National Statistics System Division (NSSD)

Measurement, Standards and Capacity Assessment

The South African Labour Force Panel Study

Methodology Document

2006

Contents

1. Introduction

In panel studies data are collected repeatedly, using the same units (e.g. individuals, households, enterprises, etc.). The rationale for longitudinal surveys has been debated at length. The first advantage of using longitudinal surveys is to account for the difference between net and gross change. Gross change is only visible in longitudinal data. For example, the decomposition of changes in the unemployment rate into contributions from inflows and outflows can only be observed using longitudinal data. In addition, the capacity to make causal inference is enhanced by temporal ordering. Some phenomena such as chronic poverty and unstable employment are inherently longitudinal. Lastly, repeated observations on individuals allow for the possibility of controlling unobserved differences between individuals by using fixed and random effect models. The most famous panel survey is the US Panel Survey of Income Dynamics (PSID) which is also one of the longest (about 35 years). Panel studies are subject to 'survivor bias' since some longitudinal processes mean that some of the respondents (e.g. weak, sick, poor, those subject to an environmental stressor such as those living in informal settlements, etc.) are disproportionately unlikely to survive long enough to be interviewed. Repeated measures take time to accumulate into longitudinal data and consequently delay the availability of the data. Achieving internal consistency between retrospective responses to similar questions is vital to the quality of the data.

In 2001 Statistics South Africa introduced a survey that was similar to a panel design when it conducted its fourth Labour Force Survey. This type of survey continued up to 2003, when the ninth Labour Force Survey was conducted. The first three LFSs were done as pilots to test the methodology and the survey instruments, and the surveys which followed LFS9 are based on a new sample. The design was neither a panel of individuals nor households as it was designed as a rotating panel of dwelling units (DUs). In principle, individuals and households are not tracked, and although every effort is made to go to the same dwelling unit (DU), the dwelling unit could be occupied by a household different to the one enumerated in the previous round. Data collection, processing and release were made as cross section for each survey and thus no direct connection was made previously. The construction of the panel was done in a subsequent exercise.

Wave 1 – September 2001 (LFS4);
Wave 2 – March 2002 (LFS5);
Wave 3 – September 2002 (LFS6);
Wave 4 – March 2003 (LFS7);
Wave 5 – September 2003 (LFS8);
Wave 6 – March 2004 (LFS9).

2. Sample design

The first wave began in September 2001 and achieved a total sample of 30 000 households. The last wave was conducted in March 2004. The sample frame consists of the 3 000 Primary Sampling Units (PSUs) of the Master Sample based on the 1996 census of Enumeration Areas (EAs) and remains fixed up to the lifetime of the Master Sample. Kish's method of combining smaller EAs has been used to form the PSUs. On average, a PSU consists of 150 to 250 dwelling units. The PSUs were explicitly

stratified by province and area type, and within each explicit stratum the PSUs were further stratified by arranging them in geographical order by District Council, Magisterial District and, within the Magisterial District, by average household income (for formal urban areas and hostels). Sample selection was done with probability proportional to size in each stratum. The measure of size was the estimated number of dwelling units in each PSU. Clusters of 10 dwelling units each were formed systematically within each PSU so that each cluster could be considered as being representative of the PSU. Five clusters in each PSU were formed to be used in the different rounds of the LFS. The survey was conducted in such a way that a rotation of 20% of dwelling units was implemented, i.e. each time the LFS was conducted, the dwelling units in the 600 PSUs were added or discarded at each wave up to the lifetime of the Master Sample frame on which the sample design was based. This means out of the original 30 000 DUs used in the first wave of the LFS (LFS September 2001), 24 000 were to be used in the following survey (LFS March 2002) and 6 000 new ones were added to make up the required 30 000 DU total sample size until there were no DUs from the original sample in wave 5. A dwelling unit selected in September 2001 was therefore maintained until September 2003. All households found in the selected dwelling units were interviewed. Indeed the Labour Force Surveys fit more closely with the methodology of repeated surveys with partial overlap rather than longitudinal studies. Repeated surveys are planned at regular intervals (6 months in our case) and include rotating panels (20% in our case). In other words, sampling units are introduced in the survey, surveyed a number of times and then rotated out of the survey. The main purpose for the overlap is to reduce the variance. There is no attempt to follow people in sample units that move or link records for identical units to make longitudinal estimates. On the other hand, longitudinal surveys with rotation are designed to follow a particular group of units for a specified period, to introduce new sample units at specified periods, to create longitudinal records for each observational unit, and to include longitudinal analysis.

The LFS design has no tracking or following rules that involve tracing of sample members wherever they go. The lack of such rule makes the design vulnerable to attrition whereas in panel studies the ultimate goal should be to reduce as much as possible the level of attrition. In addition, attrition is severely amplified due to the lack of proper sample data management, lack of rotation monitoring process and lack of proper individual identifiers. Also, records were not linked and assembled immediately to create the panels although longitudinal analyses were part of the estimation plan. Panel studies are also very particular in respondent rules (self-reporting versus proxy information) that can seriously affect the quality of the data. We do not use answers given in previous interviews to assist respondents to recall their previous answers or for the interviewer in the field to check whether he is at the right dwelling unit in case of overlap. There are differences in response patterns if the respondent or the proxy used changes across the waves. Indeed it has significant implications in our case since our mechanism of identifying individuals across waves depends on their response to some of their particulars (e.g. age, gender, etc.). Data collection was made by face-to-face interviews with heads of selected households.

Figure 1 shows how the rotation patterns were conceptualised. Two clusters would be sufficient for the surveys, one cluster serving as the original sample and the other one used in the rotation. In this way, out of the original 30 000 DUs, only 24 000 DUs would be used in Round 1 and 6 000 new DUs would be introduced from the second

cluster. In this way, in Round 5 we would be left with no DUs from the original cluster and the three previous rotations would have accumulated 18 000 DUs. Many problems were experienced with the implementation of the sample in the field.

					Balance
Round 0	30 000				30 000
Round 1	24 000			6 000	30 000
Round 2	18 000		6 000	6 000	30 000
Round 3	12 000	6 000	6 000	6 000	30 000
Round 4	6000	6 000	6 000	6 000	6 000	30 000

Figure 1: Rotation of DUs within PSUs

Unfortunately no record was kept of how the rotation scheme was applied or of the actual dwelling units rotated per panel owing to a lack of data management and proper methodological documentation. Table 1 represents an attempt at recreating the rotation pattern, and describes how actual implementation of the sample was done. The table was constructed after analysis of the sampled files.

Table 1: Implementation of rotation

Wave	Households	Dwelling units (DU)	PSUs	Rotated DUs		% overlap
				in	out	
1	32 831	30 011	3 004	-	-	-
2	33 509	30 014	3 000	5 758	5 755	80
3	31 342	29 980	2 998	7 912	7 943	74
4	31 587	30 015	3 003	12 141	12 137	60
5	31 393	29 879	2 988	17 656	17 524	40
6	31 779	30 176	3 003	23 077	23 242	20

As described in Table 1, the rotation between waves 1 and 2 was 20% where 5 758 DUs were replaced by 5 755 newly selected DUs. These make up approximately 617 PSUs. In the second round, only about 7 900 DUs were replaced instead of the planned 40% (about 12 000 DUs) between the initial wave and wave 3**.** This creates an overlap of 74% instead of the expected 60%. This finding raises a strong suspicion that there was no rotation between LFS5 and LFS6. It seems rather that re-sampling within PSUs possibly took place between these two surveys as new listing information became available. Re-listing shifts the numbering of the dwelling units in the PSU. This effectively destroys the panels in the particular PSU. The original clusters should have remained fixed within each PSU for the lifetime of the Master Sample. Growth should have been taken care of in the weighting. This is probably the reason why we actually visited the last 20% of the original dwelling units in LFS9.

