Statistics South Africa

Preferred supplier of quality statistics

Reweighting of the GHS 2002–2008 data series

Social Analysis
August 2010

Table of contents **Page**

1. Introduction

During 2009, all the historical GHS data were recalibrated and reweighted. This was necessary because of the major revisions that the population estimates underwent after the release of the Community Survey 2007 results and updates to the estimates of the impact of HIV/Aids on demographic trends in South Africa.

Several activities took place during the reweighting process:

1) Missing values for the demographic variables age, sex or population group were imputed for all historical datasets.
2) The demographic estimates published in 2009 were used for benchmarking all the historical data and the house files were weighted independently of the person files using estimates of the number of households in South Africa.
3) Provincial boundary adjustments were made to the historical datasets to adjust the historical data to the December 2005 geographical boundaries.

Each of these activities will be briefly discussed in this report.

2. Imputation of missing values for demographic variables

The new programs that were introduced for weighting from 2008 onwards, discard all records with missing values for age, sex or population group. Therefore it became necessary to impute missing values for the key demographic variables of the historical series (GHS 2002–2008). Since an automated editing system was developed for the GHS 2009, the imputation of demographic variables will now be done as a rule from 2009 onwards.

Imputations usually used a combination of logical and hot-deck imputation techniques. Hot-deck imputation was only used to deal with missing values after logical imputations were applied. The emphasis was on obtaining reliable imputations rather than a 100% imputation rate.

Population group imputation

A type of hot deck, the nearest neighbour imputation method, was used for the imputation of population group. The household was used as the first reference level and by virtue of its proximity/'nearest neighbour' nature, the PSU was the second level of reference used for imputation. The code used is based on a macro developed by Chien[1].

Age imputation

Logical imputation

The average age difference between children and their father/mother is calculated by age group of the household head and race of the father. The age of children with missing ages is derived from the average. This is also done if the age of the child is known but that of the parent is not known.

Individuals that are married or live together as partners are identified. The average age difference of partners by population group is used to impute the age of the partner that has an unspecified age.

The ages of children currently attending school are imputed using their highest level of education.

[1] Chien, L. and Weaver, M. A macro for nearest neighbour imputation. Sesug Paper CC-016, http://analytics.ncsu.edu/sesug/2008/CC-016.pdf, accessed April 21, 2009.

In cases where Q111EduI in '3', '4', and '5' (i.e. attending tertiary education institutions), the average age of persons are calculated and used to impute ages of persons attending similar institutions.

Hot-deck imputation

The age of the household head is imputed using a hot deck composed of the number of persons in the household, sex and population group of the household head. For members of the household that are not the household head the hot deck was composed of the sex of the household head, the age group of the household head and the relationship with the household head. If the conditions for these two decks were not satisfied because of missing relationship to the household head data, age was further imputed using total number of persons in the household and population group.

Age imputation is further refined by building decks for workers (15 years and older) separately from non-workers (younger than 15 years).

Sex imputation

Logical imputation

Individuals identified as the mother or father of someone in the household are extracted. If their sex is missing, the sex of their mother or father role is allocated to them.

Additional imputations were also made using manual assessments of sex based on the names of individuals. If a name can be male or female, a sex value is allocated randomly to that individual. In the case of the GHS 2008, a combination of the scanned images and the GHS 2009 questionnaires (which were conducted at the same dwelling units where GHS 2008 was conducted) were used. These assessments of sex were incorporated through hard coding in the imputation program.

Hot-deck imputation

Sex was imputed for the household head using a deck comprising the household head's age and population group. For household members that are not the household head, age group, population group and relationship to the household head were used. If all these failed because of missing relationship information, the last deck was used without the relationship variable.

Headship imputation

Unspecified household head status is normally the result of the original head being disqualified at question B4, i.e. not meeting the 4x4 rule. Household head status was imputed based on the fact that household members are listed from oldest to youngest and that spouse/partner of the household head usually follows the household head. Thus, if the household head was disqualified as a result of not meeting the 4x4 rule, the person listed after the household head in the household replace him/her as the household head for weighting purposes.

Imputation rates

Table 1: Number of missing values (NM), rates of missingness[2] (RM) and imputation rates[3] (IR) for age, sex and population group for GHS 2002 to GHS 2008

GHS year	No. of records in person file	Age Before imputation		Age After imputation		Population group Before imputation		Population group After imputation		Sex Before imputation		Sex After imputation	
		NM	RM	NM	IR	NM	RM	NM	IR	NM	RM	NM	IR
2002	102 471	106	0,10	99	0,10	21	0,02	0	0,02	25	0,02	21	0,00
2003	99 428	43	0,04	39	0,04	0	0,00	0	0,00	13	0,01	13	0,00
2004	97 197	39	0,04	33	0,03	14	0,01	0	0,01	9	0,01	9	0,00
2005	108 002	79	0,07	69	0,06	11	0,01	0	0,01	14	0,01	14	0,00
2006	105 727	87	0,08	79	0,07	101	0,10	0	0,10	30	0,03	30	0,00
2007	110 114	130	0,12	109	0,10	35	0,03	0	0,03	30	0,03	19	0,01
2008	94 897	167	0,18	149	0,16	75	0,08	0	0,08	495	0,52	184	0,33

Table 2: Number of missing values (NM), rates of missingness[4] (RM) and imputation rates[5] (IR) for age, sex and population group for GHS 2002 to GHS 2008 after additional hot-deck imputations

