

https://dspace.org/

An investigation of the consistency of Statistics South Africa's employment data between surveys

Lukhwareni, Joseph

2012-01-31
http://hdl.handle.net/10539/11204

Downloaded from a DSpace repository. https://dspace.org/

An investigation of the consistency of Statistics South Africa's employment data between surveys

Joseph Lukhwareni

Student number: 0400363F

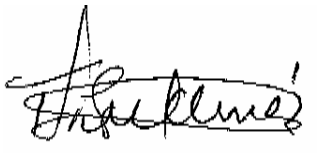
A research report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science.

Supervisor: Prof. J. Galpin

Johannesburg, 2011

DECLARATION

I declare that this research proposal is my own, unaided work. It is being submitted for the Degree of Master of Science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.



_____

(Signature of candidate)

10 October 2011

ABSTRACT

The purpose of the study is to investigate possible reasons as to why different surveys conducted by Statistics South Africa (Stats SA) give different estimates of the percentages in the different employment categories. In order to investigate the different sources of variability, that is, surveys done in different years, surveys using different questionnaires, different sample designs and different employment profiles, the following comparisons were done for Gauteng and the Eastern Cape:

- To compare estimates of employment status over time for the March Labour Force Survey (LFS) 2006 and 2007; September LFS 2006 and 2007; and General Household Survey (GHS) September 2006 and July 2007.
- To compare estimates of employment status across surveys for LFS September 2006; GHS September 2006; and LFS September 2007, July GHS 2007 and Community Survey (CS) October 2007.

In order to generate a set of comparable estimates across surveys and within surveys over time, this study identifies and addresses the various sources of potential non-comparability. The methodologies utilised are Chi-squared Automatic Detection (CHAID) and multinomial logistic regression. These statistical techniques were used to identify variables which are associated with employment status.

The predictor variables included in the analysis are age group, highest level of education, marital status, population group, sex and source data. The results from CHAID for all data sets show that age group is the most significant predictor on which data on employment status can be segmented. At the root node (the first level of the CHAID tree), data was partitioned by the categories of age group. Highest level of education, sex, population group and province were significant within the categories of age group. Either province or population group was significant within the age group 20–29 years old depending on the data that is being analysed. Sex was most significant within the age group 50–65 years old.

The results of multinomial regression show several significant interactions involving from five to seven factors for different data sets. The logistic regression results were not as good as those of the CHAID analyses, but both techniques give us an indication of the relationships between the predictor variables and employment.

The analysis of the CS, LFS and GHS in 2007, when explaining employment status, split on age group. Highest level of education was the most significant predictor when comparing the three data sets. There are differences among the three data sets when explaining employment status. This is due to the use of different mid-year population estimates, differences in the instructions given in the questionnaire for CS 2007 and other surveys, as well as the sample size of the surveys. There are indeed significant differences between Gauteng and Eastern Cape in relation to employment status.

DEDICATION

This report is dedicated to my two precious and loving daughters:
Nduvho and Prudent Lukhwareni

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Jacky Galpin, from the School of Statistics and Actuarial Science at Wits University for her guidance and encouragement. Many thanks to Statistics South Africa for providing such an opportunity, as well as to colleagues and friends for their support.

Lastly, my warm gratitude to my wife Patience, my two daughters for their support and understanding during my study period.

CONTENTS

DECLARATION ..... **ii**

DEDICATION ..... **iv**

ACKNOWLEDGEMENTS ..... **v**

CONTENTS ..... **vi**

LIST OF TABLES ..... **viii**

LIST OF FIGURES ..... **x**

ACRONYMS ..... **xi**

CHAPTER 1: Introduction and Problem Statement **1**

1.1 Introduction 1

1.2 Definitions of concepts 2

1.3 Data sources 3

1.4 Objectives of the study 7

1.5 Structure of the research report 8

CHAPTER 2: Literature Review **9**

2.1 Data quality issues 9

2.2 Specific issues for surveys of employment 11

2.3 Literature on the methodology used to compare surveys 15

2.4 Summary 16

CHAPTER 3: Methodology and data analysis **18**

3.1. Data 18

3.2. CHAID 20

3.2.1 Introduction 20

3.2.2 Basic tree-building algorithm 21

3.2.3 Testing the significance of each predictor variable 21

3.2.4 Illustration of the CHAID methodology using March LFS 2006 and 2007 22

3.3 Multinomial logistic regression 28

3.3.1 Introduction 28

3.3.2 Multinomial logistic regression model 28

3.3.3 Fitting the multinomial regression model 30

3.3.4 Interpretation of the fitted model 31

3.3.5 Assessing the adequacy of the predicted model 33

3.3.6 Illustration of logistic analysis methodology using March LFS 2006 and 2007 33

3.4 Discussion of the CHAID and logistic regression results 37

CHAPTER 4: Results **39**
4.1 September LFS 2006 and 2007 39
4.1.1 Results for CHAID (employment status as response variable) 39
4.1.2 Results for multinomial logistic regression (employment status as response variable) 43
4.1.3 Discussion of the CHAID and logistic regression results (employment status as response variable) 47
4.1.4 March LFS 2006 and 2007 and September LFS 2006 and 2007 (source data as response variable) 49
4.1.5 Discussion of the CHAID and logistic regression results (March LFS 2006 and 2007; September LFS 2006 and 2007) 51
4.2 September GHS 2006 and July GHS 2007 52
4.2.1 Results for CHAID (employment status as response variable) 52
4.2.2 Results for multinomial logistic regression (employment status as response variable) 55
4.2.3 Discussion for CHAID and logistic regression (employment status as response variable) 59
4.2.4 September GHS 2006 and July GHS 2007 (source data as response variable) 60
4.3 September LFS 2006 and September GHS 2006 by employment status 61
4.3.1 Results for CHAID (employment status as response variable) 61
4.3.2 Results for logistic regression (employment status as response variable) 65
4.3.3 Discussion of CHAID and logistic regression (employment status as response variable) 69
4.3.4 September LFS 2006 and GHS 2006 September GHS 2006 (source data as response variable) 70
4.4 September LFS 2007, July GHS 2007 and October CS 2007 71
4.4.1 Results for CHAID (employment status as response variable) 71
4.4.2 Results for multinomial logistic regression (employment status as response variable) 75
4.4.3 Discussion of CHAID and logistic regression results (employment status as response variable) 78
4.4.4 September LFS 2007, July GHS 2007 and October CS 2007 (source data as response variable) 79
4.4.5 Summary and comparison of the results of all the sections 82
4.4.5.1 Summary of the comparison of employment estimates from surveys over-time 82
4.4.5.2 Summary of the comparison of employment estimates across surveys 84
CHAPTER 5: Overall discussion, Recommendations and Conclusions **86**
5.1 Discussion of the CHAID and logistic regression results 86
5.2 Limitations of the study 87
5.3 Summary and conclusions 87
References **88**
APPENDIX A: Comparisons of questions from and within surveys **91**
APPENDIX B: SAS Program **98**

LIST OF TABLES

Table 1.3.1: Available data and times of data collection for 2006 – 2007……..…….....4
Table 1.3.2: Estimated percentages in the different employment categories and the 95% confidence intervals ….....4
Table 1.3.3: Estimated percentages of total the population in the different employment categories and the 95% confidence intervals for GHS 2006 – 2007, LFS 2006 – 2007 and CS 2007 by province……..……6
Table 3.1.1: Sample sizes in the different categories of employment status (March LFS 2006 and 2007)……..…….....18
Table 3.1.2: Percentage of rotation sample (March LFS 2006 and 2007)…….....20
Table 3.1.3: Variables and their response categories for CHAID analysis……....20
Table 3.2.4.1: List of significant predictors of employment status (March LFS 2006 and 2007) ..24
Table 3.2.4.2: Age group by employment status (March LFS 2006 and 2007) ….……...…25
Table 3.2.4.3: Age group categories by most significant predictors involved in the first order interactions and their p-values……...…….…….……..26
Table 3.2.4.4: Profiles of each subgroup formed by CHAID analysis (March LFS 2006 and 2007) ……...…….…….....26
Table 3.3.2.1: List of variables for multinomial logistic regression analysis...…….....30
Table 3.3.9.1: Model Fit Statistics (March LFS 2006 and 2007) ……..……...33
Table 3.3.9.2: Testing Global Null Hypothesis: BETA=0 (March LFS 2006 and 2007)……...33
Table 3.3.9.3: Parameter estimates from the logistic regression model (March LFS 2006 and 2007)……..……..……...34
Table 3.3.9.4: Sex by marital status (March LFS 2006 and 2007)……36
Table 3.3.9.5: Assessment of the adequacy of the model in percentages (March LFS 2006 and 2007)……...……..……....37
Table 4.1.1.1: List of significant predictors (September LFS 2006 and 2007) …….....39
Table 4.1.1.2: Employment status per categories of age group (September LFS 2006 and 2007) ……..……...42
Table 4.1.1.3: Age group categories by predictors involved in the first order interactions and their p-values (September LFS 2006 and 2007) …….……..……43
Table 4.1.1.4: Profiles of each subgroup formed by CHAID analysis (September LFS 2006 and 2007) ……..……43
Table 4.1.2.1: Model Fit Statistics (March LFS 2006 and -2007)……..……..43
Table 4.1.2.2: Testing Global Null Hypothesis: BETA=0 (March LFS 2006 and 2007) …….....44
Table 4.1.2.3: Parameter estimates from the logistic regression model (September LFS 2006 and 2007) ……...……44
Table 4.1.2.4: Assessment of the adequacy of the model in percentages (September LFS 2006 and 2007) ……..…..…47
Table 4.1.4.1: List of significant predictors (March LFS 2006 and 2007)……....49
Table 4.1.4.2: Percentage distribution of source data by highest level of education (March LFS 2006 and 2007)……..……..50
Table 4.1.4.3: Profiles of each subgroup formed by CHAID analysis (September LFS 2006 and 2007)……..……....50
Table 4.1.4.4: List of significant predictors (September LFS 2006 and 2007)…….……51
Table 4.1.4.5: Profiles of each subgroup formed by CHAID analysis (September LFS 2006 and 2007) .....……...51
Table 4.2.1.1: List of significant predictors (September GHS 2006 and July GHS 2007)…….....52
Table 4.2.1.2: Employment status per categories of age group (September GHS 2006 . and July GHS 2007)……..…..…….……...54
Table 4.2.1.3: Age group categories by predictors involved in the first order interactions (September GHS 2006 and July GHS 2007)……..……...…….....55
Table 4.2.1.4: Profiles of each subgroup formed by CHAID analysis (September GHS 2006 and July GHS 2007).....……...55
Table 4.2.2.1: Model Fit Statistics (September GHS 2006 and July GHS 2007) ……....55
Table 4.2.2.2: Testing Global Null Hypothesis: BETA=0(September GHS 2006 and July GHS 2007) ……...…….....55

Table 4.2.2.3: Parameter estimates from the logistic regression model (September GHS 2006 and July GHS 2007)……..…….....56
Table 4.2.2.4: List of significance five factor interactions (September GHS 2006 and July GHS 2007)…….…..…58
Table 4.2.2.5: Assessment of the adequacy of the model in percentages (September GHS 2006 and July GHS 2007).....…..58
Table 4.2.4.1: List of significant predictors (September GHS 2006 and July GHS 2007)…....….61
Table 4.2.4.2: Profiles of each subgroup formed by CHAID analysis (September GHS 2006 and July GHS 2007)……....61
Table 4.3.1.1: List of significant predictors (September LFS 2006 and September GHS 2006)…62
Table 4.3.1.2: Age group by employment status (September LFS 2006 and September GHS 2006).....64
Table 4.3.1.3: Age group categories by predictors involved in the first order interactions and their p-values (September LFS 2006 and September GHS 2006)……....65
Table 4.3.1.4: Profiles of each subgroup formed by the CHAID analysis (September LFS 2006 and September GHS 2006)……..……...65
Table 4.3.2.1: Model Fit Statistics (September LFS 2006 and September GHS 2006) ……..... 66
Table 4.3.2.2: Testing Global Null Hypothesis: BETA=0 (September LFS 2006 and September GHS 2006) .....….....66
Table 4.3.2.3: Parameter estimates from the logistic regression model (September LFS 2006 and September GHS 2006) …….....66
Table 4.3.2.4: Assessment of the adequacy of the model in percentages (September LFS 2006 and September GHS 2006).....…...68
Table 4.3.4.1: List of significant predictors (September LFS 2006 and July GHS 2006).…….....70
Table 4.3.4.2: Profiles of each subgroup formed by CHAID analysis (September LFS 2006 and September GHS 2006) ……....71
Table 4.4.1.1: List of significant predictors (September LFS 2007, July GHS 2007 and October CS 2007)……..…….…..72
Table 4.4.1.2: Age group by employment status (LFS September 2007, GHS July 2007 And CS October 2007)……..……..……...74
Table 4.4.1.3: Age group categories by predictors involved in the first order interactions and their p-values (September LFS 2007, July GHS 2007 and October CS 2007)…...75
Table 4.4.1.4: Profiles of each age group by CHAID (September LFS 2007, July GHS 2007 and October CS 2007)……..…….....75
Table 4.4.2.1: Model Fit Statistics (September LFS 2007, September GHS 2007, and October CS 2007)……..……..……75
Table 4.4.2.2: Testing Global Null Hypothesis: BETA=0 (September LFS 2007, September GHS 2007, and October CS 2007)……76
Table 4.4.2.3: Parameter estimates from the logistic regression model (September LFS 2007, July GHS 2007 and October CS 2007)…….....76
Table 4.4.4.1: List of significant predictors (LFS September 2007, GHS July 2007 and CS October 2007) ……..……..……80

LIST OF FIGURES

Figure 3.2.4.1: Classification tree diagram for March LFS 2006 and 2007……..……..…23
Figure 3.2.4.2: Level of education among people in age group 15-19 years old by employment status (before merge)……..……..27
Figure 3.2.4.3: Level of education among people in age group 15-19 years old by employment status (after merge)……..……27
Figure 4.1.1.1: Classification tree diagram for September LFS 2006 and 2007……..……41
Figure 4.1.4.1: Classification tree diagram for March LFS 2006 and 2007 (source data as response variable)……49
Figure 4.1.4.2: Classification tree diagram for September LFS 2006 and 2007 (source data as response variable)……...……..50
Figure 4.2.1.1: Classification tree diagram for September GHS 2006 and July GHS 2007……....53
Figure 4.2.4.1: Classification tree diagram for September GHS 2006 and July GHS 2007 (source data as response variable)……..……..…60
Figure 4.3.1.1: Classification tree diagram for September GHS 2006 and September LFS 2006 ……...63
Figure 4.3.4.1: Classification tree diagram for September LFS 2006 and September GHS 2006 (source data as response variable)…..…..……...……..70
Figure 4.4.1.1: Classification tree diagram for September LFS 2007, July GHS 2007and October CS 2007…..…..……...……...73
Figure 4.4.4.1: Classification tree diagram for September LFS 2007, July GHS 2007 and October CS 2007(source data as response variable)…….……81

ACRONYMS

BLS	Bureau of Labour Statistics
CES	Current Employment Statistics
CHAID	Chi- squared Automatic Detection
Chisq	Chi-square
CS	Community Survey
CPS	Current Population Survey
DF	Degree of freedom
EA	Enumeration area
EC	Eastern Cape
XAID	Extended Automatic Interaction Detection
Exp(Est)	Exponentiation of the beta coefficient
FESAC	Federal Economic Statistics Advisory Committee
FS	Free State
GHS	General Household Survey
GP	Gauteng
ILO	International Labour Office
KZN	KwaZulu-Natal
LP	Limpopo
LFS	Labour Force Survey
MLE	Maximum Likelihood Estimation method
MP	Mpumalanga
NSO	National Statistics Office
-2LOGL	Negative two times the log-likelihood
NW	North West
NC	Northern Cape
OHS	October Household Survey
ONS	Office for National Statistics
PES	Post Enumeration Survey
p-value	Probability value
QLFS	Quarterly Labour Force Survey
SA	South Africa
Stats SA	Statistics South Africa
UI	Unemployment Insurance
UK	United Kingdom
USA	United States of America
WC	Western Cape

CHAPTER 1: Introduction and Problem Statement

1.1 Introduction

Availability of good official statistics has become fundamental to the effective functioning of a democratic society. These are statistics that have been designated as official statistics by a National Statistics Office (NSO) to the extent that deductions can be made from them, and are 'fit for use' for the purpose they were designed for. The need for official statistics has recently increased rapidly, as a result of growing demand from users for new and better statistics to describe new phenomena and to monitor the development of various programmes and projects (Holt, 2008). In South Africa, the government's monitoring and evaluation programme uses official statistics to promote a culture of evidence-based planning and decision-making in pursuit of socio-economic development and good governance. The collection of reliable information enhances the monitoring of progress in service delivery, in that it forms a basis for informed decision-making, the development of policies, or the planning required for a massive programme of social transformation (Stats SA[1], 2008a).

Prior to 1994, there were few surveys in South Africa (SA) measuring the social characteristics of people. In particular, in the area of economic statistics, more effort was focused on the direct collection of employment data from formal industry and commerce, rather than the estimation of the unemployment rate through surveys of households. Since the first democratic election in 1994, we have seen more and more social data being collected. There are a number of organisations in SA producing data. Data from different organisations or departments within organisations have been collected and processed using different methods and procedures, and these sometimes lead to contradictory results, resulting in a quality gap. Some of these differences are caused by known differences in the statistics, for example in the section of the population for which separate statistics are needed, but other differences are more difficult to explain. The country faces a challenge to close this quality gap in terms of common standards, including concepts, definitions, classifications, methodologies and sample frames.

This is not a problem unique to SA. The literature cites many possible reasons for inconsistencies in survey data (Krosnick, 1989; Brackstone, 1999; Collins and Sykes, 1999; Haworth and Caplan, 1999; Nardone, Bowler, Kropf, Kirkland and Wetrogan, 2003 and Fu, 2004). Problems with the quality of the survey process, surveys being done at different times of the year (which may result in inclusion or exclusion of seasonal workers), and changes in the questionnaire, are some of the most common causes cited. For SA, the literature notes that data collected in the early post-apartheid period are problematic for various reasons such as differing sampling schemes, non-coverage (failure to

[1] Statistics South Africa (Stats SA)

adequately cover all components of the population being studied), and small samples (Klasen and Woolard, 2000; Casale and Posel, 2002; Kingdon and Knight, 2007; Yu, 2009).

Stats SA conducted the October Household Survey (OHS) annually from 1994 until 1999. This survey was discontinued in 1999 due to the re-prioritisation of surveys in the face of changing data needs and financial constraints. It was replaced by two surveys, the Labour Force Survey (LFS) and the General Household Survey (GHS). The first round of the LFS was conducted in 2000, while the first GHS was conducted in July 2002. The LFS covers some areas previously covered by the OHS but not all, since it is a specialised survey principally designed to measure the dynamics in the labour market. The LFS of September each year included a section designed to measure social indicators such as access to infrastructure. Again, this section did not go into as much depth as the OHS used to. As a result of this, a need to measure the level of the country's development and performance of government programmes and projects arose. The GHS was specifically designed for such purposes.

In order to generate a set of sufficiently comparable estimates across surveys and within surveys over time, it is necessary to identify and address the various sources of potential non-comparability. According to Casale and Posel (2002), comparability between OHS and LFS over time is undermined by both changes of questions between surveys and changes in the way employment and unemployment are derived from the questions in the different surveys.

Kingdon and Knight (2007) indicated that SA has an interesting labour market as compared to other countries. They state that its sharp segmentation, high unemployment and low non-farm informal sector make SA different from other countries. They further note inconsistencies as to how Stats SA defined and derived statistics on employment and unemployment from the various questions relating to employment status over the years. The changes in questions, definitions and sampling, and by reweighting in the light of new census data over the years make it difficult to compare these estimates.

1.2 Definitions of concepts

The main indicators used are the labour force, employment rate, unemployment rate, not economically active population, the employed and unemployed. In defining labour force and unemployment, the International Labour Office (ILO) definitions were used. The following definitions were extracted from the Stats SA LFS 2007 report (Stats SA, 2007a).

Population of working age: All persons living in SA aged 15–65 inclusive at the time of the survey.

Employed: Persons aged 15–65 who did any work or who did not work but had a job or business in the seven days prior to the survey interview.

Unemployed (official definition): Persons aged 15–65 who did not have a job or business in the seven days prior to the survey interview but had looked for work or taken steps to start a business in the four weeks prior to the interview and were available to take up work within two weeks of the interview.

Unemployed (expanded definition): Persons aged 15–65 who did not have a job or business in the seven days prior to the survey interview but had not taken active steps to find work in the four weeks prior to the interview (i.e. discouraged work-seekers).

Labour force: The sum of employed and unemployed persons.

Not in the labour force (not economically active): Persons who are neither employed nor unemployed.

Discouraged work-seekers: Persons who want to work and are available to work but who say that they are not actively looking for work.

Unemployment rate: The percentage of the economically active population that is unemployed.

Labour absorption rate: The proportion of the working-age population that is employed.

Labour participation rate: The percentage of the working-age population that is economically active (employed and unemployed), i.e. labour force/labour market.

1.3 Data sources

Economic data obtained from household interviews by Stats SA include those from population censuses, the OHS up to 1999, the GHS from 2002, and the LFS between 2000 and 2007. In 2005, Stats SA undertook a major revision of the LFS. This revision resulted in changes to the survey methodology, the survey questionnaire, the frequency of data collection (from two to four surveys per year – resulting in what was now renamed as the Quarterly Labour Force Survey), and data releases, and the survey data capture and processing systems (Stats SA, 2008b). The first estimates from the Quarterly Labour Force Survey (QLFS) were published in 2008.

The questionnaires used in both LFS and GHS surveys from 2002 to 2005 have constantly been reviewed, in the light of changes to international standards, concepts and methodologies of the ILO, accommodation of national requirements in terms of providing information to inform policymakers, as well as streamlining of questionnaires to improve respondent understanding. These improvements mean that data from different years are sometimes not completely comparable. Since 2006, changes in both GHS and LFS have been stabilised, and not many differences should be envisaged. The study by Yu (2009) also indicated that data inconsistency in the LFS series has become less of a problem with more recent data sets.

For these reasons, this study will compare data from the GHS and LFS over the period 2006–2007. The 2007 Community Survey (CS) also included some economic questions, and it is included as it utilised a much shorter questionnaire, but had a much bigger sample size. The available data and times of data collection are listed in Table 1.3.1.

Table 1.3.1: Available data and times of data collection for 2006–2007

	LFS	GHS	CS 2007
2006	March September	September	
2007	March September	July	October

Employment status with the response categories; employed, unemployed and not economically active; will be used as the dependent variable. The independent variables include age, gender, level of education, marital status and population group. This information is obtained directly from respondents, with only employment status being a derived variable (i.e. computed from the individual responses). Although there are many other variables available, these were chosen as they appear in all of the data sets, and are mainly demographic variables that are used to profile the attributes of employment status. Table 1.3.2 lists the estimated percentages in the different employment categories and the 95% confidence intervals (as calculated from the data).

Table 1.3.2: Estimated percentages in the different employment categories and the 95% confidence intervals

Survey	Employed	Unemployed	Not economically active	Population size
LFS March 2006	43.97 (43.43-44.51)	14.32 (13.95-14.69)	41.71 (41.16-42.26)	29 851 888
LFS September 2006	42.71 (42.15-43.26)	14.63 (14.24-15.02)	42.66 (42.08-43.24)	30 005 828
GHS September 2006	43.70 (43.21-44.29)	14.30 (13.97-14.74)	41.80 (41.33-42.44)	29 934 992
LFS March 2007	43.75 (43.22-44.29)	14.36 (13.98-14.74)	41.89 (41.33-42.44)	30 195 309
GHS July 2007	44.22 (43.67-44.78)	13.84 (13.47-14.21)	41.93 (41.35-42.51)	30 333 844
LFS September 2007	43.52 (42.96-44.08)	12.97 (12.60-13.34)	43.51 (42.93-44.10)	30 413 283
CS October 2007	41.13 (40.98-41.27)	22.48 (22.36-22.60)	36.39 (36.25-36.53)	25 733 145

Stats SA (2007a, p. i) describes the LFS as “a biannual household survey, specifically designed to measure the labour market”. The survey is more focused on labour issues. The major objective is to measure the extent of trends and levels of unemployment in the country using two indicators: the official unemployment definition and the expanded definition of unemployment.

According to Stats SA (2007b, p. i), the GHS is “an annual household survey, specifically designed to measure various aspects of the living circumstances of South African households”. The key

findings reported focus on the five broad areas, namely: education, health, activities related to work and unemployment, housing and household access to services and facilities. The GHS was not specifically designed for labour market issues and a number of labour related questions that are included in the LFS are not in the GHS.

Stats SA (2007c, p.10) also describes the CS as "the largest survey that has ever been carried out by Stats SA. The survey collected data on population size, composition and distribution; migration, fertility and mortality; disability and social grants; school attendance and educational attainment; labour force; and income. The main objective of the CS release is to provide key results emanating from the analysis of the data that were collected during the survey. The specific objectives are:

- To provide emerging trends and differentials with regard to demographic, socio-economic and social profiles of the population of South Africa.
- To highlight some of the successes that have been achieved to date and the challenges that need to be addressed in meeting the development goals that government has set".

Differences are expected for surveys done in different seasons, in different years, and surveys using different questionnaires or different sample designs. As we have discussed in the second paragraph of this section, changes in questions always entail comparability and continuity problems. Such changes mainly result in a break in the time series. Appendix A provides tables highlighting areas of differences in the questions asked in March LFS 2006 and LFS 2007, September LFS 2006 and LFS 2007, September GHS 2006 and July GHS 2007, and October CS 2007.

According to the literature review, comparability of surveys over time is undermined by changes of questions. Even a small change in the phrasing of a question may lead to different interpretations. Words, phrases and items used in a questionnaire are subject to misunderstanding, as in any form of communication. Appendix A shows that almost all the surveys used the same questions. There were some differences in the details of the questions asked, for example the marital status in CS 2007 contains more response categories than other surveys. For the labour force questions, the CS data collectors were instructed to probe as to the employment status, to ensure that work for short periods of time was correctly included. Most of the questions asked in the CS questionnaire were accompanied by detailed instructions to data collectors.

Another source of variability is in areas with different employment profiles. Employment rate information is released at provincial level, and there are major differences between the provinces (see Table 1.3.3 below). The province abbreviations are listed in the acronyms.

Table 1.3.3: Estimated percentages of the total population in the different employment categories and the 95% confidence intervals for GHS 2006–2007, LFS 2006–2007 and CS 2007 by province

Employment status	Province	LFS March 2006	LFS Sept 2006	GHS Sept 2006	LFS March 2007	LFS Sept 2007	GHS July 2007	CS Oct 2007
Employed	WC	3.49 (3.26-3.72)	3.45 (3.23-3.66)	3.49 (3.27-3.71)	3.45 (3.22-3.68)	3.43 (3.20-3.65)	3.66 (3.43-3.89)	2.95 (2.90-3.01)
	EC	6.37 (6.14-6.61)	7.17 (6.93-7.42)	7.56 (7.31-7.82)	6.43 (6.19-6.67)	6.79 (6.53-7.06)	7.03 (6.76-7.31)	5.72 (5.65-5.78)
	NC	0.85 (0.80-0.89)	0.78 (0.74-0.82)	0.82 (0.77-0.86)	0.94 (0.88-0.99)	1.01 (0.95-1.06)	0.91 (0.86-0.96)	0.84 (0.82-0.86)
	FS	2.76 (2.63-2.9)	2.78 (2.65-2.95)	2.67 (2.53-2.81)	2.85 (2.71-2.99)	2.75 (2.62-2.87)	2.59 (2.46-2.71)	2.06 (2.02-2.10)
	KZN	9.86 (9.57-10.14)	9.42 (9.14-9.69)	9.77 (9.48-10.05)	9.36 (9.08-9.65)	9.67 (9.39-9.95)	10.23 (9.94-10.52)	8.44 (8.36-8.52)
	NW	3.82 (3.63-4.01)	3.70 (3.52-3.88)	3.87 (3.68-4.06)	3.41 (3.24-3.57)	3.40 (3.24-3.57)	3.21 (3.06-3.36)	2.59 (2.54-2.63)
	GP	7.31 (6.97-7.63)	7.13 (6.79-7.46)	6.47 (6.14-6.80)	6.92 (6.56-7.28)	7.39 (7.03-7.75)	7.16 (6.82-7.50)	5.83 (5.76-5.89)
	MP	2.97 (2.82-3.11)	2.88 (2.74-3.01)	2.98 (2.84-3.12)	3.10 (2.95-3.25)	2.96 (2.81-3.11)	3.22 (3.08-3.37)	2.90 (2.85-2.95)
	LP	6.54 (6.29-6.79)	6.44 (6.20 -6.67)	6.35 (6.09-6.60)	6.25 (6.01-6.48)	6.12 (5.89-6.34)	6.21 (5.99-6.46)	5.07 (5.00-5.13)
Unemployed	WC	1.17 (1.04 -1.29)	1.27 (1.13-1.40)	1.29 (1.15-1.42)	1.09 (0.97-1.20)	1.23 (1.07-1.34)	1.40 (1.26-1.55)	2.03 (1.99-2.07)
	EC	1.55 (1.44-1.66)	1.58 (1.46-1.71)	1.71 (1.58-1.83)	2.13 (1.98-2.28)	1.47 (1.37-1.58)	1.64 (1.52-1.76)	2.67 (2.62-2.71)
	NC	0.26 (0.23-0.28)	0.31 (0.28-0.34)	0.32 (0.29-0.35)	0.41 (0.37-0.45)	0.35 (0.32-0.38)	0.36 (0.33-0.39)	0.45 (0.44-0.47)
	FS	1.03 (0.94-1.11)	0.95 (0.87-1.03)	1.18 (1.09-1.27)	0.94 (0.86-1.02)	0.88 (0.80-0.95)	1.00 (0.92-1.08)	1.47 (1.44-1.51)
	KZN	3.13 (2.95-3.31)	3.13 (2.96-3.31)	3.70 (3.51-3.90)	2.94 (2.76-3.12)	3.08 (2.89-3.28)	2.66 (2.53-2.80)	4.94 (4.88-5.01)
	NW	1.37 (1.27-1.47)	1.42 (1.31-1.54)	1.32 (1.22-1.42)	1.13 (1.03-1.23)	0.89 (0.79-0.98)	1.07 (0.97-1.17)	1.58 (1.54-1.61)
	GP	3.35 (3.12-3.57)	3.33 (3.08 -3.58)	3.94 (3.67-4.21)	3.62 (3.37-3.86)	3.04 (2.83-3.26)	3.49 (3.24-3.74)	5.63 (5.57-5.70)
	MP	1.02 (0.93-1.09)	1.00 (0.92-1.08)	1.08 (1.00-1.16)	1.17 (1.08-1.27)	0.96 (0.87-1.04)	1.01 (0.93-1.09)	1.65 (1.62-1.69)
	LP	1.44 (1.32-1.55)	1.36 (1.26-1.45)	1.47 (1.34-1.60)	1.23 (1.11-1.32)	1.09 (0.99-1.19)	1.21 (1.11-1.31)	2.06 (2.02-2.10)
Not economically active	WC	6.18 (5.87-6.48)	6.11 (5.81 -6.40)	5.97 (5.67-6.28)	6.16 (5.81-6.50)	5.92 (5.60-6.24)	5.87 (5.58-6.16)	6.40 (6.33-6.48)
	EC	5.46 (5.25-5.67)	4.62 (4.41-4.82)	4.12 (3.93-4.31)	4.53 (4.31-4.74)	4.90 (4.57-5.24)	4.69 (4.34-5.04)	3.63 (3.58-3.68)
	NC	0.84 (0.79-0.88)	0.87 (0.82-0.91)	0.79 (0.75-0.84)	1.01 (0.96-1.07)	1.00 (0.95-1.05)	0.98 (0.93-1.04)	0.90 (0.87-0.92)
	FS	2.60 (2.45-2.74)	2.65 (2.51-2.79)	2.52 (2.38-2.67)	2.61 (2.46-2.75)	2.73 (2.58-2.88)	2.64 (2.50-2.78)	2.30 (2.25-2.34)
	KZN	7.35 (7.07-7.63)	7.60 (7.32-7.87)	6.94 (6.65-7.24)	8.12 (7.82-8.42)	7.51 (7.23-7.79)	7.61 (7.32-7.90)	7.34 (7.26-7.42)
	NW	2.95 (2.76-3.13)	3.02 (2.83-3.21)	3.08 (2.89-3.29)	2.67 (2.50-2.84)	2.79 (2.63-2.96)	2.77 (2.58-2.95)	2.61 (2.56-2.66)
	GP	11.03 (10.61-11.45)	11.39 (10.93-11.84)	11.57 (11.13-12.01)	11.96 (11.48-12.43)	2.56 (2.08-3.04)	11.95 (11.49-12.40)	12.27 (12.17-12.37)
	MP	2.69 (2.55-2.82)	2.79 (2.65-2.93)	2.43 (2.30-2.56)	3.02 (2.86-3.17)	3.23 (3.07-3.39)	2.91 (2.76-3.06)	2.97 (2.92-3.02)
	LP	2.61 (2.46-2.76)	2.84 (2.67-3.01)	2.59 (2.43-2.75)	2.59 (2.43-2.75)	2.86 (2.69-3.02)	2.52 (2.37-2.68)	2.72 (2.67-2.76)

This study will investigate the sources of variability, using data from only the Gauteng (GP) and Eastern Cape (EC) provinces. The two provinces were chosen to represent the different socio-economic characteristics and employment patterns. Gauteng is one of the leading provinces in SA

where the basic needs of the community are mostly provided e.g. highest percentage of formal dwelling units, education levels are high, unemployment rates are low etc. The province serves as the economic engine room of the country. It is the most densely populated province in SA, as compared to EC. It is also much more rural and one of the poorest provinces in SA as compared to GP. Most of the infrastructure such as schools, roads and houses are not in place in some parts of the EC. There is also a high rate of unemployment in EC as compared to GP.