3. Following rules

'Following rules' is required to maintain representativeness of the original population and their descendants – these specify who would be eligible to be interviewed at each wave. Since the units of the sample are dwelling units, we 'follow' dwelling units rather than individuals or households. There is also no tracing of individuals or households. Consequently we enumerate households that currently occupy the selected dwelling unit. There is no guarantee that the same households or individuals occupy that particular dwelling unit. In recreating the panel, we are matching

individuals rather than other units, and our panel data is a panel of individual members of households.

- The Longitudinal Sample consists of members of original households, their natural descendants born since the start of the panel, and co-residents of original sample members as long as they belong to the same household, stay at the same address (dwelling unit), and were matched at least twice during the lifetime of the panel.
- Sample members are eligible for an interview across each wave as long as they remain in that particular address for four nights (the household membership rule strictly applies).
- Sample members are not followed as they move.
- The original sample included only members of private households and hostels.
- At each wave the households that occupied the selected dwelling unit was also eligible even if it differed from the previous household. In addition, 20% of newly selected dwelling units were added to the sample owing to rotation.

The panel sample will be reduced by:

- Attrition – refusal, non-contact and rotation
- Moves out of scope (resample within PSU which destroyed the structure of the panel, new re-listing also altered the numbering of the original DU listing sometimes)
- Poor record-keeping of DUs that were rotated in or out made it difficult to pinpoint the exact structure of the sample
- Constant change of the household structure
- Non-sampling errors encountered during data processing

Due to the unavailability of the 'General' files for each survey, it is impossible to reconcile what has been sampled and what actually was enumerated. Thus we cannot distinguish which sample members have been lost due to rotation (owing to the design of our sample) and those lost due to follow-up (either owing to the destruction of the panel, the household moving out, or poor fieldwork). This makes it very difficult to use the data set for the analysis of changes in household characteristics over time.

4. Data processing

The LFS data sets have been collected, processed and published independently as cross sections. Therefore the panels had to be recreated. It took a project of its own to recreate the panels and involved further data processing as adjacent LFSs were matched with one another and ambiguity was found. This problem arose because the data sets were always treated as cross sections, and at the time data processing was only concerned in correcting the right ranges of age or gender, and consistency of the variables within the data set itself. Records that came back from fieldwork operations at each LFS were not compared to previous LFSs –

- to confirm whether the rotational scheme was working correctly; and
- as a simple measure of data quality checking for consistency of the data across LFSs .

There are no ID numbers in any of the LFSs that could assist with the matching of persons within households. Program codes had to be developed to match individuals based on their characteristics in previous panels. Fieldwork in the LFSs also seemed of questionable quality as persons in the same household ended up with different ages, genders, etc. in ensuing panels. Person numbers in adjacent panels were also not consistent and the formation of unique numbers for persons, which could have been used for matching purposes, were not consistent across panels because persons received different person numbers in adjacent panels. It also seems that some household members migrated to other dwelling units or fieldworkers were using incorrect information.

However, the task at hand is now to correct the consistency of the variables among the different waves. This is very difficult as many discrepancies are observed at least between the few variables that were checked. The job is divided into two components. The first component involves matching the data set and the second is to check the data set for consistency.

In the matching process, two procedures are followed.

- First round of matching: manual matching of persons by reverting to the initial information in the questionnaires.
- Second round of matching: computer matching of person records of the cross section the published LFS data sets once manual matching has been established.
- First round matching: computer matching of persons for whom manual matching could not be done.

The procedures used for manual matching can be outlined as follows:

1. Check geographical addresses of dwelling units by comparing the primary sampling unit (PSU) from each LFS. (e.g. LFS4 and LFS5).
2. Match the dwelling unit numbers of the same PSU. If they do not match, the PSU was rotated.
3. If the dwelling units match, compare the physical addresses for the two surveys. Here the linking variables are addresses (street addresses, or settlement description). If similar, consider the dwelling unit a match.
4. Now compare each household composition in each dwelling unit by looking at respondents' names, surnames, age, gender, and other variables like education level or employment status.
5. If the information is the same for both surveys, consider it a match and write down the unique number, person number, age and code for gender in one row for each survey.

Table 2: Manual matching process

	LFS5	LFS6	LFS7	LFS8	LFS9
LFS4	Yes	Yes	Yes	Yes	No
LFS5		Yes	No	No	No
LFS6			Yes	Yes	No
LFS7				Yes	No
LFS8					No

As indicated in the above table, manual matching could not be carried out in all cases for the following reasons:

1. Questionnaires for LFS5 were damaged owing to poor storage facilities.
2. Owing to a lack of proper documentation, it was not clear whether rotation took place during the implementation of LFS9.

The available information refers to a rotation of 20% of dwelling units each time the LFS was conducted (starting from LFS4). This leads us to believe that the rotation ended at LFS8. After discovering that the last 20% rotation actually took place in the last survey (LFS9), we decided not to pursue manual matching for LFS9 because it would be too costly and time-consuming. Consequently we switched to computer matching for all matches between LFS9 and all the surveys; and between LFS5 and those for which we did not have manual matches. The computer match was based on the PSU and person numbers only and is therefore less precise than the manual matching.

Generally the rate of matching was poor for the following possible reasons:

1. Lack of an identity number that could uniquely identify households between waves. This happened where we had multiple households within the same dwelling unit.
2. Lack of an identity number that could uniquely identify individuals between waves.
3. During enumeration the general rule is to record individuals by age from eldest to youngest. But this rule was not strictly observed and consequently individuals might end up with different person numbers between waves even if the household composition remained the same.
4. New person numbers were given to each person every time the household was enumerated, and these numbers did not necessarily correspond with previous ones.
5. Owing to poor data management of the samples for each survey it was unclear which dwelling units had been rotated in or out, or re-sampled. Consequently it was very difficult to actually retrace the original clusters.
6. Owing to a lack of information on the sample maintenance methodology used for the Master Sample, it was not clear whether changes in the dwelling unit numbers were the result of re-sampling or the result of shifts in the numbering of the dwelling units due to re-listing of the PSU. This did, however, destroy the panel.

7. Owing to the unavailability of the ‘General’ files for each survey, it was impossible to reconcile what had been sampled and what actually was enumerated.

We could not distinguish which sample members were lost due to rotation, (owing to the design of our sample) and those lost due to follow-up (owing to the destruction of the panel).

Although manual matching is more tedious than computer matching, it was more reliable since we could work with addresses and names and surnames of individuals. The identifying variables were the names and surnames of individuals. Those that had matching addresses, names and surnames but different genders were considered a match after correcting the gender variable. However, those that had matching addresses, names and surnames but different ages were further investigated in the LFS subsequent to the current LFS being investigated. If the following LFS showed a similar age as that of the previous LFSs that were being investigated, this age was assigned to both data sets after making the necessary adjustments for the period that had lapsed between surveys.

Once manual matching had been done, we performed the second round of matching. The computer matching process can be described as follows:

1. We worked on a few selected variables from workers’ and persons’ files.
2. We considered the LFS4 file at person record level and created an indicator variable for the LFS4 record.
3. We then considered the LFS5 file at person record level and created an indicator variable for the LFS5 record.
4. Taking the manual matching record file into consideration, we merged the LFS4, LFS5 and matching record file by unique number and person numbers.
5. We did the same for LFS4 and LFS6, then LFS4 and LFS7 and LFS8.
6. Where no manual match data were available, we did an electronic match as by merging LFS4 and LFS9 by unique household number and person number.
7. We then created the final data set for all LFS4 matches, merging them by LFS4 unique household number and person number.
8. We created a unique identifying variable to link all matched persons.
9. Age and gender differences between the surveys were tested. These formed some of the key primary identification variables that were used to further identify individuals.
10. Unacceptable age and gender differences were deleted.
11. The final matched data sets were then created.
12. The final unmatched data sets were also created (we are keeping the unmatched people to identify who they are and try to explain why they did not match).

