Coverage evaluation of South Africa's last census

*Jeremy Gumbo
RMPRU, Chris Hani Baragwaneth Hospital, Johannesburg, South Africa
Clifford Odimegwu
Demography and Population Studies; Wits Schools of Public Health and Social Sciences
*corresponding author
jeremy.d.gumbo@gmail.com

Introduction

South Africa's census 2011 recorded an undercount estimate of 14.6%, suggesting that about one person among each seven people was missed. This was despite that census 2011 was better funded than South Africa's other previous censuses, and in particular census 1996 which had a lower undercount of 10.6% (Berkwotz, 2013). Funding for census 1996 was about 5 times less than that for census 2011. Moreover, neighboring Mozambique known to have small budgets for census taking still managed to record an undercount rate of 6% for census 1996 (StatsSA, 2011). Expectations are that increase in census funding should lead to improvement in quality of census enumeration (Gernertzky, 2012). For instance, better funding in South Africa's latest census resulted in the incorporation of modern census techniques like GIS mapping and cartography (Onsembe and Ntozi, 2006), as well as use of video training and barcodes for questionnaires (Cronje and Budlender, 2004). Such improvements are expected to translate into better census coverage.

The measurement and adjustment of coverage errors in census 2011 StatsSA who are the custodians of census processes used Post Enumerative Survey (PES). However, some researchers have questioned the accuracy of counts that were arrived through this process. For instance, Dorrington and Moultrie (2012) noted that the publishing of census 2011 results was too early to have allowed accurate processing of such huge data. For this reason the researchers, argue that there are strong grounds for doubting the accuracy of the census counts published by StatsSA. Other researchers also brought forward some evidence to indicate inaccuracy of this census' counts. For example, the increase in fertility suggested by census 2011 was interpreted as inconsistent with prior fertility trends that have prevailed in South Africa for many decades (Berkowitz, 2013). The argument is that South Africa has consistently experienced declining fertility as early as 1960s; therefore increase in fertility suggested by census 2011 indicates inaccuracy of counts especially for children ever born. There has also been an increase in counts for the female whites aged 20-24 years, which other census analysts believed can neither be traced from previous censuses nor migration records (Gernertzky, 2012).

Even some members of the public have also indicated their doubts over figures obtained from the census, and have expressed their views through social media. For instance, one twitted that the published counts for census 2011 were a mere political gimmick. The view expressed was that, the counts were reflective of political interests rather than what should have come

from a genuine census process. Commenting on the increase of white females aged 20-24, another member from the public mockingly twitted; Hahahahahaa the invasion of the young white women (rnoliphant, 2012). The twit was meant to express doubts the writer had on the accuracy of counts for this particular sub population group. Commenting on the increase in fertility and population of young white women, another twit went as: "The odd baby boom and the strange influx of young white women" (sarahhemilyduff, 2013).

The views summarized above indicate that among both researchers and members of the public there are some who are convinced that 2011 census counts are largely inaccurate. The concern raised cannot be under played as census counts are partly used for resource allocations, and service delivery planning in South Africa. Taking into cognizance the various concerns raised by census stakeholders across the South African population on the subject matter, this study therefore investigated the extent of accuracy of 2011 census counts. The actual count for any given population is always unknown; hence expecting an actual population count from a census is unrealistic.

Data Sources

We used census 2011's 10% samples for both Households and Individuals' data, and the full census data. After completing the processing of results for the census, Statistics South Africa (StatsSA) sampled 10% of the data which they availed for public access. The 10% sampled data is representative of the entire population. The full 2011 census data (100%) is only available in SuperCross, and in this form it can only be analysed through tabulations. For the purpose of this research we needed actual counts of people for the area covered by Agincourt Health and Demographic Surveillance Site (HDSS. We overlaid Agincourt HDSS area's digital boundary using Arc GIS to map out actual counts for 2011 census from this area from SuperCross data. In this paper we termed these counts, actual census counts.

We also used the 2011 General Household Survey (GHS) data obtained from StatsSA, and is collected annually. This data is nationally representative at House hold level. The other data set was drawn from National Community Survey (NCS) 2012, which was collected by Johns Hopkins Health and Education South Africa (JHHESA). This is also a national representative survey that collected population characteristic on individuals aged 24-55 years in South Africa. At small area level we drew data from Agincourt HDSS. This is longitudinal data drawn from Agincourt surveillance site area of about 420 square kilometres, covering about

28 villages, and close to 15 000 homesteads. The base population was in 1992, and since then the data has been regularly updated after each three month.