1.4 Objectives of the study

The main objective of the study is to investigate possible reasons as to why the seven surveys give different estimates of the percentages in the different employment categories, as shown in Table 1.3.2. In order to investigate the different sources of variability, that is, surveys done in different years, surveys using different questionnaires, different sample designs and different employment profiles, the following comparisons will be done:

1. To compare estimates of employment status over time, where the surveys are run in the same months of consecutive years, use the same questionnaire and the same sample design:
    (a) LFS March 2006 and 2007
    (b) LFS September 2006 and 2007
2. To compare estimates of employment status over time, where there is a difference in the months as well as the years, but the same questionnaire and the same sample design are used:
    (a) GHS September 2006 and July 2007.
3. To compare estimates of employment status across surveys, where the surveys are run in the same months and years, use the same questionnaire and sample design:
    (a) LFS September 2006 and GHS September 2006
4. To compare estimates of employment status across surveys, where there is a difference in the months, questionnaire and sample design, but the same year:
    (a) LFS September 2007, GHS July 2007 and CS October 2007.

As a way of checking to what extent the combinations of predictors identified by the models are stable, the comparison in 1(b) will be used to check the differences found in 1(a). The following aspects of changes in employment questions from LFS, GHS and CS over the 2006–2007 period will be considered:

(a) how questions have been asked within and across surveys over time
(b) how the response categories have changed within and across surveys over time
(c) what changes arise as a result of different reference periods.

The target population for all the surveys is all households in SA, excluding institutions such as old age homes, hospitals, prisons and military barracks. For a person to be included in the survey he/she would have stayed in the selected household for at least four nights on average per week during the

last four weeks. Those who are not household members (those who haven't spent at least four nights per week) would be eliminated as the instruction in this question is to end the interview for those who have answered "No" to the question that is asked to establish this. It is through this question that eligible household members are identified within the selected dwelling. This is an important question for all these household surveys, because it determines who should be included in the survey.

Employment questions cover only the population aged 15 years and above. The samples for LFS and GHS were obtained from a master sample based on population Census 2001 enumeration areas, and the design of the CS sample was based on the 2001 population census enumeration areas. The GHS, LFS and CS asked if a person was employed in the last seven days (before the interview date) and if that person has worked at least one hour during that period, or was absent from work during these seven days but has some work to which to return. Questionnaires from all three surveys were administered through face-to-face interviews for each household visited.

After the South African Statistics Council had evaluated the results of the CS, it was found that some of the measures were not comparable with Census and other survey data sources. A concern was raised regarding the interpretation of the results for certain variables (including unemployment, access to social grants, and income) in the CS, and the South African Statistics Council issued the following warning to caution users. "The measure of unemployment in the Community Survey is higher and less reliable due to the differences in questions asked relative to the normal Labour Force Surveys" (Stats SA, 2007c, p.5). The reason for the difference was mainly related to the differences in the details of the questions asked. The LFS questionnaire was designed to measure employment trends in the country and the survey can afford to include more prompts to clarify some questions, which was not possible during CS enumeration.

1.5 Structure of the research report

The report has five chapters.

Chapter 1: contains the background, definitions, concepts, data sources, and the objectives of the study.

Chapter 2: contains the literature review of other studies relevant to this paper. This chapter is divided into general issues regarding data quality, specific issues for surveys of employment, methodologies for comparing surveys, as well as providing a summary.

Chapter 3: presents the theoretical details of the methodologies used in this report, and discusses the output for the comparison of the March LFS 2006 and 2007 in detail.

Chapter 4: provides the results for the other comparisons, and discusses and compares the results.

Chapter 5: the last chapter provides the summary of the findings, conclusions and recommendations.

CHAPTER 2: Literature Review

There is an increasingly growing literature covering problems of the quality of the data, generally, in relation to employment or labour market indicators. NSOs and other data producing agencies are in the process of putting quality management systems in place (Lyberg, Biemer, Collins, deLeeuw, Dippo, Schwarz and Trewin, 1997 and Collins and Sykes, 1999). The issues on data quality range from general to more specific based on the particular data set. Stats SA has been collecting labour market data in a fairly comparable format since 1993.

2.1 Data quality issues

In order to understand these quality issues, it is important that the characteristics of data quality should be analysed. Brackstone (1999) suggests three characteristics of data quality as comparability of statistics, coherence and completeness. Comparability refers to the ability to make reliable comparisons over time; coherence refers to the ability of the statistical data programme to maintain common definitions, classifications, and methodological standards when data originates from several sources; and completeness is the ability of the statistical data collection to provide statistics for all domains identified by the user community. Measures such as the existence and degree of use of standard frameworks, concepts, variables and classification systems; the existence and degree of use of common tools and methodologies for survey design and implementation; and the incidence and size of inconsistencies in published data should be in place to assess success in achieving all characteristics of data quality.

According to Keating (2007), statistics should be examined to ensure consistency of reporting across statistical surveys and other data sources, including administrative sources. There were various reasons identified as the causes for data inconsistency. They range from simple timing effects to misinterpretation of statistical revisions (i.e. the situation where the statistics are released on an ongoing basis) or, in the most detailed cases, to a complex arrangement of data transfer from one agency to another resulting in the data being recorded on different files.

Fu (2004) notes several other data quality issues in addition to the existing data gaps in many areas that constantly cause questions about the accuracy and reliability of the data. Three types of data inconsistencies have been identified. These are the inconsistency between national and international data, inconsistency between data series and differences in timing, and the frequency of data revisions. The quality of data can be affected by different sources of bias in surveys, which together are referred to as the ‘Total survey error’. Total survey error consists of both random and systematic errors, including sampling errors and a range of other types of non-sampling errors, including coverage error, non-response error, and measurement error. The choice of data collection mode (i.e. the method used to collect data, for example collection via the

Web, telephone, personal interviewing, mobile surveys, and paper questionnaires) influences the extent to which the data are affected by each type of non-sampling error (Roberts, 2007).

Sampling error is an error that results from sampling. It arises because observations are made on the basis of a sample rather than on a whole population under study. It describes the variation of the estimates calculated from the possible samples. In the design of the sample selection procedure for a specific survey, a sampling scheme is desired under which the sampling error would be as small as possible. Lehtonen and Pahkinen (2004) state that the standard error of an unbiased estimate is used as a measure of the sampling error, and the comparison of the sampling errors under various sampling schemes is carried out using the design-effect statistic. The standard error of an estimator is the square root of its sampling variance. This measure provides an indication of sampling error using the same scale as the estimate, whereas the variance is based on squared differences.

Non-sampling errors are errors that are not due to sampling. Non-sampling errors can occur from interviewer errors, non-response, coding errors, computer processing errors, errors in the sampling frame, and reporting errors. Various techniques such as data editing (ways to minimise data problems), weighting (the use of probability theory to estimate population parameters) and improvement in various quality control procedures, are available for correcting the undesirable effects of this non-ignorable non-sampling error.

Measurement errors may arise in answers to survey questions for a variety of reasons, including respondents misunderstanding the meaning of the question, failure to recall the information correctly, failure to construct a response correctly and refusal to respond. Measurement error comes from the following four primary sources: questionnaire, data collection method, interviewer and respondent (Biemer, Groves, Lyberg, Mathiowetz and Sudman, 1991). Measurement error is the degree to which observed values are not representative of the true values.

Tarozzi (2007) noted that comparisons of data over time are meaningful only in so far as the necessary data are collected consistently across different rounds of surveys. Statistical agencies often introduce questionnaire changes that can raise doubts on the comparability of data. There are many reasons why a questionnaire has to change, among others these include changes in societal attitudes, or as systems become outdated or methodology becomes in need of redesign. It is however important for a statistical agency to keep quality assurance in mind at the forefront of all its activities in order to minimise the relevance gap and prevent a significant decrease in quality over time. The questionnaire is designed to communicate with the respondent in an unambiguous manner. It represents the survey designer’s request for information. Words, phrases and items used in a questionnaire are subject to misunderstanding as in any form of

communication. There are many potential errors, for example even if the concept to be measured is clearly formulated; it may not be clearly represented by the question.

When analysing time series from multiple cross-sectional data, it is important to check whether changes in the questionnaire design lead to data inconsistency. Question formats in which respondents are asked to respond using a specified set of options (closed format) may yield different responses to those when respondents are not given categories (open format) (Bishop, Hippler, Schwarz and Strack, 1988). It is well known that even slight changes in the wording of questions can result in different answers (Krosnick, 1989).

2.2 Specific issues for surveys of employment

In May 2005, the United States of America (USA) Bureau of Labor Statistics (BLS) asked the Federal Economic Statistics Advisory Committee (FESAC) to investigate discrepancies between the employment trends reported by the Current Population Survey (CPS) and Current Employment Statistics (CES) (FESAC, 2005). The CPS is a household survey conducted by the BLS, and the CES is a survey of business enterprises, conducted by state employment security agencies in cooperation with the BLS. They indicated that one of the difficulties is the fact that the CES is benchmarked each year to universe counts derived from administrative files of employees covered by unemployment insurance (UI) in order to manage possible survey error.

For example, the original sample-based estimates are replaced with benchmark data from the previous year. Statistical benchmarking is the way of using auxiliary information about the population structure from which sampling weights can be adjusted in estimation processes, in order to yield more accurate estimates of totals. Some of the changes in employment sections were reported to be quite large. Age limitations, differences in survey coverage and periods, and different sampling and estimation procedures are some of the factors resulting in the discrepancies between the employment series from the two different surveys. The two surveys (CPS and CES) measure employment in the USA, but they have different definitions of employment, along with different samples, estimation procedures, and concepts. They track well together over a long period of time, even though their rates of growth or decline differ significantly. It has been indicated that the differences between the two surveys can be measured and that it is also possible to adjust them to a relatively similar concept for comparison (Bowler and Teresa, 2006).

Nardone *et al.* (2003) examined the discrepancy in employment growth between the CPS and the CES. The CPS has a broad definition of employment and provides detailed demographic characteristics of individuals by their labour force status. Both the CPS and CES surveys publish data on a seasonally adjusted basis; that is, data adjusted for normal seasonal variations that regularly occur in certain months during the year, for example in agricultural and construction

employment. The findings indicate that the estimates of over-the-month employment change from these two surveys usually do not match in size or even direction. This is due to many differences in the surveys, including sample size, estimation procedures, coverage, and definitions. In examining the discrepancies between the two surveys, they consider the impact of several factors including differences in the universe and concepts in the surveys, under-coverage of certain population groups, and reporting issues. Understanding these differences is a key to attempt to explain the discrepancies. The analyses, however, did not provide any clear-cut answers.

In measuring labour force flows (i.e. persons whose employment status changed from outside the labour force at the start of the year to those in the labour force by the end of the year or vice versa), serious problems of inconsistency between the change in the published labour force levels and the change obtained by balancing out the inflows and the outflows in the monthly gross flow table, have been noted. Several factors including “response variability in the CPS, the effects of conditioning on responses to non-interviewer and mover effects, and clerical errors, as the possible reasons for inconsistency between the gross flows and the net changes and for the possible overstatement of flows”, have been identified (Flaim and Hogue, 1985, p.10).

“Over the course of the year, employment levels undergo sharp fluctuations due to seasonal changes in weather, reduced or expanded production, major holidays, and opening and closing of schools.” .. “The purpose of most labour market analyses, however, is to identify the underlying trends in the data, apart from normal seasonal movements. For this reason, employment data are seasonally adjusted a process that smooths out the normal seasonal shifts that can obscure underlying economic trends” (Rydzewski, Deming, and Rones, 1993, p.3).

Haworth and Caplan (1999), in their paper on time series and cross-sectional analysis and modelling in monitoring of the United Kingdom (UK) labour market, indicated that, within the constraint of the established design, the UK is using the continuous survey as a source of monthly series constructed from three month moving averages. They indicated that sampling variability, particularly for estimates of changes in levels of unemployment, but also for employment and inactivity, are large. Sampling variability is a more serious issue at regional and small area levels. They identified three main components as trend analysis, small area estimation and the construction of longitudinally linked data sets and analyses. The first two components seek to deal with the LFS sampling variability issue. The longitudinal methodology work is seeking to exploit the panel element of the LFS for estimating changes in gross flows between key labour market status categories over time. The report did not concentrate much on the latter components, since it is looking at cross-sectional studies.

The initial findings of the investigation into the effects of response error (e.g. individuals reporting a status change when the status has remained the same) suggest that this is likely to affect the longitudinal data sets, probably in the direction of an upward bias in estimates of gross flows between different economic activity categories. They conclude by recommending that extensive user education and documentation must be provided to support the programme and gain user confidence in the results. The data inconsistency problem may undermine the statistical credibility of any organisation and affect the effectiveness of policy discussion and weaken the policy dialogues within the government. According to Blair (1999), having access to official statistics which can be trusted is essential in any healthy society. He further states that for official statistics to play that key role effectively in democracy, we need to have confidence in the figures themselves.

Brook and Barham (2005), in assessing the reliability of the two-quarter longitudinal LFS flow data, also indicated that the longitudinal data sets are subject to two sources of possible error, non-response bias and misclassification error. The estimated flows are adjusted for non-response bias through calibration weights, which are included in the longitudinal data sets and are also used to weight estimates to UK population totals. The unemployment rate derived from the flow variable, which is expressed as a percentage of working age, will also be lower and is not consistent with the unemployment rate given in the Office for National Statistics (ONS) Online, which is defined according to ILO definitions in terms of all adults aged 16 and over as a percentage of the economically active (employed and unemployed). The study concluded that the Maximum Likelihood Estimation method (MLE) is the most suitable method for estimating the effect of response error in the flows. The MLE method uses the observed flows to drive the adjusted flows, subject to the specified misclassification matrix.

Artola and Bell (2001) evaluated the appropriateness of the standard methodologies and the quality of the data used to analyse labour market dynamics in Europe. A matched file approach was used and it was reported to suffer at the outset from a progressive loss of the panel component over time. The approach has been subjected to a number of specific problems, such as sample attrition and misclassification errors. Their results indicated that, due to recall error and heterogeneous survey design, the retrospective approach tends to result in a considerable number of spurious transitions being recorded. They have recommended the use of quasi-longitudinal data to overcome such problems. This approach entails collecting data at different points in time, from different individuals. The samples at each time point are representative, such that changes can be compared at aggregate level.

Pirouz (2004) examined the household size and the structures in OHS 1995, 1997and 1999; and the LFS September 2001 and 2002. The objective of the research was to investigate whether labour market outcomes affected the structure of households in SA. The results indicated an

average decrease in household size in SA by 0.4 household members. This reflects changes in the proportion of individuals with certain characteristics in each labour market state. It is noted that some of the factors affecting the decline of households' size are the living arrangements such that single households or two-member households work, or opt to work instead of having children.

Kapsos (2007) identified the main sources of non-comparability of labour force estimates as survey type, age group coverage, geographic coverage, inclusion or non-inclusion of military conscripts, variations in the definitions of the economically active population, particularly with regard to the statistical treatment of contributing family workers and unemployed, not looking for work; and differences in survey reference periods.

Kingdon and Knight (2007) examined changes in the incidence of unemployment in SA across different worker groups defined by age, gender, education, race and location. Their paper considered three aspects of the change in unemployment over the 1995–2003 period:

(a) how the distribution of unemployment has changed across worker groups
(b) how incidence of entry into unemployment from the employed status has changed
(c) how duration of unemployment has changed.

The distributions of unemployment were examined by looking at descriptive statistics. However, since race, education and location are highly correlated in SA; descriptive statistics were supplemented with simple binary probit models of unemployment. The probit model provides the researcher with predicted probabilities of various outcomes. This model has two outcomes, 0 and 1, representing unemployed and employed respectively, unlike the multinomial logit model which allows employment to be further split into employee employment and self-employment. The OHS 1995 did not explicitly ask whether workers were self-employed or employed by a third party. Both logit and probit regression are used when the response variables are categorical in nature. "The probit model is derived from the assumption that the error terms are normally distributed. This model assumes that the predictions for the dependent variable will always fall between 0 and 1" (Hair, Anderson, Tatham and Black, 1995, p.131).

Bignami –Van Assche (2003), in their article on the individual consistency in survey response in rural Malawi, reported that, in order to gain a deeper understanding of the nature and implications of inconsistencies in the survey response, some covariates of individual consistency should be analysed and the implications of inconsistencies for the univariate and multivariate data analyses should be evaluated. The paper identified the extent of individual consistency in response to questions about HIV/AIDS and other topics. The outcomes of the investigation revealed that individual inconsistency does not affect the conclusions that can be drawn from the survey.

“Although household surveys are the best source of data for the measurement of inequality, it is crucial to remember when using these numbers, that there is an important story behind each number. Despite much progress in recent years, the data inconsistency problems remain substantial and they call into question the quality of international statistics” (Fu, 2004, p.7). Data producing agencies should consider providing the margins of uncertainty for key statistical indicators. It is important for the user to understand any uncertainty that may impact final estimates.

2.3 Literature on the methodology used to compare surveys

This section reviews the literature on the methodology used to compare survey estimates. According to Stoker (1988), the relationship between the response behaviour and the predictor variables can be determined using various methods such as Chi-squared Automatic Detection (CHAID) and logistic regression. It is important to identify auxiliary variables which are associated with response behaviour. Both techniques can consider all the predictor variables simultaneously in respect of their separate or joint relationship with the response variable. It follows that the predictor variables are being ordered according to the importance in predicting the response variable. He indicated that CHAID is applicable to a fairly large data set (a sample size of 500 or more). It is necessary to rescale the sample weights so that they add to the sample size, since using data weighted up to the population results in large chi-square values, and hence very small p-values, which reflect the population size not the sample size.

According to Cox (1970), logistic regression is a technique that can be used when the dependent variable is categorical. When the predictor variable has only two categories, such a model is known as binary logistic regression. A model with predictor variables containing more than two response categories is known as multinomial logistic regression. Logistic regression has the ability to predict a response variable on the basis of categorical or continuous predictor variables, and to determine the percent of variation of the response variable explained by the predictor variables. The model ranks predictor variables according to their importance in explaining the response variable. This technique can also reveal any interactions in the data (the differing effect of one predictor variable at the different levels of another predictor variable). An example of such an interaction is a differing response to gender in different age groups when the response is the employment status. According to Hair *et al*. (1995), the logistic regression model regresses a function of the probability that a case falls in a certain category of the dependent variable on a linear combination of the independent variables. The slope coefficient informs us of the effect of changing from the base category to another category, or the effect of a unit of change in a continuous variable.

Stats SA (2004) used Extended Automatic Interaction Detection (XAID) during the adjustment of the undercount for Census 2001 instead of CHAID. This technique was used since the dependent variable was continuous. One of the aims of analysing data using this technique was to understand the extent of undercount or over-count in the census data, and to adjust for this where and if there is an evidence of under/over coverage. An analysis variable was created where people who were counted by both census and post enumeration survey (PES) were coded as 1, people who were missed by census and covered by the PES are coded as 0, and people who were counted in the census but not in the PES were assigned a continuous random value from a uniform (0,1) distribution. Predictor variables such as geography type, sex, age group, and population group were used. Examples of geography type used were tribal settlement, commercial farm, smallholding, urban settlement, and informal settlement.

XAID identified subgroups defined by combinations of these variables, according to different categories of the analysis variable. After the subgroups had been identified, Stats SA then estimated the undercount for each group, and adjusted the census data.

These methods have also been used in non-survey sampling data sets. Antipov and Pokryshevskaya (2009) applied CHAID and logistic regression diagnosis and classification accuracy improvement. CHAID was used to evaluate classification accuracy across segments of observations. The dependent variable used in the study was whether the client churned (de-activated their account) or not. The results of the CHAID analysis were used to split the data set into four parts, and a separate logistic model was developed for each segment. In the end, the results gave a better insight into factors influencing customer behaviour.

Bakır, Batmaz, Güntürkün, İpekçi, Köksal and Özdemirel (2006) studied the causes of defects by using Decision-Tree and Regression analysis. The main aim of their study was to identify the most influential variables that cause defects in the items produced by the Casting Company, located in Italy. Logistic regression was used to develop the model with two-way interactions. The overall fitted model was significant, but none of the parameters was found to be significant. They had to conclude that this model does not fit the data. When using the Decision-Tree method, 91.93% of the responses were correctly classified. In the end, nine process variables were found to be influencing defects.

2.4 Summary

A study of the literature shows the importance of understanding data inconsistency. For instance, Flaim and Hogue (1985), Brackstone (1999) and Bowler and Teresa (2006), address factors affecting the comparability of data, such as the universe, concepts and definitions, classification, differences in question wording, methodologies and periods. Nardone *et al.* (2003) found that, in the USA, due to many differences in the surveys from different sources, employment estimates

over the months from different surveys do not always match in size, or follow the same estimation procedures, nor have the same coverage and definitions of concepts. Varying reasons for inconsistencies in the data have been found, ranging from simple errors or timing effects to misinterpretation of statistical estimates. Other factors, such as seasonality and changes in average household size, as discussed by Artola and Bell (2001) and Pirouz (2004), could have a major effect on the direction of estimates. Seasonal patterns also influence the instability of employment data (Rydzewski *et al.* 1993).

Factors that affect data consistency should be analysed and made known to data users. It is of great importance to understand the quality of data before doing any analysis on such data. According to this literature review, data tends to be skewed due to these factors. Haworth and Caplan (1999) argue for the use of time series analysis and modelling to determine sampling variability, particularly for estimates of employment and unemployment status. The effects of response error are likely to affect longitudinal data sets, probably in the direction of an upward bias in the estimates.

Various methods provide ways of using inconsistent survey estimates, but do not necessarily provide an adequate way of summarising the nature of the inconsistencies. There are many ways to examine the factors, which can significantly impact on inconsistencies in the final estimates. Some of the popular ways are to look at the stability of questionnaires, sampling errors, and non-sampling errors.

The literature shows methodological interest in the possible biasing effects of the data collection process, particularly in the designing of the questionnaire and sample, individual response error and response effects. A careful review of survey findings against the background of previous knowledge and relationships observed in similar circumstances, as well as checks on internal data consistency, provides the most valuable indication of incoherent survey results.

According to Stoker (1988) CHAID and logistic regression are useful techniques for comparing survey estimates for a categorical response. According to Cox (1970), logistic regression is a technique that can be used for modelling when the dependent variable is categorical. Logistic regression ranks the predictor variables according to their importance. Both CHAID and logistic regression identify the most influential variables predicting the response variable, and can be used to evaluate the adequacy of the model (Antipov *et al*. 2009 and Stats SA, 2004).

.

CHAPTER 3: Methodology and data analysis

In this study, we will be using CHAID and multinomial logistic regression to meet the objectives given in Section 1.3. This chapter will outline the abovementioned methodologies. We will provide the theoretical background to these two techniques, and the interpretation of the results. In order to illustrate the interpretation of the models and the model diagnostics, we will show the detailed analysis of the March LFS data for 2006 and 2007.

The two data sets will be discussed in detail in Section 3.1, where details of the sample design, weighting of the sample to the entire population and list of predictor variables will be given. Section 3.2 will provide the theoretical details of CHAID, and discuss the results of the application to the March LFS data in detail. Section 3.3 will do the same for multinomial logistic regression. Section 3.4 will discuss the results.

3.1. Data

To illustrate the approach, we use the data sets from the March LFS 2006 and the March LFS 2007. The results of this comparison will be validated using the September LFS 2006 and the September LFS 2007 later in Chapter 4. These data sets were mainly collected to provide information on the nature and pattern of employment status in SA. The dependent variable is employment status (whether people are employed, unemployed or not economically active). In order to profile people's employment status, certain explanatory variables were considered. Table 3.1.1 shows the sample sizes in the different categories of employment status for GP and EC.

Table 3.1.1: Sample sizes in the different categories of employment status by province (March LFS 2006 and 2007)

Survey	Province	Employed	Not economically active	Unemployed	Sample size
March 2006	Eastern Cape	1400	1894	271	3565
	Gauteng	590	2280	303	3173
	Total	1990	4174	574	6738
March 2007	Eastern Cape	1453	1889	289	3631
	Gauteng	605	2364	305	3274
	Total	2058	4253	594	6905
March 2006 and 2007		4048	8427	1168	13643

The sample for each data set was drawn from the master sample, which Stats SA uses to draw samples for its surveys. The master sample is drawn from the database of enumeration areas as was established during the demarcation phase of census 2001. For the analysis, the samples from both surveys were weighted up to the entire population in order to get estimates of the employment status for the population. However, the sample size for March LFS 2006 for the EC and GP is not 10 470 588, but 6 738; and the sample size for March LFS 2007 was 6 905 for these two provinces. Since the significance of the chi-squared tests is highly dependent on the

sample size, it is necessary to rescale the weights to provide an approximation of the true sample size. This was done by dividing the sum of the weights by the average of the weights, giving sample normalised weights. The use of a weight variable gives unequal treatment to the cases in the data set.

The weights relate to the population as estimated by the various mid-year estimates - i.e., the surveys are not weighted to the same population in all cases. The weights of the March LFS 2006 were based on the results of the 2006 mid-year population estimates while those of the March LFS 2007 were based on the 2007 mid-year estimates. The 2006 and 2007 mid-year estimates were estimated using assumptions based on parameters derived from the 2001 Census (based on migration patterns and the contribution of each population group to the overall birth and death rate for both years). These assumptions will become less reliable as the time goes on, since the most recent census reflected increases, meaning that the estimates used for March LFS 2007 are less reliable than those used for March LFS 2006.

Stats SA uses a rotation panel methodology for both the LFS and GHS surveys, in order to get a better picture of the movement of people within the labour market over time. This method involves visiting the same dwelling units on a number of occasions for a particular survey (e.g. LFS). Rotation is designed to keep the sample up-to-date and it is done for a random sample of households. For the GHS and LFS, 20% of the dwelling units are dropped at each successive survey and replaced with a new sample within the selected clusters. This implies that there are cases where you could find the same group of people in both March LFS 2006 and 2007. This is not the same for GHS and LFS as each use different sections of the master sample, so there are no overlaps of respondents between say the September LFS 2006 and the September GHS 2006. The CS does not use the same master sample as GHS and LFS. The CS sampling frame was based on the population census 2001 enumeration areas (EAs). Only 796 466 EAs were considered in the frame.

Table 3.1.2 summarises the percentage of the sample that was common between the March LFS 2006 and 2007, as well as the new sample from each data set. It follows from the table that 74.8% of the sample from the EC and 66.6% of the sample from GP were not affected by rotation. Less than 20% of the sample has been rotated from both data sets. The information from the overlapping respondents can be used to examine movement in and out of employment or labour market flow (Jenkins and Chandler, 2010). However, the aim of this research is to investigate differences in the estimates of employment from different surveys and surveys at different times, so this overlap must be ignored.

Table 3.1.2: Percentage of rotation sample (March LFS 2006 and 2007)

Province	March LFS 2006 and 2007	March LFS 2006	March LFS 2007	Grand total
Eastern Cape	12.1	13.0	74.8	7196
Gauteng	16.0	17.6	66.4	6447
Total	**14.0**	**15.2**	**70.8**	**13643**

Table 3.1.3 lists all the variables and their response categories that will be used in this study.

Table 3.1.3: Variables and their response categories for the CHAID analysis

Variable name	Abbreviation	Response category
1. Province	Prov1	1= Gauteng 2= Eastern Cape
2. Gender	Sex1	1= Female 2= Male 9= Unspecified
3. Age group	Aggrp1	1=15-19 2=20-29 3=30- 39 4=40-49 5=50-65
4. Highest level of education	Educgrp1	1=Grade 8 - 11 2= Less than grade 1/no schooling 3= Grade 1 – grade 7 4=Grade 12 5=Certificate/diploma 6=Degree and higher 9=Unspecified
5. Population group	Race1	1=African black 2=Coloureds 3=Indian/Asian 4=Whites 9=Unspecified
6. Marital status	Marital1	1=Never married 2=Married 3=Living together like husband and wife 4=Widow/widower 5=Divorced/separated 9 = Unspecified
7. Employment status	Empl_sta	1=Employed 2=Not employment active (NEA) 3=Unemployed
8. Source_data	Source	1=March LFS 2006 2=March LFS 2007 3=September LFS 2006 4=September LFS 2007 5=July GHS 2006 6=September GHS 2007 2=October CS 2007

3.2. CHAID

3.2.1 Introduction

The first part of this section provides a theoretical discussion of the CHAID method and then we will illustrate this method using part of the output from the March LFS 2006 and 2007 analyses, as well as performing further analyses to check the accuracy of the main output.

CHAID is one of the oldest classification tree methods, developed by Kass (1980). It is a technique that repeatedly splits a sample into unique sub-groups or segments, predictive of the categorical response variable. This technique can detect interactions between predictor variables. The results of the CHAID analysis will enable us to gain an understanding of the importance of, and interrelationships between predictors, as well as making predictions.

3.2.2 Basic tree-building algorithm

CHAID is a non-parametric algorithm, such that no distributional assumptions about the data have to be made. The only condition for CHAID to work effectively is that the data set used is large (Stoker, 1988). It selects appropriate combinations of variables that can describe the features of the response variable (Diepen and Franses, 2006). This is mainly due to its non-linear approach, i.e. different sub-groupings of the features are used at different levels of the tree instead of using the complete set of features jointly to make a decision. For example, the combinations predictive of the employment status could differ for different education groups. The identification of the interaction between independent variables happens automatically. CHAID partitions a contingency table produced from cross-tabulation of more than two predictor variables.

CHAID determines the relationship between the response and the predictor variables. Each predictor can be categorical or be on a continuous scale. Continuous predictor variables need to be categorised before the analysis can be done. The variable that gives the greatest reduction of the variation of the predictor variable is chosen as the splitting variable. The same splitting process is repeated on each of the sub-groups formed. Each sub-group is treated as a new sub-sample defined by the values of another predictor variable. The splitting process is repeated until the minimum variability within groups and the maximum variability between the groups is met, providing the split is significant, and the size of the group is large enough. It is up to the user to define the significance levels for the splitting and merging of categories. The size of the group needs to be specified by the user. Setting this too small may result in a split as a result of the predictor variable being significant due to one or two points in the group being different from the others.

The following paragraph describes the algorithm as outlined by Kass (1980) to be followed when performing CHAID analyses. "Let the dependent variable have d≥2 categories, and a particular predictor under analysis have c≥2 categories. A subproblem in the analysis is to reduce the given c x d contingency table to the most significant j x d table by combining (in an allowable manner) categories of the predictor. Conceptually, we may first calculate statistics $T_j^{(i)}$, the usual χ^2 statistics for the *i*th method of forming a j x d table (j=2,3,…,c; the range of i depending on type of the predictor). Then, if $T_j^{(*)} = \max_i T_j^{(i)}$ is the χ^2 statistic for the best j × d table, choose the most significant $T_j^{(*)}$" (Kass, 1980, p. 120).

3.2.3 Testing the significance of each predictor variable

The algorithm above requires a test of significance of the reduced contingency table. The Chi-Square (χ^2) test is performed to test the independence of predictor variables from each sub-group that was formed by CHAID. Hawkins and Kass (1982) noted the difficulties of establishing

the significance of the association between the k-way grouping of the categories of the predictor and the dependent variables. Bonferroni inequalities are used to determine the number of ways the c initial categories for a predictor of a given type can be reduced to r groups such that $1 \le r \le c$. This gives a bound for the significance level between these groupings and the dependent variables. Kass (1980) outlined three types of predictors, as well as providing the formulae for calculating these multipliers for the three types of predictors in CHAID.

Consider B as the number of ways in which k groups can be formed from the initial c categories. B depends on the type of predictor.

(a) Monotonic predictors

The categories for these predictors are considered to be ordered with respect to the category number.

$$B = \binom{c-1}{k-1}$$

(b) Free predictors

There are no restrictions placed on the possible ordering of the predictor's categories

$$B = \sum_{i=0}^{k-1} (-1)^i \frac{(k-i)^c}{i!(k-i)!}$$

(c) Floating predictors

These predictors contain categories that are ordered except for one single category, which is allowed to float up or down the ordered scale.

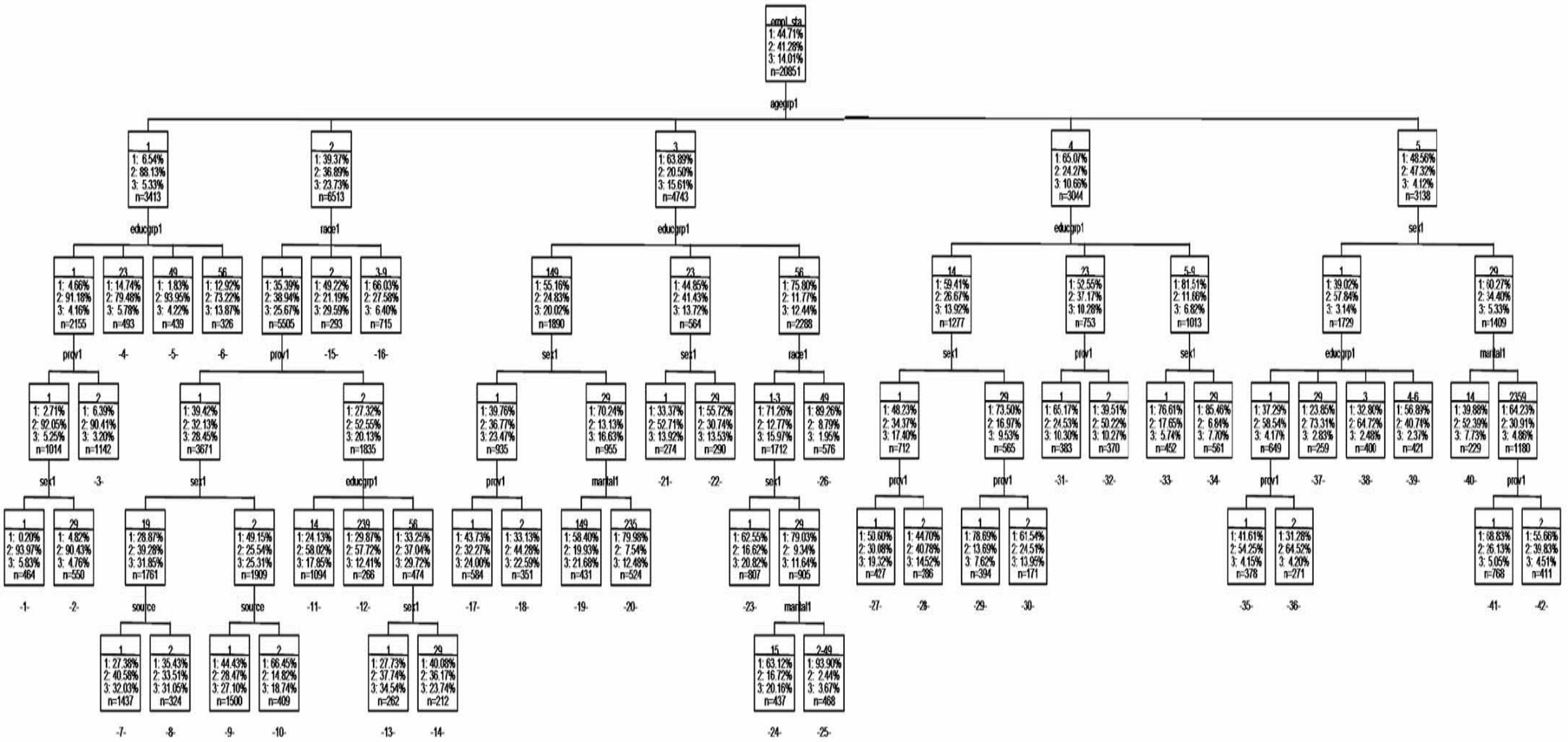
$$B = \frac{k-1+k(c-k)}{c-1}\binom{c-1}{k-1}.$$

3.2.4 Illustration of the CHAID methodology using March LFS 2006 and 2007

The March LFS 2006 and 2007 data sets were initially combined using SAS® 9.2, by the common identification variables of year of survey and province, which were used during data concatenation (see program in Appendix B). The data set was then exported to SPSS® in order to run CHAID.