In areas where there was a lot of movement, this inevitably implied that there would be less matching of either households or individuals. This phenomenon was more prominent in informal settlements. We considered a person part of the panel data sets

when we had at least one valid match. All 6 waves of raw data were created. The number of records for the raw data set is given below:

Wave 1 ---> Number of records: 60 985
Wave 2 ---> Number of records: 72 000
Wave 3 ---> Number of records: 68 522
Wave 4 ---> Number of records: 67 852
Wave 5 ---> Number of records: 59 818
Wave 6 ---> Number of records: 46 885

What the computer did not consider a match:

- Age differences larger than 5 between consecutive LFSs
- Gender differences between consecutive LFSs

A unique 13-digit number (individual identifier) has been created in order to link individuals between waves. This identifier differs from the usual unique numbers that we normally release with Stats SA data sets. The first digit is Province, followed by Type of Area, and the rest of the digits comprises the unique person number.

Table 3: The distribution of matches between consecutive LFSs

LFS	Number of persons
4_5	109 410
4_6	102 480
4_7	100 834
4_8	98 748
4_9	98 256
5_6	102 480
5_7	100 834
5_8	98 748
5_9	98 256
6_7	100 834
6_8	98 748
6_9	98 256
7_8	98 748
7_9	98 256
8_9	98 256

We will have to edit the data so that it conforms to panel format. This means we have to check and maintain consistency for each variable between waves (e.g. we have seen an increase in qualifications from grade 12 to a Master's Degree between two consecutive waves). In addition, some matches are duplicates and need to be cleaned. This is a time-consuming process and requires the development of a computer code for edit and some methodology for imputation. We can give a proper account of the quality of the data only once the edit has been performed.

5. Edits

We use the minimum processability criteria by first cleaning records for individuals who have a false match in the whole panel and then cleaning individuals across panels. The next step of the edit is to smooth the data across the waves to make them more meaningful.

5.1 Minimum processability of individual records in waves

- Person records in panels must have only one unique identifier (UQ). Cases were found during manual matching when sizes of matched households were different between waves and consequently the matched persons identifiers were not the same. This led to a person being matched more than once (duplicate matches) and a single individual having more than one unique identifier (UQ). If a person had more than one identifier (UQ), we discarded the duplicate and corrected the UQ in all the waves where the individual had been identified.
- Matched persons from waves must have equal to 5 or less than 5 years' difference in age and have the same gender, otherwise the records were discarded and reviewed (flag = 1).
- Matched persons from waves cannot have more than 5 years' difference in age (gender may be the same). They were discarded and reviewed (flag = 2).
- Where the difference in level of education was too big (more than 3 years) and the gender was not the same, the cases were discarded and reviewed (flag = 3).
- In cases where we had a difference of less than five years in age and the gender remained the same, but the language and population group were different, the persons were removed from the panel for review (flag = 4).
- If the check for the minimum age for education showed a misbalance, we reviewed the reason (flag = 5), most of them are children attending school.
- In cases where we had difference of less than five years in age, but not the same gender and a difference of more than three years in the level of education, we removed the persons from the panel for review (flag = 6).
- In cases where the difference in age was less than 5 years , and the gender, language or population group were not the same, we discarded the cases and reviewed these (flag = 6).
- Where the difference was less than five years in age, and a different population group was indicated for the same person (except for code 2 – coloured), the cases were reviewed (flag = 7).
- In cases with less than 5 years difference in age: if marital status code 4 (never married) followed by code 2 (widow/widower) or code 3 (divorced) in other waves, we removed the person and reviewed when population group or language were not the same (flag = 8)
- Children aged 12 or younger in the first wave were deleted from the panel if they were left as the only member of the former household.

5.2 Edit of data across waves

The age variable was collected by asking the following question:

D	**How old is .....?**	*(in completed years - In figures only)* *Less than 1 year = 00*

Valid range: 0 – 126
Unspecified: 999
Missing values: (.)
Dots (.) were replaced with 999

Unspecified: (999)

Rule: If a person belongs to the panel and age is missing in one wave and other fixed variables as gender, population group, language (and education +3) are the same, validate the missing age using the individual's age in the nearest other wave. Otherwise reject the record during the structural edit.

Rules used for editing age for panel (after structural edit):

1. Take the highest age if age difference is 1 (for cases when next consecutive wave age is less than the previous one)
2. Take the average age if the age difference is more than 1 (in consecutive waves)
3. For non-consecutive waves apply the rule:
   - If waves' number difference is 2, increase age by 1 year
   - If waves' number difference is 3 or 4, increase age by 2 years
   - If waves' number difference is 5, increase age by 3 years

The gender variable was collected by asking the following question:

C	**Is ..... a male or a female?** 1 = Male 2 = Female

Valid range: 1 – 2
Unspecified: 9
Missing values: (.)
Dots (.) were replaced with 9

Rules used to edit gender for panel, including unspecified (9):
If there is a difference in gender for the same individual in the panel, but age, population group, language and education level match, change gender and make it similar in all waves (we agree it is riskier for a match of two waves).

The population group variable was collected by asking the following question:

E	**What population group does ..... belong to?** 1 = African/Black 2 = Coloured 3 = Indian/Asian 4 = White 5 = Other (*Specify*)

Valid range: 1 – 4
Unspecified: 9
Missing values: (.)
Dots (.) were replaced with 9
Code 5 was replaced with code 2

Rules used to edit population group for panel (perplexity in coloured population definition):
When the population group in the panel is different, but person age, gender match, population group has codes 1, 2 or 4, and language has codes 1 or 2, we can assume the record reflects the same person and thus assign the most frequent population group

value in the waves. In cases where the person matches in two waves only, record the population group value that is shown in the previous wave.

The marital status variable was collected by asking the following question:

1.1.a	**What is ……'s present marital status?** 1 = Married or living together as husband and wife 2 = Widow/Widower 3 = Divorced or Separated → *Go to Q 1.2* 4 = Never married

Valid range: 1 – 4
Unspecified: 9
Missing values: (.)
Dots (.) were replaced with 9

Rules used to edit marital status in panel:
When marital status in the panel is different from wave to wave, but the person's record is acceptable for age, gender, language and population group, marital status code 4 – never married – cannot follow marital status codes 2 or 3, therefore the person is assigned a marital status the same as the first of the matched waves.

The language variable was collected by asking the following question:

1.2	**Which language does …… speak most often at home?** 01 = Afrikaans 02 = English 03 = IsiNdebele/South Ndebele/North Ndebele 04 = IsiXhosa/Xhosa 05 = IsiZulu/Zulu 06 = Sepedi/Northern Sotho 07 = Sesotho/Southern Sotho/Sotho 08 = Setswana/Tswana 09 = SiSwati/Swazi 10 = Tshivenda/Venda 11 = Xitsonga/Tsonga 12 = Other (*Specify)*

Valid range: 1 – 12
Unspecified: 99
Missing values: (.)
Dots (.) were replaced with 99
Codes 12 and 99 were replaced with codes that show most frequently in the waves.

Rules used to edit language variable in the panel:
When the language in the panel is not the same for the same person, but the person record is acceptable for age and gender, and population group code is 1, 2 or 4 and the language code is 1 or 2, the person is assigned a language that is common across the waves. For a match of 2 waves only, the language of the first of the matched waves is assigned to the person. When the language in the panel is different for the same person, but the person record is acceptable for age and gender, and the population group code is 1 and the language code is between 3 and 11, the person is assigned a language that is common across the waves. For a match of 2 waves only, the language

of the first of the matched waves is assigned to the person. When the language in the panel is different for the same person, but the person record is acceptable for age and gender, and the population group code is 3 and at least one language is English (code 2), change the language to English in all the waves.

