living in backyard shacks had higher weights. By the definition of a backyard shack, these households were located at dwelling units with more than one household present, so if the sampling method is correct then the weights should again be adjusted upwards when a household living in a backyard shack was enumerated. Columns 3-4 in Table 1 show the results of regressions of the household weight on a backyard shack dummy while allowing for cluster fixed effects. The results show that these households did not have higher weights - in fact in OHS 97 and 98 backyard households even have slightly lower weights.

It is clear from this analysis that, despite a sampling method that systematically decreased the probability of households at multiple household dwelling units being sampled and further decreased the probability of small households at multiple household dwelling units

being sampled, there was no corresponding increase in the weights of the households at multiple household dwelling units that were sampled. Together with possible under-enumerating of dwelling units containing multiple households, mentioned above, this led to an undercount of households (particularly small households) at multiple household dwelling units prior to the introduction of the master sample. The combined effect is shown in the last column of Table 2. The weighted percentage of households coming from multiple household dwelling units is 2.4% in OHS 95, only 1.4% in OHS 96 and an average of 12% in OHS 99 and the LFSs, with a maximum of 19% in the March 2002 LFS. This suggests that a very different set of households makes up the population in the early OHSs, on the one hand, and OHS 99 and the LFSs on the other.

Applications: Increases in single person households and employment

Clearly the changes in sampling methodology and fieldwork practice that we have described may have important consequences for discussions of trends in several important series derived from the survey data. Here we link our analysis above with observed changes in household size in the OHS and LFS survey data and with increases in employment.

Between October 1997 and February 2000 OHS and LFS survey data show an increase from about 10 percent to 20 percent in the weighted proportion of single person households. This represents an improbable increase from 1.3 million to 2.3 million one person households in two and a half years and contributed to a large decrease in average household size, as Table 2 shows.

Wittenberg and Collinson (2007) show that the growth rate in single person households was much lower in the Agincourt Demographic Surveillance Site and suggested this might be due to changes in sample frames or household definitions in the OHSs and LFSs. The census data also suggests the increases in single person households in the OHS and LFS are too large. Table 2 shows that the 1996 and 2001 Censuses found roughly 1.5 and 2.5 million single person households respectively and by 2001 the number of households using the Labour Force Survey data was roughly the same as the census, suggesting a large undercount in the earlier OHSs. We argue here that this was partly the result of the sampling methods employed in the early OHSs- where smaller households were much less likely to be chosen at multiple household dwelling units and, as we showed above, not weighted up to compensate for this method of selection.

It should be pointed out that the increase in single person households actually started between OHS 1997 and OHS 1998- a year earlier than the change in the sample methods- this is likely due to the lack of coverage of hostels in OHS 1996 and 1997 that is mentioned in the OHS 1998 release (Statistics South Africa, 2000b), which meant a very large

undercount of hostel dweller households, which are much more likely to be single person households. To illustrate this, Census 1996 found 567 000 hostel dwellers who were single person households (those living in hostels in households of size larger than 1 were enumerated through the household questionnaire and cannot be directly identified as living in a hostel). In contrast, the OHS 1996 data suggests there were only 88 000 single person households living in a "room in a hostel or compound", the figure that is comparable to the Census 1996 figure above, implying an undercount by a factor of six in the 1996 OHS. Thus not only were early OHS sampling methods undercounting small households, fieldwork difficulties further contributed to the undercount of mostly small, hostel dwelling households in OHS 1996 and 1997.

The rapid increase in single person households also contributes to increases in total employment. Total employment rose very dramatically in the late 1990s- growing by 33% between October 1997 and September 2000, as shown in Table 3. Table 3 also shows that total employment growth in single person households was 112% over this period, compared to a growth rate for households larger than 1 person of 25%. The growth in employment in 1 person households was around 30% of the total growth in employment over the period, despite the employed in single person households only making up 14% of the employed at the end of the period and 8.6% at the start. Interestingly, though the growth rate of employment in single person households was high for both men and women, it was much higher for women- employment growth for women living by themselves was 240% compared to 90% growth for men living by themselves (see Table 3). Clearly the sampling strategy in the early OHSs that under-sampled small households affected estimates of employment as well, but further investigation of this issue is left for a future research project.