Employment status was used as the target (response variable) and the variables defined in Table 3.1.3 above were selected as predictor variables. The sample normalised weight was specified as the weight variable. All predictors were treated as free variables to avoid any ordering of the predictor's categories. SPSS CHAID® gives options for setting the number of branch levels, which in this analysis was set to 5. This was chosen in order to allow the CHAID tree to grow such that we can examine a number of significant splits. To obtain segments large enough for the subsequent analysis, the minimum size of the subgroup before merging was set to 500 and the size of subgroup after merging to 250. The merging level was set to be 0.05. The subgroups or nodes of the tree represent the segments which differ by the correct classification rate. The SPSS CHAID® program scanned the data and thereafter generated the tree diagram shown in Figure 3.2.4.1.

Figure 3.2.4.1: Classification tree diagram for March LFS 2006 and 2007



At each stage of the analysis, CHAID splits the tree on the predictor variable having the lowest probability value (p-value), which represents the probability that the observed sample relationship between the predictor and response variable would occur if the two variables were statistically independent. Table 3.2.4.1 lists all predictors according to the level of significance. CHAID, as an explanatory technique, ranks the predictors according to the adjusted p-value. It gives the most statistically significant predictor the highest ranking. In this case, age group has the smallest p-value (4.1e-1316) which implies that age group explains more of the variation in employment status than any other predictor in the analysis. Age is therefore used as the splitting variable at the first depth of the CHAID tree. The next most predictive variable is marital status (p-value1.2e-440), giving a ratio of 3.4e-449, indicating that age is a much stronger predictor than marital status. The largest (significant) p-value is for the source of the data (3.7e-33), giving a ratio of 1.1e-1283, indicating that there is indeed a significant difference between the 2006 and 2007 employment figures. Although all p-values in the table are extremely small, it is important to note that CHAID is partitioning the data to find the most predictive tree. Only significant p-values are listed (i.e. the null hypothesis is rejected in all cases). The table shows an ordering of predictive ability for all variables listed.

Some categories of the predictor variables were merged into one composite class, reducing the number of categories as each category fails to be significant at 5% significance level. For age group, all categories were significantly different. The categories of marital status were reduced from six to five. The fourth category (widow/widower) has a similar profile to that of category nine (unspecified) and they were merged into one composite class. Highest level of education was also reduced from seven to six response categories. People with no formal education (category 2) have a similar employment profile to those who did not specify their highest level of education (category 9).

Category 3 (Indian/Asian) of population group has a similar profile to that of category 9 (unspecified). The two categories were merged to form one composite class. The last predictor with reduced categories was sex. Female (category 1) has a similar profile to people who did not specify their sex. Source data (1=March LFS 2006 and 2=March LFS 2007) is significant but much less significant than age, indicating that one needs to look at source data within the different age groups, to see where the significance comes from.

Table 3.2.4.1: List of significant predictors of employment status (March LFS 2006 and 2007)

Predictor	p-value	Levels	Groups
Age group	4.1e-1316	5	1 2 3 4 5
Marital status	1.2e-449	6->5	1 2 3 49 5
Highest level of education	2.5e-339	7->6	1 29 3 4 5 6
Province	2.9e-193	2	1 2
Population group	9.2e-132	5->4	1 2 39 4
Gender	5.1e-120	3->2	19 2
Source	3.7e-33	2	1 2

At the root node (the first level of the CHAID tree) the most significant split is obtained by segmenting the cases containing employment status into 5 different age groups (see Figure 3.2.4.1). Table 3.2.4.2 summarises trends with regard to age group categories in relation to the employment status.

Table 3.2.4.2: Age group by employment status (March LFS 2006 and 2007)

Agegrp1	Age group	Employed	Not economically active	Unemployed	Sample size
1	15-19	6.54	88.13	5.33	3413
2	20-29	39.37	36.89	23.73	6513
3	30-39	63.89	20.50	15.61	4743
4	40-49	65.07	24.27	10.66	3044
5	50-65	48.56	47.32	4.12	3138

It follows from Table 3.2.4.2 that the probability of a person aged 15-19 years old being employed is less than that of a person aged 20-29 years old and higher age groups. 88.13% of people in this category were not economically active; probably because the majority of them should still be full-time students. Only 6.54% were employed and 5.33% were unemployed. The results also show that the percentage of unemployed people increases among the age group 20-29 years old. This category is likely to include a number of people who have just completed their studies and who were still looking for work. 23.73% of this age group were unemployed; 39.37% were employed and 36.89% were not economically active.

The percentage of people aged 30-39 years old who were employed increased to 63.89%, while there has been a proportional decline in the other categories of employment status. 20.50% of people in this age category were not economically active and 15.61% were unemployed. There has been a slight increase in the percentage of people who were employed (65.07%) and not economically active (24.27%) in the age group 40-49 years old. The percentage of people who were unemployed has decreased to 10.66%. For the age group 50-65 years old, the patterns for employment status change significantly. The percentage of people who were employed has decreased to 48.56%; a large increase in the percentage of people who were not economically active to 47.32% and a drop to 4.12% among people who were unemployed. This result indicates the comparative relationships between age and employment status. The chances of a younger person to be employed are less than for higher age groups. It also shows that people's chance of being employed when they are approaching their retirement age becomes less, possibly due to people taking early retirement.

CHAID then takes each remaining predictor in turn to determine the next segmenting variable. Following the employment status of the different age groups (shown in the tree in Figure 3.2.4.1) one is able to see which of the combinations of attributes comprise each of the 41 terminal nodes. The results were used to understand the predictive power of the predictor variables used and their inter-relationships. At the second level of partitioning it was found that highest level of education, population group and sex were the most significant predictors. The three predictors

are competing with each other within the categories of age group. Table 3.2.4.3 indicates that highest level of education for age group 30-39 years old has the most significant p-value. The predictor variable with the most significant p-value for age group 20-29 years old was population group. Sex was the most predictive variable for the age group 40-49 years old. The least significant predictor variable was that for highest level of education for the age group 15-19 years old.

Table 3.2.4.3: Age group categories by most significant predictors involved in the first order interactions and their p-values

Age group	1st order interactions	Likelihood ratio chi-square	Degrees of freedom	p-value
15-19	Highest level of education	160.42	6	1.7e-29
20-29	Population group	236.35	4	1.4e-48
30-39	Highest level of education	337.31	4	2.9e-69
40-49	Sex	234.18	4	5.0e-47
50-65	Highest level of education	187.61	2	5.5e-41

The age group 20-29 years old was partitioned by population group, whereas the age group 50-65 years old was partitioned by sex. This is probably due to the minimum retirement age for females being lower than that for males. All other age groups were partitioned by highest level of education. Table 3.2.4.4 lists the predictors used at different branch levels of each subgroup identified by CHAID.

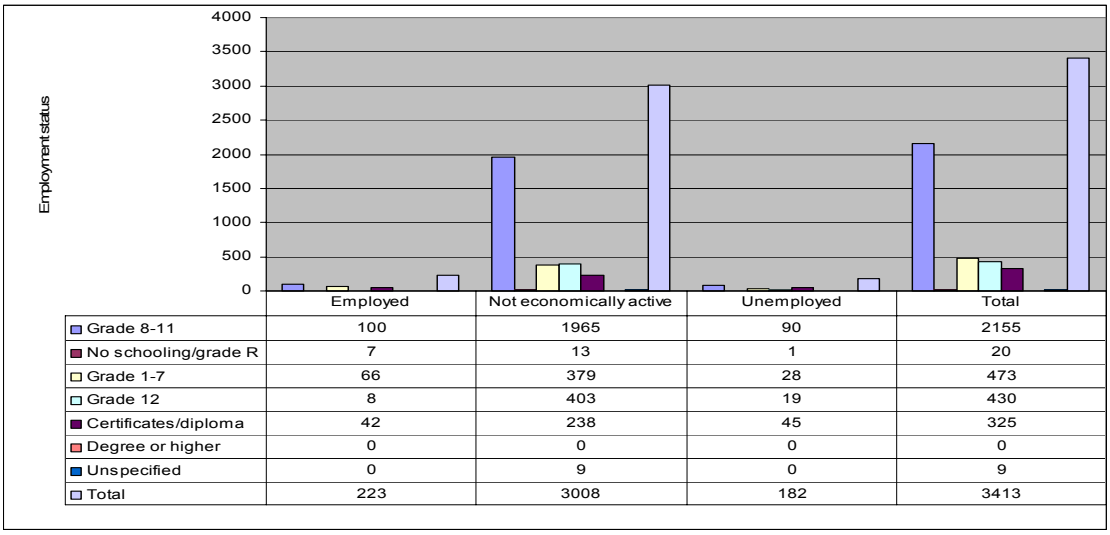
Table 3.2.4.4: Profiles of each subgroup formed by the CHAID analysis (March LFS 2006 and 2007)

Age group	Other predictors involved in the interactions			
Level 1	Level 2	Level 3	Level 4	Level 5
15-19	Highest level of Education	Education group 1: Province	Province 1: Sex	Source data
20-29	Population group	Population group 1: Province	Prov 1: Sex	Source data
			Prov 2: Highest level of education	Education levels 5 and 6: Sex
30-39	Highest level of Education	Education groups 1, 4, 9: Sex	Province	
			Marital status	
		Education groups 2, 3: Sex		
		Education groups 5,6: race		
40-49	Highest level of Education	Sex	Province	
		Province		
50-65	Sex	Highest level of education	Province	
		Marital status		

The age group 15-19 years old was split by highest level of education, and then further split by province and sex. People who have completed any of the grades 8-11 were further split by province (and then by sex); whereas the other categories couldn't split further. Some of the categories of the highest level of education were merged into one composite class, reducing the

number of categories as each category fails to be significant at 5% significance level. The results of merging the predictor variable categories (before merge and after merge) are presented in Figure 3.2.4.2 and Figure 3.2.4.3 respectively. 91.18% of people who have completed any of grades 8-11 (category 1) were not economically active; 4.66% were employed and 4.16% were unemployed. Category 1 (Gauteng) of province was further split by sex.

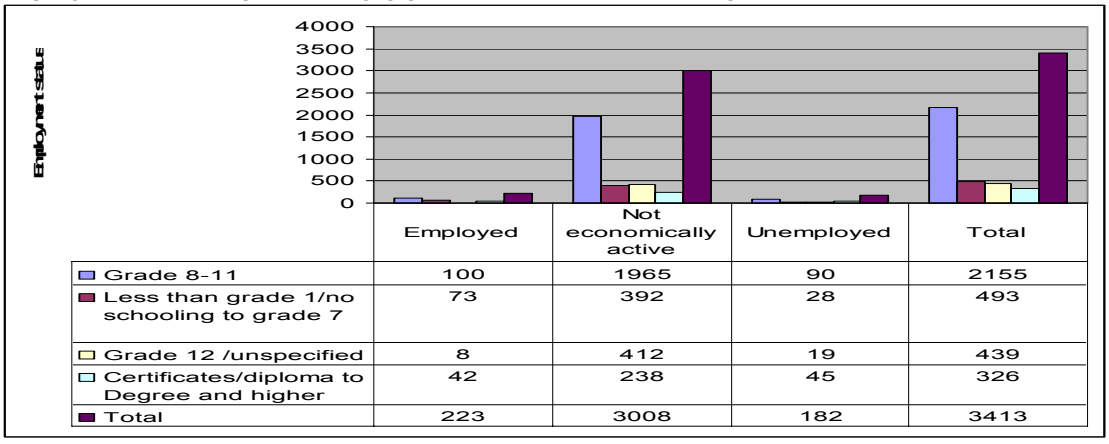
Figure 3.2.4.2: Level of education among people in age group 15-19 years old by employment status (before merge) (March LFS 2006 and 2007)



	Employed	Not economically active	Unemployed	Total
Grade 8-11	100	1965	90	2155
No schooling/grade R	7	13	1	20
Grade 1-7	66	379	28	473
Grade 12	8	403	19	430
Certificates/diploma	42	238	45	325
Degree or higher	0	0	0	0
Unspecified	0	9	0	9
Total	223	3008	182	3413

* LR chi-square=167.53 df=12 prob=1.5e-29

Figure 3.2.4.3: Level of education among people in age group 15-19 years old by employment status (after merge) (March LFS 2006 and 2007)



	Employed	Not economically active	Unemployed	Total
Grade 8-11	100	1965	90	2155
Less than grade 1/no schooling to grade 7	73	392	28	493
Grade 12 /unspecified	8	412	19	439
Certificates/diploma to Degree and higher	42	238	45	326
Total	223	3008	182	3413

* LR chi-square=160.42 df=6 prob=1.7e-29 (adj.)

It follows from the tree diagram that the age group 20-29 years old was further split by population group, sex, highest level of education, and source data. An African black was the only category of population group which was involved in a further split. It was further split by sex and level of education, which is an indication of overlap in their analytic information. In other words, sex and level of education were both influential with respect to related indicators of employment status.

The profiles of the people aged 30-39 years old and 40-49 years old are similar. Predictor variables that are involved in the interactions among the age group 30-39 years old are highest level of education, sex and population group, province and marital status. Province and marital status were competing with one another in the fourth level of interaction for the age group 30-39 years old.

The predictor variables that are involved in the interactions among 40-49 year old are highest for level of education, sex and province. In the 3rd order interactions for age group 40-49 years old, sex and province were competing with one another. The last age group category, age 50-65 years old, splits further by sex, highest level of education, marital status and province. It is interesting to note that 39.02% of females were employed as compared with 60.27% of the combination of males and people who did not specify their sex. Females were further split by highest level of education, whereas the composite class (male and unspecified) was split by marital status. Both groups were further split by province.

3.3 Multinomial logistic regression

3.3.1 Introduction

This section provides a theoretical discussion of multinomial logistic regression, together with an illustration of the technique using the 2006 and 2007 March LFS analysis. We will also perform different analyses to check the adequacy of the model obtained. Logistic regression will be used as an alternative method to CHAID. As discussed in section 3.1, CHAID is a totally non-parametric method, in which there could be several competing splits for any node. Having obtained a 'good' predictive model, we would then want to see if the results can be approximated well using a more parametric model. This technique can detect interactions between predictor variables. The results of the logistic regression will enable us to gain an understanding of the strengths of, and interrelationships between predictors, as well as making predictions.

3.3.2 Multinomial logistic regression model

This technique can be seen as an extension of standard regression analysis, a general statistical technique used to analyse the relationship between a single dependent variable and several independent variables, to generalised linear models. The models are appropriate for the analysis of a dichotomous response variable as well as a categorical response variable. Hosmer and Lemeshow (2000) outlined the notation for the analysis of a dichotomous response variable and a single independent variable, which they later expanded to multinomial logistic analysis (for a categorical response variable with two or more categories).

In logistic regression, the dependent variable is the result of a transformation of an underlying response variable to linearity. Like any regression problem, the key quantity is the mean value of the response variable, given the value of the predictor variable. This quantity is expressed as

E(Y | x), where Y denotes the response variable and x denotes the value of the predictor variable. E(Y | x) is the expected value of Y given the value of x. We will use the notation $\pi(x) = P(Y = 1 | x)$ for simplicity, assuming that the binary response is coded as 1 for yes, and 0 for no. The logistic regression is given by:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \text{ where } 0 \leq \pi(x) \leq 1$$

In logistic regression model the link function is the logit transformation of $\pi(x)$ which has the linear form:

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x$$

According to Hosmer and Lemeshow (2000), multiple logistic regression generalises the binomial logistic model to the case of more than one independent variable. For the purposes of illustration, let us consider a collection of p independent variables (say x_1 = gender; x_2 = province; … x_p = highest level of education) denoted by the vector $x' = (x_1, x_2, ..., x_p)$. The logit of the multiple logistic regression model is given by the equation $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ such that the logistic regression is

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

The slope coefficient (β_i) represents the change in the logit corresponding to a change of one unit in the independent variable x_i [i.e. $\beta_i = g(x_i + 1) - g(x_i)$], with the intercept being given by β_0.
If some of the independent variables are discrete, such as race or sex, they are transformed into dummy variables. In general, if a categorical variable has p possible values, then p -1 dummy variables will be needed. For example, a three category variable would be transformed into two dummy variables, d_1 and d_2, where the base category has the values (0, 0) and the other two categories have (1, 0) and (0, 1).

Table 3.3.2.1 lists all the variables and their response categories that will be used in this study. In most analyses, employment status was used as the target (response) variable and the remaining variables were used as predictor variables. In other cases, the survey (source of the data) was used.

It is possible that it is not just the individual variables as listed below that are predictive of the employment status, but that there could be interactions. The results from CHAID in Table 3.2.4.4 indicate that there are indeed further important predictors which vary over the age categories, so that the interactions should be put into the model as well.

Table 3.3.2.1: List of variables for multinomial logistic regression analysis

Variable	Abbreviation	Response category
1. Employment status	EM	1=Not employment active 0 =Employed 2= Unemployed
2. Province	PR	0= Gauteng 1= Eastern Cape
3. Gender (sex)	SE	0= Male 1= Female 2= Unspecified
4. Age group	AG	1=15-19 0=20-29 2=30- 39 3=40-49 4=50-65
5. Highest level of education	ED	1= Less than grade 1/no schooling 2= Grade 1 – grade 7 0=Grade 8 -11 3=Grade 12 4=Certificate/diploma 5=Degree and higher 6=Unspecified
6. Population group (race)	RA	0=African black 1=Coloureds 2= Indian/Asian 3=Whites 4=Unspecified
7. Marital status	MA	1= Married 2=Living together like husband and wife 3=Widow/widower 4= Divorced/separated 0= Never married 5=Unspecified
8. Source_data	DS	1=March LFS 2006 2=March LFS 2007 3=September LFS 2006 4=September LFS 2007 5=July GHS 2006 6=September GHS 2007 2=October CS 2007

3.3.3 Fitting the multinomial regression model

Suppose we have a sample of n independent observations of the response and predictor variables (y_i, x_{1i}, x_{2i}, …, x_{pi}), i=1,2,…,n. The model requires that we obtain the estimates of regression coefficients. We will use the maximum likelihood method to obtain the estimated coefficients.

The fitted model may or may not be the appropriate one, so we will need to check /evaluate the overall model fit. This is done by testing the null hypothesis that the full set of predictors are all unnecessary, i.e. testing (β_1, β_2, … β_p) = 0. The likelihood ratio test statistic is used for testing this hypothesis. If this is significant, then at least one of the variables is predictive of the response. However, it is very possible that some of the predictor variables are not significant.

We will then look at the significance of each of the predictor variables, to determine which individual variables are significant in the model, and which could be dropped. This is done using the Wald chi-square statistic for the individual predictor variables. This test statistic gives the importance of the contribution of each predictor to the model.

A higher value of the Wald estimate indicates the importance of such a predictor variable. The probability value (p-value) indicates whether each of the predictors significantly improves the predictive ability of the model, given the other variables that are in the model. Each p-value will be compared with the given threshold level (0.05) for dropping the corresponding predictor from the model. If the p-value is less than 0.05, we will reject the hypothesis that the parameter is equal to zero. It should be noted that the aim is to find the most efficient combination of the predictor variables to use in the model (Hosmer and Lemeshow, 2000).

In order to obtain the most efficient model, there are four selection methods available in the literature: forward selection, backward elimination, stepwise selection, and best subset selection of including predictor variables in the model. The best subset selection is based on the likelihood score statistic. This method identifies a specified number of best models containing one, two, three effects, and so on, up to a single model containing effects for all the explanatory variables.

A forward selection method involves starting with no predictor variables in the model. The procedure tries out the predictor variables one by one and includes them if they are statistically significant. Once a predictor variable is entered in the model it remains in the model. The backward elimination method starts with all possible predictor variables in the model. Each predictor variable is tested for its statistical significance, and any predictor that is not significant is deleted. Once the variable has been excluded from the model, it will remain excluded. The stepwise method is similar to the forward selection method except that there is a chance that predictor variables already in the model can be removed. The process of looking at variables entering into and being removed from the model is similar to a forward selection step followed by a backward elimination procedure. These methods may not necessarily give the same results. The results in this report are based on backward elimination analysis.

It is important that we choose the reference group/category for each variable. Note that this group is necessary, as each observation must belong to one of the categories of each variable. Having the category with the largest sample size as the base group gives the most efficient test. For example, the employed category from employment status was chosen as reference category. This means that regression coefficients of all other employment dummy variables will be evaluated against this reference group.

3.3.4 Interpretation of the fitted model

In order to interpret the fitted model, it is important to consider what the estimated coefficients tell us about the relationships between predictor variables and employment status. According to Hosmer and Lemeshow (2000) the estimated coefficient represents the rate of change of the response variable which compares the changes in the unit of

predictor variable. In the case of a dichotomous variable, we assume that the predictor variable, x, is coded as either zero or one. The difference in the logit is:

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$$

The first step in interpreting the effect of a covariate in a model is to express the desired logit difference in terms of the model, which is in this case is equal to β_1. In order to interpret this result we need to introduce and discuss a measure of association termed the odds ratio. The odds for x=0 amongst individuals with x = 1 is defined as $\pi(0)/[1-\pi(0)]$. The odds ratio, denoted by OR, is defined as the ratio of the odds for x = 1 to the odds for x = 0, and is given by the equation

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \qquad 3.3.4.1$$

and thus

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\Big/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\Big/\left(\frac{1}{1+e^{\beta_0}}\right)}$$

$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$

$$= e^{(\beta_0+\beta_1)-\beta_0}$$

$$= e^{\beta_1}$$

This shows the relationship between the odds ratio and the regression coefficient for logistic regression with a dichotomous independent variable coded 1 or 0. We can expand this relationship to more categories by keeping the x values constant. There are two important aspects to consider when interpreting a logistic model; the relationships between the response and the predictors; and the determination of the changes in the response to a unit of change in the model. The odds ratio shows the strength of association between a predictor and the response of interest. The coefficients can be positive or negative. A positive coefficient indicates an increase in the odds of a person being in the higher employment status category, whereas, the negative coefficient will result in reduced odds.

We will also use the confidence intervals, which determine the significance of the predictor variables. In the case where the confidence limits do not include zero, the effect is said to be statistically significant. This interval gives the estimated range of values, which is likely to include the estimated population parameter. This gives the percentage (e.g. 95%) of confidence as to how uncertain we are about the estimated parameter.

3.3.5 Assessing the adequacy of the predicted model

The classification table will be used to assess how well the model predicts when comparing the actual events and the predicted values. This table cross-classifies the true state and that predicted by the model. This shows events that were correctly predicted or where misclassification occurs. The number on the diagonal of the matrix represents the number of individuals who were correctly classified. According to Hair *et al.* (1995) the overall percentage of people correctly classified is given by (100*(the sum of the number on the diagonal matrix))/(the actual total).

3.3.6 Illustration of logistic analysis methodology using March LFS 2006 and 2007

For the purpose of illustration of the multinomial regression analysis, the March LFS 2006 and 2007 data will be used. SAS® 9.2 was used to develop the model. The objective of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. In order to achieve this, we have built a model that includes all useful predictor variables in predicting employment status.

The output in Table 3.3.9.1 gives the model fit statistics, while Table 3.3.9.2 tests whether any of the terms in the model are needed. In order to be confident that the multinomial logistic regression gave the correct model, the overall relationship should at least be statistically significant. It follows from the results that all three tests yield similar conclusions. Since our test statistics are significant at 0.05, we reject the null hypothesis that there are no relationships between employment status and the set of predictor variables.

Table 3.3.9.1: Model Fit Statistics (March LFS 2006 and 2007)

	Intercept Only	Intercept and Covariates
AIC	63358.749	57485.705
SC	63375.479	58564.808
-2 Log L	63354.749	57227.705

Table 3.3.9.2: Testing Global Null Hypothesis: BETA=0 (March LFS 2006 and 2007)

	Chi-Square	DF	Pr> ChiSq
Likelihood Ratio	6127.0440	127	<.0001
Score	5402.4604	127	<.0001
Wald	4757.3644	127	<.0001

Table 3.3.9.3 lists the output of the effect of each predictor variable and their interactions contributed to the model. The results show several significant interactions of up to the seven factors. Since the highest order interaction is significant, this means that the change over the provinces differs over the combinations of (interactions between) the other 6 variables. Since the reference categories are unemployed and EC, this implies that the odds of being unemployed is 1.027 times higher for people in the EC than in GP, if the other variables are held constant in the model.

Table 3.3.9.3: Parameter estimates from the logistic regression model (March LFS 2006 and 2007)

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
								Lower	Upper
Intercept	0	1	-0.742	0.124	35.562	<.0001	0.476	-0.986	-0.498
Intercept	1	1	1.487	0.125	142.046	<.0001	4.424	1.243	1.732
AG		1	0.015	0.050	0.089	0.766	1.015	-0.083	0.113
ED		1	-0.059	0.053	1.229	0.268	0.943	-0.163	0.045
AG*ED		1	0.099	0.023	18.782	<.0001	1.104	0.054	0.143
MA		1	0.284	0.166	2.926	0.087	1.329	-0.041	0.610
AG*MA		1	-0.024	0.042	0.331	0.565	0.976	-0.107	0.059
ED*MA		1	-0.035	0.075	0.220	0.639	0.966	-0.181	0.111
AG*ED*MA		1	-0.008	0.020	0.141	0.707	0.992	-0.047	0.032
RA		1	-0.069	0.272	0.064	0.800	0.933	-0.602	0.464
AG*RA		1	0.038	0.085	0.201	0.654	1.039	-0.129	0.206
ED*RA		1	0.080	0.083	0.911	0.340	1.083	-0.084	0.243
AG*ED*RA		1	-0.026	0.027	0.889	0.346	0.975	-0.079	0.028
MA*RA		1	-0.513	0.262	3.833	0.050	0.599	-1.027	0.001
AG*MA*RA		1	0.083	0.061	1.849	0.174	1.086	-0.037	0.202
ED*MA*RA		1	0.269	0.096	7.839	0.005	1.309	0.081	0.458
AG*ED*MA*RA		1	-0.054	0.023	5.554	0.018	0.947	-0.099	-0.009
SE		1	-0.222	0.174	1.626	0.202	0.801	-0.563	0.119
AG*SE		1	0.136	0.077	3.127	0.077	1.145	-0.015	0.286
ED*SE		1	0.069	0.076	0.822	0.365	1.071	-0.080	0.217
AG*ED*SE		1	-0.077	0.035	4.876	0.027	0.926	-0.145	-0.009
MA*SE		1	0.326	0.357	0.835	0.361	1.386	-0.373	1.026
AG*MA*SE		1	-0.078	0.089	0.772	0.380	0.925	-0.253	0.096
ED*MA*SE		1	0.476	0.185	6.624	0.010	1.610	0.114	0.838
AG*ED*MA*SE		1	-0.086	0.047	3.422	0.064	0.917	-0.178	0.005
RA*SE		1	-0.008	0.344	0.001	0.981	0.992	-0.682	0.665
AG*RA*SE		1	0.082	0.131	0.386	0.535	1.085	-0.176	0.339
ED*RA*SE		1	0.055	0.110	0.256	0.613	1.057	-0.160	0.270
AG*ED*RA*SE		1	0.004	0.041	0.011	0.918	1.004	-0.076	0.085
MA*RA*SE		1	1.466	0.756	3.760	0.053	4.332	-0.016	2.948
AG*MA*RA*SE		1	-0.320	0.189	2.860	0.091	0.726	-0.691	0.051
ED*MA*RA*SE		1	-0.475	0.248	3.675	0.055	0.622	-0.960	0.011
AG*ED*MA*RA*SE		1	0.096	0.060	2.589	0.108	1.101	-0.021	0.213
PR		1	-0.117	0.026	20.557	<.0001	0.889	-0.168	-0.067
AG*PR		1	0.043	0.010	18.699	<.0001	1.044	0.023	0.062
ED*PR		1	0.018	0.010	3.648	0.056	1.019	0.000	0.037
AG*ED*PR		1	-0.013	0.004	10.181	0.001	0.987	-0.021	-0.005
MA*PR		1	-0.019	0.029	0.419	0.518	0.981	-0.076	0.038
AG*MA*PR		1	-0.001	0.008	0.037	0.848	0.999	-0.016	0.014
ED*MA*PR		1	0.009	0.012	0.586	0.444	1.009	-0.015	0.033
AG*ED*MA*PR		1	0.001	0.003	0.094	0.759	1.001	0.006	0.008
RA*PR		1	0.000	0.044	0.000	0.992	1.000	-0.088	0.087
AG*RA*PR		1	0.000	0.014	0.000	0.989	1.000	-0.028	0.028
ED*RA*PR		1	0.006	0.013	0.182	0.669	1.006	-0.021	0.032
AG*ED*RA*PR		1	-0.001	0.004	0.013	0.910	0.999	0.009	0.008
MA*RA*PR		1	0.159	0.048	11.152	0.001	1.173	0.066	0.253
AG*MA*RA*PR		1	-0.032	0.011	8.420	0.004	0.968	-0.054	-0.010
ED*MA*RA*PR		1	-0.042	0.016	6.818	0.009	0.959	-0.074	-0.011
AG*ED*MA*RA*PR		1	0.009	0.004	5.447	0.020	1.009	0.001	0.017
SE*PR		1	-0.742	0.124	35.562	<.0001	0.476	-0.986	-0.498

Table 3.3.9.3: Parameter estimates from the logistic regression model (March LFS 2006 and 2007) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
							Lower	Upper
AG*SE*PR	1	-0.012	0.015	0.672	0.412	0.988	-0.041	0.017
ED*SE*PR	1	-0.008	0.013	0.335	0.563	0.992	-0.034	0.018
AG*ED*SE*PR	1	0.008	0.006	1.599	0.206	1.008	0.004	0.020
MA*SE*PR	1	0.184	0.062	8.673	0.003	1.201	0.061	0.306
AG*MA*SE*PR	1	-0.033	0.016	4.104	0.043	0.968	-0.064	-0.001
ED*MA*SE*PR	1	-0.076	0.030	6.647	0.010	0.927	-0.134	-0.018
AG*ED*MA*SE*PR	1	0.017	0.008	5.038	0.025	1.018	0.002	0.033
RA*SE*PR	1	0.016	0.057	0.076	0.782	1.016	-0.096	0.127
AG*RA*SE*PR	1	-0.019	0.022	0.740	0.390	0.981	-0.061	0.024
ED*RA*SE*PR	1	-0.012	0.018	0.447	0.504	0.988	-0.047	0.023
AG*ED*RA*SE*PR	1	0.002	0.007	0.124	0.725	1.002	-0.011	0.015
MA*RA*SE*PR	1	-0.212	0.129	2.685	0.101	0.809	-0.465	0.042
AG*MA*RA*SE*PR	1	0.045	0.031	2.116	0.146	1.046	-0.016	0.106
ED*MA*RA*SE*PR	1	0.099	0.042	5.678	0.017	1.105	0.018	0.181
AG*ED*MA*RA*SE*PR	1	-0.021	0.010	4.764	0.029	0.979	-0.040	-0.002
DS	1	-0.084	0.176	0.225	0.635	0.920	-0.429	0.262
AG*DS	1	-0.078	0.071	1.231	0.267	0.925	-0.217	0.060
ED*DS	1	-0.113	0.075	2.274	0.132	0.893	-0.260	0.034
AG*ED*DS	1	0.036	0.032	1.278	0.258	1.037	-0.026	0.098
MA*DS	1	-0.340	0.234	2.108	0.147	0.712	-0.798	0.119
AG*MA*DS	1	0.076	0.059	1.638	0.201	1.079	-0.040	0.192
ED*MA*DS	1	0.206	0.107	3.699	0.054	1.229	0.004	0.416
AG*ED*MA*DS	1	-0.055	0.029	3.731	0.053	0.946	-0.111	0.001
RA*DS	1	-0.208	0.381	0.297	0.586	0.813	-0.954	0.539
AG*RA*DS	1	0.069	0.121	0.323	0.570	1.071	-0.169	0.306
ED*RA*DS	1	0.121	0.119	1.030	0.310	1.129	-0.113	0.355
AG*ED*RA*DS	1	-0.034	0.039	0.743	0.389	0.967	-0.110	0.043
MA*RA*DS	1	1.332	0.440	9.157	0.003	3.789	0.469	2.195
AG*MA*RA*DS	1	-0.271	0.102	7.075	0.008	0.762	-0.471	-0.071
ED*MA*RA*DS	1	-0.331	0.153	4.657	0.031	0.718	-0.632	-0.030
AG*ED*MA*RA*DS	1	0.078	0.037	4.458	0.035	1.081	0.006	0.150
SE*DS	1	-0.090	0.247	0.134	0.715	0.914	-0.574	0.393
AG*SE*DS	1	-0.074	0.108	0.471	0.492	0.928	-0.287	0.138
ED*SE*DS	1	0.114	0.107	1.136	0.287	1.121	-0.096	0.324
AG*ED*SE*DS	1	0.025	0.049	0.264	0.608	1.025	-0.071	0.121
MA*SE*DS	1	0.116	0.497	0.054	0.816	1.123	-0.859	1.090
AG*MA*SE*DS	1	0.027	0.124	0.047	0.829	1.027	-0.216	0.270
ED*MA*SE*DS	1	-0.280	0.252	1.227	0.268	0.756	-0.774	0.215
AG*ED*MA*SE*DS	1	0.037	0.064	0.330	0.566	1.037	-0.089	0.163
RA*SE*DS	1	0.568	0.498	1.299	0.254	1.765	-0.409	1.545
AG*RA*SE*DS	1	-0.024	0.180	0.017	0.896	0.977	-0.376	0.329
ED*RA*SE*DS	1	-0.081	0.165	0.244	0.622	0.922	-0.405	0.242
AG*ED*RA*SE*DS	1	-0.014	0.058	0.057	0.811	0.986	-0.127	0.100
MA*RA*SE*DS	1	-1.720	0.975	3.113	0.078	0.179	-3.630	0.191
AG*MA*RA*SE*DS	1	0.320	0.233	1.890	0.169	1.377	-0.136	0.776
ED*MA*RA*SE*DS	1	0.667	0.352	3.584	0.058	1.948	-0.024	1.357
AG*ED*MA*RA*SE*DS	1	-0.133	0.081	2.722	0.099	0.876	-0.291	0.025
PR*DS	1	-0.008	0.036	0.051	0.821	0.992	-0.080	0.063

Table 3.3.9.3: Parameter estimates from the logistic regression model (March LFS 2006 and 2007) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits Lower	Upper
AG*PR*DS	1	0.007	0.014	0.223	0.637	1.007	-0.021	0.034
ED*PR*DS	1	0.032	0.014	5.367	0.021	1.032	0.005	0.058
AG*ED*PR*DS	1	0.003	0.006	0.213	0.645	0.997	-0.014	0.009
MA*PR*DS	1	0.073	0.041	3.205	0.073	1.075	0.007	0.152
AG*MA*PR*DS	1	-0.015	0.011	1.867	0.172	0.986	-0.035	0.006
ED*MA*PR*DS	1	-0.046	0.017	7.128	0.008	0.955	-0.080	-0.012
AG*ED*MA*PR*DS	1	0.008	0.005	3.098	0.078	1.008	0.001	0.018
RA*PR*DS	1	0.055	0.062	0.765	0.382	1.056	-0.068	0.177
AG*RA*PR*DS	1	-0.016	0.020	0.655	0.418	0.984	-0.055	0.023
ED*RA*PR*DS	1	-0.028	0.019	2.078	0.150	0.973	-0.065	0.010
AG*ED*RA*PR*DS	1	0.007	0.006	1.306	0.253	1.007	0.005	0.020
MA*RA*PR*DS	1	-0.265	0.075	12.630	0.000	0.767	-0.411	-0.119
AG*MA*RA*PR*DS	1	0.055	0.017	10.034	0.002	1.056	0.021	0.089
ED*MA*RA*PR*DS	1	0.064	0.025	6.663	0.010	1.066	0.015	0.112
AG*ED*MA*RA*PR*DS	1	-0.015	0.006	6.197	0.013	0.985	-0.026	-0.003
SE*PR*DS	1	0.039	0.049	0.629	0.428	1.039	-0.057	0.134
AG*SE*PR*DS	1	0.026	0.021	1.573	0.210	1.026	-0.015	0.067
ED*SE*PR*DS	1	-0.014	0.019	0.559	0.455	0.986	-0.051	0.023
AG*ED*SE*PR*DS	1	-0.010	0.009	1.320	0.251	0.990	-0.027	0.007
MA*SE*PR*DS	1	-0.120	0.085	2.000	0.157	0.887	-0.286	0.046
AG*MA*SE*PR*DS	1	0.011	0.022	0.247	0.619	1.011	-0.032	0.054
ED*MA*SE*PR*DS	1	0.063	0.040	2.468	0.116	1.065	-0.016	0.141
AG*ED*MA*SE*PR*DS	1	-0.009	0.011	0.713	0.399	0.991	-0.030	0.012
RA*SE*PR*DS	1	-0.104	0.082	1.628	0.202	0.901	-0.264	0.056
AG*RA*SE*PR*DS	1	0.007	0.030	0.054	0.816	1.007	-0.051	0.065
ED*RA*SE*PR*DS	1	0.012	0.026	0.195	0.659	1.012	-0.040	0.063
AG*ED*RA*SE*PR*DS	1	0.002	0.009	0.056	0.814	1.002	-0.016	0.021
MA*RA*SE*PR*DS	1	0.251	0.161	2.423	0.120	1.286	-0.065	0.567
AG*MA*RA*SE*PR*DS	1	-0.048	0.038	1.627	0.202	0.953	-0.122	0.026
ED*MA*RA*SE*PR*DS	1	-0.128	0.056	5.108	0.024	0.880	-0.238	-0.017
AG*ED*MA*RA*SE*PR*DS	1	0.027	0.013	4.360	0.037	1.027	0.002	0.052

The cross tabulation of some of the categories of the predictor variables has zero respondents, as they are a linear combination of other variables. The model does not enter such interactions (see Table 3.3.9.4 below for an illustration of a zero cell). This occurs mainly for interactions including the 'not specified' category.