Education variable was collected by asking the following question:

1.3.a	**What is the highest level of education that …… has completed?** 00 = No schooling 01 = Grade 0 02 = Sub A/Grade 1 03 = Sub B/Grade 2 04 = Grade 3/Standard 1 05 = Grade 4/Standard 2 06 = Grade 5/Standard 3 07 = Grade 6/Standard 4 08 = Grade 7/Standard 5 09 = Grade 8/Standard 6/Form 1 10 = Grade 9/Standard 7/Form 2 11 = Grade 10/Standard 8/Form 3 12 = Grade 11/Standard 9/Form 4 13 = Grade 12/Standard 10/Form 5/Matric 14 = NTC l 15 = NTC II 16 = NTC III **17** = Diploma/certificate with less than Grade 12/Std 10 **18** = Diploma/certificate with Grade 12/Std 10 **19** = Degree **20** = Postgraduate degree or diploma 21 = Other (*Specify in column)* 22 = Don't know

Valid range: 0 – 22
Unspecified: 99
Missing values: (.)
Dots (.) were replaced with 99

The level of education must be consistent with the person's age:

Minimum age level	
5	00
5	01
6	02
7	03
8	04
9	05
10	06
11	07
12	08
13	09
14	10
15	11
16	12
16	13
16	14

16	15
16	16
18	17
18	18
20	19
21	20

Codes 21, 22 and 99 were replaced with codes that show most frequently in the waves.

Rules used to edit education in the panel:
For non-consecutive waves and for persons aged 20 or less and education has more than 3 levels of difference:

- If waves' number difference is 2, education increases by one level
- If waves' number difference is 3 or 4, education increases by 2 levels
- If waves' number difference is 5, education increases by 3 levels

For consecutive waves, if the education level has more than 1 level difference but all fixed variables are the same, change education level to the nearest valid education code. When the person is more than 65 years old and education level is not the same, adopt the education level shown in the first wave where the person appears.

No editing was done for the rest of the variables.

5.3 Renaming of variables

To match the same questions correctly for consecutive surveys, some changes of variable names were done as follow:

From LFS(8) Person file for Wave 5 :
Q16aHiEd to Q13aHiEd
Q16bArea to Q13bArea
Q17SklTr to Q14SklTr
Q18LngTr to Q15LngTr
Q19FldTr to Q16FldTr
Q110aRea to Q17aRead
Q110bWri to Q17bWrit
Q111EduI to Q18EduIn
Q112Full to Q19FulPa
Q113DLrn to Q110dLrn
Q114FetW to Q111FetW
Q115HrsW to Q112HrsW
Q116FetD to Q113FetD
Q117HrsD to Q114HrsD

From LFS(1) and LFS(2) Worker File for Wave 1 and Wave 2:
Q32Accep to Q35Accep
Q33Whnst to Q36Whnst
Q34aLook to Q37Look
Q34bBgnB to Q37BgnB
Q35seekh to Q38seekh

Q36TimSe to Q39TimSe
Q37YnotS to Q310RsnN
Q38EverW to Q311EverW
Q39TmGap to Q312TmGap
Q310Occu to Q313Occu
Q311Indu to Q314Indu
Q312pens to Q315pens

5.4 Table for matching variables in LFSs

Name of variable	Flag of match	Name of variable	Flag of match
Age	111111	Q23rsnab	000001
Agegrp	111111	Q24Start	111111
B_4Night	001110	Q310Occu	110000
Gender	111111	Q310RsnN	001111
IncmGrp	111111	Q311Ever	001111
Indus	111111	Q311Indu	110000
Indusp	110111	Q312TmGa	001111
Occup	111111	Q312char	110000
Occuprev	010111	Q312inhh	110000
PSU	000111	Q312noth	110000
PersonNr	111111	Q312oddj	110000
Popgrp	111111	Q312oths	110000
Prov	111111	Q312pens	110000
Q13aHiEd	001101	Q312savi	110000
Q13bArea	111101	Q312uifs	110000
Q14SklTr	111101	Q313Occu	001111
Q16aHiEd	000010	Q314Indu	001111
Q16bArea	000010	Q315Char	001111
Q17SklTr	000010	Q315OddJ	001111
Q20SelfR	111111	Q315Othr	001111
Q21aOwnB	111111	Q315Pens	001111
Q21bPaid	111111	Q315Savi	001111
Q21cDome	111111	Q315UIFS	001111
Q21dUnPd	111111	Q315inHH	001111
Q21eFarm	111111	Q315notH	001111
Q21fCons	111111	Q31YnotW	111111
Q21gCtch	111111	Q32Accep	110000
Q21hBeg	111111	Q32JobOf	001111
Q22HaveW	111111	Q33Whnst	110000
Q23Mainr	000010	Q33WrkRe	001111
Q23RsnAb	001100	Q34BgnBu	010000
Q23RsnaB	110000	Q34Lookw	010000
Q34WhyNo	001000	Q426HntX	110000
Q34aLook	100000	Q426Seek	001100
Q34bBgnB	100000	Q426Seekk	000011
Q34notwi	000100	Q427TypX	111111
Q34whyno	000011	Q42Indus	111111
Q35Accep	001111	Q43MWork	001111
Q35seekh	110000	Q43Mwork	110000
Q36TimSe	110000	Q44NrEmp	111111
Q36WhnSt	001111	Q45MnthS	001111
Q37BgnBs	001010	Q45Mnths	110000

Q37BgnBu	000101	Q45Years	111111
Q37LookW	001111	Q46LengJ	111111
Q37YnotS	110000	Q47Tools	111111
Q38EverW	110000	Q48WrtnC	001111
Q38SeekH	001100	Q48Wrttn	110000
Q38SeekHh	000010	Q49Super	111111
Q38Seekhh	000001	Q51FarmA	001111
Q39TimSe	001100	Q51FarmY	110000
Q39TimSee	000011	Q52Apr01	110000
Q39TmGap	110000	Q52Apr02	001100
Q410WhoP	111111	Q52Apr03	000011
Q411Pens	111111	Q52Aug01	110000
Q412Leav	111111	Q52Aug02	001100
Q413TrdU	111111	Q52Aug03	000011
Q414TypB	111100	Q52Dec00	100000
Q414TypBb	000011	Q52Dec01	011000
Q415aSal	111111	Q52Dec02	000110
Q415bSal	111111	Q52Dec03	000001
Q415cSal	111111	Q52Feb01	100000
Q416NrWo	111111	Q52Feb02	011000
Q417aReg	111111	Q52Feb03	000110
Q417bUIF	111111	Q52Feb04	000001
Q417cMed	111111	Q52Jan01	100000
Q417dVat	111111	Q52Jan02	011000
Q418Sect	001111	Q52Jan03	000110
Q418sect	110000	Q52Jan04	000001
Q419Loca	111100	Q52Jul01	110000
Q419Locaa	000011	Q52Jul02	001100
Q41Occup	111111	Q52Jul03	000011
Q420aHrs	111111	Q52Jun01	110000
Q420bHrs	111111	Q52Jun02	001100
Q420cHrs	111111	Q52Jun03	000011
Q421aHrs	111111	Q52Mar01	110000
Q421bHrs	111111	Q52Mar02	001100
Q421cHrs	111111	Q52Mar03	000011
Q422Fixe	001111	Q52May01	110000
Q422Flex	110000	Q52May02	001100
Q423More	111111	Q52May03	000011
Q424Star	001111	Q52Nov00	100000
Q424WhnX	110000	Q52Nov01	011000
Q425HntX	110000	Q52Nov02	000110
Q425look	001111	Q52Nov03	000001
Q52Oct00	100000	Q55aVday	101111
Q52Oct01	011000	Q55bMedC	101111
Q52Oct02	000110	Q55cTrai	101111
Q52Oct03	000001	Q55dLawk	101111
Q52Sep00	100000	Q55eComu	001111
Q52Sep01	011000	Q55eMain	100000
Q52Sep02	000110	Q55fOrgE	001111
Q52Sep03	000001	Q55fOrge	100000
Q53FarmW	001111	Q55gColl	101111
Q53YFarm	110000	Q55hFund	101111
Q54Volun	111111	Q55iOthr	101111
Q55Collm	010000	Sector	111111