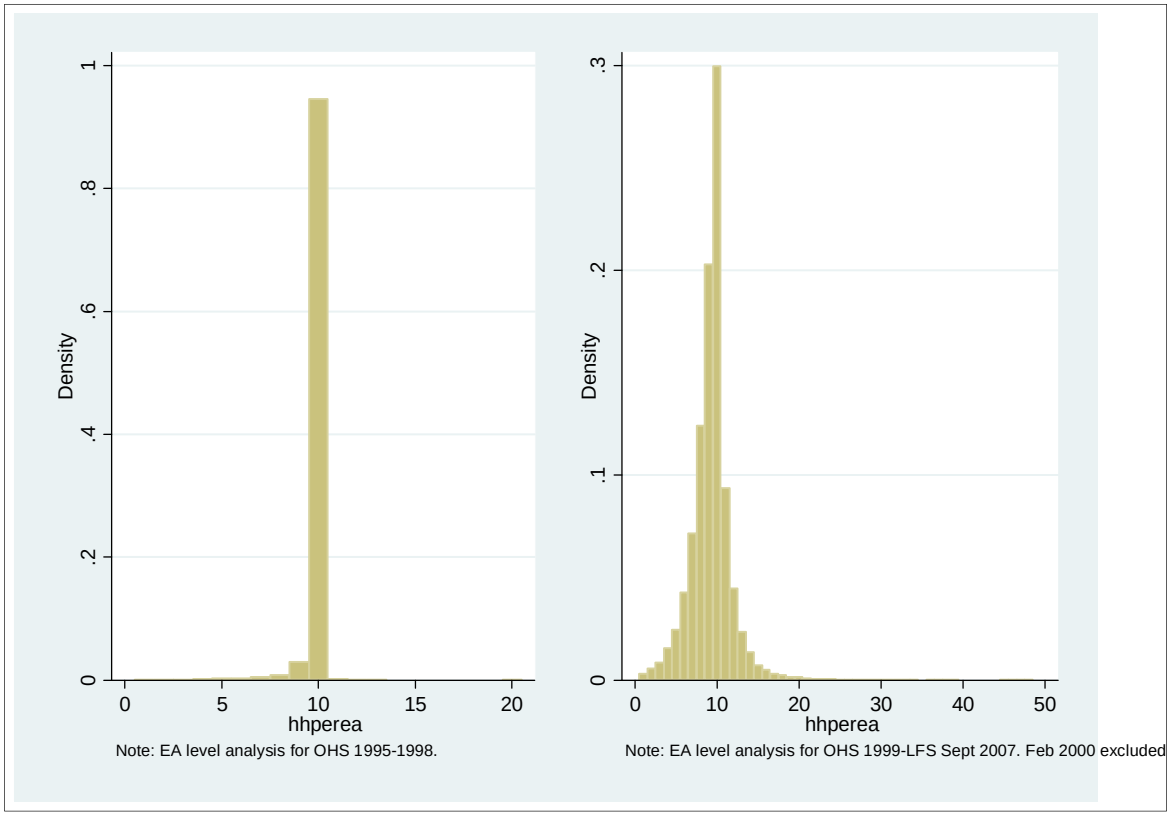
Conclusion

In this research note we have highlighted important changes made by Statistics South Africa in its sampling methods and fieldworker practice that first occurred in the 1999 October Household Survey. The introduction of a master sample led to a change in control over sample selection from fieldworkers to Stats SA office staff, possibly leading to increased reporting of dwelling units with multiple households by fieldworkers. The introduction of the master sample also meant substitutions were no longer allowed and that all households at multiple household dwelling units were sampled. This last change, together with the underweighting of households at multiple household dwelling units pre-master sample that we have shown above, likely explains a substantial part of the implausibly large increase in single person households from October 1997 to February 2000. We have also shown that it explains some of the increases in employment around the late 1990s and early 2000s. The under sampling of hostels in OHS 1996 and 1997 also contributes to the rapid increase in

one person households and employment over this period. The sampling changes first introduced with the master sample in OHS 1999 were an improvement on the methods used in the earlier OHSs but make comparisons between the two periods more difficult. More detailed work is required to understand how these changes may have affected other important trends derived from the OHS and LFS data.

Figures

Figure 1: The distribution of households per EA in OHS1995-1998 and OHS 199-LFS September 2007



Tables

Table 1: Household weight regressions

	OHS 95	OHS 96	OHS 97	OHS 98
Multi hh Dummy	5.75	-		
	(4.95)			
backyard dummy			-1.393	-5.455**
			(1.609)	(2.362)
Const.	305.6167***	569.523	310.544***	489.422***
	(0.52)		(0.197)	(0.292)
N	29700	15921	29811	18981
R^2	0.747	1	0.945	0.976

Notes: The dependent variable is the household weight released by Statistics South Africa. Standard errors in parentheses. **indicates significance at the 5% level, *** indicates significance at the 1% level. Cluster fixed effects included in all regressions. The variation in weights in OHS 96 is fully explained by the cluster fixed effects, ie weights are constant within enumeration areas.