Table 3.3.9.4: Sex by marital status (March LFS 2006 and 2007)

Sex	Marital status 1	2	3	4	5	6	Total
1	7684.07	4958.96	1576.27	1021.75	605.359	4.79506	15851.2
2	8871.47	4822.64	1603.12	210.37	366.016	7.12046	15880.7
3	2.05689	0	0	0	0	0	2.05689
Total	16557.6	9781.6	3179.39	1232.1	2 971.375	11.9155	31734

The predicted probabilities were validated by using a classification table. Table 3.3.9.5 shows the classification matrix used to determine the model accuracy. The results give the classification

accuracy of 45.5% for employed, 39.6% for not economically active and 15.1% for unemployed people. The overall percentage correctly classified is 100*[(9742.8+9748.3+4.76)/31771] = 61.4%.

Table 3.3.9.5: Assessment of the adequacy of the model in percentages (March LFS 2006 and 2007)

LFS Predicted	Employed	Not Economically active	Unemployed	LFS profile
Employed	70.0	35.3	23.3	45.33
Not economically active	20.3	73.9	35.2	39.57
Unemployed	9.7	26.1	41.5	15.1
Predicted total	100.0	100.0	100.0	100.0
Regression profile	43.8	41.5	14.7	

Although 61.4% appears reasonably good, the logistic regression is predicting the employed group total to give about the correct percentage, but is predicting neither the not economically active nor the unemployed groups well.

CHAID models the structure of the data, by breaking this down into homogeneous subgroups. CHAID gives the prediction of the employment profile of a subgroup, not a prediction for the individual observations. CHAID finds that source data is only significant in age group 2, race 1, and province 1. It is interesting that the difference between the sources is opposite for the two sexes.

3.4 Discussion of the CHAID and logistic regression results

The results of both techniques give divisions into subgroups. Age group was the most significant predictor on which data on employment status was segmented. Highest level of education was the predictor mostly involved in the first order interactions. Other predictors that took part in the first order interactions were population group and sex. Province and marital status were only involved in the second and third order interactions, and source data was involved in the fourth order interactions.

CHAID partitioned the data into five different subgroups as per categories of age group. Age group is a categorical variable such that the older a person is the better the chances of being employed as opposed to unemployed or not economically active (until you get close to the retirement age). Young people between the ages of 15 and 19 years old are often still busy with their studies and are largely not economically active – the data shows that 88.13% of people in this age group are not economically active. Employment status in this age group can be further explained in terms of highest level of education attained, province and sex.

The results show that persons from the age group 20–29 years old, who are likely to come directly from completing their studies, only 39.37% were employed; 36.89% were not economically active and 23.73% were unemployed. It is possible that some of those who were declared not economically active may still further their higher education studies. Employment

status in this age group can further be explained in terms of province, sex, highest level of education and source data. People aged within the age groups 30-39 and 40-49 years old have almost the same characteristics such as highest level of education, sex, province and marital status. The last subgroup of age group 50-65 years old revealed that 60.27% of females and unspecified were still employed as compared to 39.02% of males. Highest level of education, marital status and province were the other predictor variables that played a significant role in explaining employment status in this subgroup.

The results of the logistic regression were not as good as those of CHAID. In the multinomial logistic regression technique, the highest order interaction was significant, meaning that the change over the provinces differs over the combinations of (interaction between) the other 6 variables.

The overall percentage correctly classified and identified by logistic regression model was 61.4%. The logistic regression model revealed several first and second order interactions as indicated by CHAID. The results from both techniques point out some similarities and differences regarding the contribution of the predictor variables in the model.

CHAPTER 4: Results

This research, as indicated in the previous chapters, used data from Stats SA household surveys to compare estimates of employment status over time and between surveys. To compare estimates of employment status over time, we will use data from September LFS 2006 and 2007; and July GHS 2006 and September GHS 2007 as stated in section 1.4. The comparison from September LFS 2006 and 2007 will be used to check the differences found in the March LFS 2006 and 2007 analyses that were discussed in Chapter 3.To compare estimates of employment status across surveys, we will use data from LFS September 2006 and GHS September 2006; and LFS September 2007, GHS July 2007 and CS October 2007.

CHAID and multinomial logistic regression will be used to determine the predictor variables that best predict employment status from each set of data. Both techniques use similar criteria as discussed in Chapter 3 as a basis of comparison. This chapter presents the results of these further analyses. The first two sections present the comparison of survey estimates over time (September LFS and GHS) and further sections present the analysis of data obtained from across surveys (LFS, GHS and CS).

4.1 September LFS 2006 and 2007

4.1.1 Results for CHAID (employment status as response variable)

The output of the CHAID analysis is presented in Figure 4.1.1.1 below, and all predictor variables which are statistically significant are listed in Table 4.1.1.1. Province was the only predictor variable not significant in predicting employment status at the first level of splitting. It follows from the tree diagram that province was not involved in any interaction with other predictor variables.

Table 4.1.1.1: List of significant predictors (September LFS 2006 and 2007)

Predictor	p-value	Levels	Groups
Age group	5.7e-2118	5	1 2 3 4 5
Marital status	5.0e-544	6->4	1 25 3 49
Highest level of education	1.8e-491	7->6	1 29 3 4 5 6
Gender	3.8e-217	3->2	1 29
Population group	3.7e-158	5->3	1 2 3-9
Source data	1.7e-19	2	1 2

The results in the table above show that age group was the most highly significant predictor. Age group (p-value = 5.7e-2118) explains more of the variation in employment status than any other predictor in the analysis. Marital status (p-value = 5.0e-544) is the second most significant predictor.

Some categories of the predictor variables were merged into one composite class, reducing the number of categories as each category fails to be significant at the 5% significance level. The

categories of marital status were reduced from six to four. Category 2 (married) and 5 (divorced/separated) have a similar profile and were merged into one composite class. Category 4 (widow/widower) and 9 (unspecified) have the same profile and were also merged into one composite class. Highest level of education was also reduced from seven to six response categories. People with no formal education (category 2) have a similar profile to those who did not specify their highest level of education (category 9).

Category 2 (male) of sex has a similar profile to that of category 9 (unspecified) and they were merged into one composite class. Categories 3 (Indian/Asian), 4 (Whites) and 9 (unspecified) for population group have the same profile and were merged to form one composite class. Source data (1=September LFS 2006 and 2=September LFS 2007) is significant but much less significant than age, indicating that one needs to take a look at the source of the data set in the different age groups, to see where the significance comes from.

The results of the CHAID analysis shows that at the root node the most significant split was obtained by segmenting the cases containing employment status into 5 different age groups (see Figure 4.1.1.1).

Figure 4.1.1.1: Classification tree diagram for September LFS 2006 and 2007

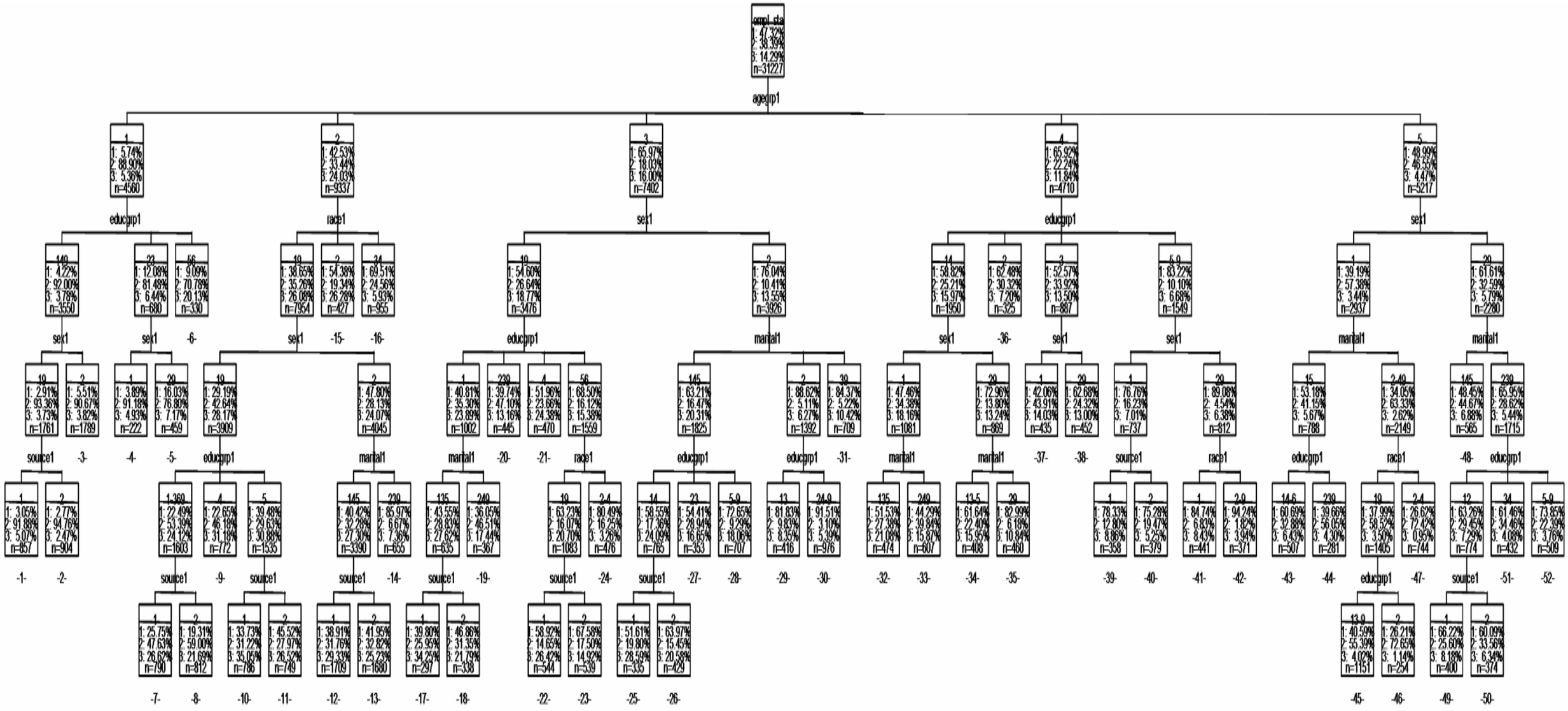


Table 4.1.1.2: Employment status per categories of age group (September LFS 2006 and 2007)

Agegrp	Age group	Employed	Not economically active	Unemployed	Sample size
1	15-19	5.74	88.90	5.36	4560
2	20-29	42.53	33.44	24.03	9337
3	30-39	65.97	18.03	14.29	7402
4	40-49	65.92	22.24	11.84	4710
5	50-65	48.99	46.55	4.47	5217

Table 4.1.1.2 summarises trends with regard to age group categories in relation to the employment status. As we have seen in Table 3.2.4.2, this subgroup (people aged 15-19 years old) comprises mainly people who are not economically active (88.90%). The chance of a person aged between 15 and 19 years old being employed is again less than that of a person aged 20-29 years old and older. The results also show that the percentages of unemployed people increase among the age group 20-29 years old; 24.03% of these people were unemployed, 42.53% were employed and 33.44% were not economically active.

The percentage of people aged 30 to 39 years old who were employed increased slightly to 65.97%, whereas we have seen a decline in the other categories of employment status. 18.03% of people in this age category were not economically active and 14.29% were unemployed. The percentage of people who were employed (65.92%) largely remains unchanged, whereas the percentage of not economically active people has decreased slightly to 22.24% while the percentage of people who were unemployed has increased slightly to 11.84%. The percentage of people who were employed has marginally increased to 48.99%; there is a decrease in the percentage of people who were not economically active to 46.55% and a marginal increase to 4.47% among people who were unemployed.

As before, CHAID then takes each remaining predictor in turn to determine the next segmenting variable. The results were used to understand the predictive power of the predictor variables used and their inter-relationships. At the second level of partitioning, it was found that highest level of education, population group and sex were still the most significant predictors. The three predictors were competing with each other within the categories of age group. Table 4.1.1.3 indicates that for age group 30-39 years old, the most predictive variable is sex. The most predictive variable for age group 40-49 years old was the highest level of education, with sex being the most predictive variable for the age group 50-65 years old. The least significant predictor variable for age group 15-19 years old was highest level of education. Age group 20-29 years old was partitioned by population group.

Table 4.1.1.3: Age group categories by predictors involved in the first order interactions and their p-values (September LFS 2006 and 2007)

Age group	First order interaction	Likelihood ratio chi-square	Degree of Freedom	p-value
15-19	Highest level of Education	207.35	4	3.0e-41
20-29	Population group	323.07	4	2.9e-67
30-39	Sex	504.12	8	1.3e-101
40-49	Highest level of Education	419.36	6	6.9e-85
50-65	Sex	348.18	2	7.4e-76

The different groups could be split further. Table 4.1.1.4 lists the predictors used at different branch levels of each subgroup identified by CHAID.

Table 4.1.1.4: Profiles of each subgroup formed by CHAID analysis (September LFS 2006 and 2007)

Age group	Other predictors involved in the interactions			
15-19	Highest level of education	Sex	Source data	
20-29	Population group	Sex	Marital status	Source data
			Highest level of education	Source data
30-39	Sex	Highest level of education	Marital status	Source data
			Population group	Source data
		Marital status	Highest level of education	Source data
40-49	Highest level of education	Sex	Marital status	
			Source data	
			Population group	
50-65	Sex	Marital status	Highest level of education	Source data

4.1.2 Results for multinomial logistic regression (employment status as response variable)

A model was developed in a similar way to that in Section 3.3.6. All the useful predictor variables and possible interactions were included in the model. Table 4.1.2.1 and Table 4.1.2.2 below give the model fit statistics, and the tests of the overall fit of the model. It follows from the results that all three tests are significant at 0.05. Hence, we reject the null hypothesis that there are no relationships between employment status and the set of predictor variables.

Table 4.1.2.1: Model Fit Statistics (March LFS 2006 and 2007)

	Intercept Only	Intercept and Covariates
AIC	62436.373	56268.215
SC	62453.071	57345.241
-2 Log L	62432.373	56010.215

Table 4.1.2.2: Testing Global Null Hypothesis: BETA=0 (March LFS 2006 and 2007)

	Chi-Square	DF	Pr> ChiSq
Likelihood Ratio	6422.1584	127	<.0001
Score	5656.2280	127	<.0001
Wald	4944.0951	127	<.0001

Table 4.1.2.3 gives the parameter estimates for the individual variables and for interactions between variables and their significance. The results show several significant interactions of up to seven factors. The highest order interaction is significant. One interpretation is that the change over the provinces differs over the combinations of (interaction between) the other 6 variables.

A negative sign of the coefficient gives us an indication of the negative contribution of the predictor variables in the model. The results show several significant interactions of up to the seven factors. Since the highest order interaction is significant, this means that the change over the provinces differs over the combinations of (interactions between) the other 6 variables. Since the reference categories are unemployed and EC, this implies that the odds of being unemployed is 0.864 times lower for people in the EC than in GP, if the other variables are held constant in the model. We are 95% confident that the true value estimated as -0.1469 is within the range (-0.2732 and -0.019).

Table 4.1.2.3: Parameter estimates from the logistic regression model (September LFS 2006 and 2007)

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits Lower	Upper
Intercept	0	1	-1.3601	0.0964	199.2361	<.0001	0.257	-1.5489	-1.1712
Intercept	1	1	0.8049	0.0961	70.1773	<.0001	2.236	0.6166	0.9932
AG		1	0.3464	0.0334	107.2621	<.0001	1.414	0.2808	0.4119
ED		1	0.1691	0.0314	28.954	<.0001	1.184	0.1075	0.2307
AG*ED		1	0.00343	0.0127	0.0732	0.7867	1.003	-0.0214	0.0283
MA		1	0.0198	0.0868	0.0521	0.8194	1.02	-0.1504	0.19
AG*MA		1	-0.0546	0.0238	5.2577	0.0219	0.947	-0.1012	-0.00792
ED*MA		1	-0.0105	0.0305	0.1192	0.7299	0.99	-0.0702	0.0492
AG*ED*MA		1	0.0111	0.00917	1.4738	0.2247	1.011	-0.00684	0.0291
RA		1	0.0285	0.0969	0.0864	0.7688	1.029	-0.1614	0.2184
AG*RA		1	0.00128	0.0398	0.001	0.9743	1.001	-0.0767	0.0793
ED*RA		1	0.1033	0.0308	11.2634	0.0008	1.109	0.043	0.1636
AG*ED*RA		1	-0.0448	0.011	16.4548	<.0001	0.956	-0.0664	-0.0231
MA*RA		1	0.1371	0.1028	1.7787	0.1823	1.147	-0.0644	0.3386
AG*MA*RA		1	-0.0411	0.0278	2.1804	0.1398	0.96	-0.0957	0.0135
ED*MA*RA		1	0.0518	0.0342	2.3009	0.1293	1.053	-0.0151	0.1188
AG*ED*MA*RA		1	-0.00485	0.0088	0.3031	0.582	0.995	-0.0221	0.0124
SE		1	0.2126	0.1249	2.8959	0.0888	1.237	-0.0323	0.4574
AG*SE		1	0.0329	0.0495	0.4419	0.5062	1.033	-0.0642	0.13
ED*SE		1	0.00832	0.0418	0.0396	0.8422	1.008	-0.0736	0.0902
AG*ED*SE		1	0.00361	0.0186	0.0378	0.8458	1.004	-0.0328	0.04
MA*SE		1	2.412	0.229	110.9379	<.0001	11.157	1.9632	2.8609
G*MA*SE		1	-0.3931	0.0584	45.2618	<.0001	0.675	-0.5077	-0.2786

Table 4.1.2.3: Parameter estimates from the logistic regression model (September LFS 2006 and 2007) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
							Lower	Upper
ED*MA*SE	1	-0.208	0.075	7.695	0.006	0.812	-0.355	-0.061
AG*ED*MA*SE	1	0.013	0.020	0.431	0.511	1.013	-0.027	0.053
RA*SE	1	0.027	0.138	0.037	0.847	1.027	-0.243	0.296
AG*RA*SE	1	-0.017	0.060	0.079	0.779	0.983	-0.134	0.101
ED*RA*SE	1	-0.012	0.041	0.083	0.773	0.988	-0.093	0.069
AG*ED*RA*SE	1	0.036	0.017	4.588	0.032	1.037	0.003	0.070
MA*RA*SE	1	-0.397	0.229	3.003	0.083	0.673	-0.845	0.052
AG*MA*RA*SE	1	0.037	0.057	0.427	0.514	1.038	-0.074	0.148
ED*MA*RA*SE	1	0.155	0.087	3.163	0.075	1.168	-0.016	0.326
AG*ED*MA*RA*SE	1	-0.029	0.020	2.145	0.143	0.971	-0.069	0.010
PR	1	0.149	0.131	1.283	0.257	1.160	-0.109	0.406
AG*PR	1	-0.250	0.050	25.185	<.0001	0.779	-0.347	-0.152
ED*PR	1	-0.145	0.050	8.392	0.004	0.865	-0.244	-0.047
AG*ED*PR	1	0.075	0.021	13.316	0.000	1.078	0.035	0.116
MA*PR	1	0.150	0.146	1.045	0.307	1.161	-0.137	0.437
AG*MA*PR	1	0.028	0.038	0.529	0.467	1.028	-0.047	0.103
ED*MA*PR	1	0.122	0.063	3.724	0.054	1.130	-0.002	0.246
AG*ED*MA*PR	1	-0.049	0.017	7.897	0.005	0.952	-0.083	-0.015
RA*PR	1	-0.145	0.250	0.337	0.561	0.865	-0.634	0.344
AG*RA*PR	1	0.046	0.079	0.331	0.565	1.047	-0.110	0.201
ED*RA*PR	1	0.064	0.075	0.721	0.396	1.066	-0.084	0.212
AG*ED*RA*PR	1	-0.021	0.024	0.775	0.379	0.979	-0.067	0.026
MA*RA*PR	1	0.056	0.280	0.040	0.841	1.058	-0.492	0.604
AG*MA*RA*PR	1	-0.005	0.065	0.006	0.940	0.995	-0.133	0.123
ED*MA*RA*PR	1	-0.046	0.092	0.244	0.621	0.955	-0.227	0.135
AG*ED*MA*RA*PR	1	0.019	0.022	0.731	0.393	1.019	-0.024	0.061
SE*PR	1	-0.156	0.176	0.788	0.375	0.856	-0.500	0.188
AG*SE*PR	1	0.109	0.074	2.216	0.137	1.116	-0.035	0.253
ED*SE*PR	1	0.119	0.069	2.946	0.086	1.126	-0.017	0.255
AG*ED*SE*PR	1	-0.100	0.031	10.287	0.001	0.905	-0.161	-0.039
MA*SE*PR	1	-1.498	0.337	19.775	<.0001	0.224	-2.159	-0.838
AG*MA*SE*PR	1	0.173	0.083	4.280	0.039	1.188	0.009	0.336
ED*MA*SE*PR	1	0.367	0.153	5.757	0.016	1.443	0.067	0.666
AG*ED*MA*SE*PR	1	-0.016	0.039	0.173	0.678	0.984	-0.092	0.060
RA*SE*PR	1	0.180	0.327	0.304	0.582	1.197	-0.460	0.820
AG*RA*SE*PR	1	0.109	0.119	0.844	0.358	1.115	-0.124	0.342
ED*RA*SE*PR	1	0.043	0.110	0.152	0.697	1.044	-0.173	0.258
AG*ED*RA*SE*PR	1	0.017	0.039	0.200	0.655	1.018	-0.059	0.094
MA*RA*SE*PR	1	0.740	0.699	1.120	0.290	2.095	-0.630	2.110
AG*MA*RA*SE*PR	1	-0.171	0.153	1.246	0.264	0.843	-0.470	0.129
ED*MA*RA*SE*PR	1	-0.383	0.206	3.480	0.062	0.682	-0.786	0.019
AG*ED*MA*RA*SE*PR	1	0.057	0.046	1.545	0.214	1.058	-0.033	0.146
DS	1	-0.174	0.134	1.693	0.193	0.840	-0.436	0.088
AG*DS	1	-0.051	0.048	1.115	0.291	0.951	-0.145	0.043
ED*DS	1	-0.078	0.044	3.173	0.075	0.925	-0.165	0.008
AG*ED*DS	1	0.001	0.018	0.002	0.962	1.001	-0.034	0.036
MA*DS	1	0.206	0.124	2.765	0.096	1.229	-0.037	0.449
AG*MA*DS	1	0.028	0.034	0.673	0.412	1.028	-0.038	0.094
ED*MA*DS	1	0.014	0.043	0.104	0.747	1.014	-0.071	0.099
AG*ED*MA*DS	1	-0.014	0.013	1.242	0.265	0.986	-0.040	0.011

Table 4.1.2.3: Parameter estimates from the logistic regression model (September LFS 2006 and 2007) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
							Lower	Upper
RA*DS	1	-0.065	0.147	0.197	0.657	0.937	-0.354	0.223
AG*RA*DS	1	-0.010	0.055	0.033	0.855	0.990	-0.117	0.097
ED*RA*DS	1	0.016	0.044	0.136	0.712	1.016	-0.070	0.103
AG*ED*RA*DS	1	0.034	0.015	5.002	0.025	1.035	0.004	0.064
MA*RA*DS	1	-0.089	0.183	0.236	0.627	0.915	-0.447	0.270
AG*MA*RA*DS	1	0.010	0.046	0.047	0.829	1.010	-0.079	0.099
ED*MA*RA*DS	1	0.018	0.055	0.105	0.746	1.018	-0.089	0.125
AG*ED*MA*RA*DS	1	-0.010	0.013	0.604	0.437	0.990	-0.037	0.016
SE*DS	1	0.413	0.175	5.593	0.018	1.512	0.071	0.756
AG*SE*DS	1	-0.003	0.071	0.002	0.967	0.997	-0.141	0.135
ED*SE*DS	1	-0.063	0.059	1.144	0.285	0.939	-0.177	0.052
AG*ED*SE*DS	1	0.021	0.026	0.630	0.427	1.021	-0.031	0.072
MA*SE*DS	1	-0.719	0.315	5.219	0.022	0.487	-1.336	-0.102
AG*MA*SE*DS	1	0.035	0.080	0.196	0.658	1.036	-0.121	0.192
ED*MA*SE*DS	1	0.196	0.106	3.457	0.063	1.217	-0.011	0.403
AG*ED*MA*SE*DS	1	-0.013	0.028	0.196	0.658	0.987	-0.068	0.043
RA*SE*DS	1	-0.122	0.196	0.384	0.536	0.886	-0.506	0.263
AG*RA*SE*DS	1	0.077	0.082	0.891	0.345	1.080	-0.083	0.237
ED*RA*SE*DS	1	0.039	0.058	0.436	0.509	1.039	-0.076	0.153
AG*ED*RA*SE*DS	1	-0.059	0.023	6.606	0.010	0.942	-0.105	-0.014
MA*RA*SE*DS	1	1.033	0.411	6.323	0.012	2.809	0.228	1.838
AG*MA*RA*SE*DS	1	-0.161	0.092	3.032	0.082	0.852	-0.341	0.020
ED*MA*RA*SE*DS	1	-0.331	0.130	6.435	0.011	0.719	-0.586	-0.075
AG*ED*MA*RA*SE*DS	1	0.070	0.029	5.836	0.016	1.072	0.013	0.127
PR*DS	1	0.264	0.184	2.072	0.150	1.302	-0.096	0.624
AG*PR*DS	1	-0.016	0.070	0.053	0.817	0.984	-0.154	0.122
ED*PR*DS	1	-0.060	0.070	0.729	0.393	0.942	-0.197	0.078
AG*ED*PR*DS	1	0.027	0.029	0.844	0.358	1.027	-0.030	0.084
MA*PR*DS	1	-0.250	0.209	1.439	0.230	0.779	-0.659	0.159
AG*MA*PR*DS	1	0.014	0.055	0.061	0.806	1.014	-0.094	0.121
ED*MA*PR*DS	1	0.082	0.094	0.766	0.381	1.085	-0.102	0.265
AG*ED*MA*PR*DS	1	-0.007	0.026	0.072	0.788	0.993	-0.057	0.043
RA*PR*DS	1	0.072	0.332	0.047	0.828	1.075	-0.579	0.723
AG*RA*PR*DS	1	0.000	0.108	0.000	0.998	1.000	-0.211	0.211
ED*RA*PR*DS	1	-0.026	0.101	0.065	0.799	0.975	-0.224	0.172
AG*ED*RA*PR*DS	1	-0.006	0.033	0.030	0.863	0.994	-0.070	0.058
MA*RA*PR*DS	1	-0.116	0.352	0.109	0.741	0.890	-0.806	0.573
AG*MA*RA*PR*DS	1	0.032	0.084	0.145	0.704	1.032	-0.132	0.196
ED*MA*RA*PR*DS	1	-0.010	0.118	0.007	0.936	0.991	-0.241	0.222
AG*ED*MA*RA*PR*DS	1	-0.006	0.029	0.038	0.847	0.994	-0.062	0.051
SE*PR*DS	1	-0.482	0.247	3.812	0.051	0.617	-0.966	0.002
AG*SE*PR*DS	1	-0.030	0.104	0.085	0.771	0.970	-0.235	0.174
ED*SE*PR*DS	1	0.067	0.097	0.470	0.493	1.069	-0.124	0.258
AG*ED*SE*PR*DS	1	-0.012	0.044	0.078	0.781	0.988	-0.098	0.074
MA*SE*PR*DS	1	0.000	0.457	0.000	1.000	1.000	-0.896	0.896
AG*MA*SE*PR*DS	1	0.148	0.115	1.657	0.198	1.160	-0.078	0.374

Table 4.1.2.3: Parameter estimates from the logistic regression model (September LFS 2006 and 2007) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
							Lower	Upper
ED*MA*SE*PR*DS	1	-0.319	0.205	2.418	0.120	0.727	-0.721	0.083
AG*ED*MA*SE*PR*DS	1	0.031	0.053	0.346	0.556	1.032	-0.073	0.135
RA*SE*PR*DS	1	0.170	0.429	0.156	0.693	1.185	-0.671	1.010
AG*RA*SE*PR*DS	1	-0.208	0.157	1.764	0.184	0.812	-0.515	0.099
ED*RA*SE*PR*DS	1	-0.1591	0.1402	1.2881	0.2564	0.853	-0.4338	0.1156
AG*ED*RA*SE*PR*DS	1	0.0385	0.0502	0.5886	0.443	1.039	-0.0599	0.137
MA*RA*SE*PR*DS	1	-1.0703	0.9032	1.4043	0.236	0.343	-2.8404	0.6999
AG*MA*RA*SE*PR*DS	1	0.2753	0.2047	1.8092	0.1786	1.317	-0.1258	0.6764
ED*MA*RA*SE*PR*DS	1	0.7105	0.2815	6.3709	0.0116	2.035	0.1588	1.2623
AG*ED*MA*RA*SE*PR*DS	1	-0.1462	0.0648	5.0871	0.0241	0.864	-0.2732	-0.0191

Table 4.1.2.4 below presents the classification accuracy percentage of the employment status categories. The following are the percentages correctly classified: 46.2% for employed, 39.8% for not economically active and 14.0% for unemployed people. The overall percentage correctly classified is 61.8%. This is similar to the results from March LFS 2006 and 2007.

Table 4.1.2.4: Assessment of the adequacy of the model in percentages (September LFS 2006 and 2007)

LFS Predicted	Employed	Not economically active	Unemployed	LFS profile
Employed	70.5	26.6	21.0	46.2
Not economically active	21.3	54.9	56.1	39.8
Unemployed	8.2	18.5	18.6	14.0
Predicted Total	100.0	100.0	100.0	100.0
Regression profile	44.7	55.2	0.1	

Although 61.8% does not sound too bad, the logistic regression is predicting the employed group total to the correct percentage, but is predicting neither the not economically active nor the unemployed groups well. CHAID models the structure of the data, by breaking this down into homogeneous subgroups. CHAID gives the prediction of the employment profile of a subgroup, not a prediction for the individual observations. CHAID finds that source data is significant in all age groups.