Q55Fundr	010000	Status1	111111
Q55Lawke	010000	Status2	111111
Q55Maint	010000	Stratum	001111
Q55MedCA	010000	Type	111000
Q55Orgev	010000	UqNr	111111
Q55Othr	010000	Worker_wg	000100
Q55Train	010000	Worker_wgt	111011
Q55Vdayh	010000	Indusp	001000
		Occuprev	101000

The column 'Name of variable' in the table above shows the name of the variable as it was found in the surveys. The column 'Flag of match' represents a flag where '0' means no such name was found in the survey and a flag '1' is the presence of variables in the survey. The position of every digit of the flag means consecutive number of the survey: 1st position – LFS survey (4), 2nd – LFS survey (5), etc. The table was used to determine variables to be included in panel.

5.5 Table for matching of data files in LFS surveys

Consecutive number of LFS survey	**Date of the survey**		**Files names, x-missing in survey**					
	Year	**Month**	**1**	**2**		**3**	**4**	**5**
4	2001	September	HOUSE	PERSON	X	WORKER	x	x
5	2002	February	X	PERSON	(Health data)	WORKER	x	DEATH
6	2002	September	HOUSE	PERSON	X	WORKER	MIGRANT	x
7	2003	March	X	PERSON	(Health data)	WORKER	x	x
8	2003	September	HOUSE	PERSON	X	WORKER	MIGRANT	x
9	2004	March	HOUSE	PERSON	X	WORKER	x	x

Only PERSON and WORKER files were chosen for panel surveys. Health data were part of the Person files in LFS(5) and LFS(7) only and were excluded from the panel.

6. The Labour Force Survey questionnaires

Data are collected around a number of main themes:

- Demographics and education
- Labour market participation, jobs and income
- Migration, health and death
- Housing and access to services

6.1 Survey instruments

Data are collected through one instrument which has the following sections:

- A household coversheet, which tracks addresses and respondent status of the household (vacant, non-contact, etc.).
- A flap or an enumeration grid, collecting individual information about all household members by main respondents and proxy information on other members of the household. The questionnaire allows for data of 10 people.

- A labour market activities section intended for all household members aged 15 and older with detailed questions on economic or non-economic activities in the last 7 days.
- A household section, for data at that level.

6.2 Questionnaire stability

In general, questions remain fixed across the panel waves, otherwise the usefulness of the panel is compromised. However, there are occasions where the questions differ. The labour market questions were kept fixed between waves 1 and 2. They remained fixed again from waves 3 to 6 after some changes were made in wave 3. Three new questions were introduced in wave 3, section 3 of the questionnaire. The new questions were:

Q3.2 Has …… been offered any job in the past 6 months and has he/she turned it down?

1 = Yes
2 = No

Q3.3 Does …… know of any available work for which he/she has the relevant qualification but is not willing to do?

1 = Yes
2 = No

Q3.4 Why would …… not be willing to do this work?

1 = Wages too low
2 = Job is not permanent
3 = Location (work too far)
4 = Working conditions unsuitable
5 = Work not legal
6 = Other

The household questions were kept fixed between waves 1, 2 and 4 except for a few additional questions introduced in 4, increasing the number of household-related questions from the original 35 questions in wave 1 to 39 questions in wave 4. In wave 6 the household-related questions were limited to 9 questions with only 7 questions from the original list of questions being kept. Four new questions were introduced in wave 4, section 7 of the questionnaire. The new questions were:

Q7.5 What is the amount of rent paid or value of rent (if rent-free) for this dwelling?

1. Total rent paid or value of rent, if rented free, for this dwelling
    - Amount paid by you excluding amount subsidised, or value of rent, if rent-free
    - Amount subsidised (e.g. by employer)
    - Amount paid for garage and/or domestic worker's room if rented separately

Q7.6 Is this a

1 = Weekly amount?
2 = Fortnightly amount?
3 = Monthly amount?
4 = Annual amount?

Q7.7 Does the rent include

1 = Electricity?
2 = Water?
3 = Garage/parking space?
4 = Refuse removal?
5 = Other, specify

Q7.8 Is this dwelling rented with or without furniture?

1 = Unfurnished
2 = Semi-furnished
3 = Furnished

The migrant worker questions were kept fixed between waves 3 and 4. In addition a question was included in wave 2 and carried through to all the waves to check whether a household member was part of the current household in the previous survey.

QB. 2 Did ..... stay in this household in September 2001?

1 = Yes
2 = No

This question applied to each person found in the selected dwelling. This question was aimed at identifying permanent household members within the selected dwelling. However, the reference period for wave 3 remained the same whereas it should have been changed to March 2002. In wave 4, the reference period changed to September 2002; in wave 5 to March 2003 and in wave 6 to September 2003.

6.3 Inconsistencies in response categories

There was some inconsistency in the responses categories between waves that had to be adjusted before comparisons were made.

- In question 1.8 a new option was introduced in wave 6: 'pre-school/crèche'.
- In question 3.1 the order of the option was changed slightly.
- In question 3.1, a new option 'cannot find work' was introduced.
- In question 3.1, the option 'recently retrenched' was changed to 'retrenched'.
- In question 3.5 (waves 1 and 2) which became question 3.8 in the rest of the waves, option 6: 'sought/underwent training' was dropped.
- In question 3.8 (waves 1 and 2) which became question 3.11 in the rest of the waves, the question: **Has …… ever worked before?** was rephrased to: **Has …… ever worked for pay, profit or family gain?**

6.4 Questions dropped as low priority

Questions for waves 1 and 6 and questions for waves 3 and 5 were fairly similar. However, some of the questions deemed as low priority were dropped from the household information in wave 6. These were questions on the type of facility and resources available to the household (questions 6.6 to 6.33 in wave 1).

- Questions on household information were dropped in waves 2 and 4.
- Questions 10 to 37 on household information were dropped in waves 6.

Additional cross-sectional information

- Questions on death in wave 5, questions on health in waves 2 and 4, questions on migrant workers in waves 3 and 5 were added.
- In addition, the following questions were posed in wave 5:

  Q 1.3 Five years ago (in September 1998), was … living in this area (i.e. this suburb, ward, village, farm, informal settlement)?

  1 = Yes
  2 = No
  3 = Born after September 1997
  4 = Don't know

  Q 1.4 Where did … move from?

  (If more than one move, give details of the last move)

  - Province
  - Main place (e.g. city, town, tribal area, administrative area, etc.)
  - If from another country, name of the country

  Q 1.5 In which year did … move to this place?

The following question was posed in wave 6 only:

Q 6.12 What is the type of these living quarters?

1 = Private dwelling
2 = Workers' hostel

7. Changes in the definition of the labour market variable

Stats SA uses two measures of unemployment, the official and the expanded definition. One of the conditions for a person to be classified as unemployed is that he/she should be available to take up employment during the reference period. Up until September 2003 (LFS8), Stats SA has been using a reference period of one week for this criterion. However, not everyone who is seeking work can be expected to take up a job immediately when it is offered. A person could be temporarily sick, or may have to make arrangements concerning childcare. Thus to be able to include more people, Stats SA increased the availability period from one week to two weeks for both the official and expanded definition of unemployment, starting from March 2004 (LFS9). However, the difference in the unemployment numbers according to the two availability reference periods is very small and thus the effect on the unemployment rate is negligible.

