SMALL AREA ESTIMATION OF UNEMPLOYMENT FOR SOUTH AFRICAN LABOUR MARKET STATISTICS

by

Jean-Marie Vianney Hakizimana

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

Supervisor: Professor Jacqueline Galpin

School of Statistics and Actuarial Science
University of Witwatersrand

Johannesburg, 2011

Declaration

I declare that this research report is my own, unaided work. It is being submitted to the Faculty of Science, University of Witwatersrand in fulfilment of the requirements for the degree of Master of Science.

The report has not been submitted before for any degree or examination at any other University.

Signature:

Name: Jean-Marie Vianney Hakizimana

Date: 28 February 2011

Abstract

The need for Official Statistics to assist in the planning and monitoring of development projects is becoming more intense, as the country shifts toward better service delivery by local government. It is evident that the demand for statistics at small area level (municipal rather than provincial) is high. However, the statistics with respect to employment status at municipal level is limited by the poor estimation of unemployment in 2001 Census and by changes in boundaries in local government areas. Estimates are judged to be reliable only at provincial level (Stats SA, 2003)

The aim of this study is to investigate possible methods to resolve the problem of the misclassification of employment status in Census 2001 by readjusting the data with respect to the classification of people as employed, unemployed or economically inactive, to that of Labour Force Survey of September 2001. This report gives an overview of the different methods of small area estimation proposed in the literature, and investigates the use of these methods to provide better estimates of employment status at a small area (municipal) level.

The application of the small area estimation methods to employment status shows that the choice of the method used is dependent on the available data as well as the specification of the required domain of estimation. This study uses a two-stage small area model to give estimates of unemployment at different small areas of estimation across the geographical hierarchy (i.e. District Council and Municipality). Even though plausible estimates of the unemployment rate were calculated for each local municipality, the study points out some limitations, one of which is the poor statistical representation (very few people) living in some specific municipalities (e.g. District Management Areas used for national parks). Another issue is the poor classification of employment status in rural areas due to poor data with respect to economic activities, mostly with respect to family businesses, and the non-availability of additional auxiliary data at municipal level, for the validation of the results. The inability to incorporate the time difference factors in the small area estimation model is also a problem.

In spite those limitations, the small area estimation of unemployment in South Africa gives the reference estimates of unemployment at municipality level for targeted policy intervention when looking at reducing the gap between those who have jobs and those who do not. Hence, the outcome of the small area estimation investigation should assist policy makers in their decision-making. In addition, the methodological approach used in this report constitutes a technical contribution to the knowledge of using Small Area Estimation techniques for South African Employment statistics.

Dedication

To my nuclear family:

Albertina, my wife

Nancy, my first daughter

Nadine, my second daughter

Acknowledgement

The completion of this research report would not have been possible were it not for the support of several people. Primarily, I would like to express my deep gratitude to Prof Jacqueline Galpin, my research supervisor. She has been patient and very generous with her time in providing invaluable advice. The same appreciation goes to Prof Paul Fatti, with whom I started the preparation for the proposal before his retirement.

Statistics South Africa has provided the data and the computer software. I am especially grateful to Mr Pali Lehohla and Prof Mbulaheni Nthangeni for having allowed me to further my studies and use Statistics South Africa resources.

I would also like to extend my heartfelt thanks to friends who have been supportive. They have given me moral support to persevere when over-commitments between office work and studies were becoming difficult to manage. I would like to thank colleagues at work who allowed the rescheduling of some requirements so that I could be able to continue on the path of personal development after so many years outside school. Going back to school following the turbulent years of migration from Rwanda was not easy.

Finally, I want to thank the members of family who have endured a reduction in family time. They have been patient, forgiving and understanding. They have given me the moral and physical support that has allowed me to persevere.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronym	Description
ARD	Absolute Relative Difference
BLUP	Best Linear Unbiased Predictor
CV	Coefficient of Variation
DSE	Dual System Estimation
DU	Dwelling Unit
EA	Enumeration Area
EB	Empirical Bayes
EBLUP	Empirical Best Linear Unbiased Predictor
GIS	Geographic Information Systems
HLFS	Household Labour Force Survey
ILO	International Labour Organisation
IPF	Iterative Proportional Fitting
ISTAT	Italy Statistics
LFS	Labour Force Survey
MSE	Mean Squared Error
OFHS	Ohio Family Health Survey
ONS	Office for National Statistics
PES	Post-Enumeration Survey
PSU	Primary Sampling Unit
RDP	Reconstruction and Development Programme
SAE	Small Area Estimation
SEE	Survey of Employment and Earnings
SPREE	Structure Preserving Regression Estimation
Stats SA	Statistics South Africa
UK	United Kingdom
UI	Unemployment Insurance
USA	United States of America

CHAPTER 1: INTRODUCTION

1.1. The context of this research study

1.1.1. Statistical information for policy making

Statistics have become fundamental to the effective functioning of a democratic society. Central and local government use official statistics to develop and monitor public policy, make decisions, allocate resources and for other administrative purposes. Hence, official statistics are vital for public policy, for monitoring the well-being of people and for the democratic process itself. The reporting of the status of economic and social changes in society is a constant feature of modern society (Holt, 2000).

With the establishment of an all-inclusive democratically governed state in 1994, the South African government inherited a country of gross inequities with high unemployment (Knight, 2001). Government policy priorities were: alleviating widespread poverty, reducing income disparities and high unemployment and supplying basic necessities to poor people (Reconstruction and Development Programme [RDP] goals).

1.1.2. The importance of the labour market statistics

In the South African context, the measurement of poverty and inequality uses different comparative indicators between people as a proxy measure for the standard of living. In the summary report to the Office of Deputy President, May, Budlender, Mokate, Rogerson and Stavrou (1998) stated that poverty is closely correlated with unemployment. The lack of a job culminates in a lack of wage income and the reduction of basic resources. The unemployment indicators can also be used to determine the level of social exclusion in the development of people.