4.1.3 Discussion of the CHAID and logistic regression results (employment status as response variable)

The results in Table 4.1.1.2 show that CHAID can be used effectively with employment data. The most important predictor of employment status is the age group. Marital status, highest level of education, sex, population group and source data have significant influences on predicting employment status by interacting with age group and some of the other predictor variables in different stages of the tree. The results show that province has no influence in predicting employment status in this study. Some categories of the predictor variables were merged to form one composite group. Category 9 (unspecified) of all the predictor variables contains one or more

zero observations when cross-tabulating with other variables. This category contains the lowest response as compared to other categories.

In order to understand the predictive power of the predictor variables used and their inter-relationships, each subset of age group was further partitioned by the remaining the predictors. The results show that highest level of education, population group and sex were the most significant predictors at the second level. The end result of the CHAID analysis gives a profile of people's employment status. It should be noted that the completed percentage in the tree trends upwards.

From the CHAID analysis the strongest predictor of employment status was age group. Age is a continuous variable such that the older a person is the better the chances of such person to be employed. Young people between the ages of 15 and 19 years old are often still busy with their studies and are largely not economically active. CHAID had revealed that 88.90% of people among this age group are not economically active. Employment status in this age group can further be explained in terms of highest level of educational attainment, sex and different data source. The results do not indicate much difference between males and females; and neither do September LFS 2006 and September LFS 2007, in terms of explaining employment status.

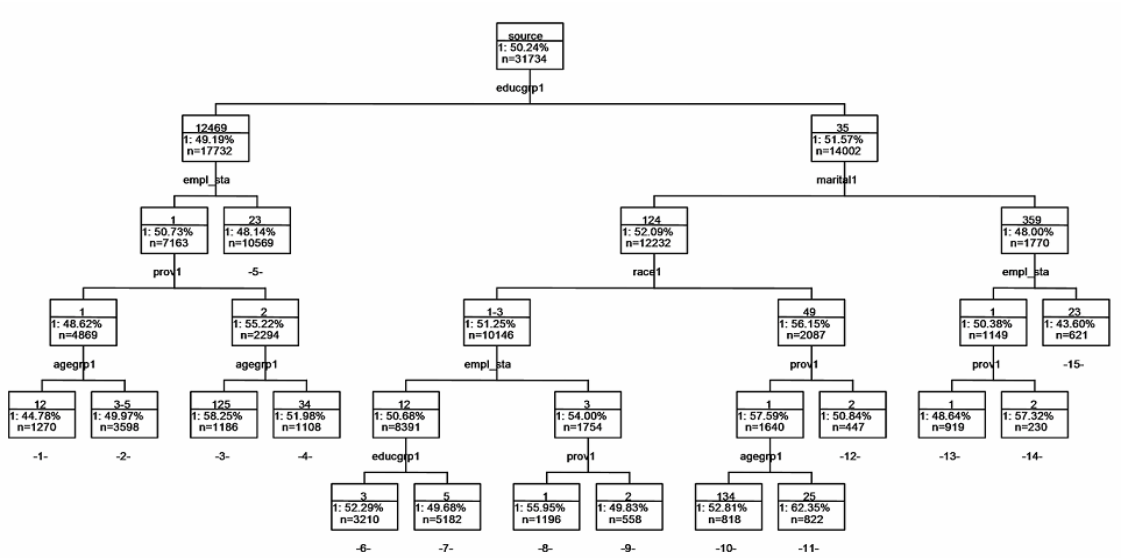
The results showed that in the age group 20-29 years old, only 42.53% were employed; 33.44% were not economically active and 24.03% were unemployed. It is possible that some of those who were declared not economically active may still further their higher education studies. Employment status in this age group can further be explained in terms of province, population group or sex, highest level of education and source data. The trends of employment for the age group 30-39 years old and 40-49 years old have significantly increased to about 65.0%. The employment status between the two age groups cannot be easily separated. The last subgroup of age group 50-65 years old revealed that 61.61% of female and unspecified were still employed as compared to 39.19% of male. Marital status, highest level of education, population and source data were the other predictor variables that played a significant role in explaining employment status in this subgroup.

The results of the logistic regression were not as good as CHAID. In the multinomial logistic regression, the highest order interaction was significant, meaning that the change over the provinces differs over the combinations of interaction between the other 6 variables. The overall percentage correctly classified and identified by logistic regression model was 61.8%. Logistic regression revealed several first and second order interactions as indicated by CHAID. The results from both techniques point out some similarities and differences regarding the contribution of the predictor variables in the model.

4.1.4 March LFS 2006 and 2007 and September LFS 2006 and 2007 (source data as response variable)

We will now look at CHAID analysis when source data (March LFS 2006 and 2007) was taken as response variable. The results will help to determine the relationships between March LFS 2006 and March LFS 2007 defined by the predictor variables. The output of the CHAID tree is shown in Figure 4.1.4.1.

Figure 4.1.4.1: Classification tree diagram for March LFS 2006 and 2007 (source data)



All significant predictors are listed in Table 4.1.4.1. Highest level of education and employment status were the only significant predictors. Highest level of education (p-value =0.0013) was the most significant predictor associated with different data sources.

Table 4.1.4.1: List of significant predictors (March LFS 2006 and 2007)

Predictor	p-value	Levels	Groups
Highest level of education	0.0013	7->2	12469 35
Employment status	0.0088	3->2	13 2

The categories of highest level of education were merged and reduced into only two categories. It is surprising to see category 6 (degree or higher) merged with categories 1 to 4 (grade 12). Category 3 (grade 1-7) was merged with the category of those people who have a certificate or diploma. Table 4.1.4.2 was constructed in order to check the uncertainty. In this case CHAID has grouped together response categories for which the 2007 survey has a higher percentage than 2006, versus those where the 2006 percentage is higher than the 2007 percentage.

Table 4.1.4.2: Percentage distribution of source data by highest level of education (March LFS 2006 and 2007)

Source data	% Highest level of education							
	1	2	3	4	5	6	9	Total
1	18.24	1.98	6.81	4.94	15.95	2.03	0.29	50.24
2	18.57	2.11	6.30	5.36	15.07	2.06	0.30	49.76
Total	36.81	4.09	13.10	10.30	31.02	4.09	0.59	100.00

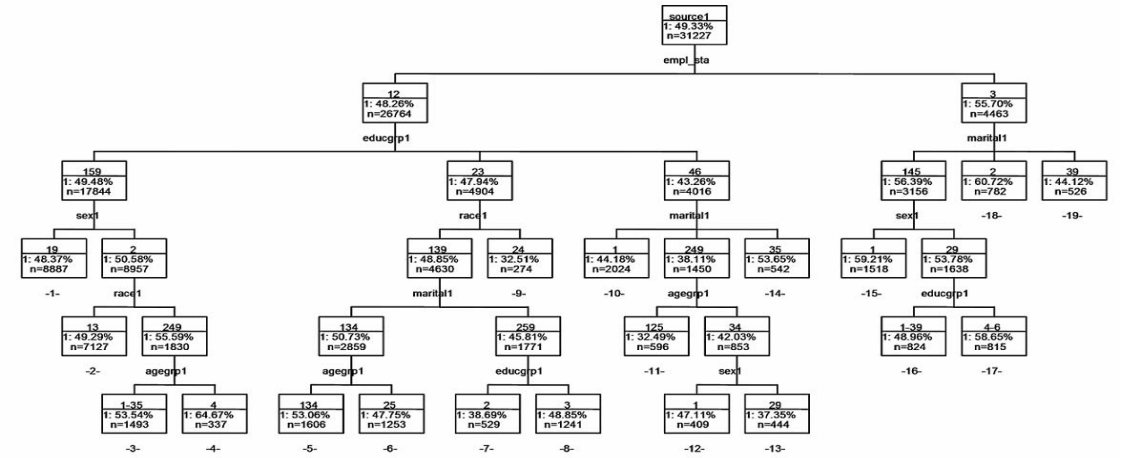
The CHAID tree shows that at the root node the most significant split was obtained by segmenting the cases containing source data into the above categories of highest level of education. The results of each sub-group formed are shown in Table 4.1.4.3.

Table 4.1.4.3: Profiles of each subgroup formed by CHAID analysis (September LFS 2006 and 2007)

	Other predictors involved in the interactions			
Highest level of education	Employment status	Province	Age group	
		Population group	Employment status	Age group
			Province	Province
				Age group
	Marital status	Employment status	Province	

The above results will be compared with the results from September LFS 2006 and 2007. The output of the CHAID tree of September LFS 2006 and 2007 results is shown in Figure 4.1.4.2.

Figure 4.1.4.2: Classification tree diagram for September LFS 2006 and 2007



All significant predictors are listed in Table 4.1.4.4. Employment status, highest level of education and age group were the only significant predictors. Employment status (p-value =4.5e-20) was now the most significant predictor associated with different data sources. This predictor variable explains more of the variation in both data sets than any other predictors in the analysis. This was followed by highest level of education with p-value of 6.7e-9 and age group (0.044). All other predictors were not significant as a single predictor variable.

Table 4.1.4.4: List of significant predictors (September LFS 2006 and 2007)

Predictor	p-value	Levels	Groups
Employment status	4.5e-20	3->2	12 3
Highest level of education	6.7e-9	7->4	159 23 4 6
Age group	0.044	5->2	124 35

Some categories of the predictor variables were merged into one composite class, reducing the number of categories as each category fails to be significant at 5% significance level. The categories of employment status were reduced from three to two. Category 1 (employed) has a similar profile as category 3 (not economically active) and they were merged into one composite class. Highest level of education was also reduced from seven to four response categories. People who have completed any secondary education level excluding grade 12 (i.e. grade 8-11), certificate/diploma and those who did not specify their highest level of education completed have similar profiles and were merged into one composite class. People with no formal education (category 2) have a similar profile to those who did complete any primary education level (category 3). The categories of age group were reduced from five to two. Category 1 (15-19 years old), 2 (20-29 years old) and 3 (40-49 years old) have a similar profile and were merged into one composite class.

The CHAID tree shows that at the root node, the data was partitioned by employment status. The analysis will then take each predictor variable in turn to determine the next segmenting variable (see Table 4.1.4.5). The data was partitioned into 19 nodes.

Table 4.1.4.5: Profiles of each subgroup formed by CHAID analysis (September LFS 2006 and 2007)

	Other predictors involved in the interactions			
Employed and not economically active	Highest level of education	Sex	Age group	Population group
		Population group	Marital status	Age group
				Highest level of education
Unemployed	Marital status	Sex	Highest level of education	

4.1.5 Discussion of the CHAID and logistic regression results (March LFS 2006 and 2007; September LFS 2006 and 2007)

Highest level of education and employment status were common in both sets of data. Age group was significant only for March LFS 2006 and 2007. It follows from the results that there is not much difference between the two sets of data. However, the two sets of data had different groupings of the category of employment status and highest level of education. Sex was not significant at all in March LFS 2006 and 2007, and province was not significant in September 2006 and 2007 data. The results gave variability between the two sets of data with regards to their relationships with the predictor variables. We could not analyse the results of logistic regression when source data was taken as the response variable since the overall model was not adequate. Hosmer and Lemeshow Goodness-of-Fit Test with Chi-Square (44.0286), degree of

freedom (8) and p-value less than 0.0001 indicate that the model does not fit. The results show that there are differences between the different data sets.

4.2 September GHS 2006 and July GHS 2007

4.2.1 Results for CHAID (employment status as response variable)

The output of the CHAID analysis is shown in Figure 4.2.1.1 and all predictor variables which are statistically significant are listed in Table 4.2.1.1. The results in Table 4.2.1.1 below indicate that age group was the most significant predictor variable. Age group (p-value =1.3e-1975) explains more of the variation in employment status than any other predictor in the analysis

Table 4.2.1.1: List of significant predictors (September GHS 2006 and July GHS 2007)

Predictor	p-value	Levels	Groups
Age group	1.3e-1975	5	1 2 3 4 5
Marital status	4.2e-593	6->5	1 2 3 49 5
Highest level of education	9.9e-577	7->6	1 2 3 4 59 6
Province	9.0e-419	2	1 2
Sex	1.2e-254	3->2	19 2
Population group	3.3e-187	5->4	1 2 3 49
Source	2.9e-5	2	1 2

Some categories of the predictor variables were collapsed into one composite class, reducing the number of categories as each category fails to be significant at the 5% significance level. The categories of marital status were reduced from six to five. The fourth category (widow/widower) has a similar profile to that of category nine (unspecified) and they were merged into one composite class. Highest level of education was also reduced from seven to six response categories. People who had completed either a certificate or diploma (category 5) had a similar profile to those who did not specify their highest level of education (category 9).

Category 1 (female) of sex had a similar profile to that of category 9 (unspecified) and these were merged into one composite class. Categories 4 (whites) and 9 (unspecified) for population group had a similar profile and they were merged to form one composite class. Source data (1=September GHS 2006 and 2=July GHS 2007) is significant but much less significant than age, indicating that one needs to take a look at the source of the data in the different age groups, to see where the significance comes from.

The CHAID tree show that, at the root node, the most significant split was obtained by segmenting the cases containing employment status into 5 different age groups (see Figure 4.2.1.1).

Figure 4.2.1.1: Classification tree diagram for September GHS 2006 and July GHS 2007

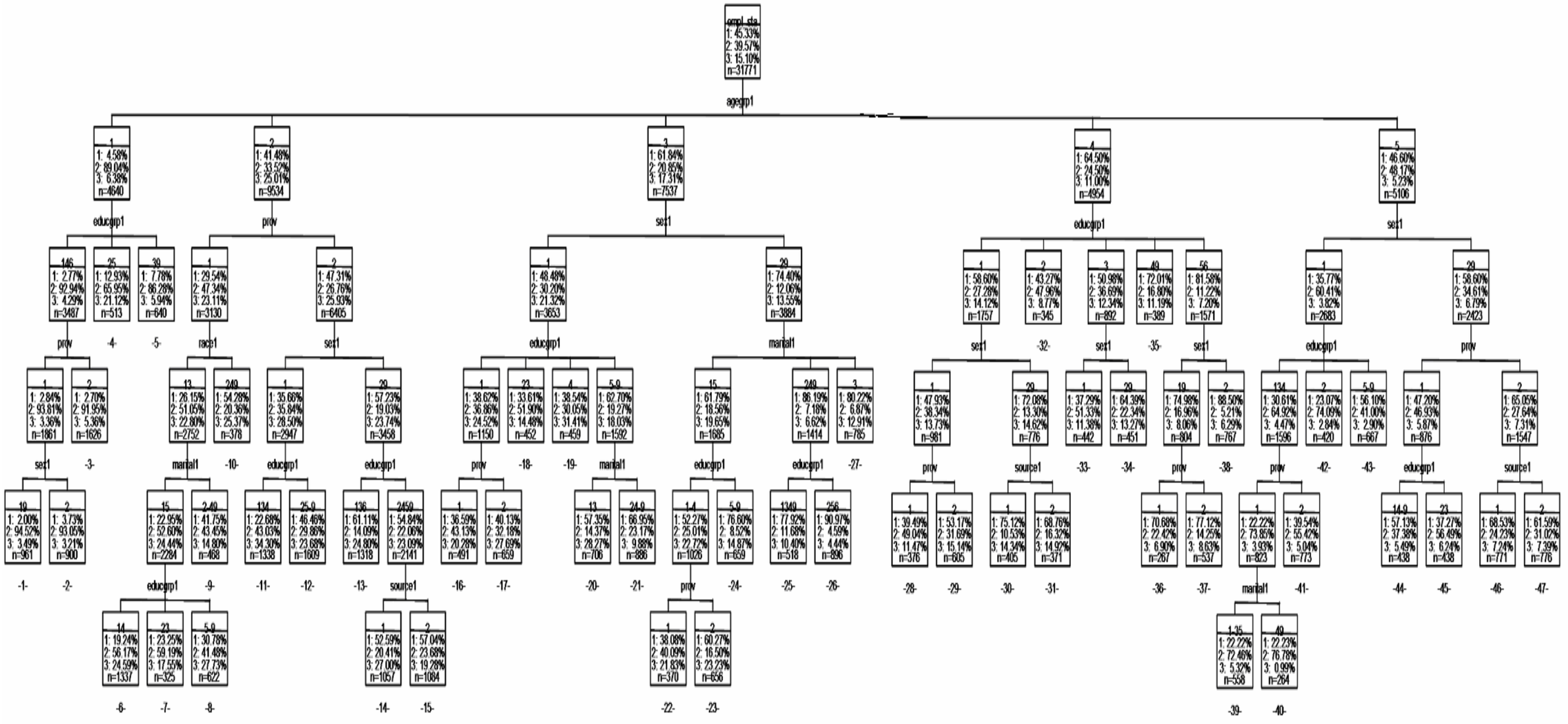


Table 4.2.1.2: Employment status per category of age group (September GHS 2006 and July GHS 2007)

agegrp	Age group	Employed	Not economically active	Unemployed	Sample size
1	15-19	4.58	89.04	6.38	4640
2	20-29	41.48	33.52	25.01	9534
3	30-39	61.841	20.85	17.31	75.37
4	40-49	64.50	24.50	11.00	4954
5	50-65	46.60	48.17	5.23	5106

Table 4.2.1.2 summarises trends with regard to age group categories in relation to the employment status. As we have seen in the previous sections, the subgroup of people aged 15-19 years old comprises mainly people who are not economically active (89.04%). The chance of a person aged between 15 and 19 years old being employed is again less than that of a person aged 20-29 years old and higher. The results also show that the percentage of unemployed people increases among the age group 20-29 years old; 25.01% of these people were unemployed, 41.48% were employed and 33.52% were not economically active.

The percentage of people aged 30-39 years old who were employed increased slightly to 61.84%, whereas there has been a proportional decline in the other categories of employment status: 20.85% of people in this age category were not economically active and 17.31% were unemployed. There has been a slight increase in the percentage of people who were employed (64.50%) and not economically active (24.50%) in the age group 40-49 years old. The percentage of people who were unemployed has decreased to 11.00%. The age group 50-65 years old changes significantly. The percentage of people who were employed has decreased to 46.60%; a large increase to 48.17% on the percentage of people who were not economically active and a significant drop to 5.23% among people who were unemployed. This result indicates the comparative relationships between age and employment status. The chances of a younger person to be employed are less compared to the higher age groups. It also shows that people's chance of being employed when they are approaching their retirement age also becomes less.

As before, CHAID then takes each remaining predictor in turn to determine the next segmenting variable. Following employment status down the tree, one is able to see which of the attributes comprise each terminal node. The results were used to understand the predictive power of the predictor variables used and their inter-relationships. At the second level of partitioning it was found that highest level of education, province and sex were the most significant predictors. The three predictors are competing with each other within the categories of age group. Table 4.2.1.3 indicates that for age group 30-39 years old, the most predictive variable is sex. The most predictive variable for age group 40-49 years old is highest level of education, with sex being the most predictive variable for the age group 50-65 years old. The least significant predictor variable for age group 15-19 years old is highest level of education. Age groups 20-29 years old and 50-65 years old are partitioned by sex.

Table 4.2.1.3: Age group categories by predictors involved in the first order interactions (September GHS 2006 and July GHS 2007)

Age group	1st order interactions	Likelihood ratio chi-square	Degree of freedom	p-value
15-19	Highest level of education	267.20	4	3.9e-54
20-29	Province	432.66	2	1.1e-94
30-39	Sex	539.56	2	2.1e-117
40-49	Highest level of education	505.24	8	7.4e-102
50-65	Sex	365.78	2	1.14e-79

The different groups could be split further. Table 4.2.1.4 lists the predictors used at different branch levels of each subgroup identified by CHAID.

Table 4.2.1.4: Profiles of each subgroup formed by CHAID analysis (September GHS 2006 and July GHS 2007)

Age group	Other predictors involved in the interactions			
15-19	Highest level of education	Province	Sex	
20-29	Province	Population group	Marital status	Highest level of education
		Sex	Highest level of education	
30-39	Sex	Highest level of education	Province	
			Marital status	Source data
		Marital status	Highest level of education	Source data
40-49	Highest level of education	Sex	Province	
			Source data	
50-65	Sex	Highest level of education	Province	Marital status
		Province	Highest level of education	
			Source data	

4.2.2 Results for multinomial logistic regression (employment status as response variable)

A model was developed in a similar way to that in the previous sections. We have included all the potentially useful predictor variables and possible interactions in the model. Table 4.2.2.1 and Table 4.2.2.2 below give the model fit statistics, and the tests of whether any predictors in the model are useful. All three tests are significant at 0.05. Hence, we reject the null hypothesis that there are no relationships between employment status and the set of predictor variables.

Table 4.2.2.1: Model Fit Statistics (September GHS 2006 and July GHS 2007)

	Intercept only	Intercept and Covariates
AIC	64242.077	58218.206
SC	64258.810	58979.540
-2 Log L	64238.077	58036.206

Table 4.2.2.2: Testing Global Null Hypothesis: BETA=0 (September GHS 2006 and July GHS 2007)

	Chi-Square	DF	Pr> ChiSq
Likelihood Ratio	6201.8713	89	<.0001
Score	5491.1279	89	<.0001
Wald	4912.1349	89	<.0001

Table 4.2.2.3 lists the output of the effect of each predictor variable and their interactions contributing in the model. The results show several significant interactions up to the five factor interactions. There were no significant six and seven factor interactions. Since the highest order interaction (five factor interaction) is significant, this means that the change over the provinces differs over the combinations of the other 4 variables.

Table 4.2.2.3: Parameter estimates from the logistic regression model (September GHS 2006 and July GHS 2007)

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits Lower	Upper
Intercept	0	1	-1.211	0.081	225.085	<.0001	0.298	-1.370	-1.053
Intercept	1	1	0.964	0.081	143.264	<.0001	2.621	0.806	1.122
AG		1	0.078	0.032	5.762	0.016	1.081	0.014	0.141
ED		1	-0.040	0.034	1.386	0.239	0.961	-0.106	0.027
AG*ED		1	0.069	0.014	26.161	<.0001	1.072	0.043	0.096
MA		1	0.107	0.088	1.488	0.223	1.113	-0.065	0.279
AG*MA		1	0.000	0.022	0.000	0.998	1.000	-0.042	0.043
ED*MA		1	0.122	0.037	11.150	0.001	1.130	0.050	0.194
AG*ED*MA		1	-0.042	0.009	19.381	<.0001	0.959	-0.060	-0.023
RA		1	-0.153	0.120	1.644	0.200	0.858	-0.388	0.081
AG*RA		1	0.071	0.038	3.401	0.065	1.073	-0.004	0.146
ED*RA		1	0.173	0.031	31.799	<.0001	1.189	0.113	0.234
AG*ED*RA		1	-0.043	0.008	29.987	<.0001	0.958	-0.058	-0.028
MA*RA		1	0.138	0.086	2.593	0.107	1.148	-0.030	0.305
AG*MA*RA		1	-0.040	0.016	6.458	0.011	0.961	-0.070	-0.009
ED*MA*RA		1	0.015	0.013	1.347	0.246	1.015	-0.011	0.041
SE		1	-0.049	0.110	0.201	0.654	0.952	-0.266	0.167
AG*SE		1	0.092	0.048	3.664	0.056	1.096	-0.002	0.186
ED*SE		1	0.082	0.048	2.913	0.088	1.086	-0.012	0.176
AG*ED*SE		1	-0.045	0.021	4.886	0.027	0.956	-0.085	-0.005
MA*SE		1	1.259	0.118	113.154	<.0001	3.522	1.027	1.491
AG*MA*SE		1	-0.263	0.019	198.847	<.0001	0.769	-0.299	-0.226
ED*MA*SE		1	0.082	0.040	4.247	0.039	1.086	0.004	0.160
RA*SE		1	0.036	0.160	0.051	0.821	1.037	-0.277	0.349
AG*RA*SE		1	0.079	0.052	2.291	0.130	1.083	-0.023	0.182
ED*RA*SE		1	0.025	0.031	0.651	0.420	1.025	-0.035	0.085
MA*RA*SE		1	0.508	0.153	11.062	0.001	1.661	0.209	0.807
AG*MA*RA*SE		1	-0.161	0.024	44.847	<.0001	0.851	-0.208	-0.114
ED*MA*RA*SE		1	0.039	0.015	6.297	0.012	1.039	0.008	0.069
PR		1	-0.246	0.119	4.252	0.039	0.782	-0.479	-0.012
AG*PR		1	0.177	0.043	16.657	<.0001	1.193	0.092	0.262
ED*PR		1	0.054	0.044	1.505	0.220	1.056	-0.033	0.141
AG*ED*PR		1	-0.045	0.017	7.001	0.008	0.956	-0.078	-0.012
MA*PR		1	-0.140	0.096	2.102	0.147	0.870	-0.329	0.049
AG*MA*PR		1	0.017	0.023	0.539	0.463	1.017	-0.028	0.062
ED*MA*PR		1	-0.061	0.037	2.658	0.103	0.941	-0.134	0.012
AG*ED*MA*PR		1	0.027	0.010	7.664	0.006	1.028	0.008	0.046
RA*PR		1	0.296	0.133	4.944	0.026	1.344	0.035	0.556
AG*RA*PR		1	-0.149	0.041	13.060	0.000	0.862	-0.230	-0.068
ED*RA*PR		1	-0.045	0.032	1.906	0.167	0.956	-0.108	0.019
AG*ED*RA*PR		1	0.026	0.008	9.514	0.002	1.026	0.009	0.043
MA*RA*PR		1	0.124	0.067	3.441	0.064	1.132	0.007	0.255

Table 4.2.2.3: Parameter estimates from the logistic regression model (September GHS 2006 and July GHS 2007) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
							Lower	Upper
SE*PR	1	0.293	0.157	3.493	0.062	1.340	-0.014	0.600
AG*SE*PR	1	0.041	0.061	0.451	0.502	1.042	-0.079	0.161
ED*SE*PR	1	0.058	0.060	0.927	0.336	1.060	-0.060	0.177
AG*ED*SE*PR	1	0.030	0.024	1.552	0.213	1.031	-0.017	0.078
MA*SE*PR	1	0.055	0.115	0.228	0.633	1.056	-0.171	0.281
ED*MA*SE*PR	1	-0.131	0.041	10.421	0.001	0.877	-0.211	-0.052
RA*SE*PR	1	-0.314	0.176	3.178	0.075	0.731	-0.658	0.031
AG*RA*SE*PR	1	0.063	0.057	1.207	0.272	1.065	-0.049	0.175
ED*RA*SE*PR	1	-0.086	0.032	7.184	0.007	0.917	-0.150	-0.023
MA*RA*SE*PR	1	-0.072	0.120	0.361	0.548	0.930	-0.308	0.163
DS	1	0.071	0.111	0.415	0.519	1.074	-0.146	0.288
AG*DS	1	-0.009	0.042	0.041	0.839	0.991	-0.092	0.075
ED*DS	1	0.011	0.046	0.053	0.818	1.011	-0.080	0.101
AG*ED*DS	1	0.006	0.018	0.118	0.731	1.006	-0.028	0.040
MA*DS	1	0.141	0.097	2.121	0.145	1.152	-0.049	0.331
AG*MA*DS	1	-0.056	0.022	6.354	0.012	0.945	-0.100	-0.013
ED*MA*DS	1	-0.061	0.032	3.640	0.056	0.941	-0.123	0.002
AG*ED*MA*DS	1	0.028	0.009	10.026	0.002	1.028	0.011	0.045
RA*DS	1	0.015	0.143	0.011	0.917	1.015	-0.264	0.294
AG*RA*DS	1	-0.035	0.043	0.667	0.414	0.965	-0.120	0.049
ED*RA*DS	1	-0.015	0.018	0.694	0.405	0.985	-0.050	0.020
MA*RA*DS	1	-0.054	0.111	0.237	0.626	0.947	-0.271	0.163
AG*MA*RA*DS	1	0.049	0.020	6.270	0.012	1.050	0.011	0.087
ED*MA*RA*DS	1	-0.029	0.015	3.868	0.049	0.972	-0.057	0.000
SE*DS	1	0.022	0.157	0.020	0.888	1.022	-0.285	0.329
AG*SE*DS	1	0.005	0.065	0.006	0.936	1.005	-0.122	0.132
ED*SE*DS	1	0.063	0.067	0.882	0.348	1.065	-0.068	0.194
AG*ED*SE*DS	1	-0.009	0.026	0.108	0.743	0.991	-0.061	0.043
MA*SE*DS	1	0.084	0.117	0.511	0.475	1.087	-0.146	0.313
ED*MA*SE*DS	1	-0.122	0.036	11.601	0.001	0.885	-0.192	-0.052
RA*SE*DS	1	-0.256	0.196	1.708	0.191	0.774	-0.641	0.128
AG*RA*SE*DS	1	0.110	0.069	2.603	0.107	1.117	-0.024	0.245
MA*RA*SE*DS	1	-0.202	0.144	1.972	0.160	0.817	-0.484	0.080
PR*DS	1	-0.244	0.156	2.427	0.119	0.784	-0.550	0.063
AG*PR*DS	1	0.064	0.054	1.422	0.233	1.066	-0.041	0.169
ED*PR*DS	1	0.117	0.058	4.025	0.045	1.124	0.003	0.230
AG*ED*PR*DS	1	-0.039	0.020	3.814	0.051	0.961	-0.079	0.000
MA*PR*DS	1	0.020	0.071	0.077	0.782	1.020	-0.119	0.158
RA*PR*DS	1	0.111	0.151	0.544	0.461	1.117	-0.184	0.406
AG*RA*PR*DS	1	0.029	0.047	0.392	0.531	1.030	-0.062	0.120
MA*RA*PR*DS	1	-0.198	0.091	4.760	0.029	0.820	-0.376	-0.020
SE*PR*DS	1	0.668	0.217	9.488	0.002	1.949	0.243	1.092
AG*SE*PR*DS	1	-0.267	0.082	10.721	0.001	0.766	-0.427	-0.107
ED*SE*PR*DS	1	-0.266	0.083	10.246	0.001	0.767	-0.429	-0.103
AG*ED*SE*PR*DS	1	0.083	0.030	7.461	0.006	1.086	0.023	0.142
MA*SE*PR*DS	1	0.199	0.133	2.250	0.134	1.220	-0.061	0.459
RA*SE*PR*DS	1	0.303	0.216	1.963	0.161	1.354	-0.121	0.727
AG*RA*SE*PR*DS	1	-0.186	0.076	5.989	0.014	0.830	-0.335	-0.037
MA*RA*SE*PR*DS	1	0.462	0.162	8.121	0.004	1.587	0.144	0.779

Table 4.2.2.4 lists the five factor interactions. A negative sign of the coefficient of the interaction among age group, population group, sex, province, and source data gives us an indication of the negative contribution of the interaction variables in the model. The interactions will decrease the odds of a person not being economically active, and unemployed compared to those who were employed by 0.83 if the other variables are held constant in the model. Marital status and highest level of education are not significant in this interaction. Marital status and highest level of education are also not significant in this model.

The interactions of AG*ED*SE*PR*DS and MA*RA*SE*PR*DS both contribute positively in the model. Marital status and population group were not significant in the interactions among age group, highest level of education, sex, province, and source data; whereas age group and highest level of education were not significant among the interactions marital status, sex, province, and source data. This is a contradiction to the results we obtained from CHAID as age group and highest level of education were always significant.

Table 4.2.2.5 below presents the classification accuracy percentage of employment status categories. The following are the percentages correctly classified: 45.3% for employed, 39.6% for not economically active and 15.1% for unemployed people. The overall percentage correctly classified is 61.95%.

Table 4.2.2.4: List of significant five factor interactions (September GHS 2006 and July GHS 2007)

Interactions	Estimate	Exp(Est)
AG*ED*SE*PR*DS	0.0826	1.086
AG*RA*SE*PR*DS	-0.186	0.83
MA*RA*SE*PR*DS	0.4617	1.587

Table 4.2.2.5 below presents the classification accuracy percentage of the employment status categories. The following are the percentages correctly classified: 45.3% for employed, 39.6% for not economically active and 15.1% for unemployed people. The overall percentage correctly classified is 61.95%.

Table 4.2.2.5: Assessment of the adequacy of the model in percentages (September GHS 2006 and July GHS 2007)

LFS Predicted	Employed	Not economically active	Unemployed	GHS profile
Employed	71.2	25.7	20.9	45,3
Not economically active	19.4	54.9	25.8	39,6
Unemployed	9.3	19.4	76.6	15,1
Predicted Total	100	25.7	20.9	100
Regression profile	43.2	56.8	0.0	

Although 61.95% appears reasonably good, the logistic regression is predicting the employed group total to give the correct percentage, but is predicting neither the not economically active nor

the unemployed groups well. CHAID models the structure of the data, by breaking this down into homogeneous subgroups. CHAID gives the prediction of the employment profile of a subgroup, not a prediction for the individual observations. CHAID finds that source data is significant in age groups 40-49 years old and 50-55 years old.

4.2.3 Discussion for CHAID and logistic regression (employment status as response variable)

The results in Table 4.2.1.2 show that CHAID can be used effectively with employment data. The most important predictor of employment status is the age group. Marital status, highest level of education, sex, population group and source data have a significant influence on predicting employment status by interacting with age group, and some of the predictor variables in different stages of the tree. The results show that province has no influence in predicting employment status in this study. Some categories of the predictor variables were merged to form one composite group. Category 9 (unspecified), of all the predictor variables, contains zero observations when cross-tabulated with other variables. This category contains the lowest responses as compared to other categories.

CHAID split the employment status data into different age groups. The results show that the highest level of education, population group, and sex were the most significant predictors at the second level. The most strongly associated predictor of employment status was age group. Highest level of education, marital status, province, sex, source data, and population group have significant influence on predicting employment status by interacting with age group and some of the predictor variables in different stages of the tree. The data was further partitioned into five different subgroups as per categories of age groups.

CHAID revealed that 89.04% of people in the age group 15-19 years old are not economically active. Employment status in this age group can further be explained in terms of highest level of education attained, province and sex. The results showed that for the group 20-29 that consisted mostly of recent graduates or people who are still studying, 41.48% were employed, 33.52% were not economically active and 25.01% were unemployed. Employment status in this age group can further be explained in terms of province, population group or sex, marital status, and highest level of education. The trends of employment for the age group 30-39 years old and 40-49 years old have increased greatly to about 62.0%. The last age group, 50-65 years old, reveals that 58.60% of females and persons of unspecified sex are still employed as compared to 35.77% of males.