Analysis Plan

Test for expected trends and patterns from accurate counts

There are certain trends and patterns expected from accurate counts, and we did the following tests for these. Firstly, the distribution of males counts relative to females' counts. Due to increased risk of mortality among males relative to females, the former's counts are expected to reduce more than the latter's counts as age increases (Keane et al, 1985). Secondly, we checked growth rates from census 2001 to census 2011. These are expected to fall within the range of 0 to 3.5, in the absence of a heavy migration effect (Moultrie et al, 2013). For Graphical Cohort Analysis, counts by 5 year age groups for censuses 1996, 2001 and 2011 were plotted. Due to mortality effect distribution of counts by age groups for census 2011 should be at the bottom, followed by those for censuses 2001 and 1996 respectively, i.e. in the absence of distortive effects like heavy migrations (Keane et al, 1985).

Matching of counts at national level

For comparison of counts from data at household level, Household headship counts' distribution by 5-year age groups from the 10% sample Household data for census 2011were matched with those from 2011 GHS data. The analyses were done for males and females separately. At individual level, population distribution by 5-year age groups from the 10% sample persons' data was matched with those from 2012 NCS data. The matching was done for 5-year age groups falling between 24 and 54 years, because the latter data was restricted to this age range. Again the analyses were carried out disaggregated according to sex. All data sets were weighted before analysis.

Matching of counts at Small area level

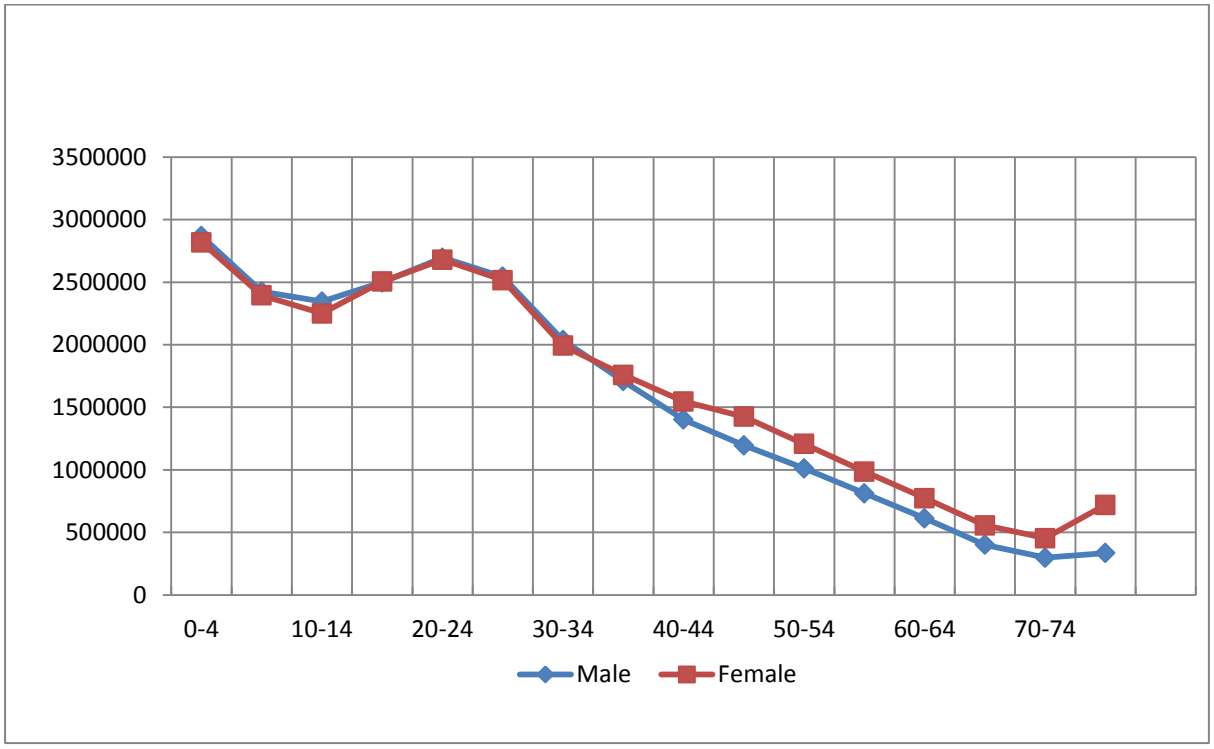
We arranged census data for the municipal area covered by Agincourt HDSS into 5 year age groups disaggregated by sex. Totals by sex, and for both sexes combined were also computed. This same procedure was done for the longitudinal data Agincourt surveillance site. We took Agincourt HDSS counts as the gold measurement for accurate counts for this area. We assumed Agincourt HDSS counts to be accurate because the data is updated on a regular basis, and is collected on a small area which should be easier to manage.