7.1 Unemployment old definition (official)

To be classified as unemployed, a person has to satisfy three conditions:

1. The person did not work during the seven days prior to the survey interview and does not have any job attachment
2. The person wants to work and is available to start within a week
3. The person has taken active steps to look for work or to start his/her own business in the four weeks prior to the interview.

7.2 Unemployment new definition (official)

To be classified as unemployed, a person has to satisfy three conditions:

1. The person did not work during the seven days prior to the survey interview and does not have any job attachment
2. The person wants to work and is available to start within two weeks
3. The person has taken active steps to look for work or to start his/her own business in the four weeks prior to the interview.

7.3 Additional changes

Minor changes were introduced in the definition of the labour market variables in March 2003 (LFS7). The first change was the introduction of an additional probe into the reason why the person did not work in the past seven days:

(09 = Cannot find any work)

The second change was to classify whoever replied 'yes' to the question below as employed.

2.2	**Even though …… did not do any of these activities in the last seven days, does he/she have a job, business, or other economic or farming activity that he/she will definitely return to?**

	For agricultural activities, the off season in agriculture is not a temporary absence. 1 = Yes 2 = No	→*Go to Section 3*

Subsequently all previous LFSs were released in September 2005 with data re-weighted to Census 2001 and with the new labour market definitions.

7.4 SAS code used for derivation of old employment status definition

Status1 codes: 0 = Not economically active
1 = Employed
2 = Unemployed

- LFS September 2001 (LFS4)–

```
status1=0;
      if Q21aOwnB=1 or Q21bPaid=1 or Q21cDome=1 or Q21dUnPa=1 or
Q21eFarm=1 or Q21fCons=1 or Q21gCtch=1 then Status1=1;
      if Q22HaveW=1 and Q23Absnt ne 99 and Q23Absnt<10 and Q23Absnt
ne 6) then Status1=1;
      if Q22HaveW = 1 and  (Q23Absnt=6 or Q23Absnt>9) then Status1=2;
      if  (1<=Q31YnotW<=2) or  (8<=Q31YnotW<=12) then Status1=2;
      if Status1=2 and (Q32Accep>=2 or Q33WhenS>1) then Status1=0;
      if Status1=2 and (Q34aLook ne 1 and Q34bBgnB ne 1) then
status1=0;
run;
```

- LFS February 2002 (LFS5)–

```
status1=0;
      if Q21aOwnB=1 or Q21bPaid=1 or Q21cDome=1 or Q21dUnPa=1 or
Q21eFarm=1 or Q21fCons=1 or Q21gCtch=1 then Status1=1;
      if Q22HaveW=1 and  (Q23Absnt ne 99 and Q23Absnt<10 and Q23Absnt
ne 6) then Status1=1;
      if Q22HaveW=1 and (Q23Absnt=6 or Q23Absnt>9) then Status1=2;
      if (1<=Q31YnotW<=2) or (8<=Q31YnotW<=12) then Status1=2;
      if Status1=2 and (Q32Accep>=2 or Q33WhenS>1) then Status1=0;
      if Status1=2 and (Q34aLook ne 1 and Q34bBgnB ne 1) then
status1=0;
      run;
```

- LFS September 2002 (LFS6)–

```
status1=0;
      if Q21aOwnB=1 or Q21bPaid=1 or Q21cDome=1 or Q21dUnPd=1 or
Q21eFarm=1 or Q21fCons=1 or Q21gCtch=1 then Status1=1;
      if Q22HaveW=1 and (Q23RsnAb ne 99 and Q23RsnAb<10 and Q23RsnAb
ne 6) then Status1=1;
      if Q22HaveW=1 and  (Q23RsnAb=6 or Q23RsnAb>9) then Status1=2;
      if (1<=Q31YnotW<=2) or  (8<=Q31YnotW<=12) then Status1=2;
      if Status1=2 and (Q35Accep>=2 or Q36Whnst>1) then Status1=0;
      if Status1=2 and (Q37LookW ne 1 and Q37BgnBs ne 1) then
status1=0;
run;
```

- LFS March 2003 (LFS7)–

```
Status1=0;
      if q21aownb=1 or q21bpaid=1 or q21cdome=1 or q21dunpd=1 or
q21efarm=1 or  q21fcons=1 or q21gCtch=1 then status1=1;
```

```
      if q22havew='1' and (q23rsnab ne 99 or q23rsnab<10 and q23rsnab
ne 6) then status1=1;
      if q22havew='1' and (q23rsnab=6 or q23rsnab>9) then status1=2;
      if (q31YnotW=1) or ( 7<=q31YnotW<=13) then status1=2;
      if status1=2 and (q35accep>=2 or q36whnst>1) then status1=0;
      if status1=2 and (q37lookw=2 and q37bgnbu=2) then status1=0;
      if q31YnotW=1 then status1=2;
run;
```

- LFS September 2003 (LFS8)–

```
      if q21aownb=1 or q21bpaid=1 or q21cdome=1 or q21dunpd=1 or
q21efarm=1 or  q21fcons=1 or q21gctch=1 then status1=1;
      if q22havew=1 and (q23mainr ne 99 or q23mainr<10) and q23mainr
ne 6  then status1=1;
      if q22havew=1 and (q23mainr=6 or q23mainr>9) then status1=2
      if (q31YnotW=1) or ( 7<=q31YnotW<=13) then status1=2;
      if status1=2 and (q35accep>=2 or q36whnst>1) then status1=0;
      if status1=2 and (q37lookw=2 and q37bgnbs=2) then status1=0;
      if q31YnotW=1 then status1=2;
run;
```

- LFS March 2004 (LFS9)–

```
reason=q31ynotw+0;
status1=0;
      if q21aownb=1 or q21bpaid=1 or q21cdome=1 or q21dunpd=1 or
q21efarm=1 or q21fcons=1 or q21gctch=1 or q22havew=1 then status1=1;
      if reason=1 or (7<=reason<=13) then status1=2;
      if status1=2 and (q35accep=>2 or q36whnst>2) then status1=0;
      if status1=2 and (q37lookw>1 and q37bgnbu>1) then status1=0;
      if reason=1 then status1=2;
run;
```

7.5 SAS code used for derivation of new employment status definition

- LFS September 2001 (LFS4)–

```
reason=q31ynotw+0;
status1=0;
if q21aownb=1 or q21bpaid=1 or q21cdome=1 or q21dunpd=1 or q21efarm=1
or  q21fcons=1 or q21gctch=1 or q22havew=1 then status1=1;
if reason=1 or reason=2 or (8<=reason<=12) then status1=2;
if status1=2 and (q32accep=>2 or q33whnst>2) then status1=0;
if status1=2 and (q34alook>1 and q34bbgnb>1) then status1=0;
if reason=1 then status1=2;
run;
```

- LFS February 2002 (LFS5)–

```
reason=q31ynotw+0;
status1=0;
     if q21aownb=1 or q21bpaid=1 or q21cdome=1 or q21dunpd=1 or
q21efarm=1 or  q21fcons=1 or q21gctch=1 or q22havew=1 then status1=1;
  if reason=1 or reason=2 or (8<=reason<=12) then status1=2;
  if status1=2 and (q32accep=>2 or q33whnst>2) then status1=0;
  if status1=2 and (q34lookw>1 and q34bgnbu>1) then status1=0;
  if reason=1 then status1=2;
run;
```