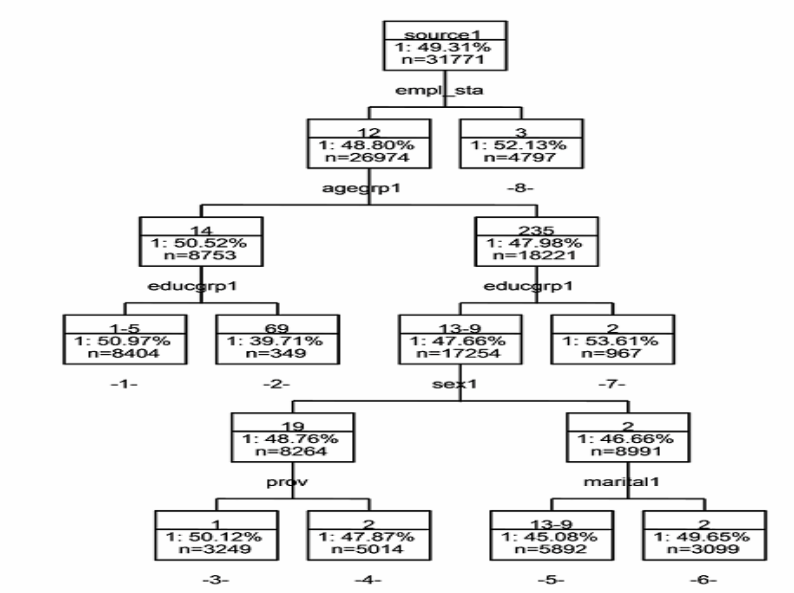
The results of the logistic regression were not as good as those of CHAID. Using the multinomial logistic regression technique, the highest order interaction (5 factor interaction) was significant, meaning that the change over the provinces differs over the combinations of the other 4 variables.

The overall percentage correctly classified and identified by the logistic regression model was 61.95%. The logistic regression model revealed several first and second order interactions as indicated by CHAID. The results from both techniques point out some similarities and differences regarding the contribution of the predictor variables in the model.

4.2.4 September GHS 2006 and July GHS 2007 (source data as response variable)

We will now consider September GHS 2006 and July GHS 2007 as response variables to determine whether there are any relationships in the data with respect to the predictor variables. The output of the CHAID tree is shown in Figure 4.2.4.1.

Figure 4.2.4.1: Classification tree diagram for September GHS 2006 and July GHS 2007 (source data as response variable)



All significant predictors are listed in Table 4.2.4.1. Employment status, highest level of education, and age group were the only significant predictors. Employment status (p-value of 2.9e-5) was the most significant predictor variable associated with different data sources. This variable explains more of the variation in both data sets than any other predictor variable in the analysis. This was followed by highest level of education with p-value of 0.00055 and age group (0.044). All other predictors were not significant.

Some categories of the predictor variables were merged into one composite class, reducing the number of categories as each category fails to be significant at 5% significance level. The categories of employment status were reduced from three to two. Category 1 (employed) has a similar profile as category 3 (not economically active) and was merged into one composite class. Highest level of education was also reduced from seven to three response categories. People who had completed any of grade 1-7, grade 8-11, grade 12 and certificate/diploma have similar profiles and they were merged into one composite class. There was no difference in the profile of

people who had completed either a degree or higher, to those who did not specify their highest level of education. The two categories were merged into one composite class.

The categories of age group were reduced from five to two. Category 1 (15-19 years old) and 4 (40-49 years old) have a similar profile and were merged into one composite class. Also, Category 2 (20-29 years old), 3 (30-39 years old) and 5 (50-65 years old) have a similar profile and were merged into one composite class.

Table 4.2.4.1: List of significant predictors (September GHS 2006 and July GHS 2007)

Predictor	p-value	Levels	Groups
Employment status	2.9e-5	3->2	12 3
Highest level of education	0.00055	7->3	13-5 2 69
Age group	0.045	5->2	14 235

The CHAID tree shows that data was first partitioned by employment status. As before, CHAID then takes each remaining predictor in turn to determine the next segmenting variable (see Table 4.2.4.2). The data was partitioned into 6 nodes.

Table 4.2.4.2: Profiles of each subgroup formed by the CHAID analysis (September GHS 2006 and July GHS 2007)

Province	Other predictors involved in the interactions			
Employed and not economically active	age group	Sex	age group	Population group
		highest level of education	Province	
			Marital status	
Unemployed				

4.3 September LFS 2006 and September GHS 2006 by employment status

4.3.1 Results for CHAID (employment status as response variable)

The output of the CHAID tree is shown in Figure 4.3.1.1 and all predictor variables which are statistically significant are listed in Table 4.3.1.1. It follows from Table 4.3.1.1 that all predictors, except source data, are highly predictive in the full data set. Age group was the most significant predictor variable with a p-value of 4.3e-2049. The other two most significant predictor variables were marital status and highest level of education.

Some categories of the predictor variables were merged into one composite class, reducing the number of categories as each category fails to be significant at the 5% significance level. The categories of highest level of education were reduced from seven to six. People who had completed either a certificate or diploma (category 5) have a similar profile to those who did not specify their highest level of education (category 9). Category 1 (female) of sex has a similar profile to that of category 9 (unspecified) and these were merged into one composite class. The

categories of the population group were reduced from five to three. Category 3 (Indian/Asian), 4 (Whites) and 9 (unspecified) have similar profiles and were merged to form one composite class.

Table 4.3.1.1: List of significant predictors (September LFS 2006 and September GHS 2006)

Predictor	p-value	Levels	Groups
Age group	4.3e-2049	5	1 2 3 4 5
Marital Status	1.5e-660	5	1 2 3 4 5
Highest level of education	3.1e-561	7->6	1 2 3 4 59 6
Province	5.9e-389	2	1 2
Gender	5.7e-216	3->2	19 2
Population group	1.3e-171	5->3	1 2 3-9

At the root node, the most significant split was obtained by segmenting the cases containing employment status into 5 different age groups (see Figure 4.3.1.1).

Figure 4.3.1.1: Classification tree diagram for September LFS 2006 and September GHS 2006

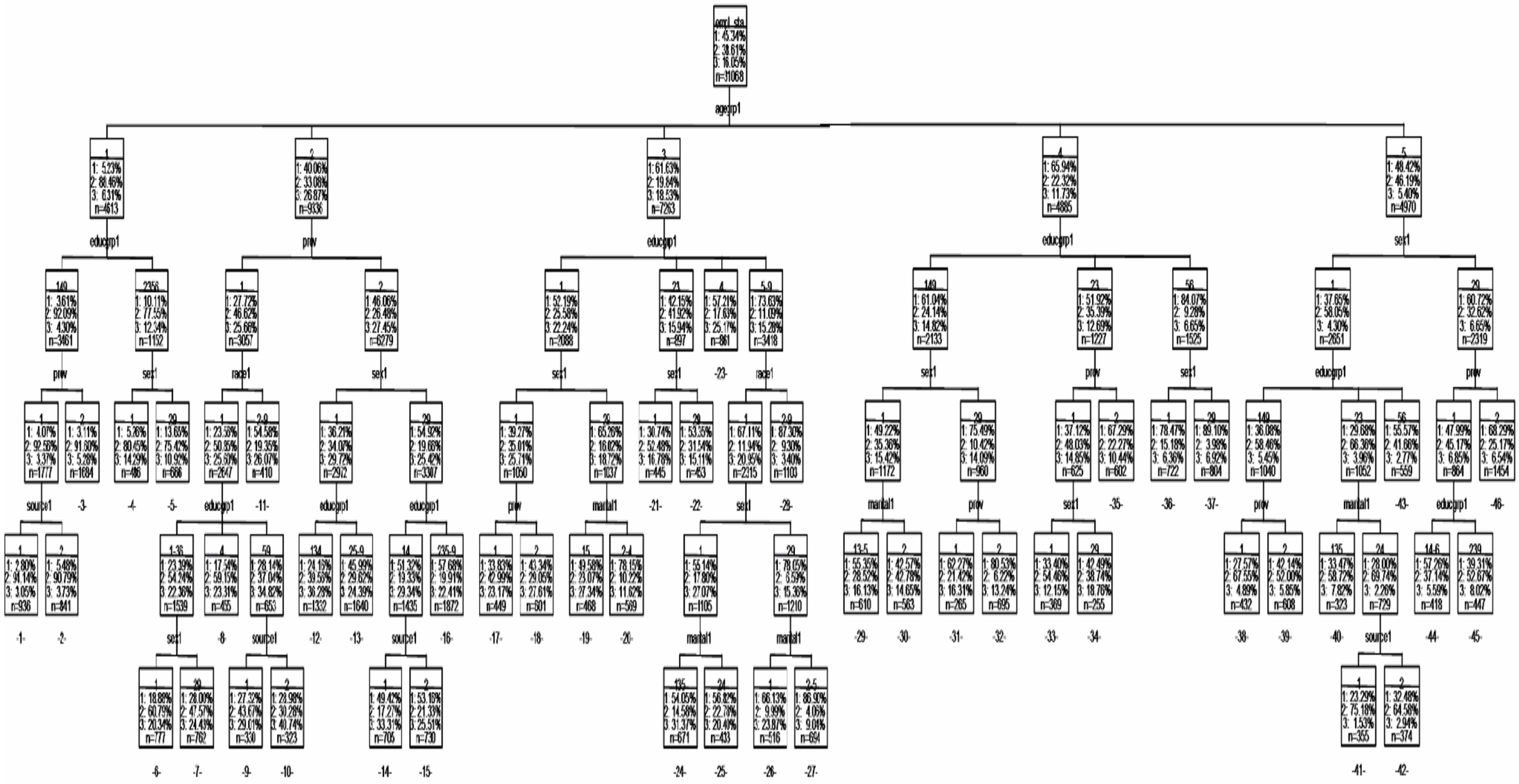


Table 4.3.1.2 below summarises the trends with regard to age group categories in relation to the employment status.

Table 4.3.1.2: Age group by employment status (September LFS 2006 and September GHS 2006)

Age grp	Age group	Employed	Not economically active	Unemployed	Sample size
1	15-19	5.23	88.46	6.31	4613
2	20-29	40.06	33.06	26.87	9336
3	30-39	61.63	19.84	18.53	7263
4	40-49	65.94	22.32	11.73	4885
5	50-65	48.42	46.19	5.40	4970

It follows from Table 4.3.1.2 that the probability that a person aged 15-19 years old and being employed is less than that of a person aged 20-29 years old and higher age groups. 88.46% of people in this category (age group 15-19 years old) were not economically active; probably because the majority of them should still be full-time students. Only 5.23% were employed and 6.31% were unemployed. The results also show that the percentage of unemployed people increases among the age group 20-29 years old. This category is likely to include a number of people who had just completed their studies and who were still looking for work. 26.87% of these people were unemployed; 40.06% were employed and 33.06% were not economically active.

The percentage of people aged 30-39 years old who were employed increased to 61.63%, whereas there was a proportional decline in the other categories of employment status. 19.84% of people in this age category were not economically active and 18.53% were unemployed. There has been a slight increase in the percentage of people who were employed (65.94%) and not economically active (22.32%) in the age group 40-49 years old. The percentage of people who were unemployed decreased to 11.73%. For the age group 50-65 years old, the patterns for employment status change significantly. The percentage of people who were employed has decreased to 48.42%; a large increase on the percentage of people who were not economically active to 46.19% and a drop to 5.40% among people who were unemployed. This result indicates the comparative relationships between age and employment status. The chance of a younger person being employed is less than that of persons in higher age groups. It also shows that people's chance of being employed when they are approaching their retirement age becomes less, possibly due to people taking early retirement.

CHAID then takes each remaining predictor in turn to determine the next segmenting variable. Following employment status of the different age groups (shown in Figure 4.3.1.1), one is able to see which of the attributes comprise each terminal node. The results were used to understand the predictive power of the predictor variables used and their inter-relationships. At the second level of partitioning it was found that the highest level of education (age group 1, 3 and 4), province (age group 2) and sex (age group 5) were significant. The three predictors are competing with each other within the categories of age group. Table 4.3.1.3 indicates that the highest level of education for age group 30-39 years old has the most significant p-value. The predictor variable

with the most significant p-value for age group 20-29 years old was province whereas sex was the most predictive variable for age group 50-65 years old. The least significant predictor variable was that for the highest level of education for the age group 15-19 years old.

Table 4.3.1.3: Age group categories by predictors involved in the first order interactions and their p-values (September LFS 2006 and September GHS 2006)

Age group	1st order interactions	Likelihood ratio chi-square	Degree of freedom	p-value
15-19	Highest level of education	216.30	4	3.5e-43
20-29	Province	434.81	2	3.8e-95
30-39	Highest level of education	625.92	8	8.7e-128
40-49	Highest level of education	472.14	4	2.1e-98
50-65	Sex	351.82	2	1.2e-76

The age group 20-29 years old was partitioned by province, whereas age group 50-65 years old was partitioned by sex. This is probably due to the minimum retirement age for females generally being lower than that for males. All other age groups were partitioned by the highest level of education. Table 4.3.1.4 lists the predictors used at different branch levels of each subgroup identified by CHAID.

Table 4.3.1.4: Profiles of each subgroup formed by the CHAID analysis (September LFS 2006 and September GHS 2006)

Age group	Other predictors involved in the interactions			
15-19	Highest level of education	Province	Source data	
		Sex		
20-29	Province	Population group	Highest level of education	Sex
				Source data
		Sex	Highest level of education	Source data
30-39	Highest level of education	Sex	Province	
			Marital status	
		Population group	Sex	Marital status
40-49	Highest level of education	Sex	Marital status	
			Province	
		Province	Sex	
50-65	Sex	Highest level of education	Province	
			Marital status	Source data
		Province	Highest level of education	

4.3.2 Results for logistic regression (employment status as response variable)

The output in Table 4.3.2.1 and Table 4.3.2.2 below describes and tests whether the predictors contribute significantly to the model. The model was developed in a similar way to that in Section 3.3.6. All the useful predictor variables and possible interactions were included in the model. It follows from the results that all three tests are significant at 0.05. Since our test statistics are significant at 0.05, we reject the null hypothesis that there are no relationships between employment status and set of predictor variables.

Table 4.3.2.1: Model Fit Statistics (September LFS 2006 and September GHS 2006)

	Intercept only	Intercept and Covariates
AIC	63365.222	57303.962
SC	63381.910	58180.075
-2 Log L	63361.222	57093.962

Table 4.3.2.2: Testing Global Null Hypothesis: BETA=0 (September LFS 2006 and September GHS 2006)

	Chi-Square	DF	Pr> ChiSq
Likelihood Ratio	6267.2603	103	<.0001
Score	5497.6411	103	<.0001
Wald	4866.7850	103	<.0001

Table 4.3.2.3 lists the output showing the effect of each predictor variable and their interactions contributed to the model. The results show several significant interactions of up to the six factor interaction. There were no significant seven factor interactions. Since the highest order interaction (six factor interaction) is significant, this means that the change over the source data differs over the combinations of the other 5 variables. Province was not significant in this model.

Table 4.3.2.3: Parameter estimates from the logistic regression model (September LFS 2006 and September GHS 2006)

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
								Lower	Upper
Intercept	0	1	-1.166	0.083	198.971	<.0001	0.312	-1.328	-1.004
Intercept	1	1	0.935	0.083	128.434	<.0001	2.547	0.773	1.097
AG		1	0.091	0.034	7.331	0.007	1.096	0.025	0.158
ED		1	-0.046	0.034	1.788	0.181	0.955	-0.113	0.021
AG*ED		1	0.061	0.014	19.539	<.0001	1.063	0.034	0.089
MA		1	0.063	0.101	0.398	0.528	1.066	-0.134	0.261
AG*MA		1	0.002	0.026	0.007	0.932	1.002	-0.049	0.053
ED*MA		1	0.139	0.040	12.001	0.001	1.149	0.060	0.217
AG*ED*MA		1	-0.042	0.011	15.849	<.0001	0.959	-0.063	-0.021
RA		1	-0.096	0.126	0.572	0.450	0.909	-0.343	0.152
AG*RA		1	-0.017	0.043	0.155	0.694	0.983	-0.102	0.068
ED*RA		1	0.198	0.039	26.240	<.0001	1.219	0.122	0.274
AG*ED*RA		1	-0.031	0.012	6.524	0.011	0.970	-0.055	-0.007
MA*RA		1	0.058	0.133	0.191	0.662	1.060	-0.203	0.319
AG*MA*RA		1	0.009	0.034	0.064	0.800	1.009	-0.058	0.076
ED*MA*RA		1	0.008	0.034	0.053	0.818	1.008	-0.059	0.075
AG*ED*MA*RA		1	-0.003	0.009	0.122	0.727	0.997	-0.021	0.014
SE		1	-0.043	0.114	0.143	0.706	0.958	-0.266	0.180
AG*SE		1	0.055	0.051	1.162	0.281	1.057	-0.045	0.155
ED*SE		1	0.087	0.049	3.216	0.073	1.091	0.008	0.183
AG*ED*SE		1	-0.030	0.021	1.948	0.163	0.971	-0.072	0.012
MA*SE		1	1.132	0.213	28.144	<.0001	3.103	0.714	1.551
AG*MA*SE		1	-0.216	0.053	16.828	<.0001	0.806	-0.319	-0.113
ED*MA*SE		1	0.099	0.065	2.370	0.124	1.104	-0.027	0.226
AG*ED*MA*SE		1	-0.013	0.015	0.744	0.389	0.987	-0.043	0.017

Table 4.3.2.3: Parameter estimates from the logistic regression model (September LFS 2006 and September GHS 2006) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
							Lower	Upper
RA*SE	1	0.027	0.160	0.028	0.868	1.027	-0.287	0.340
AG*RA*SE	1	0.177	0.057	9.544	0.002	1.193	0.065	0.289
ED*RA*SE	1	-0.038	0.038	0.992	0.319	0.963	-0.113	0.037
AG*ED*RA*SE	1	-0.015	0.014	1.171	0.279	0.985	-0.043	0.012
MA*RA*SE	1	0.752	0.332	5.123	0.024	2.121	0.101	1.403
AG*MA*RA*SE	1	-0.252	0.075	11.314	0.001	0.777	-0.399	-0.105
ED*MA*RA*SE	1	0.216	0.083	6.809	0.009	1.242	0.054	0.379
AG*ED*MA*RA*SE	1	-0.027	0.019	2.073	0.150	0.973	-0.064	0.010
PR	1	-0.316	0.125	6.392	0.012	0.729	-0.561	-0.071
AG*PR	1	0.178	0.047	14.632	0.000	1.195	0.087	0.269
ED*PR	1	0.084	0.045	3.541	0.060	1.088	0.004	0.172
AG*ED*PR	1	-0.045	0.017	6.922	0.009	0.956	-0.079	-0.012
MA*PR	1	-0.014	0.119	0.013	0.908	0.986	-0.246	0.219
AG*MA*PR	1	-0.010	0.031	0.093	0.761	0.991	-0.071	0.052
ED*MA*PR	1	-0.103	0.039	6.856	0.009	0.902	-0.180	-0.026
AG*ED*MA*PR	1	0.036	0.010	12.238	0.001	1.037	0.016	0.056
RA*PR	1	0.298	0.128	5.434	0.020	1.347	0.048	0.549
AG*RA*PR	1	-0.080	0.038	4.342	0.037	0.923	-0.155	-0.005
ED*RA*PR	1	-0.092	0.033	7.647	0.006	0.912	-0.158	-0.027
AG*ED*RA*PR	1	0.023	0.009	7.592	0.006	1.024	0.007	0.040
MA*RA*PR	1	0.186	0.106	3.076	0.080	1.204	-0.022	0.394
AG*MA*RA*PR	1	-0.036	0.027	1.840	0.175	0.964	-0.089	0.016
SE*PR	1	0.375	0.165	5.157	0.023	1.454	0.051	0.698
AG*SE*PR	1	0.038	0.067	0.324	0.570	1.039	-0.093	0.169
ED*SE*PR	1	0.019	0.061	0.094	0.759	1.019	-0.100	0.137
AG*ED*SE*PR	1	0.028	0.025	1.342	0.247	1.029	-0.020	0.076
MA*SE*PR	1	0.026	0.231	0.013	0.910	1.026	-0.427	0.479
AG*MA*SE*PR	1	0.007	0.057	0.015	0.902	1.007	-0.105	0.120
ED*MA*SE*PR	1	-0.119	0.042	8.058	0.005	0.888	-0.202	-0.037
RA*SE*PR	1	-0.374	0.137	7.451	0.006	0.688	-0.642	-0.105
AG*RA*SE*PR	1	0.002	0.044	0.002	0.964	1.002	-0.084	0.088
MA*RA*SE*PR	1	-0.616	0.289	4.535	0.033	0.540	-1.183	-0.049
AG*MA*RA*SE*PR	1	0.130	0.064	4.140	0.042	1.138	0.005	0.255
DS	1	0.062	0.116	0.283	0.595	1.064	-0.166	0.289
AG*DS	1	-0.066	0.046	2.084	0.149	0.936	-0.157	0.024
ED*DS	1	-0.051	0.048	1.131	0.288	0.951	-0.144	0.043
AG*ED*DS	1	0.041	0.019	4.911	0.027	1.042	0.005	0.078
MA*DS	1	0.164	0.132	1.554	0.213	1.179	-0.094	0.423
AG*MA*DS	1	-0.006	0.034	0.036	0.850	0.994	-0.073	0.061
ED*MA*DS	1	-0.013	0.039	0.118	0.731	0.987	-0.090	0.063
AG*ED*MA*DS	1	0.000	0.011	0.000	0.991	1.000	-0.022	0.022
RA*DS	1	-0.039	0.156	0.063	0.802	0.962	-0.345	0.267
AG*RA*DS	1	0.079	0.046	2.930	0.087	1.083	-0.012	0.170
ED*RA*DS	1	-0.050	0.047	1.130	0.288	0.952	-0.141	0.042
AG*ED*RA*DS	1	-0.002	0.014	0.027	0.870	0.998	-0.029	0.024
MA*RA*DS	1	-0.137	0.154	0.791	0.374	0.872	-0.438	0.165
AG*MA*RA*DS	1	-0.010	0.039	0.065	0.799	0.990	-0.087	0.067
ED*MA*RA*DS	1	0.046	0.047	0.938	0.333	1.047	-0.047	0.139
AG*ED*MA*RA*DS	1	-0.010	0.012	0.658	0.417	0.990	-0.033	0.014
SE*DS	1	0.074	0.163	0.207	0.649	1.077	-0.246	0.394

Table 4.3.2.3: Parameter estimates from the logistic regression model (September LFS 2006 and September GHS 2006) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits Lower	Upper
AG*SE*DS	1	0.037	0.070	0.279	0.597	1.038	-0.100	0.173
ED*SE*DS	1	0.008	0.069	0.014	0.905	1.008	-0.128	0.144
AG*ED*SE*DS	1	-0.043	0.028	2.290	0.130	0.958	-0.099	0.013
MA*SE*DS	1	-1.058	0.276	14.701	0.000	0.347	-1.598	-0.517
AG*MA*SE*DS	1	0.204	0.070	8.520	0.004	1.227	0.067	0.341
ED*MA*SE*DS	1	0.001	0.082	0.000	0.989	1.001	-0.159	0.162
AG*ED*MA*SE*DS	1	0.015	0.022	0.421	0.517	1.015	-0.029	0.058
RA*SE*DS	1	0.066	0.198	0.112	0.738	1.069	-0.322	0.454
AG*RA*SE*DS	1	-0.148	0.067	4.948	0.026	0.863	-0.278	-0.018
ED*RA*SE*DS	1	0.038	0.052	0.529	0.467	1.039	-0.064	0.140
AG*ED*RA*SE*DS	1	0.005	0.019	0.069	0.794	1.005	-0.033	0.043
MA*RA*SE*DS	1	0.290	0.358	0.655	0.418	1.336	-0.412	0.991
AG*MA*RA*SE*DS	1	0.045	0.082	0.310	0.578	1.046	-0.114	0.205
ED*MA*RA*SE*DS	1	-0.331	0.116	8.207	0.004	0.718	-0.557	-0.105
AG*ED*MA*RA*SE*DS	1	0.053	0.026	4.259	0.039	1.055	0.003	0.104
PR*DS	1	-0.056	0.168	0.109	0.742	0.946	-0.386	0.275
AG*PR*DS	1	0.089	0.061	2.116	0.146	1.093	-0.031	0.208
ED*PR*DS	1	0.092	0.059	2.409	0.121	1.097	-0.024	0.209
AG*ED*PR*DS	1	-0.053	0.021	6.493	0.011	0.949	-0.093	-0.012
MA*PR*DS	1	-0.037	0.144	0.067	0.795	0.963	-0.320	0.245
AG*MA*PR*DS	1	-0.007	0.038	0.034	0.855	0.993	-0.081	0.067
RA*PR*DS	1	-0.202	0.115	3.081	0.079	0.818	-0.427	0.024
ED*RA*PR*DS	1	0.064	0.032	3.966	0.046	1.067	0.001	0.128
SE*PR*DS	1	0.171	0.229	0.560	0.454	1.187	-0.277	0.620
AG*SE*PR*DS	1	-0.089	0.089	0.996	0.318	0.915	-0.264	0.086
ED*SE*PR*DS	1	-0.151	0.084	3.200	0.074	0.860	-0.316	0.014
AG*ED*SE*PR*DS	1	0.063	0.031	4.237	0.040	1.065	0.003	0.123
MA*SE*PR*DS	1	1.607	0.310	26.908	<.0001	4.990	1.000	2.215
AG*MA*SE*PR*DS	1	-0.354	0.078	20.429	<.0001	0.702	-0.508	-0.201
RA*SE*PR*DS	1	0.251	0.124	4.118	0.042	1.285	0.009	0.494

Table 4.3.2.4 below presents the classification accuracy of the model in predicting the percentage in the different employment status categories. The following are the percentages correctly classified: 45.3% for employed, 39.6% for not economically active and 15.1% for unemployed people. The overall percentage correctly classified is 61.11%.

Table 4.3.2.4: Assessment of the adequacy of the model in percentages (September LFS 2006 and September GHS 2006)

LFS Predicted	Employed	Not Economically active	Unemployed	LFS profile
Employed	70.9	26.1	8.7	45,3
Not economically active	18.5	53.8	79.2	39,6
Unemployed	10.6	20.1	5.1	15,1
Predicted Total	100.0	100.0	100.0	100.0
Regression profile	43.0	56.9	0.1	

Although 61.11% appears reasonably good, the logistic regression is predicting the employed group total to give the correct percentage, but is predicting neither the not economically active nor the unemployed groups well.

4.3.3 Discussion of CHAID and logistic regression (employment status as response variable)

From the CHAID analysis, the most strongly associated predictor of employment status was age group. Marital status, highest level of education, province, sex, and population group have significant influence on predicting employment status by interacting with age group and some of the predictor variables in different stages of the tree. The results show that source data has no influence in predicting employment status in this study.

Data was partitioned into five different subgroups as per categories of age groups. Age is a continuous variable such that the older a person is, the better the chances of such person to be employed. Young people between the ages of 15 and 19 years old are often still busy with their educational studies and are largely not economically active. CHAID revealed that 88.46% of people among this age group are not economically active. Employment status in this age group can further be explained in terms of highest level of educational attainment, sex or province, and different data source. The results do not give us much difference between males and females; and LFS 2006 and GHS 2006 in terms of explaining employment status.

The results showed that the age group 20-29 years old might have recently completed their studies, and only 40.06% were employed; 33.06% were not economically active and 26.87% were unemployed. It is possible that some of those who were declared not economically active may still further their higher education studies. Employment status in this age group can further be explained in terms of province, population group or sex, highest level of education and source data. People belonging to the age groups 30-39 and 40-49 years old have almost the same characteristics such as highest level of education, sex, population group, province and marital status. The last subgroup of age group, 50-65 years old revealed that 60.72% of females were still employed as compared to 37.65% of males. Highest level of education, province, marital status and source data were the other predictor variables that played a significant role in explaining employment status in this subgroup.

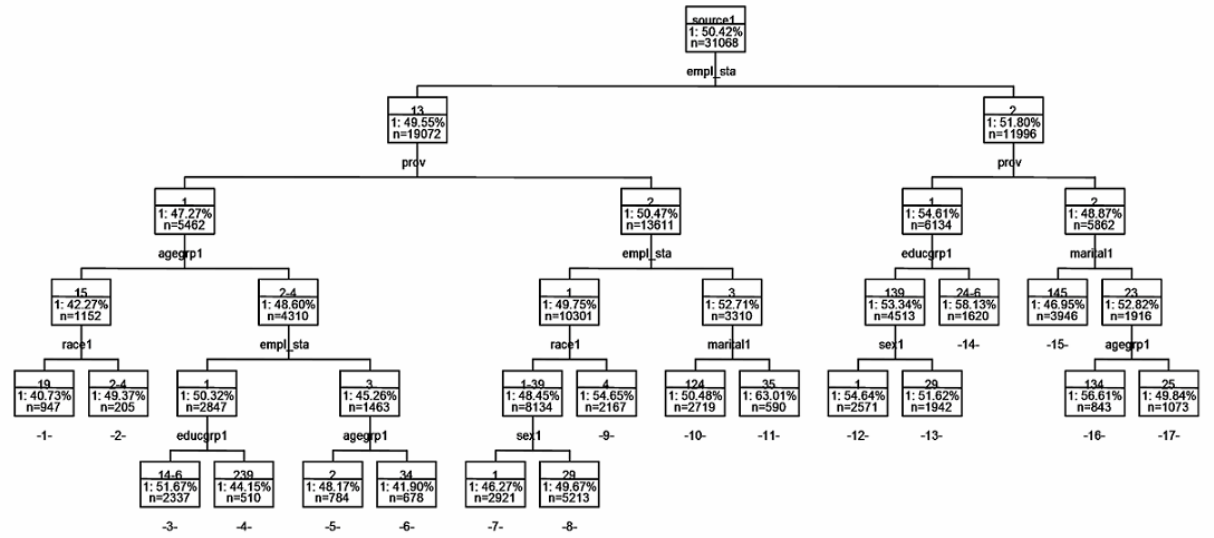
The results of the logistic regression were not as good as those of CHAID. In the multinomial logistic regression model, the highest order interaction (6 factor interaction) was significant, meaning that the change over the provinces differs over the combinations of (interaction between) the other 5 variables. The overall percentage correctly classified and identified by logistic regression model was 61.11%. Logistic regression model revealed several first and second order interactions as indicated by CHAID. The results from both techniques point out

some similarities and differences regarding the contribution of the predictor variables in the model.

4.3.4 September LFS 2006 and GHS 2006 September GHS 2006 (source data as response variable)

We will now look at the CHAID analysis when source (September GHS 2006 and July GHS 2007) was taken as the response variable. The output of the CHAID tree is shown in Figure 4.3.4.1.

Figure 4.3.4.1: Classification tree diagram for September LFS 2006 and September GHS 2006 (source data as response variable)



All significant predictors are listed in Table 4.3.4.1. Employment status was the most significant predictor associated with different data sets, but it was not that significant comparing the number of tests.

Table 4.3.4.1: List of significant predictors (September LFS 2006 and July GHS 2006)

Predictor	p-value	Levels	Groups
Employment status	0.00026	3->2	13 2
Marital status	0.00030	5->2	14 235
Population group	0.0022	5->2	19 2-4
Highest level of education	0.019	7->2	1359 246
Province	0.041	2	1 2

Some categories of the predictor variables were merged into one composite class, reducing the number of categories as each category fails to be significant at 5% significance level. The categories of employment status were reduced from three to two. Category 1 (employed) has a similar profile to category 3 (unemployed) and hence were merged into one composite class. The categories of marital status were reduced from five to two. Category 1 (never married) has a similar profile to category 4 (widow/widower) and they were merged into one composite class. Category 2 (married), 3 (living together like husband and wife), and 5 (divorced/separated) have similar profiles and were merged into one composite class.

The categories of the population group were reduced from five to two. Category 1 (African black) and category 9 (unspecified) have a similar profile and were merged into one composite class. Category 2 (Coloureds), 3 (Indian/Asian) and 4 (Whites) have similar profiles and were merged to form one composite class. The categories of highest level of education were reduced from seven to two. The profile of people who had completed any of these levels: Grade 1-7, grade 8-11, certificate/diploma and those who did not specify their highest level completed were similar and were merged into one composite class. The category for people who did not have any formal education, completed grade 12 or degree and higher have a similar profile and were merged into one composite class.

The analysis then takes each predictor variable in turn to determine the next segmenting variable (see Table 4.3.4.2). The data was partitioned into 17 nodes.

Table 4.3.4.2: Profiles of each subgroup formed by the CHAID analysis (September LFS 2006 and September GHS 2006)

Province	**Other predictors involved in the interactions**			
Employed and unemployed	Province	Age group	Population group	Population group
			Employment status	Highest level of education
		Employment status	Province	
			Marital status	
Unemployed	Province	Highest level of education	Age group	
		Marital status		

4.4 September LFS 2007, July GHS 2007 and October CS 2007

4.4.1 Results for CHAID (employment status as response variable)

The output of the CHAID analysis is presented in Figure 4.4.1.1 below, and all predictors which are statistically significant are listed in Table 4.4.1.1. All predictors are highly predictive in the full data set. Age group was the most significant predictor variable with a p-value of 2.5e-5587.

Some categories of the predictor variables were merged into one composite class, reducing the number of categories as each category fails to be significant at 5% significance level. The categories of highest level of education were reduced from seven to six. Category 3 (grade 1 – grade 7) has a similar profile to those who did not specify their highest level of education (category 9). The categories of marital status were reduced from six to five. Category 4 (widow/widower) and 9 (unspecified) have similar profiles and were also merged into one composite class. The categories of the population group were reduced from five to three. Category 3 (Indian/Asian), 4 (Whites) and 9 (unspecified) have similar profiles and were merged to form one composite class. Category 1 (female) of sex has a similar profile with that of category 9 (unspecified) and these were also merged into one composite class.

Table 4.4.1.1: List of significant predictors (September LFS 2007, July GHS 2007 and October CS 2007)

Predictor	p-value	Levels	Groups
Age group	2.5e-5587	5	1 2 3 4 5
Highest level of education	3.6e-1945	7->6	1 2 39 4 5 6
Marital status	1.8e-1857	6->5	1 2 3 49 5
Population group	1.5e-820	5->3	1 2 3-9
Source data	3.2e-629	3	1 2 3
Sex	1.0e-283	3->2	1 29
Province	1.1e-188	2	1 2

The CHAID tree shows that at the root node the most significant split was obtained by segmenting the cases containing employment status into 5 different age groups (see Figure 4.4.1.1).