Results

Test for expected trends & patterns

The distribution of female counts relative to males counts did not deviate from expected trend as age increased. Males' counts were slightly higher than female counts at age groups 0-4; 5-9, and 10-14 years, indicating as expected the effect of high sex ratio at birth. Counts for both sexes were almost the same for age groups 20-24 to 35-39 years. Thereafter, indications are that female counts were always higher than respective males' counts [Fig 1]

Fig 1 Population distribution by age group and sex



Growth rates for both females and males were largely consistent with expected patterns. For females, only age groups 0-4, 5-9, and 10-14 years had growth rates that were outside the expected range. Yet for males virtually all age groups except the 10-14 years had growth rates that were within the expected range. All growth rates that were outside the expected range were below 0, and none was above 3.5.

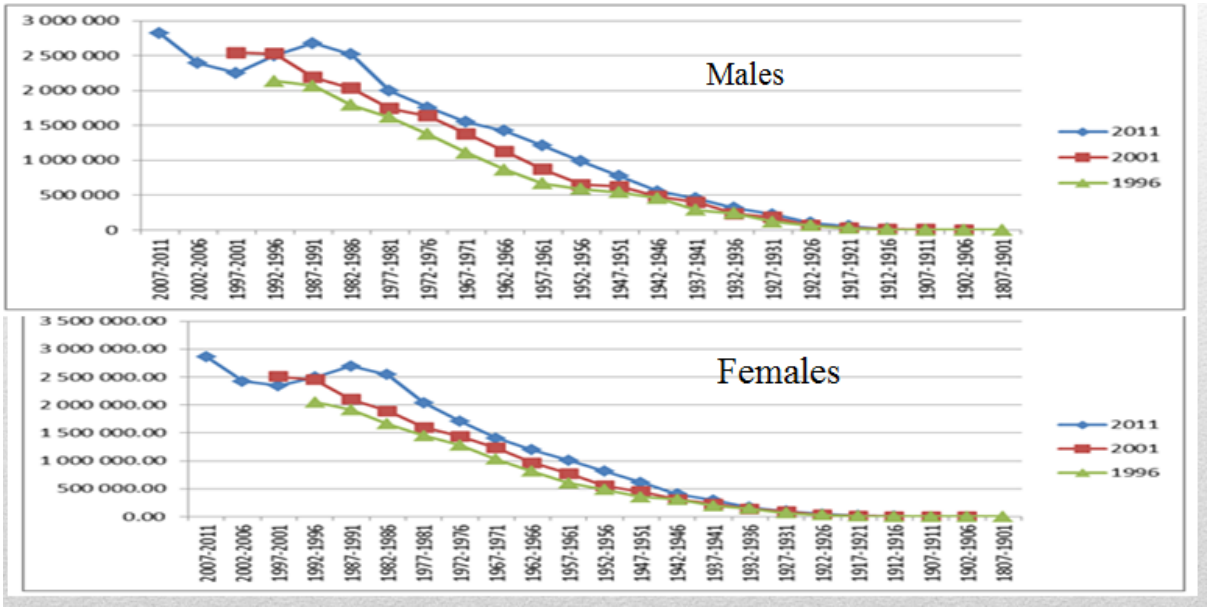
Table 6 a: Growth rates by age group, for females and males

Age group	Census 2001	Census 2011	Growth	Census	Census	Growth

(Years)	females	females	Rate (r)	2001 males	2011 males	rate (r)
0-4	2215008	2817867	0.02407	2214369	2867584.9	0.02585
5-9	2425994	2394570	-0.0013	2423906	2425181	0
10-14	2541811	2250611	-0.01217	2510361	2344275	-0.0069
15-19	2527782	2504905	-0.00091	2454284	2498572	0.00179
20-24	2189344	2679896	0.02022	2100064	2694646	0.02493
25-29	2034172	2516635	0.02128	1893200	2542681.7	0.0295
30-34	1741231	1992804	0.0135	1596760	2036206	0.02431
35-39	1635554	1758420	0.00724	1438418	1709346.5	0.01726
40-44	1376879	1546291	0.0116	1230423.1	1402328	0.01308
45-49	1125861	1424543	0.02353	962657.87	1195740	0.02168
50-54	870990.936	1206940	0.03262	770704.03	1011349	0.02717
55-59	650859.782	985458.39	0.04148	551102.11	811949.96	0.03875
60-64	622622.77	773404.01	0.02169	447549.3	612363.96	0.03135
65-69	483069.23	556256.456	0.01411	305168.98	401548.2	0.02745
70-74	396651.2	453343.42	0.01336	230192.45	297144.509	0.02553
75-79	231978.27	317675.03	0.03144	136967.29	163690.73	0.01782
80-84	179941.73	222072.27	0.02104	91981.021	100128.35	0.00849
85-89	65320.607	102683.16	0.04523	30519.573	43720.4802	0.03594
90+	47908.833	77394.255	0.04796	17077.66	30334.111	0.05745
Total	23362979.1	26581769.3	0.01291	21405705.2	25188790.9	0.01627