Table 2: Single person households, average household size and multiple household dwelling units

Survey wave	Estimated number of 1 person households	Average household size (weighted)	Estimated number of 1 person households from the Census	Percentage of dwelling units with more than 1 household reported (unweighted)	Percentage of households at multiple household dwelling units (weighted)
OHS 1994	999504	4.49			
OHS 1995	1141363	4.45		2%	2.4%
OHS 1996	1063911	4.44	Census 1996: 1480217	1.3%	1.4%
OHS 1997	1287104	4.40			
OHS 1998	1626600	4.19			
OHS 1999	2088252	4.00		4.0%	8.9%
LFS 00:1	2353726	3.90		5.7%	13.1%
LFS 00:2	2415377	3.87		5.1%	11.3%
LFS 01:1	2541079	3.81		7.1%	16.5%
LFS 01:2	2564525	3.83	Census 2001: 2479259	7.7%	16.6%
LFS 02:1	2750219	3.77		9.0%	19.2%
LFS 02:2	2602528	3.82		3.2%	7.1%
LFS 03:1	2771083	3.73		3.5%	7.6%
LFS 03:2	2861516	3.69		3.9%	8.4%
LFS 04:1	2919406	3.66		3.4%	7.6%
LFS 04:2	2822535	3.63		4.9%	10.4%
LFS 05:1	2834873	3.65		6.1%	13.2%
LFS 05:2	2863295	3.64		6.4%	14.5%
LFS 06:1	2743018	3.69		5.8%	12.7%
LFS 06:2	2952101	3.61		6.2%	13.8%
LFS 07:1	3065676	3.55		6.6%	15.2%
LFS 07:2	2947933	3.60		4.3%	10.6%

Note: OHS and LFS data weighted using cross entropy weights discussed in Branson and Wittenberg (2011). 10% Census samples weighted using Stats SA household weights.

Table 3: Employment in the OHSs and LFSs

Survey	Total employment	Employment in 1 person households		
		Total	Female	Male
OHS 1994	9635642	644114	67186	576927
OHS 1995	10296924	860598	106158	754440
OHS 1996	9537055	671088	102874	568214
OHS 1997	9643678	837025	118309	718716
OHS 1998	10204242	1089430	174701	914274
OHS 1999	11192713	1456399	339866	1116533
LFS 00:2	12777248	1774785	399092	1375693
LFS 01:2	11783379	1800188	484460	1315728
LFS 02:2	11880779	1753604	500008	1253596
LFS 03:2	11963947	1961255	510772	1450483
LFS 04:2	12290751	1867264	488329	1378935
LFS 05:2	13082463	1977816	475534	1501441
LFS 06:2	13528143	2054500	514941	1539559
LFS 07:2	14031007	2077648	547832	1529817

Note: Cross entropy weights used to weight the data- see Branson and Wittenberg (2011).

References

Branson and Wittenberg (2011). Re-weighting South African National Household Survey Data to create a consistent series over time: A cross entropy estimation approach. SALDRU Working Paper Number 54.

Buwembo, P (2010). Factors Associated with under-5 Mortality in South Africa: Trends 1997-2002. Unpublished MSocSci thesis, University of Pretoria.

Buwembo, P (2012). Personal communication with the authors.

Hodge, D. (2009). Growth, Employment and Unemployment in South Africa. *Unisa Working Paper Number 119.*

Pirouz, F. (2005) 'Have Labour Market Outcomes Affected Household Structure in South Africa?', *DPRU Working Paper No 05/100*, Cape Town: Development Policy Research Unit.

Statistics South Africa, 1996. October Household Survey 1996 Manual for Fieldworkers. Electronic copy obtained from Peter Buwembo at Statistics South Africa.

Statistics South Africa, 2000a. Statistical release P0317, October Household Survey 1999. Downloaded from http://www.datafirst.uct.ac.za/catalogue3/index.php/ddibrowser/64/download/539

Statistics South Africa, 2000b. Statistical release P0317, October Household Survey 1998. Downloaded from http://www.datafirst.uct.ac.za/catalogue3/index.php/ddibrowser/63/download/519

Statistics South Africa, 2005. Statistical release P0210, Labour Force Survey September 2005. Downloaded from www.statssa.gov.za/publications/P0210/P0210September2005.pdf

Wittenberg and Collinson (2007). Household transitions in rural South Africa, 1996–2003. *Scandinavian Journal of Public Health Supplement*. 69:130-137.

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