GHS year	No. of records in person file[6]	Age Before imputation		Age After imputation		Population group Before imputation		Population group After imputation		Sex Before imputation		Sex After imputation	
		NM	RM	NM	IR	NM	RM	NM	IR	NM	RM	NM	IR
2002	102 471	106	0,10	23	0,08	21	0,02	0	0,02	25	0,02	1	0,02
2003	99 428	47	0,05	8	0,04	0	0,00	0	0,00	13	0,01	0	0,01
2004	97 197	39	0,04	12	0,03	0	0,00	0	0,00	9	0,01	4	0,01
2005	108 002	79	0,07	10	0,06	11	0,01	0	0,01	14	0,01	5	0,01
2006	105 727	87	0,08	19	0,06	101	0,10	0	0,10	30	0,03	0	0,03
2007	110 114	130	0,12	29	0,09	35	0,03	0	0,03	30	0,03	12	0,02
2008	94 895	165	0,17	41	0,13	73	0,08	0	0,08	493	0,52	8	0,51

[2] The rate of missingness refers to the percentage of values that are missing in relation to the number of records available for weighting.
[3] The imputation rate is the percentage of values that were imputed in relation to the total number of records that were available for weighting.
[4] The rate of missingness refers to the percentage of values that are missing in relation to the number of records available for weighting.
[5] The imputation rate is the percentage of values that were imputed in relation to the total number of records that were available for weighting.
[6] Bogus records, records that do not have result codes 1, 4 or 5 and individuals who do not meet the 4x4 rule were excluded from the imputation process.

Table 3: Headship imputations

Indicator	Year						
	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
No of heads before imputation	26 306	26 384	26 204	28 150	28 007	29 328	24 347
No of heads after imputation	26 366	26 400	26 220	28 291	28 196	29 481	24 371
Number of cases imputed	60	16	16	141	189	153	24
Imputation rate	0,23	0,06	0,06	0,50	0,67	0,52	0,10

3. Independent weighting of the house files using estimates of the number of households in South Africa

Some of the problems experienced during the GHS 2008 weighting procedure were a low sample yield, poor execution of the sample and unexplained low numbers of households. Until 2009, the person weights for individual household members were used as the weights for the household files.

It was decided to adopt the same kinds of tables for the benchmarking of the person file to the house file. Two kinds of tables were generated with summaries of household estimates for each year. Firstly, tables with household numbers for age against population group and sex of the household head were produced and secondly, tables for age against province and sex of the household head.

The following methodology was used to produce these household estimates:

1) The number of households (excluding institutions) and population numbers for Census 2001 and CS 2007, LFS 2003, LFS 2005 and QLFS 2009 were used to establish independent points over time for the calculation of headship ratios.
2) Headship ratios were calculated as the percentage heads in the population calculation for a specific combination of demographic characteristics.
3) Conversion ratios for each year were then calculated based on the trend lines calculated between independent points. It was assumed that change took place exponentially.
4) Finally, household estimates were produced using the mid-year population estimates published in July 2009 and the conversion ratios.

   The formula used was: (cell population estimate*cell conversion ratio)/100.
5) The estimates produced for the population group table were considered to be the most reliable, and small cosmetic adjustments were done for the provincial tables using the totals of the population group tables for age and gender. This was necessary to align the totals for the two tables.

Table 5: Person file

Year	Original number of records sent for weighting	Number of records previous release	Number of records Re-weighted data
2002	102 471	102 461	102 359
2003	99 428	99 428	99 345
2004	97 197	97 197	97 091
2005	108 002	107 987	107 889
2006	105 727	105 727	105 708
2007	110 114	109 975	109 824
2008	94 895	94 097	94 744
2009	94 497	Not applicable	94 263

Table 6: House file

Year	Original number of records sent for weighting	Number of records previous release	Number of records Re-weighted data New method
2002	26 247	26 243	26 218
2003	26 398	26 398	26 371
2004	26 214	26 214	26 190
2005	28 132	28 129	28 102
2006	28 002	28 002	27 999
2007	29 311	29 280	29 236
2008	24 370	24 222	24 333
2009	25 349	Not applicable	25 303

In the absence of official estimates for the number of households in South Africa, various role players have developed their own estimates. The stakeholders of the GHS have generally been using the estimates published in the GHS as a result of the default generation of weights borrowed from the person file. The new weighting system for households generally yields higher estimates for the number of households in South Africa than previously published. However, it does smoothe the annual differences between years and provinces considerably.

Tables 5 and 6 indicate that some record loss in the person and house files as per previous releases took place. This was unavoidable as the new programs used for weighting need non-missing values for all the demographic variables (age, sex and population group). Even though record loss has been significantly reduced by the demographic imputations that were developed for the historical data sets, the reweighted data still have fewer records than previous releases.

Another characteristic of the reweighted data is that there is a difference in the number of records for households as derived from the person file and the number of records in the house file. This is the result of the independent weighting systems used for the two files. Households where there was missing demographic information for any of the three variables used, were not weighted. If other individuals in the household had valid demographic information, they would have received weights in the person file, but as a result of incomplete information for the household head, the whole household would be excluded from the house file.

4. Provincial boundary adjustments were made to the historical datasets to adjust the historical data to the December 2005 geographical boundaries

During reweighting, the provincial boundaries as adjusted in December 2005 were used to standardise provincial units across time. This was needed to make the data more relevant to current boundaries, but also to make it comparable with the master sample first used in 2008, which was designed based on these boundaries.