However, results from Graphical Cohort Analyses from both sexes were inconsistent with trends expected from accurate counts. Counts by age for census 2011 were more than respective counts from compared censuses. As result, the lines describing distribution of counts for census 2011 for both males and females were at the top, instead of being at the bottom. Besides, both lines also crossed the ones for census 2001 at age group 15-19 years, further suggesting inaccuracy of investigated counts.

Fig 2 Graphical Cohort Analyses



Comparisons at National level

Male Household headship counts, 2011 census and 2011 GHS

Estimated counts of households headed by males were slightly higher from census data for age groups between 10 to 29 years. The biggest difference was noted among age groups that ranged between 30 and 49 years, where counts from census data were evidently lower than those from GHS data. However, counts for the rest of the age groups ranging from 50 years to last age group examined i.e. 90-95 years generally matched each other. The counts for female headed households overall matched across all age groups except for age group 40-44 and 45-49, where counts from census data were clearly lower than those from GHS data. The compared trends of household headship for both males and females were very consistent with each other [refer to Figs 1 & 2 below]

Fig 1: Male Household headship counts by age group Census 2011 and GHS 2011

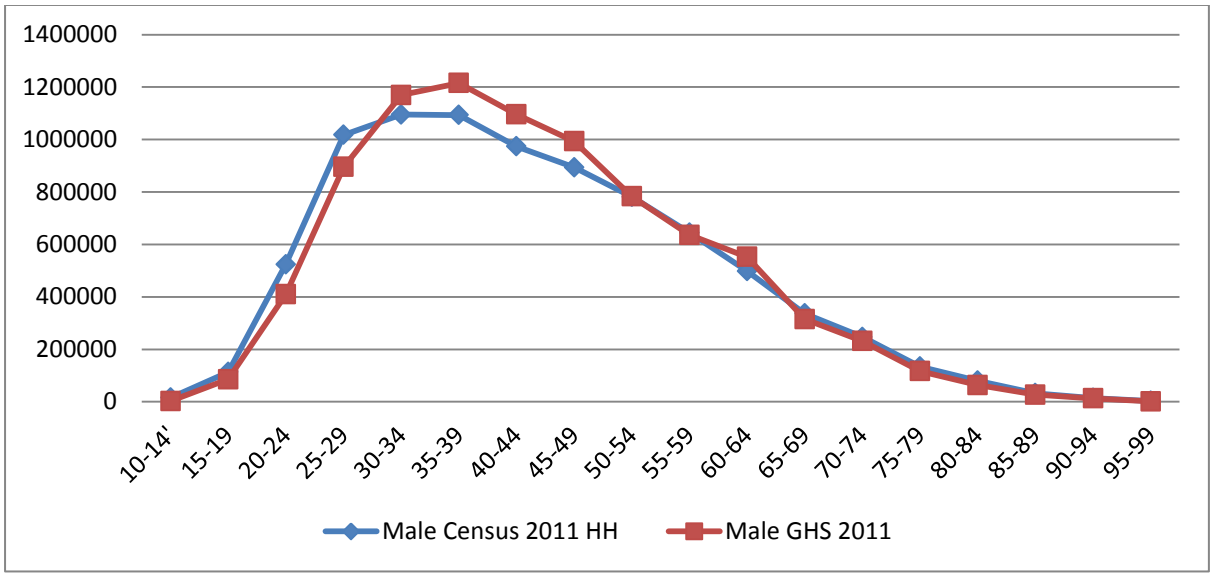
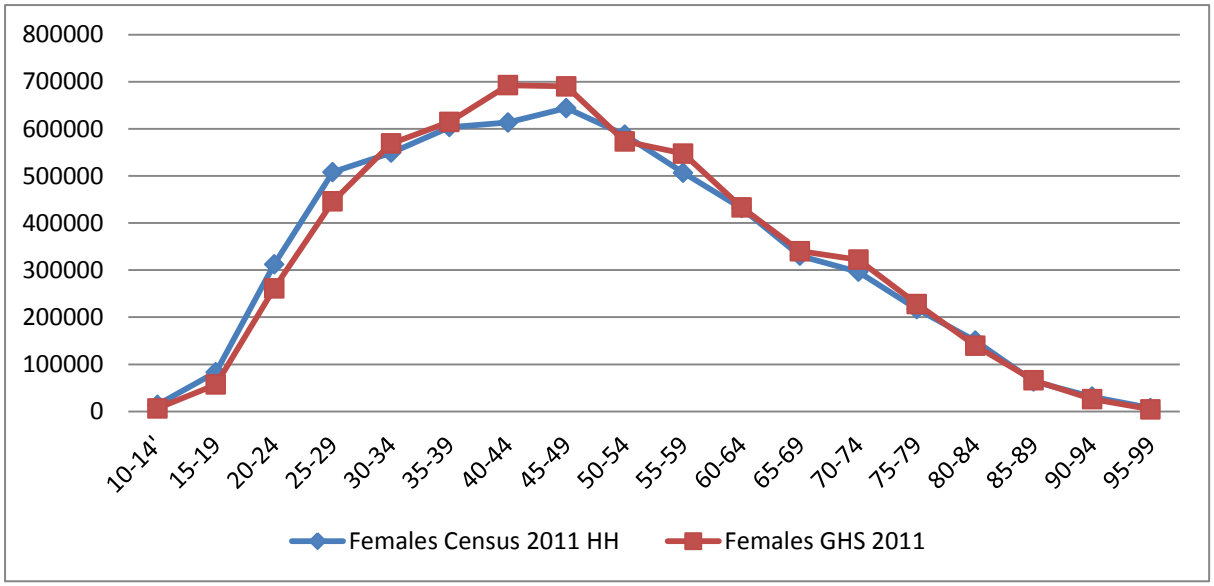