- LFS September 2002 (LFS6)–

```
reason=q31ynotw+0;
status1=0;
    if q21aownb=1 or q21bpaid=1 or q21cdome=1 or q21dunpd=1 or
q21efarm=1 or q21fcons=1 or q21gctch=1 or q22havew=1 then status1=1;
    if reason=1 or (7<=reason<=13) then status1=2;
    if status1=2 and (q35accep=>2 or q36whnst>2) then status1=0;
    if status1=2 and (q37lookw>1 and q37bgnbs>1) then status1=0;
    if reason=1 then status1=2;
run;
```

- LFS March 2003 (LFS7)–

```
reason=q31ynotw+0;
status1=0;
     if q21aownb=1 or q21bpaid=1 or q21cdome=1 or q21dunpd=1  or
q21efarm=1 or  q21fcons=1 or q21gctch=1 or q22havew=1 then status1=1;
     if reason=1 or (7<=reason<=13) then status1=2;
     if status1=2 and (q35accep=>2 or q36whnst>2) then status1=0;
     if status1=2 and (q37lookw>1 and q37bgnbu>1) then status1=0;
     if reason=1 then status1=2;
run;
```

- LFS September 2003 (LFS8)–

```
reason=q31ynotw+0;
status1=0;
     if q21aownb=1 or q21bpaid=1 or q21cdome=1 or q21dunpd=1 or
q21efarm=1 or  q21fcons=1 or q21gctch=1 or q22havew=1 then status1=1;
     if reason=1 or (7<=reason<=13) then status1=2;
     if status1=2 and (q35accep=>2 or q36whnst>2) then status1=0;
     if status1=2 and (q37lookw>1 and q37bgnbs>1) then status1=0;
     if reason=1 then status1=2;
run;
```

- LFS March 2004(LFS9)–

```
reason=q31ynotw+0;
status1=0;
     if q21aownb=1 or q21bpaid=1 or q21cdome=1 or q21dunpd=1 or
q21efarm=1 or  q21fcons=1 or q21gctch=1 or q22havew=1 then status1=1;
     if reason=1 or (7<=reason<=13) then status1=2;
     if status1=2 and (q35accep=>2 or q36whnst>2) then status1=0;
     if status1=2 and (q37lookw=>2 and q37bgnbu=>2) then status1=0;
     if reason=1 then status1=2;
run;
```

8. Data structures and files

Each wave has data about:

- respondent individuals
- children
- other non-respondent individuals

The list of variables provided by the Harvard group about individuals is stored in two files at each wave:

- **w_person**: Data from Flap and Section 1
- **w_employment**: Data from Sections 2, 3, 4 and 5

The key index variables for matching and identifying individuals when using them as *cross-sectional* data are (UqNr) which constitutes a unique household identifier and Person number (PersonNr). The key index variable for matching and identifying individuals when using *longitudinal* data is UQ, which has been created subsequently.

The data files contain no variables representing sampling weights for the panel since at this stage it is not possible to determine meaningful sampling weights for the panel sample. Anyone who is interested in arriving at estimates for finite populations of individuals and households can use the entire sample for each cross section.

The complexity of the household panel design means that we cannot effectively include all data in a single flat file (it would be too large and unwieldy to handle anyway), nor does it fit into a straightforward hierarchical pattern. Therefore the data are included in a set of separate files for each wave, reflecting the person and workers files. These files would be linked by matching the UQ variable. The matching operations are also facilitated by some basic naming conventions:

1. Wave-specific variables and files have a wave specific prefix and suffix. Thus the root file and variable name is constant across waves (say w1_person, w2_person, …, for each file; and say age1, age2,…for the variables).
2. Data are released using the same file structures for all software (AscII, STATA).
3. In addition to variables from the questionnaire, there are some derived variables.

9. Charts to compare LFS data files and Panel data files

Chart 1: LFS Statistical Release P0210, Figure 3: Unemployment rates by gender (official definition of unemployment):

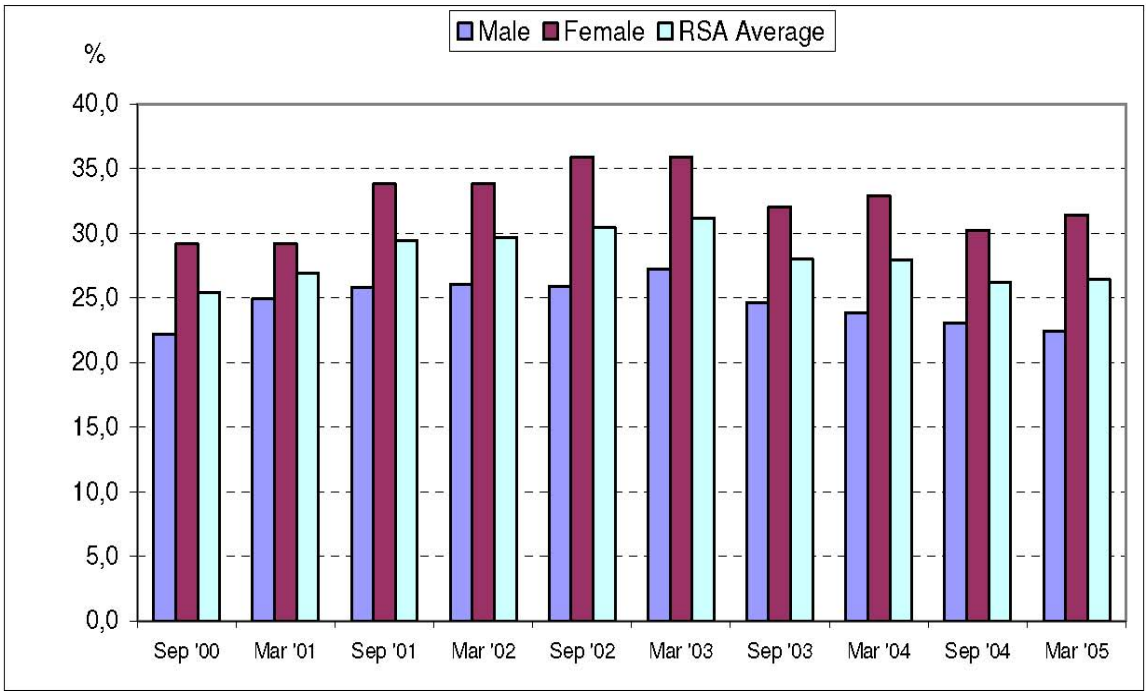


Chart 2: LFS Panel, Unemployment rates by gender (official definition of unemployment) for waves 1 to 6