Figure 4.4.1.1: Classification tree diagram for September LFS 2007, July GHS 2007and October CS 2007

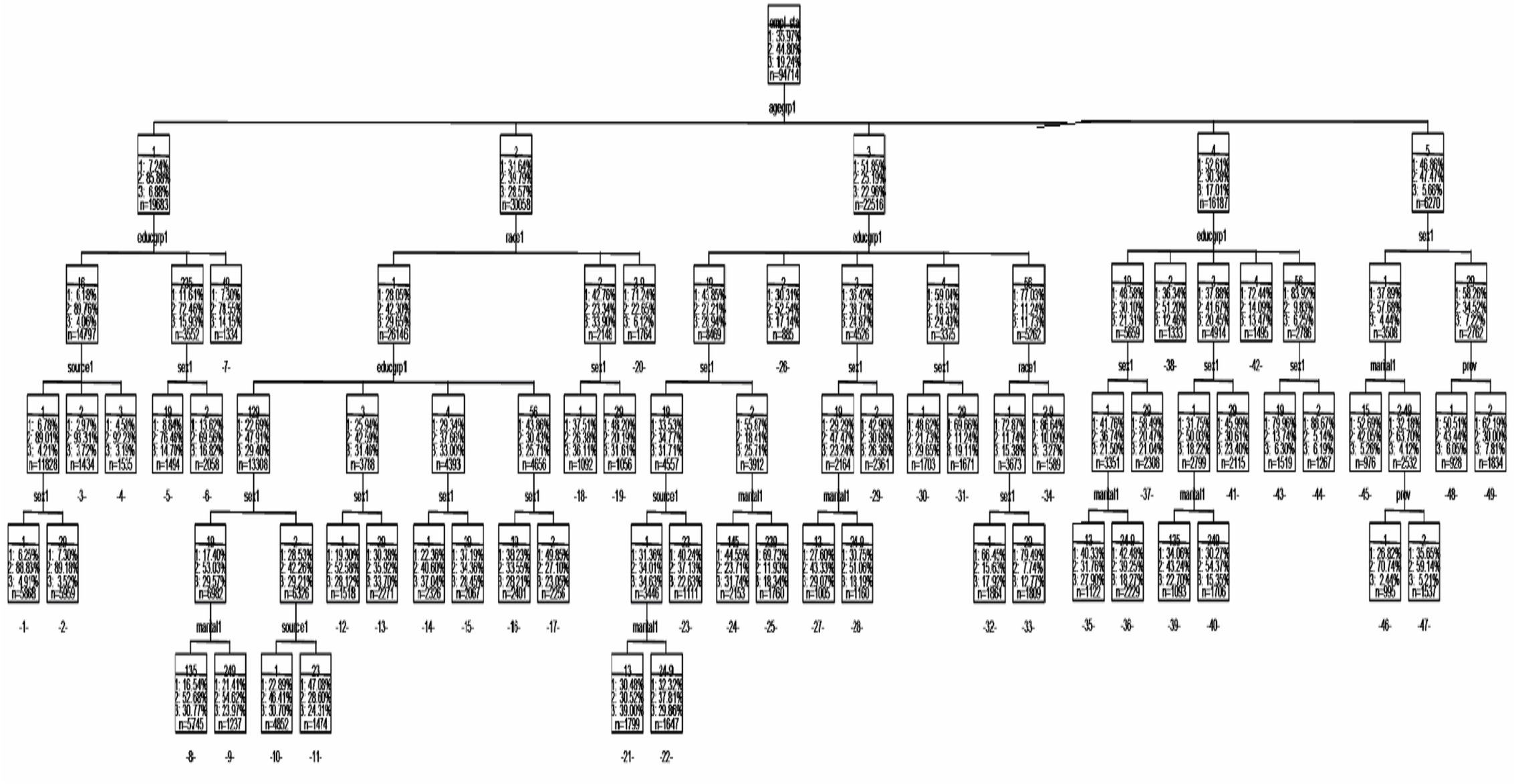


Table 4.4.1.2: Age group by employment status (LFS September 2007, GHS July 2007 and CS October 2007)

Age grp	Age group	Employed	Not economically active	Unemployed	Sample size
1	15-19	7.24	85.88	6.88	19683
2	20-29	31.64	39.79	28.57	30058
3	30-39	51.85	25.19	22.96	22516
4	40-49	55.61	30.38	17.01	16187
5	50-65	46.86	47.47	5.66	6270

Table 4.4.1.2 summarises trends with regard to age group categories in relation to the employment status. As we have seen in the previous sections, this subgroup (people aged 15-19 years old) comprises mainly people who are not economically active (85.88%). The chance of a person aged 15-19 years old being employed is again less than that of a person aged 20-29 years old and higher. The results also show that the percentages of unemployed people increase among the age group 20-29 years old; 28.57% of which were unemployed; 31.64% were employed and 39.79% were not economically active.

The percentage of people aged 30 to 39 years old who were employed increased significantly to 51.85%, whereas there has been a proportional decline in the other categories of employment status. 25.19% of people in this age category were not economically active and 22.96% were unemployed. We have seen a slight increase in the percentage of people who were employed (55.61%) and not economically active (30.38%) in the age group 40-49 years old. The percentage of people who were unemployed has decreased to 17.01%. Employment trends on age group 50-65 years old changes significantly. The percentage of people who were employed has decreased to 46.86%; while there was an increase on the percentage of people who were not economically active to 47.47% and a significant drop to 5.66% among people who were unemployed.

As before, CHAID then takes each remaining predictor in turn to determine the next segmenting variable. The results were used to understand the predictive power of the predictor variables used and their inter-relationships. At the second level of partitioning it was found that highest level of education, population group, and sex, were the most significant predictors. The three predictors were competing with each other within the categories of age group. Table 4.4.1.3 indicates that for age group 20-29 years old, the most predictive variable is population group. The most predictive variable for age group 50-65 years old is sex. Highest level of education was the most predictive variable for age group 15-19 years old, 30-39 years old and 40-49 years old. The least significant predictor variable was sex for age group 50-65 years old.

Table 4.4.1.3: Age group categories by predictors involved in the first order interactions and their p-values (September LFS 2007, July GHS 2007 and October CS 2007)

Age group	1st order interactions	Likelihood ratio chi-square	Degree of Freedom	p-value
15-19	Highest level of education	924.00	4	3.2-196
20-29	Population group	1452.02	4	5.5e-312
30-39	Highest level of education	2585.37	8	2.0e-551
40-49	Highest level of education	2585.60	8	2.1e-550
50-65	Sex	373.95	2	1.9e-81

The different groups could be split further. Table 4.4.1.4 lists the predictors used at different branch levels of each subgroup identified by CHAID. This table shows the relationships between the predictor variables when predicting employment status.

Table 4.4.1.4: Profiles of each age group by CHAID (September LFS 2007, July GHS 2007 and October CS 2007)

Age group	Other predictors involved in the interactions			
15-19	Highest level of education	Source data	Sex	
		Sex		
20-29	Population group	Highest level of education	Sex	
		Sex	Source data	Marital status
			Marital status	
30-39	Highest level of education	Sex	Province	
			Marital status	
		Population group	Sex	
40-49	Highest level of education	Sex	Marital status	
50-65	Sex	Marital status	Province	
		Province		

4.4.2 Results for multinomial logistic regression (employment status as response variable)

The outputs in Table 4.4.2.1 and Table 4.4.2.2 below describe and test the overall fit of the model. In order to be confident that the multinomial logistic regression gave the correct model, the overall relationship should be statistically significant. It follows from the results that the three tests yield similar conclusions. Since our test statistics are significant at 0.05, we reject the null hypothesis that there are no relationships between employment status and the set of predictor variables.

Table 4.4.2.1: Model Fit Statistics (September LFS 2007, September GHS 2007, and October CS 2007)

	Intercept only	Intercept and Covariates
AIC	197879.80	182076.51
SC	197898.83	183132.39
-2 Log L	197875.80	181854.51

Table 4.4.2.2: Testing Global Null Hypothesis: BETA=0 (September LFS 2007, September GHS 2007, and October CS 2007)

	Chi-Square	DF	Pr> ChiSq
Likelihood Ratio	16021.2940	109	<.0001
Score	13972.2907	109	<.0001
Wald	12594.1327	109	<.0001

Table 4.4.2.3 lists the output of the effect of each predictor variable and their interaction contributing in the model. The model was built with all the possible interactions included in the model. The results show several significant interactions up to the six factor interactions. There were no significant seven factor interactions. Since the highest order interaction (five factor interactions) is significant, this means that the change over the source data differs over the combinations of the other 5 variables.

Table 4.4.2.3: Parameter estimates from the logistic regression model (September LFS 2007, July GHS 2007 and October CS 2007)

			Standard	Wald			95% confidence limits	
Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Exp(Est)	Lower	Upper
Intercept 0	1	-0.995	0.190	27.430	<.0001	0.370	-1.367	-0.622
Intercept 1	1	1.252	0.190	43.453	<.0001	3.497	0.880	1.624
AG	1	-0.042	0.093	0.207	0.649	0.959	-0.225	0.140
ED	1	-0.145	0.075	3.758	0.053	0.865	-0.292	0.002
AG*ED	1	0.152	0.036	18.159	<.0001	1.164	0.082	0.222
MA	1	0.300	0.230	1.706	0.192	1.350	-0.150	0.750
AG*MA	1	-0.072	0.074	0.933	0.334	0.931	-0.218	0.074
ED*MA	1	0.092	0.098	0.884	0.347	1.096	-0.099	0.283
AG*ED*MA	1	-0.030	0.032	0.883	0.347	0.970	-0.093	0.033
RA	1	-0.364	0.262	1.932	0.165	0.695	-0.877	0.149
AG*RA	1	0.233	0.038	38.109	<.0001	1.262	0.159	0.307
ED*RA	1	0.152	0.083	3.346	0.067	1.164	-0.011	0.314
AG*ED*RA	1	-0.076	0.010	59.331	<.0001	0.927	-0.095	-0.057
MA*RA	1	0.318	0.190	2.818	0.093	1.374	-0.053	0.690
AG*MA*RA	1	-0.099	0.023	19.164	<.0001	0.906	-0.143	-0.055
ED*MA*RA	1	-0.086	0.067	1.666	0.197	0.918	-0.216	0.045
AG*ED*MA*RA	1	0.029	0.008	12.247	0.001	1.029	0.013	0.045
SE	1	-0.371	0.250	2.192	0.139	0.690	-0.862	0.120
AG*SE	1	0.277	0.117	5.630	0.018	1.319	0.048	0.505
ED*SE	1	0.325	0.094	11.856	0.001	1.383	0.140	0.509
AG*ED*SE	1	-0.159	0.026	37.429	<.0001	0.853	-0.211	-0.108
MA*SE	1	1.436	0.418	11.825	0.001	4.204	0.618	2.255
AG*MA*SE	1	-0.345	0.117	8.775	0.003	0.708	-0.574	-0.117
ED*MA*SE	1	-0.136	0.107	1.600	0.206	0.873	-0.346	0.075
AG*ED*MA*SE	1	0.013	0.014	0.890	0.346	1.013	-0.014	0.040
RA*SE	1	0.005	0.376	0.000	0.990	1.005	-0.733	0.742
AG*RA*SE	1	-0.081	0.044	3.416	0.065	0.922	-0.167	0.005
ED*RA*SE	1	-0.164	0.124	1.766	0.184	0.849	-0.407	0.078
AG*ED*RA*SE	1	0.072	0.015	22.413	<.0001	1.074	0.042	0.101
MA*RA*SE	1	-0.332	0.333	0.990	0.320	0.718	-0.984	0.321
AG*MA*RA*SE	1	0.068	0.044	2.356	0.125	1.070	-0.019	0.154
ED*MA*RA*SE	1	0.291	0.115	6.420	0.011	1.337	0.066	0.515
AG*ED*MA*RA*SE	1	-0.070	0.016	18.516	<.0001	0.933	-0.102	-0.038

Table 4.4.2.3: Parameter estimates from the logistic regression model (September LFS 2007, July GHS 2007 and October CS 2007) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
							Lower	Upper
PR	1	-0.394	0.191	4.274	0.039	0.674	-0.768	-0.021
AG*PR	1	0.127	0.094	1.830	0.176	1.135	-0.057	0.311
ED*PR	1	0.107	0.075	2.030	0.154	1.113	-0.040	0.255
AG*ED*PR	1	-0.037	0.036	1.056	0.304	0.964	-0.107	0.033
MA*PR	1	-0.438	0.231	3.587	0.058	0.646	-0.891	0.015
AG*MA*PR	1	0.147	0.075	3.796	0.051	1.158	0.001	0.294
ED*MA*PR	1	0.068	0.098	0.477	0.490	1.070	-0.124	0.260
AG*ED*MA*PR	1	-0.026	0.033	0.635	0.426	0.974	-0.090	0.038
RA*PR	1	0.304	0.259	1.383	0.240	1.355	-0.203	0.811
AG*RA*PR	1	-0.082	0.023	12.258	0.001	0.921	-0.128	-0.036
ED*RA*PR	1	0.040	0.082	0.234	0.629	1.040	-0.121	0.200
MA*RA*PR	1	-0.018	0.180	0.010	0.920	0.982	-0.371	0.335
ED*MA*RA*PR	1	0.017	0.063	0.071	0.790	1.017	-0.107	0.141
SE*PR	1	0.432	0.251	2.957	0.086	1.540	-0.060	0.924
AG*SE*PR	1	-0.215	0.117	3.404	0.065	0.807	-0.443	0.013
ED*SE*PR	1	-0.274	0.094	8.450	0.004	0.760	-0.459	-0.089
AG*ED*SE*PR	1	0.110	0.024	21.361	<.0001	1.116	0.063	0.157
MA*SE*PR	1	-0.743	0.417	3.179	0.075	0.476	-1.559	0.074
AG*MA*SE*PR	1	0.171	0.116	2.160	0.142	1.186	-0.057	0.398
ED*MA*SE*PR	1	0.148	0.102	2.089	0.148	1.159	-0.053	0.348
RA*SE*PR	1	0.247	0.372	0.439	0.508	1.280	-0.483	0.976
ED*RA*SE*PR	1	0.076	0.122	0.387	0.534	1.079	-0.164	0.316
MA*RA*SE*PR	1	0.322	0.314	1.047	0.306	1.379	-0.295	0.938
ED*MA*RA*SE*PR	1	-0.180	0.106	2.903	0.088	0.836	-0.386	0.027
DS	1	-0.164	0.121	1.849	0.174	0.849	-0.401	0.072
AG*DS	1	0.105	0.059	3.146	0.076	1.111	-0.011	0.221
ED*DS	1	0.107	0.048	5.011	0.025	1.113	0.013	0.202
AG*ED*DS	1	-0.039	0.022	2.973	0.085	0.962	-0.083	0.005
MA*DS	1	-0.042	0.146	0.081	0.776	0.959	-0.329	0.245
AG*MA*DS	1	0.009	0.047	0.032	0.857	1.009	-0.084	0.101
ED*MA*DS	1	0.008	0.063	0.015	0.902	1.008	-0.116	0.131
AG*ED*MA*DS	1	-0.006	0.021	0.088	0.766	0.994	-0.047	0.035
RA*DS	1	0.168	0.181	0.865	0.352	1.183	-0.186	0.523
AG*RA*DS	1	-0.110	0.028	15.845	<.0001	0.896	-0.164	-0.056
ED*RA*DS	1	-0.020	0.056	0.126	0.723	0.980	-0.130	0.091
AG*ED*RA*DS	1	0.010	0.008	1.567	0.211	1.010	0.006	0.027
MA*RA*DS	1	-0.095	0.133	0.507	0.477	0.910	-0.355	0.166
AG*MA*RA*DS	1	0.034	0.020	2.873	0.090	1.034	0.005	0.073
ED*MA*RA*DS	1	0.060	0.046	1.744	0.187	1.062	-0.029	0.150
AG*ED*MA*RA*DS	1	-0.013	0.007	3.617	0.057	0.987	-0.026	0.000
SE*DS	1	0.254	0.157	2.606	0.106	1.289	-0.054	0.561
AG*SE*DS	1	-0.082	0.072	1.330	0.249	0.921	-0.223	0.058
ED*SE*DS	1	-0.128	0.059	4.641	0.031	0.880	-0.244	-0.012
AG*ED*SE*DS	1	0.043	0.013	11.884	0.001	1.044	0.019	0.068
MA*SE*DS	1	-0.263	0.260	1.016	0.314	0.769	-0.773	0.248
AG*MA*SE*DS	1	0.042	0.074	0.323	0.570	1.043	-0.103	0.186
ED*MA*SE*DS	1	0.088	0.071	1.525	0.217	1.092	-0.052	0.228
AG*ED*MA*SE*DS	1	0.003	0.012	0.044	0.833	1.003	-0.022	0.027
RA*SE*DS	1	0.186	0.252	0.543	0.461	1.204	-0.309	0.681

Table 4.4.2.3: Parameter estimates from the logistic regression model (September LFS 2007, July GHS 2007 and October CS 2007) (continued)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)	95% confidence limits	
							Lower	Upper
AG*RA*SE*DS	1	0.055	0.040	1.852	0.174	1.056	-0.024	0.134
ED*RA*SE*DS	1	0.141	0.087	2.628	0.105	1.151	-0.029	0.311
AG*ED*RA*SE*DS	1	-0.021	0.013	2.847	0.092	0.979	-0.046	0.003
MA*RA*SE*DS	1	0.068	0.239	0.082	0.775	1.071	-0.400	0.537
AG*MA*RA*SE*DS	1	-0.033	0.039	0.735	0.391	0.967	-0.109	0.043
ED*MA*RA*SE*DS	1	-0.167	0.081	4.267	0.039	0.846	-0.326	-0.009
AG*ED*MA*RA*SE*DS	1	0.029	0.014	4.654	0.031	1.030	0.003	0.056
PR*DS	1	0.104	0.127	0.668	0.414	1.110	-0.145	0.353
AG*PR*DS	1	0.099	0.061	2.585	0.108	1.104	-0.022	0.219
ED*PR*DS	1	0.013	0.049	0.067	0.796	1.013	-0.084	0.110
AG*ED*PR*DS	1	-0.025	0.023	1.205	0.272	0.976	-0.069	0.019
MA*PR*DS	1	0.163	0.151	1.168	0.280	1.177	-0.133	0.460
AG*MA*PR*DS	1	-0.089	0.049	3.240	0.072	0.915	-0.185	0.008
ED*MA*PR*DS	1	-0.105	0.064	2.700	0.100	0.900	-0.231	0.020
AG*ED*MA*PR*DS	1	0.046	0.021	4.689	0.030	1.047	0.004	0.088
RA*PR*DS	1	-0.084	0.175	0.231	0.631	0.919	-0.427	0.259
ED*RA*PR*DS	1	-0.011	0.055	0.038	0.845	0.989	-0.119	0.097
MA*RA*PR*DS	1	-0.007	0.120	0.004	0.952	0.993	-0.242	0.228
ED*MA*RA*PR*DS	1	-0.021	0.042	0.247	0.619	0.979	-0.103	0.061
SE*PR*DS	1	0.002	0.163	0.000	0.992	1.002	-0.319	0.322
AG*SE*PR*DS	1	0.016	0.072	0.051	0.821	1.016	-0.125	0.157
ED*SE*PR*DS	1	0.092	0.060	2.361	0.124	1.096	-0.025	0.209
MA*SE*PR*DS	1	1.189	0.269	19.546	<.0001	3.285	0.662	1.717
AG*MA*SE*PR*DS	1	-0.218	0.076	8.177	0.004	0.804	-0.368	-0.069
ED*MA*SE*PR*DS	1	-0.228	0.065	12.281	0.001	0.796	-0.356	-0.101
RA*SE*PR*DS	1	-0.368	0.247	2.229	0.135	0.692	-0.851	0.115
ED*RA*SE*PR*DS	1	-0.118	0.086	1.892	0.169	0.889	-0.285	0.050
MA*RA*SE*PR*DS	1	-0.184	0.205	0.800	0.371	0.832	-0.586	0.219
ED*MA*RA*SE*PR*DS	1	0.167	0.069	5.864	0.016	1.182	0.032	0.302

4.4.3 Discussion of CHAID and logistic regression results (employment status as response variable)

The results in Table 4.4.1.2 show that both techniques can be used effectively with employment data. The trend of employment status is influenced by the complexity of the data and other factors such as demographic variables (sex, age, population group, marital status), and economic variables (level of education, province). This research identified the relationships between these variables and the outcome of employment status. Both techniques have identified the predictor variables in the order of strength of association with employment status. The results of CHAID give more divisions in subgroups. All the predictor variables are significant except source data.

From the CHAID analysis, the most strongly associated predictor of employment status was age group. Marital status, highest level of education, sex, population group and province have significant influence on predicting employment status by interacting with age group and some

of the predictor variable in different stages of the tree. The results show that source data has no influence in predicting employment status in this study.

Data was partitioned into five different subgroups as per categories of age groups. Age is a continuous variable such that the older a person is, the better the chances of such a person to be employed. Young people between the ages of 15 and 19 years old are often still busy with their educational studies and are largely not economically active. CHAID had revealed that 85.88% of people among this age group are not economically active. Employment status in this age group can further be explained in terms of highest level of educational attainment, sex and different data source. The results do not indicate much difference between males and females; and LFS 2007, GHS 2007 and CS 2007 in terms of explaining employment status.

The results showed that the age group 20-29 years old recently completed their studies, and only 31.64% were employed; 39.79% were not economically active and 28.57% were unemployed. It is possible that some of those who were declared not economically active may still further their higher education studies. Employment status in this age group can further be explained in terms of population group, highest level of education, sex, marital status or source data. The subgroup that consisted of age groups 30-39 and 40-49 years old have almost the same characteristics with highest level of education, sex and marital status. The last subgroup (i.e. age group 50-65 years old) revealed that 58.26% of females were still employed as compared to 37.89% of males. Marital status and province were the other predictor variables that played a significant role in explaining employment status in this subgroup.

The results of the logistic regression were not as good as those of CHAID. In the multinomial logistic regression technique, the highest order interaction (which is a 6 order interaction in this case) was significant, meaning that the change over the provinces differs over the combinations of the other 5 variables. The logistic regression model revealed several first and second order interactions as indicated by CHAID. The results from both techniques point out some similarities and differences regarding the contribution of the predictor variables in the model.

4.4.4 September LFS 2007, July GHS 2007 and October CS 2007 (source data as response variable)

We will now look at the CHAID analysis when source data (September LFS 2007, July GHS 2007 and October CS 2007) were taken as the response variable. The CHAID tree is given in Figure 4.4.4.1. All the significant predictors are listed in Table 4.4.4.1. Highest level of education was the most significant predictor associated with data sources. Other significant predictors were age group, province, employment status, marital status, population group and sex.

Table 4.4.4.1: List of significant predictors (LFS September 2007, GHS July 2007 and CS October 2007)

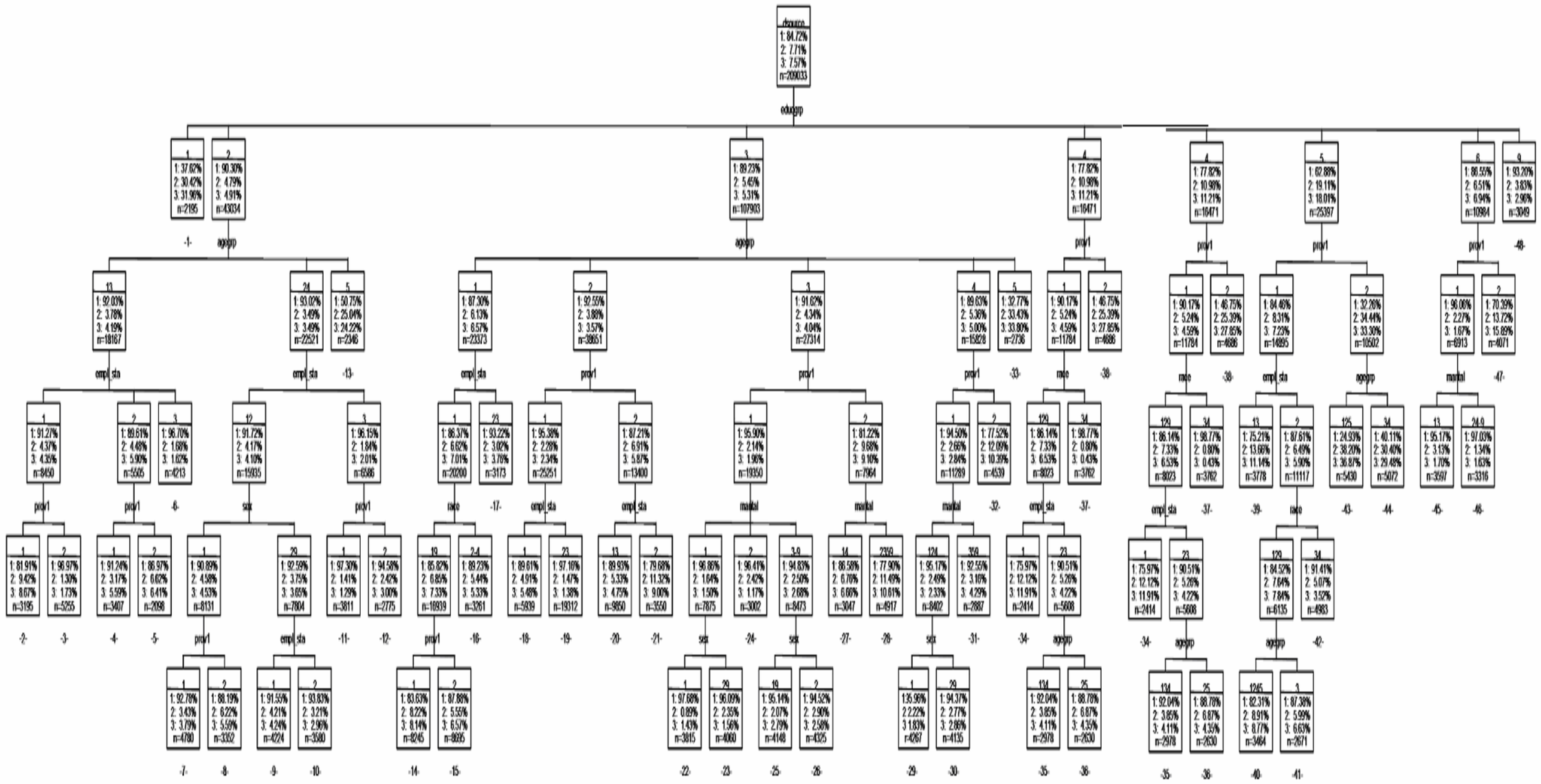
Predictor	p-value	Levels	Groups
Highest level of education	4.7e-2858	7	1 2 3 4 5 6 9
Age group	1.8e-2528	5->3	1-3 4 5
Province	6.7e-2053	2	1 2
Employment status	8.2e-415	3	1 2 3
Marital status	6.50E-235	6->4	1 24 39 5
Population group	0.00026	5->3	13 2 49
Sex	0.00043	3->2	19 2

Some categories of the predictor variables were merged into one composite class, reducing the number of categories as each category failed to be significant at the 5% significance level. The categories of age group were reduced from five to three. The categories of marital status were reduced from six to four. Category 3 (living together like husband and wife) has a similar profile to people who did not specify their marital status. Category 4 (widow/widower) and 2 have similar profiles and were merged into one composite class.

The categories of population group were reduced from five to three. Categories 1 and 3 have the same profile and so were merged to form one composite class. Category 4 of population group has a similar profile to that of category 9 (unspecified) and these were merged into one composite class.

The categories of sex were reduced from three to two. Categories 1 (males) and 9 (unspecified) have the same profile and were merged to form one composite class. The CHAID tree shows that at the root node the most significant split was obtained by segmenting the cases containing source data into the above categories of highest level of education. The results of each subgroup formed are shown in Table 4.4.4.2. The data was partitioned into 48 nodes.

Figure 4.4.4.1: Classification tree diagram for September LFS 2007, July GHS 2007 and October CS 2007 (source data as response variable)



4.4.5 Summary and comparison of the results of all the sections

After the results of each section have been analysed, the overall results are discussed in terms of employment status and source data as response variables respectively.

4.4.5.1 Summary of the comparison of employment estimates from surveys over-time

The discussions in Section 3.4 of the results of the March LFS 2006 and 2007 comparisons reveal the existence of relationships between the age group, highest level of education, marital status, population group, sex, province and the outcome employment status. Age group was by far the most significant predictor variable. At the root node, employment status data was partitioned into the 5 different categories of age group.

The highest level of education for age group 30-39 years old has the most significant p-value. The predictor variable with the most significant p-value for age group 20-29 was population group, whereas sex was the most predictive variable for the age group 40-49 years old. The least significant predictor variable was that for highest level of education for the age group 15-19 years old. Young people between the ages of 15 and 19 years old are often still busy with their educational studies and are largely not economically active – the data shows that 88.13% of people in this age group are not economically active. Employment status in this age group can be further explained in terms province and sex.

The results show that the age group 20-29 years old contains people who are likely to come directly from school, only 39.37% were employed; 36.89% were not economically active and 23.73% were unemployed. People who belong to the age groups 30-39 and 40-49 years old have almost the same characteristics - such as highest level of education, sex, province and marital status. More than 63% of people were employed in the two age groups. The last subgroup (age group 50-65 years old) shows that 60.27% of females and unspecified were still employed as compared to 39.02% of males.

In the multinomial logistic regression technique, the highest order interaction was significant, meaning that the change over the provinces differs over the combinations of the interaction between the other 6 variables. The overall percentage correctly classified and identified by the logistic regression model was 61.4%.

The comparison from September LFS 2006 and 2007 was used to check the differences found in the March LFS 2006 and 2007 analyses. Age group was still by far the most significant predictor variable. The results show that province has no significant influence in this data set. At the second level of partitioning it was found that highest level of education, population group and sex, were still the most significant predictors. The most predictive variable in age group 30-39 years old and 50-65 years old is sex. Highest level of education is

the most predictive variable for age group 15-19 years old and 40-49 years old. Population group is still the most predictive variable within the age group 15-19 years old.

There was quite a difference between the source data analysis for the March LFS 2006 and 2007, and the September LFS 2006 and 2007. This is mainly due to the fact that these surveys were weighted to different mid-year population estimates. The fact that they have used different segments of the different master sample contributes some variability. The difference is significant at the 5% level. The comparison of the September LFS 2006 and 2007 shows that there is a real change in the employment status.

In the multinomial logistic regression analyses of the March and September LFS, and September LFS 2006 and 2007 data sets , the highest order interaction was significant, meaning that the change over the provinces differs over the combinations of (interaction between) the other 6 variables. The overall percentage correctly classified and identified by the logistic regression model was 61.8%.

The results from the September GHS 2006 and July GHS 2007 analyses show that age group was by far the most significant predictor variable. At the root node, employment status data was partitioned into different categories of age group. Highest level of education for age group 40-49 years old has the most significant p-value. The predictor variable with the most significant p-value for age group 20-29 was province, whereas sex was the most predictive variable for the age group 30-39 years old and 50-65 years old. Highest level of education was the most predictive variable for the age group 15-19 years old.

Young people between the ages of 15 and 19 years old are often still busy with their educational studies so that they are largely not economically active – the data shows that 89.04% of people in this age group are not economically active. Employment status in this age group can be further explained in terms province and sex. The results show that age group 20-29 years old contains people who are likely to have recently completed their studies, hence only 41.48% were employed; 33.52% were not economically active and 25.01% were unemployed. People within the age groups 30-39 and 40-49 years old have almost the same characteristics such as highest level of education, sex, province and marital status and source data. More than 61% of people were employed in the two age groups. The last subgroup comprising people in the age group 50-65 years old revealed that 58.60% of females and unspecified were still employed as compared to 35.71% of males.

In the multinomial logistic regression technique, the highest order interaction (5 factor interaction) was significant, meaning that the change over the provinces differs over the combinations of the other 4 variables. The overall percentage correctly classified and identified by the logistic regression model was 61.95%.

The results of the analyses of the September GHS 2006 and the July GHS 2007 confirm that there is a real change in the employment status as indicated by the September LFS 2006 and 2007 results. Stats SA, (2007a) reported that there was a decline in the total number of persons employed for the second consecutive year up to, and in 2007. This was due to the discouraged people who have given up looking for work.

4.4.5.2 Summary of the comparison of employment estimates across surveys

The results from September LFS 2007 and September GHS 2007 analysis show that age group was by far the most significant predictor variable. The results from CHAID show that source data was not significant as a single predictor variable. At the root node, employment status data was partitioned into different categories of age group. Highest level of education for age group 30-39 years old has the most significant p-value. The predictor variable with the most significant p-value for age group 20-29 was province, whereas sex was the most predictive variable for the age group 50-65 years old. Highest level of education was the most predictive variable for the age group 15-19 years old. Source data was significant within the age group 15-19 years old, 20-29 years old and 50-65 years old.

The data shows that 88.46% of young people between the ages of 15 and 19 years old are not economically active. Employment status in this age group can be further explained in terms highest level of education, sex, province and source data. The results show that age group 20-29 years old contains people who are likely to come directly from school, only 40.06% were employed; 33.06% were not economically active and 26.87% were unemployed. People in the age groups 30-39 and 40-49 years old have almost the same characteristics such as highest level of education, sex, province and marital status and source data. More than 61% of people were employed in the two age groups. The last subgroup comprising people of age group 50-65 years old revealed that 60.72% of females and unspecified were still employed as compared to 37.65% of males.

In the multinomial logistic regression technique, the highest order interaction (6 factor interactions) was significant, meaning that the change over the provinces differs over the combinations of (interaction between) the other 5 variables. Province was not significant in these six factor interactions. The overall percentage correctly classified and identified by logistic regression model was 61.11%.

The results from September LFS 2007, GHS 2007 and CS 2007 analysis show that age group was the most significant predictor variable. At the root node, employment status data was partitioned into different categories of age group. The results show that source data has no influence in predicting employment status in this data sets. Data was partitioned into five different subgroups as per category of age groups.

Population group for age group 20-29 years old has the most significant p-value. The predictor variable with the most significant p-value for age group 30-39 years old and 40-49 years old was highest level of education, whereas sex was the most predictive variable for the age group 50-65 years old. Highest level of education was the most predictive variable for the age group 15-19 years old.