Fig 1: Female Household headship counts by age group Census 2011 and GHS 2011



Persons Counts distribution by age group, 2011 Census and 2012 NCS

The compared individuals' counts from the two data sets, for males were evidently wide apart though generally their trends across age groups looked identical. Except for age group 25-29, the male counts from census data were higher than the respective counts from NCS data. As for females, again counts from census data were higher than those from NCS across all age groups except for age group 25-29 years. However, the difference between the counts from

the two data sets, were closer to each other for age groups 30-34, 35-39, and 50-54. [refer to figs 3 & 4 below].

Table 3 Census 2011 and NCS 2012 counts, Males

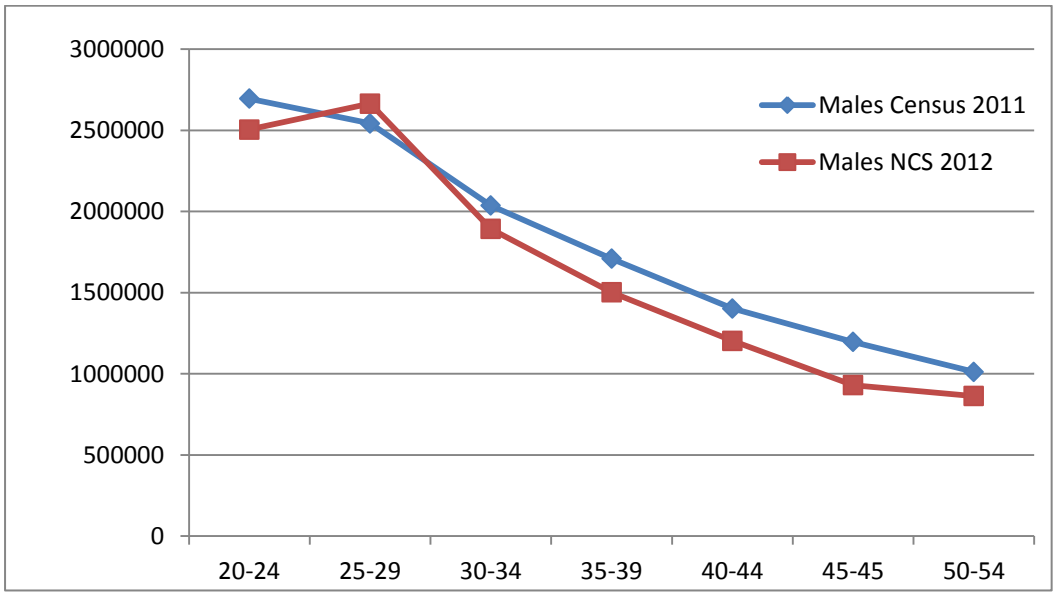
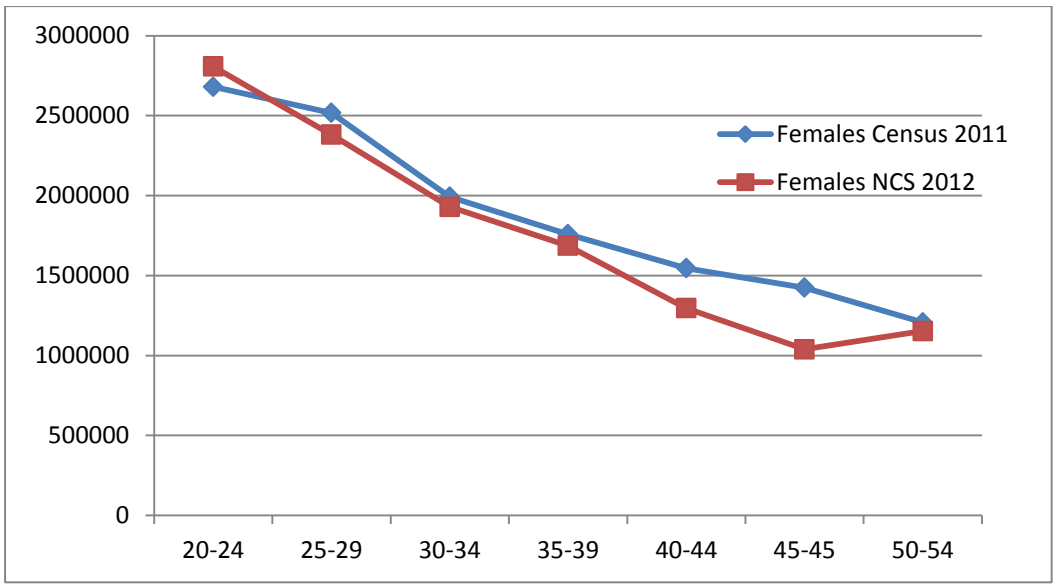


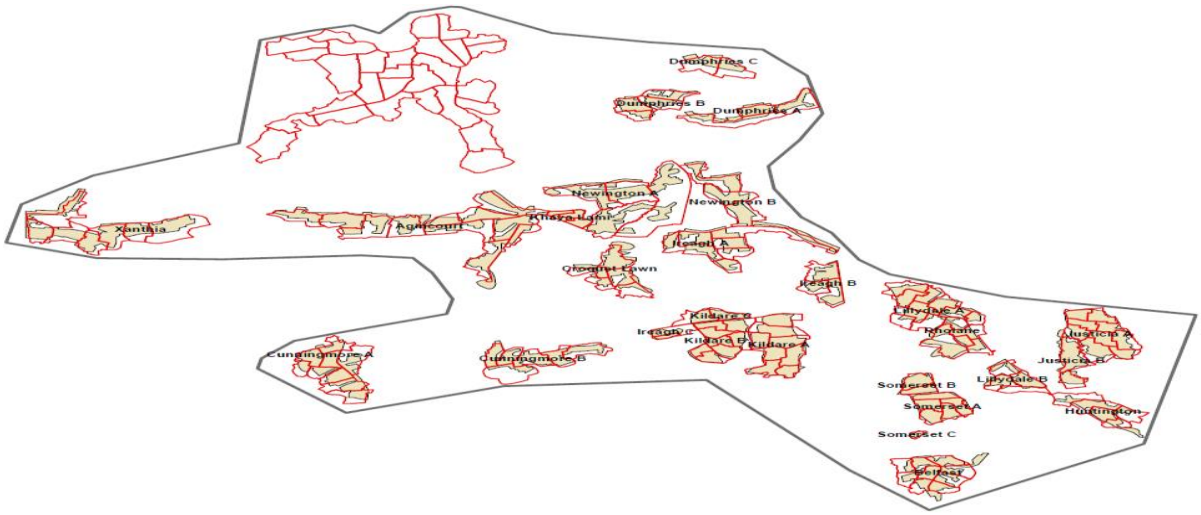
Table 4 Census 2011 and NCS 2012 counts, females



Comparisons at small area level

The matching of coinciding small areas boundaries and villages' boundaries for the area covered by Agincourt HDSS, shown in Fig 5 below indicates no perfect match. Indications are that small areas boundaries shown in red often exceeded the boundaries for Agincourt villages. Though counts from both data shall have same reference date of collection, we should expect census counts to be slightly higher across all age groups, for both sexes.