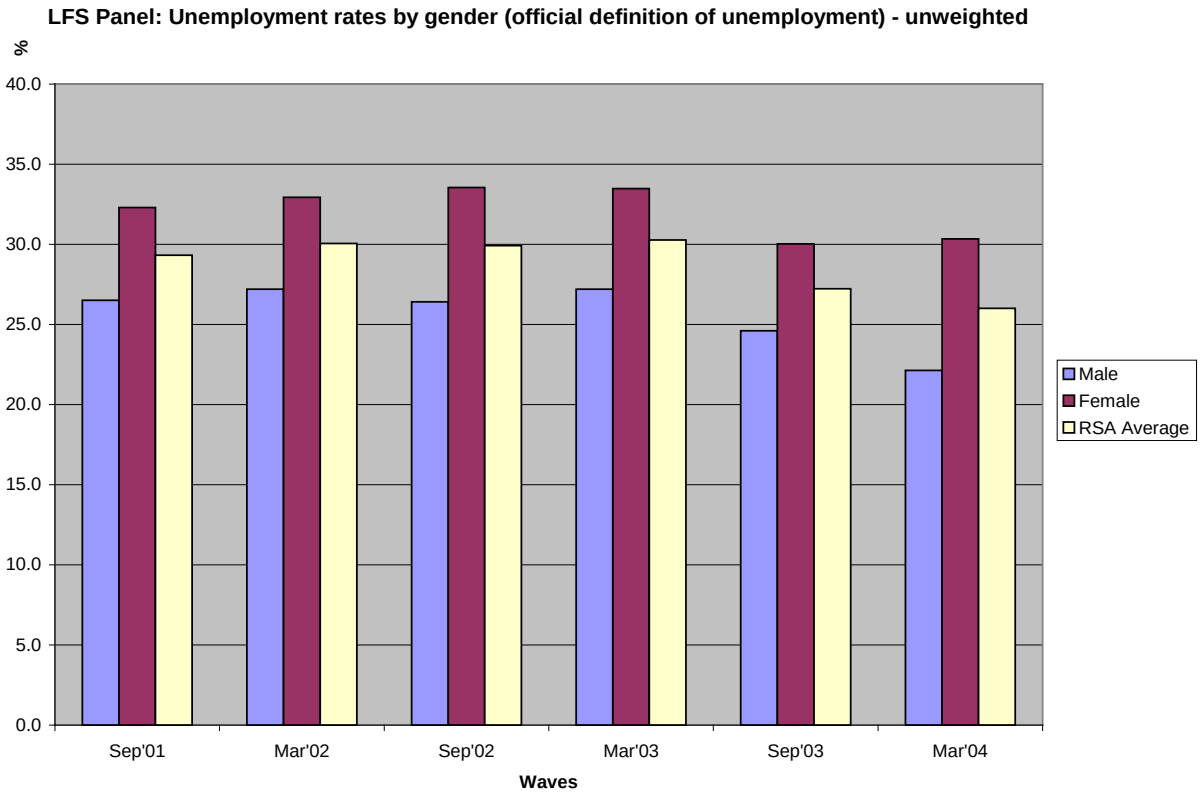


Chart 3: Age distribution in LFS surveys from September 2001 to March 2004 – unweighted

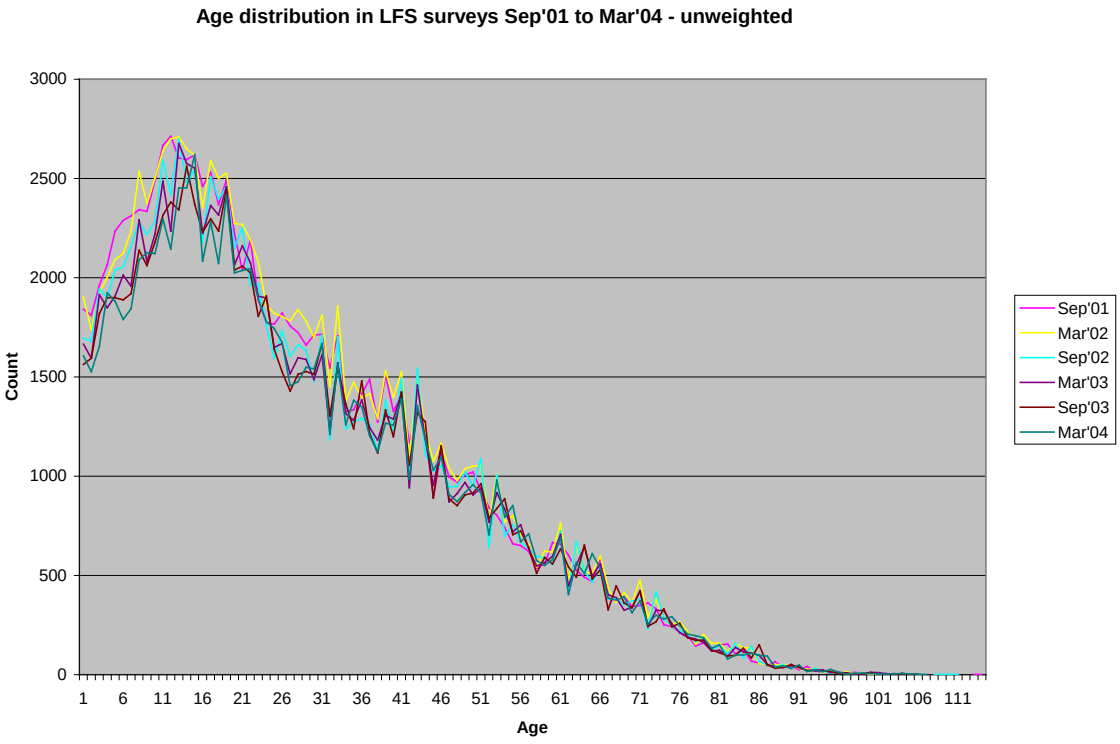
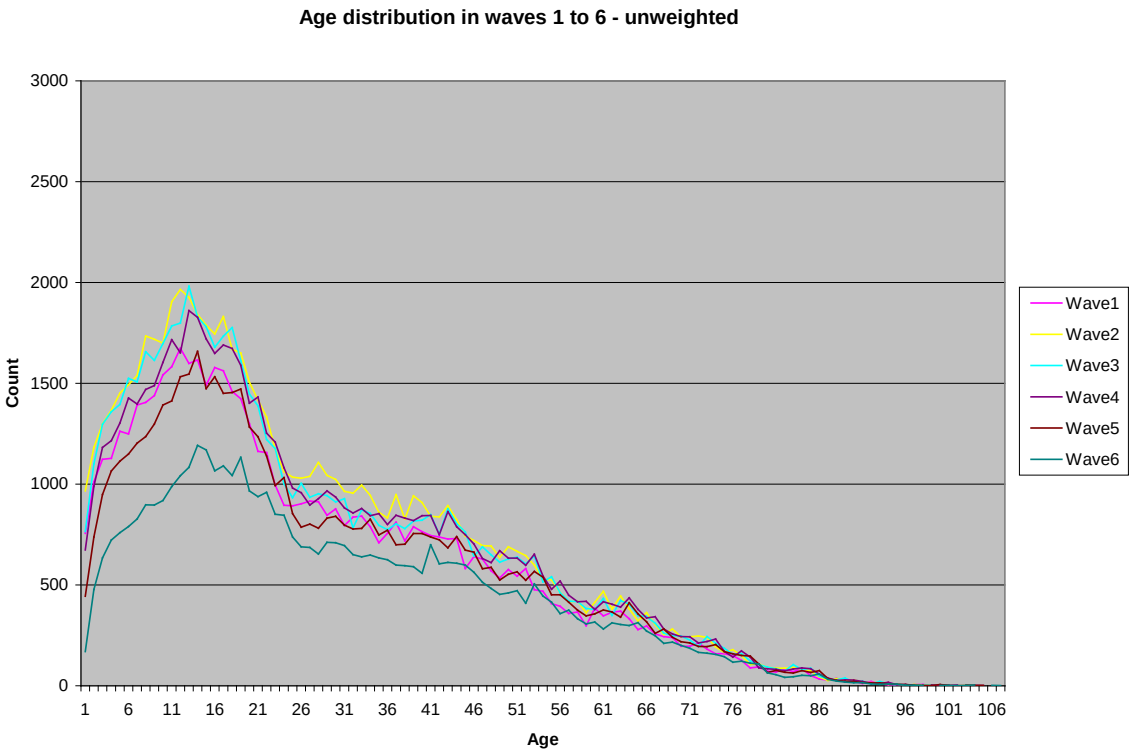


Chart 4: Age distribution in waves 1 to 6 – unweighted



10. Annexure

10.1 Other information

Other important information on the cross-sectional Labour Force Surveys is found in the:

- Questionnaire
- The metadata for each cross section and additional code lists (occupation, industry)
- Previous publications for each cross section
- Stats SA website: (http://www.statssa.gov.za/)

For additional information contact:

Tel: +27 (12) 310-8600
Fax: +27 (12) 310-8500
Email: info@statssa.gov.za
Website: www.statssa.gov.za