The data shows that 85.88% of young people between the ages of 15 and 19 years old are not economically active. The results do not indicate much difference between males and females; and LFS 2007, GHS 2007 and CS 2007 in terms of explaining employment status. The results showed that the age group 20-29 years old come directly from school, and only 31.64% were employed; 39.79% were not economically active and 28.57% were unemployed. The subgroup that consisted of age groups 30-39 and 40-49 years old have almost the same characteristics with highest level of education, sex and marital status. The last subgroup age group 50-65 years old revealed that 58.26% of females were still employed as compared to 37.89% of males. Marital status and province were the other predictor variables that played a significant role in explaining employment status in this subgroup.

The results of the logistic regression were not as good as those of CHAID. In the multinomial logistic regression technique, the highest order interaction (6 order) was significant, meaning that the change over the provinces differs over the combinations of the other 5 variables.

CHAPTER 5: Overall discussion, Recommendations and Conclusions

This study used economic data obtained from household interviews conducted by Stats SA, to compare estimates across surveys and within surveys over time. The study compared data from the General Household Survey (GHS) and Labour Force Survey (LFS) over the period 2006–2007 as well as from the 2007 Community Survey (CS). Data from the two provinces, GP and EC were used to represent the socio-economic characteristics and employment patterns in SA. The main objective of this study was to identify inconsistencies between surveys and within a survey over time. In order to generate a set of sufficiently comparable estimates over time, the study of the literature recommended the need to identify and address various sources of potential non-comparability such as:

- how questions have been asked within and across surveys over time
- how the response categories have changed within and across surveys over time
- changes as a result of different reference periods.

In order to achieve the objectives, the predictor variables such as age group, sex, population group, marital status, highest level of education and province, were chosen to profile the attributes of employment status. CHAID and Logistic Regression were used to identify important predictor variables associated with the employment status. Both Cox (1970) and Stoker (1980) had indicated that both CHAID and logistic regression order the predictor variables according to the importance of predicting the response variable. The findings of this study were presented in two parts – the discussion of the CHAID and multinomial logistic regression results, and the overall conclusions. This chapter will also present the limitations of this study and recommend any future work that should be done.

5.1 Discussion of the CHAID and logistic regression results

The results show that both CHAID and multinomial regression can be used effectively with employment data. The trends of employment status are influenced by the complexity of the data and other factors such as demographic variables (sex, age, population group), and economic variables (e.g. level of education). Both techniques have identified predictor variables in the order of strength of association with employment status.

The results of both techniques, more especially the CHAID analyses, give more divisions in subgroups. Age group was by far the most significant predictor on which the data on employment status was segmented. CHAID partitioned data into five different subgroups as per categories of age groups. Employment status changes with the increase of age until a person reaches the retirement age. Different sets of data gave the following trends within the categories of age group:

- More than 80% of people within age group 15-19 years old were not economically active. It has been noted that many of these people are still full-time students.

- Approximately 40% of people in age group 20-29 years old were employed; about 25% were not employed.
- More than 60% of people in both LFS March 2006 and 2007, September 2006 and 2007; and GHS 2006 and 2007 within age group 30-39 and 40-49 years old were employed. The results do not show much difference between them. Highest level of education attainment was the major determining factor in these age groups when predicting employment status. About 49% of people were employed when we used the combination of CS 2007, GHS 2007 and LFS 2007 data sets.
- The percentages of people who were employed and not economically active were almost the same in age group 50-65 years old. About 47% of people were employed.

The results of the logistic regression were not as good as those of CHAID. The results of multinomial regression show several significant interactions of up to seven factors for different sets of data. Logistic regression has the ability to assess the effect of each predictor variable and their interactions contributing in the model. The results show that the predictor variables have a significant role in predicting people's employment status.

The results were not consistent across and within surveys over time. Employment status changes over time and across surveys. This is, among other reasons, due to the use of different mid-year estimates, differences in the instructions given in the questionnaire for CS 2007 and other surveys, as well as the sample size of the surveys. There are major differences between GP and EC in relation to employment status.

5.2 Limitations of the study

This study is limited to only two provinces in South Africa. It would be interesting to find out what the relationships are for the other seven provinces. It would be also of interest to analyse more data sets, more especially within survey over time. This study did not consider the sample rotation, and carrying out separate analyses of portions where there is overlap and those with no overlap.

5.3 Summary and conclusions

We have observed a similar pattern when comparing results of a survey over time. For instance, age group, sex and highest level of education were highly significant when using March LFS 2006 and 2007; September LFS 2006 and 2007 and September GHS 2006 and July 2007. The results, when comparing LFS September 2006 and GHS 2006; CS 2007, GHS 2007 and LFS September 2007, show a different pattern. The overall results suggest that relationships exist between employment status and the predictor variables. The differences noticed may be due to the way questions were asked, more especially between CS 2007 and other surveys. We could not expect any differences due to seasonality for these latter studies since we are comparing surveys conducted at almost the same period.

References

Antipov, E. and Pokryshevskaya, E. (2009). *Applying CHAID for logistic regression diagnostics and classification accuracy improvement*, The State University Higher School of Economics and The Center for Business Analysis, Russia; http://mpra.ub.uni-muenchen.de/21499/MPRA Paper No. 21499, posted 27 March 2010 / 02:08.

Artola, C.and Bell, U.-L. (2001). *Identifying labour market dynamics using labour force survey data,* Discussion Paper No. 01-44, Centre for European Economic Research, Mannheim.

Bakır, B., Batmaz, I., Güntürkün, F. A., İpekçi, İ. A., Köksal, G. and Özdemirel, N. E. (2006). *Defect Cause Modeling with Decision Tree and Regression Analysis,* World Academy of Science, Engineering and Technology, Turkey.

Biemer, P.P, Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. and Sudman, S. (eds.) (1991). *Measurement Errors in Surveys*. New York: John Wiley & Sons, pp. 171-192.

Bignami –Van Assche, (2003). Individual consistency in survey response in rural Malawi, *Special Collection 1: Article 3 - Social Interactions and HIV/AIDS in Rural Africa*.

Bishop, G.F, Hippler, H.J, Schwarz, N. and Strack, F. (1988). *A comparison of Response Effects in Self-administered and Telephone Surveys*. In R.M. Groves *et al*. Telephone Survey Methodology, New York: John Wiley & Sons, pp. 321- 340.

Blair, T. (1999). *Introduction to "Building trust in statistics–The White Paper on Statistics*."[http:ww.statistics.gov.uk/about/national_statistics/downloads/WhitePaperText1.pdf] Accessed July 2008.

Bowler, M and Teresa, L. (2006). Understanding the Employment Measures from the CPS and CES Survey. *Monthly Labour Review*, Vol. 129(2), pp. 23-38; www.bls.gov/opub/mlr/2006/02/art2full.pdf.

Brackstone, G. (1999). Managing Data Quality in a Statistical Agency. *Survey Methodology,* Statistics Canada. Vol. 25(2), pp.139-149.

Brook, K. and Barham, C (2005). *Reliability of the two-quarter longitudinal LFS flows data*, Labour Market Division, ONS National Statistics Online, http://www.statistics.gov.uk/events/gss2005/downloads/PaperC3.doc.

Casale, D. and Posel D. (2002).The Continued Feminisation of the Labour Force in South Africa: An Analysis of Recent Data and Trends. *South African Journal of Economics.* Vol. 70(1), pp.156–184.

Collins, M. and Sykes, W. (1999). Extending the definition of Survey Quality. *Journal of Official Statistics.* Vol. 15(1), pp. 57-66.

Cox, D. R. (1970), The Analysis of Binary Data. *Journal of the Royal Statistical Society, Series C (Applied Statistics).* London: Methuen, pp 113-120.

Diepen, V.M. and Franses, P.H.(2006) , Evaluating Chi-squared automatic interaction detection*, Information systems*, ISSN 0306-4379, Vol. 31(8), pp 814-831.

FESAC (Federal Economic Statistics Advisory Committee) (2005). *Report of the FESAC Subcommittee on the Discrepancy in CPS-CES.*

Flaim, P. and Hogue, C (1985). Measuring labor force flows: a special conference examines the problems. *Monthly Labour Review*. Vol. 108(7), pp. 7-17.

Fu, H. (2004). *Data Inconsistency, Statistical Credibility and the Human Development Report.* Conference on Data Quality for International Organisations. Wiesbaden, Germany.

Hair, J.F, Anderson, R.E, Tatham , R.L and Black, W.C. (1995). *Multivariate Data Analysis*, Fourth edition, Prentice Hall, USA, pp 78-165.

Hawkins, D.M. and Kass, G.V. (1982). *Automatic Interaction Detection*. In D.M. Hawkins (ed) Topics in Applied Multivariate, Cambridge University Press, United Kingdom, pp. 269-302.

Haworth, M, and Caplan, D. (1999). *Time Series and Cross-Sectional Analysis and Modelling in the Monitoring of UK Labour Market*, Office for National Statistics, UK. http://www.fcsm.gov/99papers/haworth.pdf.

Holt, D.T. (2008). Official statistics, public policy and trust, *Journal of the Royal Statistical Society, Series A, Statistics in Society,* Vol. 171(2), pp 323-346.

Hosmer, D. W. and Lemeshow , S. (2000). *Applied logistic regression*. 2nd Edition. New York: Series in Probability and Statistics, pp.31-32.

Jenkins, J and Chandler, M (2010) 'Labour market gross flows data from the Labour Force Survey' *Economic and Labour Market Review* 4(2), pp. 25-30.

Kapsos, S. (2007). *World and regional trends in labour force participation: Methodologies and key results*, Geneva, ILO.

Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, Vol. 29 (2), pp. 119-127.

Keating, B. (2007). *The consistency of data for multi-national enterprises*, Central Statistics Office, Ireland, 93rd DGINS Conference, Hungary.

Kingdon, G. and Knight, J. (2007). *Unemployment in South Africa*, 1995 -2003: causes, problems and policies. *Journal of African Economies*, Vol. 16 (5), pp. 813-848.

Klasen, S and Woolard, I. (2000). *Unemployment and Employment in South Africa, 1995 – 1997*, Report to the Department of Finance, South Africa.

Krosnick, A.J. (1989). Question Wording and Reports of Survey Results. *Public Opinion Quarterly:* Vol. 53(1), p. 107. American Association for Public Opinion Research.

Lehtonen, R. and Pahkinen, E.J.,(2004). *Practical Methods for Design and Analysis of Complex* Surveys, 2nd Edition, Wiley. Longford, NT, pp 257-292.

Lyberg, L., Biemer, P., Collins, M., deLeeuw E., Dippo, C., Schwarz, N. and Trewin, D. (1997). *Survey Measurement and Process Quality*. New York: John Wiley & Sons, pp. 87-114.

Nardone, T., M. Bowler, J. Kropf, K. Kirkland, and Wetrogan, S. (2003). *Examining the Discrepancy in Employment Growth between the CPS and the CES.* Paper presented to the Federal Economic Statistics Advisory Committee, October 17.

Pirouz, F. (2004). *Have Labour Market Outcomes Affected Household Structure in South Africa? A Preliminary Descriptive Analysis of Households, African Development and Poverty Reduction*: The Macro-Micro Linkages Forum paper, South Africa.

Roberts, C. (2007). *Mixing modes of data collection in surveys: A methodological review.* Discussion paper, ESRC National Centre for Research Methods, Centre of Comparative Social Surveys, City University, London.

Rydzewski, L.G., Deming, W.G., and Rones, P.L. (1993), Seasonal employment falls over past three decades, *Monthly Labor Review*, Vol. 118, July, pp. 3 - 14.
SAS Institute Inc., 2002-2008, Version 8, Cary, NC: SAS Institute Inc., 2000, USA.

Stats SA (2004). *Census 2001: Post-enumeration survey: Results and methodology* Report no. 03-02-17 (2001), Statistics South Africa, Pretoria.

Stats SA, (2006a). *Labour Force Survey (LFS), March 2006*, Statistical Release P0210, Statistics South Africa, Pretoria.

Stats SA, (2006b). *Labour Force Survey (LFS), September 2006*, Statistical Release P0210, Statistics South Africa, Pretoria.

Stats SA, (2006c). *General Household Survey (GHS), September 2006*, Statistical Release P0318, Statistics South Africa, Pretoria.

Stats SA, (2007a). *Labour Force Survey (LFS), March 2007*, Statistical Release P0210, Statistics South Africa, Pretoria.

Stats SA, (2007b). *Labour Force Survey (LFS), September 2007*, Statistical Release P0210, Statistics South Africa, Pretoria.

Stats SA, (2007c). *General Household Survey (GHS), July 2007*, Statistical Release P0318, Statistics South Africa, Pretoria.

Stats SA, (2007d). *Community Survey (CS),* Statistical Release P0301, Statistics South Africa, Pretoria.

Stats SA, (2008a). *South African Statistical Quality Assessment Framework (SASQAF),* Statistics South Africa, Pretoria.

Stats SA, (2008b). *Guide to the Quarterly Labour Force Survey*, Statistics South Africa, Pretoria, Report 02-11-01.

Stoker, D.J. (1988). *The Analysis of Complex Sample Data*. Report WS-40, HSRC, Pretoria.

Tarozzi, A. (2007). Calculating comparable statistics from incomparable surveys, with an application to poverty in India. *Journal of Business and Economic Statistics*, Vol. 25, pp. 314-336.

Yu, D. (2009).The comparability of Labour Force Survey (LFS) and Quarterly Labour Force Survey (QLFS), *Working Paper Series No. WP08/2009.* Department of Economics, University of Stellenbosch.

APPENDIX A: Comparisons of questions from and within surveys

The questions below are similar for March LFS 2006 and 2007; September LFS 2006 and 2007. Age groups were derived from the questions on age in completed years, province was derived from the first digit of unique number and employment status were derived from a series of employment question. The questions for September GHS 2006 and July GHS 2007 were similar.

	Questions					
Variable	**March LFS 2006 -2007 and September LFS 2006-2007**		**September GHS 2006 and July GHS 2007**		**October CS 2007**	
	Question	**Response category**	**Question**	**Response category**	**Question**	**Response category**
Sex	**Is ..... a male or a female?** 1 = MALE 2 = FEMALE	☐ 1 ☐ 2	**Is ..... a male or a female?** 1 = MALE 2 = FEMALE	☐ 1 ☐ 2	**Is (the person) male or female?** **Mark the appropriate box with an X.** 1 Male 2 Female Transcribe the answer to F-03 on the flap	☐ 1 ☐ 2
Age	**How old is.....? (In completed years -** *LESS THAN 1 YEAR = 00)*		**How old is.....? (In completed years - In whole numbers)** *Less than 1 year = 00)*		**What is** (the person)**'s age in completed years?** If age not known ask for an estimate of age. If no one is able to estimate, write 998. For babies less than 1 year write 000 for age. For a person 7 years and 10 months write 007 for age.	
Population group (race)	**What population group does .....** *1 = AFRICAN/BLACK* *2 = COLOURED* *3 = INDIAN/ASIAN* *4 = WHITE* *5 = OTHER, SPECIFY IN THE BOX AT THE BOTTOM ……*	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	**What population group does ..... belong to?** *1 = AFRICAN/BLACK* *2 = COLOURED* *3 = INDIAN/ASIAN* *4 = WHITE* *5 = OTHER*	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	**How would** (the person) **describe himself/herself in terms of population group?** Write code in the box. *1 BLACK* *2 COLOURED* *3 INDIAN OR ASIAN* *4 WHITE*	☐
Marital	**What is ……'s present marital**	☐ 1	**What is ……'s present marital status?**		**What is** (the person)**'s PRESENT marital**	

status	**status?** *1 = MARRIED* *2 = LIVING TOGETHER LIKE HUSBAND AND WIFE* *3 = WIDOW/WIDOWER* *4 = DIVORCED OR SEPARATED* → *GO TO Q 1.2* *5 = NEVER MARRIED*	☐ 2 ☐ 3 ☐ 4 ☐ 5	1 = MARRIED 2 = LIVING TOGETHER LIKE HUSBAND AND WIFE 3 = WIDOW/WIDOWER 4 = DIVORCED OR SEPARATED → GO TO Q 1.2 5 = NEVER MARRIED	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	**status?** Write only one code per person. If both civil/religious & traditional indicate civil/religious. READ OUT: 1 Married civil/religious 2 Married traditional/customary 3 Polygamous marriage 4 Living together as married partners 5 Never married 6 Widower/widow 7 Separated 8 Divorced If 5 to 8, **Go to P-10**	☐
Highest level of education	**What is the highest level of education that …… has successfully completed?** 00 = No schooling 01 = Grade R/0 02 = Grade 1/ Sub A 03 = Grade 2 / Sub B 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 I 15 = NTC II 16 = NTC III 17 = Certificate with less than Grade 12/Std 10 18 = Diploma with less than Grade 12/Std 10 19 = Certificate with Grade 12/Std 10	☐ **00** ☐ **01** ☐ **02** ☐ **03** ☐ **04** ☐ **05** ☐ **06** ☐ **07** ☐ **08** ☐ **09** ☐ **10** ☐ **11** ☐ **12** ☐ **13** ☐ **14** ☐ **15** ☐ **16** ☐ **17** ☐ **18** ☐ **19** ☐ **20** ☐ **21** ☐ **22** ☐ **23** ☐ **24** ☐ **25** ☐ **26**	**What is the highest level of education that …… has successfully completed?** 00 = No schooling 01 = Grade R/0 02 = Grade 1/ Sub A 03 = Grade 2 / Sub B 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 I 15 = NTC II 16 = NTC III 17 = Certificate with less than Grade 12/Std 10 18 = Diploma with less than Grade 12/Std 10 19 = Certificate with Grade 12/Std 10 20 = Diploma with Grade 12/Std 1 21 = Bachelors Degree 22 = Bachelors Degree and diploma 23 = Honours Degree	☐ **00** ☐ **01** ☐ **02** ☐ **03** ☐ **04** ☐ **05** ☐ **06** ☐ **07** ☐ **08** ☐ **09** ☐ **10** ☐ **11** ☐ **12** ☐ **13** ☐ **14** ☐ **15** ☐ **16** ☐ **17** ☐ **18** ☐ **19** ☐ **20** ☐ **21** ☐ **22** ☐ **23** ☐ **24** ☐ **25** ☐ **26**	**What is the highest level of education that** (the person) **has completed?** For a person with grade 12, probe whether he/she has a university exemption or not. For a person with a certificate or a diploma, probe whether he/she has grade 12 (std 10) or not.7 00 Grade 0 01 Grade 1 02 Grade 2 03 Grade 3/ Std 1/ ABET 1 04 Grade 4/ Std 2 05 Grade 5/ Std 3/ ABET 2 06 Grade 6/ Std 4 07 Grade 7/ Std 5/ ABET 3 08 Grade 8/ Std 6 09 Grade 9/ Std 7/ ABET 4 10 Grade 10/Std 8/ NTCI 11 Grade 11/ Std 9/ NTCII 12 Attended Grade 12, but not completed Grade 12 13 Grade 12 / Std 10/ NTCIII (without university exemption) 14 Grade 12/ Std 10 (with university exemption) 15 Certificate with < Std10/Gr.12 16 Diploma with < Std 10/Gr. 12 17 Certificate with Std 10/Gr.12 18 Diploma with Std 10 /Gr.12 19 Bachelors degree 20 BTech 21 Post graduate diploma	☐☐

	20 = Diploma with Grade 12/Std 1 21 = Bachelors Degree 22 = Bachelors Degree and diploma 23 = Honours Degree 24 = Higher degree (Masters, Doctorate) 25 = Other, specify in the box at the bottom 26 = Don't know Diplomas or certificates should be of at least six months study duration full time (or equivalent). If code 17-24 → Go to Q 1.3.b, If other code → Go to Q 1.4		24 = Higher degree (Masters, Doctorate) 25 = Other, specify in the box at the bottom 26 = Don't know Diplomas or certificates should be of at least six months study duration full time (or equivalent). If code 17-24 → Go to Q 1.3.b, If other code → Go to Q 1.4		22 Honours degree 23 Higher degree (Masters/PhD) 24 No schooling 98 Out of scope (children under five years of age) Write code in the box.	
Province	1= WESTERN CAPE 2= EASTERN CAPE 3= NORTHERN CAPE 4= FREE STATE 5=KWAZULU –NATAL 6= NORTH WEST 7=GAUTENG 8=MPUMALANGA 9=LIMPOPO		1= WESTERN CAPE 2= EASTERN CAPE 3= NORTHERN CAPE 4= FREE STATE 5=KWAZULU –NATAL 6= NORTH WEST 7=GAUTENG 8=MPUMALANGA 9=LIMPOPO		1= WESTERN CAPE 2= EASTERN CAPE 3= NORTHERN CAPE 4= FREE STATE 5=KWAZULU –NATAL 6= NORTH WEST 7=GAUTENG 8=MPUMALANGA 9=LIMPOPO	
Employment status	**SECTION 2. This section covers activiti** ***Try to ask these questions of each person themselves if at all possible.*** *Read out*: **Now I am going to ask some questions about activities in the last seven days for each household member aged 10 and above**		**Section 2. This section covers activities in the last seven days for all household members aged 15 and above** **Try to ask these questions of each person themselves if at all possible.** *Read out:* **now I am going to ask some questions about activities in the last seven days for each household member aged 10 and above**		**SECTION E: EMPLOYMENT AND ECONOMIC ACTIVITIES - ASK OF ALL PERSONS 15 YEARS AND OLDER LISTED ON THE FLAP** **READ OUT: I am now going to ask you for information on employment of each person 15 years and older.**	
1.	**In the last seven days, did …… do** **the following activities, even for only** **hour?** **a) Run or do any kind of business, big or small, for himself/herself or with one or more partners?** *Examples: Selling things, making things for sale, repairing things, guarding cars,*	YES NO ☐ 1 ☐ 2	**In the last seven days, did …… do any of the following activities, even for only one hour? Show prompt card 2.** **a) Run or do any kind of business, big or small, for himself/herself or with one or more partners?** *Examples: Selling things, making things for sale, repairing things, guarding cars, brewing beer, hairdressing, crèche businesses, taxi or other transport business, having a legal or medical practice, etc.?* **b) Do any work as a domestic worker for a wage, salary, or any payment in kind?**	YES NO ☐ 1 ☐ 2 ☐ 1 ☐ 2	**P-30a) In the last 7 days, did** (the person) **run or do any kind of business, big or small, for himself/herself or with one or more partners even for only one hour?** Examples: Selling things, making things for sale, repairing things, guarding cars, brewing beer, hairdressing, crèche business, taxi or other transport business, having a legal or medical practice, etc. 1 Yes 2 No	☐ 1 ☐ 2 ☐ 3

	brewing beer, hairdressing, crèche businesses, taxi or other transport business, having a legal or medical practice, etc. **b) Do any work for a wage, salary, any payment in kind (excl. domestic work)?** *Examples: a regular job, contract, usual* *piece work for pay, work in exchange* *housing.* **c) Do any work as a domestic worker for a wage, salary, or any payment in kind?** **d) Help unpaid in a household business of any kind?** *Examples: Help to sell things, make things for sale or exchange, doing the accounts, cleaning up for the business, etc. Don't count normal housework.* **e) Do any work on his/her own or the household's plot, farm, food garden, cattle post or kraal, or help in growing farm produce or in looking after animals for the household?** *Examples: ploughing, harvesting, looking after livestock.* **f) Do any construction or major repair work on his/her own home, plot, cattle post or business or those of the household?** **g) Catch any fish, prawns, shells, wild animals or other food for sale or household food?** **h) Beg for money or food in public?**	☐ 1 ☐ 2 ☐ 1 ☐ 2 ☐ 1 ☐ 2 ☐ 1 ☐ 2 ☐ 1 ☐ 2 ☐ 1 ☐ 2 ☐ 1 ☐ 2	**c) Help unpaid in a household business of any kind?** *Examples: Help to sell things, make things for sale or exchange, doing the accounts, cleaning up for the business, etc. Don't count normal* **d) Do any work on his/her own or the household's plot, farm, food garden, cattle post or kraal or help in growing farm produce or in looking after animals for the household?** *Examples: ploughing, harvesting, looking after livestock.* **e) Do any construction or major repair work on his/her own home, plot, cattle post or business or those of the household?** **f) Catch any fish, prawns, shells, wild animals or other food for sale or household food?** **g) Beg for money or food in public?**	☐ 1 ☐ 2 ☐ 1 ☐ 2 ☐ 1 ☐ 2 ☐ 1 ☐ 2 ☐ 1 ☐ 2	3 Do not know **P-30b) In the last 7 days, did** (the person) **do any work for a wage, salary, commission or payment in kind (excluding domestic worker) even for only one hour?** Examples: a regular job, contract, casual or piece work for pay, work in exchange for food or housing. 1 Yes 2 No 3 Do not know **P-30c) In the last 7 days, did** (the person) **do any work as a domestic worker for a wage, salary or payment in kind even for only one hour?** 1 Yes 2 No 3 Do not know **P-30d) In the last 7 days, did** (the person) **help unpaid in a household business of any kind even for only one hour?** Examples: Help to sell things, make things for sale or exchange, doing the accounts, cleaning up for the business, etc. Do not count normal housework. 1 Yes 2 No 3 Do not know **P-30e) In the last 7 days, did** (the person) **do any work on his/her own or the household's plot, farm, food garden, cattle post or kraal, or help in growing farm produce or in looking after animals for the household even for only one hour?** Examples: Ploughing, harvesting, looking after livestock. 1 Yes 2 No 3 Do not know	☐ 1 ☐ 2 ☐ 3 ☐ 1 ☐ 2 ☐ 3 ☐ 1 ☐ 2 ☐ 3 ☐ 1 ☐ 2 ☐ 3

					P-30f) In the last 7 days, did (the person) **do any construction or major repair work on his/her own home, plot, cattle post or business even for only one hour?** 1 Yes 2 No 3 Do not know **P-30g) In the last 7 days, did** (the person) **catch any fish, prawns, shell fish, wild animals either as food for sale or for household use, even for only one hour?** 1 Yes 2 No 3 Do not know	☐ 1 ☐ 2 ☐ 3 ☐ 1 ☐ 2 ☐ 3

2.	*If* "YES" *for a person to any part of Question 2.1* → ***Go to Section 4 for that person.*** **Why did …… not work during the** 01 = HAS FOUND A JOB, BUT IS ONLY STARTING AT A DEFINITE DATE IN THE FUTURE → ***Go to Q 3.12*** 02 = SCHOLAR OR STUDENT AND PREFERS NOT TO WORK 03 = HOUSEWIFE/HOMEMAKER AND PREFERS NOT TO WORK 04 = RETIRED AND PREFERS NOT TO SEEK WORK 05 = ILLNESS, INVALID, DISABLED OR UNABLE TO WORK (HANDICAPPED) 06 = TOO YOUNG OR TOO OLD TO WORK 07 = SEASONAL WORKER, E.G. FRUIT PICKER, WOOL- SHEARER 08 = LACK OF SKILLS OR QUALIFICATIONS FOR AVAILABLE JOBS 09 = CANNOT FIND ANY WORK 10 = CANNOT FIND SUITABLE WORK (SALARY, LOCATION OF WORK OR CONDITIONS NOT SATISFACTORY) 11 = CONTRACT WORKER, E.G. MINE WORKER RESTING ACCORDING TO CONTRACT 12 = RETRENCHED 13 = OTHER REASON	☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12 ☐ 13	**Why did …… not work during the past seven days?** 01 = HAS FOUND A JOB, BUT IS ONLY STARTING AT A DEFINITE DATE IN THE FUTURE → ***Go to Q 2.17*** 02 = SCHOLAR OR STUDENT AND PREFERS NOT TO WORK 03 = HOUSEWIFE/HOMEMAKER AND PREFERS NOT TO WORK 04 = RETIRED AND PREFERS NOT TO SEEK FORMAL WORK 05 = ILLNESS, INVALID, DISABLED OR UNABLE TO WORK (HANDICAPPED) 06 = TOO YOUNG OR TOO OLD TO WORK 07 = SEASONAL WORKER, E.G. FRUIT PICKER, WOOL-SHEARER 08 = LACK OF SKILLS OR QUALIFICATIONS FOR AVAILABLE JOBS 09 = CANNOT FIND ANY WORK 10 = CANNOT FIND SUITABLE WORK (SALARY, LOCATION OF WORK OR CONDITIONS NOT SATISFACTORY) 11 = CONTRACT WORKER, E.G. MINE WORKER RESTING ACCORDING TO CONTRACT 12 = RETRENCHED 13 = OTHER REASON	☐☐	**Why did** (the person) **not work during the past seven days?** 01 Has found a job, but is only starting at a definite date in the future 02 Scholar/student and prefers not to work 03 Housewife/homemaker and prefers not to work 04 Retired and prefers not to seek formal work 05 Invalid, ill, disabled or unable to work (handicapped) 06 Too young or too old to work 07 Seasonal worker, e.g. fruit picker, wool-shearer 08 Lack of skills or qualifications for available jobs 09 Cannot find work 10 Cannot find suitable work (salary, location of work or conditions not satisfactory) 11 Contract worker, e.g. mine worker resting according to contract 12 Retrenched 13 Other reason Write code in the box.	☐☐
3.	**If a suitable job is offered, will …… accept it?** 1 = YES 2 = NO 3 = DON'T KNOW → ***Go to Q 3.12***	☐ 1 ☐ 2 ☐ 3	**If a suitable job is offered, will …… accept it?** 1 = YES 2 = NO 3 = DON'T KNOW → ***Go to Q 2.17***	☐ 1 ☐ 2 ☐ 3	**If a suitable job is offered, how soon can** (the person) **start work?** 1 Within a week 2 Within two weeks 3 Within four weeks 4 More than four weeks from now 5 Not interested 6 Not able (health or disability) 7 Do not know Write code in the box.	☐

4.	**How soon can …… start work?** 1 = WITHIN A WEEK 2 = WITHIN TWO WEEKS 3 = WITHIN FOUR WEEKS 4 = LATER THAN FOUR WEEKS FROM NOW 5 = DON'T KNOW	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	**How soon can …… start work?** 1 = WITHIN A WEEK 2 = WITHIN TWO WEEKS 3 = WITHIN FOUR WEEKS 4 = LATER THAN FOUR WEEKS FROM NOW 5 = DON'T KNOW	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5		
5.	**During the past four weeks, has ……** a) to look for any kind of work b) to start any kind of business *If "No" to <u>both</u> a) and b)* ***→ Go to Q 3.11***	YES NO ☐ 1 ☐ 2 ☐ 1 ☐ 2			**During the past four weeks, has** (the person) **taken any action to look for any kind of work?** 1 Yes 2 No 3 Do not know Mark appropriate box with an X. **During the past four weeks, has** (the person) **taken any action to start any kind of business?** If P-34 & P-35 completed **Go to P-40** 1 Yes 2 No 3 Do not know Mark appropriate box with an X.	☐ 1 ☐ 2 ☐ **3** ☐ 1 ☐ 2 ☐ 3

APPENDIX B: SAS Program

The program below allows us to set to source data to form a new file. The March 2006 data was concatenated with the March 2007 data. Each set of data contains unique number (UQNR) as the identifier variable. The code was also used to determine the sampling rotation waves over time.

```
Data MARCHLFS67;
  Set MARCH2006(in=x)
      MARCH2007(in=z);
   by uqnr;
   if x=1 then source=1;Else /* 1= LFS March 2006*/
   source=2;
   If z=1 then source1=3; Else /* 2= LFS March 2007*/
   source1=4;

   if source=1 and source1=3 then wave=1;
   if source=1 and source1=4 then wave=2;
   if source=2 and source1=3 then wave=3;
   if source=2 and source1=4 then wave=4;

   if weight1=" then weight=weight1;
   if weight2=" then weight=weight2;
run;
```

The code below was used during the rescaling of weight (when we normalise the weight to sample population).

```
Data Survey;
set MARCH671;

weight_1=weight1+0;
weight_2=weight2+0;
weight_13=weight_1+weight_2;
weight13=(weight_13)/2;

if weight_2 =(.) then weight=weight_1 ; else
if weight_1 =(.) then weight=weight_2 ;else
weight=weight13;
run;

Proc summary data=survey mean nway;
class source;
var weight;
output out=RS sum=;
run;

data Tot;
set RS;
S_weight=weight;
drop _type_ weight;
run;
proc sort data=Tot;
by source;
run;

proc sort data=survey;
by source;
run;

data march67_tot;
merge survey(in=a) Tot(in=b);
by source;
if a and b;
run;
```

```
Data March67_tot1;
set march67_tot;
by source;
norm_wgt=weight/(S_weight/_freq_);
Keep uqnr source wave sex agegrp educgrp race prov Empl_Status
marital norm_wgt weight;
run;
```

The next code was used in preparation for the setting of base (reference category) for logistic regression analysis. A variable category with the highest estimated population was set to zero.

```
 Data March20067;
 Set March67_tot1;

IF Empl_Status=2 then Empl_Status=0;else
IF Empl_Status=1 then Empl_Status=1;else
IF Empl_Status=3 then Empl_Status=2;

if source=1 then dsource=0;
if source=2 then dsource=1;

if agegrp=2 then agegrp=0;
if agegrp=1 then agegrp=1;
if agegrp=3 then agegrp=3;
if agegrp=4 then agegrp=4;
if agegrp=5 then agegrp=5;

if educgrp=3 then educgrp=0;
if educgrp=1 then educgrp=1;
if educgrp=2 then educgrp=2;
if educgrp=4 then educgrp=3;
if educgrp=5 then educgrp=4;
if educgrp=6 then educgrp=5;
if educgrp=9 then educgrp=6;

if race=1 then race1=0;
if race=2 then race1=1;
if race=3 then race1=2;
if race=4 then race1=3;
if race=9 then race1=4;

Prov1=0;
if prov=2 then prov1=1;

if sex=2 then sex1=0;
if sex=1 then sex1=1;
if sex=9 then sex1=2;

if marital=5 then marital1=0;
if marital=1 then marital1=1;
if marital=2 then marital1=2;
if marital=3 then marital1=3;
if marital=4 then marital1=4;
if marital=9 then marital1=5;
run;
```

The code below was used used to generate logistic regression output.

```
Proc logistic data=March20067;
class Empl_Status;
model Empl_Status = Agegrp|educgrp|marital|race|sex|prov1|dsource
@7/link=logit selection=backward
                slentry = 0.05  slstay = 0.05  ctable CL EXPB
CLPARM=WALD CLODDS=BOTH/* details */ ;
  output out=predicted1 p=phat lower=lcl upper=ucl  ;
   weight norm_wgt;
ods graphics off;
 Run;
```