Fig 5 Census 2011 overlay for entire Agincourt HDSS area



Actual census counts relative to Agincourt HDSS counts

For males, females, and combined counts actual census counts overall remained higher than Agincourt HDSS counts. Only few exceptions were noted e.g. age group 0-4 and 5-9 for males; 0-4 for both females and for the combined. Males counts had slight differences with respective counts among young age groups (from 0-24 years), and among the old (60 years and above). The differences were quite wide for the in between age groups ranging between 18.6% and 20.6%. Differences between females' counts were consistently slight across all age groups with only two age groups; 35-39 and 45-49 years having double digit percentage differences. The differences for combined counts from the compared data sets were a mere compromise of what came from males' counts, and what came from females' counts. The total counts for males, females, and for combined still maintained slightly higher census counts over Agincourt HDSS counts [Refer to Table 1 and Fig 6 below]

Table 1 Actual census counts relative to Agincourt HDSS's, by age group

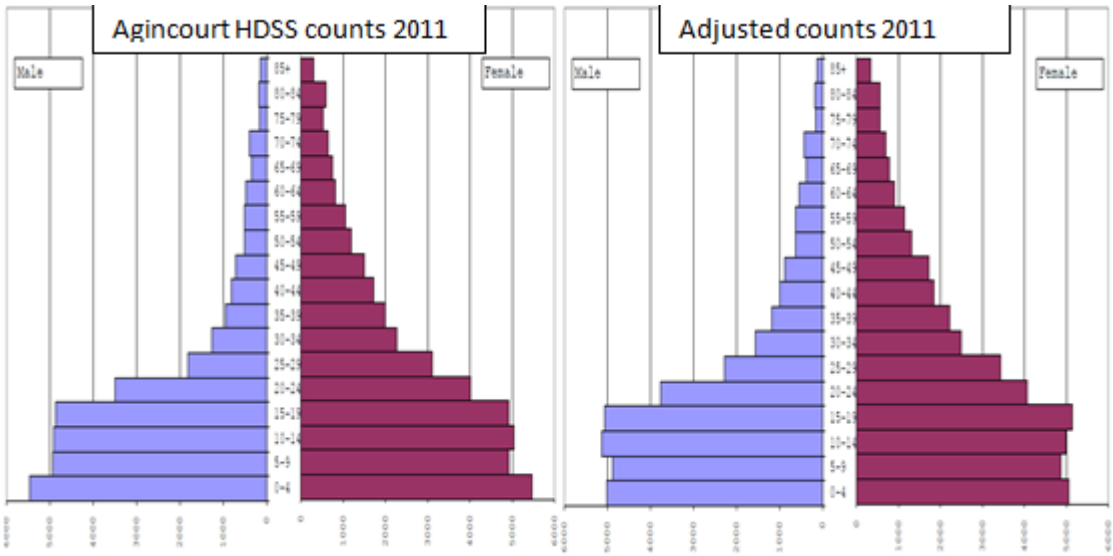
Age group	Census 2011			Agincourt HDSS 2011			% differences between compared counts		
	Males	Females	Total	Males	Females	Total			
							Males	Females	Totals
0-4	5014	5052	10066	5466	5465	10931	9	8.2	8.6
5-9	4880	4864	9744	4929	4928	9859	1	1.3	1.1
10-14	5123	4996	10119	4907	5037	9944	4.2	0.8	1.7
15-19	5058	5147	10205	4855	4918	9773	4	4.4	4.2

20-24	3754	4075	7829	3501	4012	7513	6.7	1.56	4
25-29	2275	3430	5705	1811	3101	4912	20.4	9.6	13.9
30-34	1568	2492	4060	1260	2277	3537	19.6	8.6	12.9
35-39	1171	2223	3394	948	1996	2944	19	10.2	13.2
40-44	996	1842	2838	810	1718	2528	18.7	6.7	10.9
45-49	877	1719	2596	706	1509	2215	19.5	12.2	14.7
50-54	623	1308	1931	507	1202	1709	18.6	8.1	11.5
55-59	631	1147	1778	504	1053	1557	20.1	8.2	12.4
60-64	541	898	1439	471	811	1282	12.9	9.7	10.9
65-69	383	794	1177	341	744	1085	11	6.3	7.8
70-74	426	697	1123	384	652	1036	9.9	6.5	7.7
75-79	165	569	734	168	528	696	1.8	7.2	5.2
80-84	172	559	731	169	597	766	1.7	6.8	4.8
85+	125	331	456	143	310	453	14.4	6.3	0.7

Matching of Population pyramids

The population pyramid from actual census counts for the investigated area of Agincourt HDSS was in most aspects identical to one based on counts from Agincourt HDSS data. A contested issue by demographers which the two pyramids concur is an increase of fertility. Both pyramids show higher population for age group 0-4 relative age group 5-9. The two pyramids also indicate a highly dwindling male population fro age group 25-29 upwards. Also both pyramids confirm that counts for age group 80-84 years especially for females, are higher than both the prior and preceding age ages. Slight differences noticeable include higher fertility increase from census relative to Agincourt HDSS data.

Fig Population Pyramids for 2011 census and Agincourt HDSS 2011



Discussion

Our study investigated accuracy of 2011 census counts. We did this firstly by checking if the counts produced trends and patterns consistent with those expected from accurate counts. Secondly we compared these counts with those from different data sources. Findings from comparisons of distributions of females' counts against respective males' counts were consistent with those expected from accurate counts. The same applies with findings from growth rate analyses. In the former case, the counts suggested a high sex ratio at birth, a pattern largely observed from most population. Less males' counts from middle aged adults onwards relative to respective female counts were also in line with expectations from accurate counts. Mortality is often higher among males than females as the former often have high risk practices. Inconsistences suggested from Graphical Cohort Analyses may not be due to inaccurate census counts, but rather due to substantial immigrations affecting South Africa over the past decade. Heavy immigrations into the country from countries like Zimbabwe and Mozambique, most likely countered the reducing effect of mortality on age cohorts. These analyses therefore largely pointed towards accurate counts

From the comparison of counts suggestions were also that census counts largely accurate. It is a known fact that the actual population count for any given population often remains unknown. From the comparison of counts at Small Areas level, the bigger surface area covered by small areas boundaries providing census counts , relative to Agincourt HDSS village boundaries as noted earlier meant that counts from the latter source should be less

than respective counts from the former source. The overlapping boundaries naturally meant more people incorporated into census data compared to those incorporated in Agincourt HDSS data. Furthermore, just like the boundary overlaps were not evidently very wide, the compared counts also did not widely deviate from each, except for a few instances in the males comparisons. The populations' distributions by age for the compared data sets were also fairly consistent with each other. At national level, comparisons of household headship for either sex suggested almost similar counts across all age groups.

However, still at national level the matching of census counts against those from NCS data produced trends that indicated coverage errors in the data. The NCS data was during mid-year of 2012, whereas census data was in beginning of October 2011. Naturally, expectations are that NCS data which was collected almost a year later should have slightly higher counts relative to respective count from census data. But even still, the mismatch between the counts from the two data sets does not necessarily point to inaccurate census counts. Surveys are prone to sampling errors which largely affect the quality of the data finally produced. Since most of the analyses carried out did not suggest such an extent of inaccurate census counts, it is more likely that NCS counts were rather the ones that were inaccurate.

Conclusion

From our findings, indications are that 2011 census counts are a largely accurate. The trends and patterns produced from these counts were largely consistent with those expected from accurate censuses. Above all, counts from most data sets that were compared against census counts closely matched. On the other hand results from Graphical Cohort Analyses, and matching of census counts with those from NCS data did not necessarily indicate inaccurate census counts, but most likely indicated the effect of heavy migration and sampling errors respectively. However, we encourage more researches around this subject, to verify our findings.

Reference

Berkowitz, P (2012) "The (incomplete) (probably inaccurate) sum of us." Daily Maverick

Cronje, M and Budlender, D. (2004) "Comparing census 1996 a with 2001: An operational persipective" Southern African Journal of Demography, 9 (1), p 67-90

Gernetzky, K (2012) "Census: Undercount still double digit." Mail & Guardian

Dorrington, R (2001) “Did they jump or were they pushed? An investigation into apparent undercount of whites in 1996 South African census.” SA Journal of Demography, 8(1), 2001-2002)

Keane, J, G, Balridge, M., Brown, C, J, and Jones, S. (1985) “Evaluating censuses of population and housing.” Resource support Services in agreement with US Agency for International Development

Moultrie, T and Dorrington R. (2012) “The census: Some more questions.”

Moultrie, T. (2013). “General assessment of age and sex data” Tools for Demographic estimation p 3-21

Onsembe, J, O and Ntozi, J, P, M. “The 2000 round of censuses in Africa: Achievements and challenges.” *The African Statistical Journal*, Vol 3.

(Statssa Census 2011). “Census undercount and Strategies: Version No 6.” Statistics South Africa.https://www.statssa.gov.za/census2011/documents/Census_2011_Undercount%20Strategies-version_6a.pdf