In response to the growing needs of monitoring the labour market, Statistics South Africa (Stats SA) has prioritised labour force surveys since 2000 as part of its household surveys, in order to obtain more accurate estimates of employment and unemployment levels. This was because estimates of employment based on the Survey of Employment and Earnings (SEE) may be inconsistent with estimates based on household surveys, as the SEE covers only the formal sector.

The Labour Force Survey (LFS) provides estimates of the numbers and proportions of unemployed persons at national and provincial level. The data cannot provide reliable or precise estimates of unemployment at the municipal level because the sample is not drawn to be representative at that level, and the size of the sample is not large enough to give precise estimates. The attempt to estimate unemployment at national and provincial levels using the 2001 Census data also gave implausible estimates of unemployment, because of shortcomings in the design of the census questions on labour force status.

1.1.3. The need for small area estimation

The demand for small area estimates of labour market statistics is increasing, particularly for use by local government services in their policy planning. This growing demand from local authorities is not only at macro level, but also at the level of social and economic dynamics of small parts of the area of administration (Gebizlioudlu, Depep and Toprak, 1996). This is prompted by a more decentralised approach in managing policy and allocating resources. In South Africa, information from Census 1996 and Census 2001 have been used in the budget allocations per province. As the quality of data improves, the decentralised approach is gaining momentum, particularly at the municipality level.

On one hand, the traditional approach of population estimation within the statistical domain of precision in regular sample surveys cannot provide sufficient data for small areas or sub-populations of interest for a valid and viable statistical inference. On the other hand, due to its prohibitive costs, the census operation cannot be conducted frequently enough to get data for monitoring pattern changes of dynamic phenomena such as employment and unemployment. However, several other methods can be used, for example, combining some survey data with the latest census results, frames or registers. A number of techniques have been developed for the estimation of small area statistics in order to respond to the growing use of small area estimates for local planning. The techniques related to Small Area Estimation (SAE) have been well documented through the seminal works of Rao (2003a) and Longford (2005a). The SAE methods have been identified as the direct domain of estimation methods (i.e. survey domain estimation), indirect domain estimation methods (i.e. synthetic or composite estimation), model-based methods (i.e. area level or unit level models) and the Bayesian approach to SAE. The techniques of SAE are a generalisation or a combination of other statistical methods, particularly structured around the regression framework (Marker, 1999). More detailed discussions of the methods are given in Chapter 2.

1.2. Motivation of the study

1.2.1. Methodological motivation

In South Africa, population censuses have been the main source of information at local area level. The experience with the last two censuses conducted by Stats SA has been that it constitutes a huge logistical exercise that had some weaknesses in terms of operational management. This has resulted in the reduction of coverage with the net undercount calculated using the post-enumeration survey (PES), estimated at 8.28% (10,69% without removal of census erroneous inclusion) in the 1996 Census, and 17.6% in the 2001 Census. However, Stats SA has been using almost the same conceptual definition of labour statistics from household based interviews. Consequently, some demographic variables from both the census and labour force surveys give a good basis for the development of model-based estimates.

The evaluation of the results of Census 2001 showed a substantial difference between the unemployment estimates from the September 2001 LFS and Census 2001. This prompted the SA Statistics Council to issue a warning statement regarding the higher unemployment measured from the Census as opposed to unemployment measured from the LFS (Stats SA, 2003). The difference might be attributable in part to the probing questions included in LFS but not in the Census questionnaire. There is also a possibility that the difference is due to dynamic changes that need to be adjusted from the LFS (Sept. 2001), as the LFS sample was designed using the sampling frame from the 1996 Census. Lastly, the differences might be caused by the large undercount rate from the 2001 Census.

Based on 2001 Census data, the spatial distribution of the census unemployment rates for the 262 municipalities (shown in Annexure I) varies from 0.71 to 84.06%. It is expected that combining the unemployment estimates from the LFS and the Census will produce better controlled estimates at the small areas of estimation. The focus of this study will be on minimising model errors, borrowing the strength from the LFS, making use of spatial relationships of neighbouring areas (using latitude and longitude as the x and y coordinates), taking advantage of the design criteria of complex samples (LFS) and looking into improvement of some other classification estimates. The municipality is currently considered as the small area of estimation at which statistical reliability is required in South African labour statistics.

1.2.2. Practical motivation

1.2.2.1. The limitation of accuracy of labour force survey

The measurement of unemployment has been incorporated into the household survey program since 1994. The October Household Surveys of 1994, 1995, 1996, 1997, 1998 and 1999 provided trends and levels of unemployment measured from the household members' point of view. It is important to note that some additional reconciliation is necessary when comparing the results from successive October household surveys and the subsequent LFSs. The questionnaires used in these surveys have constantly been reviewed, while keeping to internationally accredited concepts and standards. These improvements in the questionnaire as well as changes in the estimates of the total population do, however, mean that data from the different years are sometimes not completely comparable.

In view of the available data from the population census (1996 Census), the October household survey was replaced by a survey of labour statistics, which was seen as a priority. The LFS was introduced in February 2000, to be conducted twice a year. The estimates of the survey have been made reliable at least within a domain of estimation such as province (February 2000 – February 2004) and later (since September 2004), after the revision of the sample design based on the 2001 Census frame, at district council level. However, the data cannot yield reliable estimates of unemployment for planning purposes at the administrative target unit of policy implementation, such as the municipality. Other sources of data on labour market statistics, such as the SEE, are limited in terms of coverage, to the formal economy or large establishments. Hence, there is a growing demand to extend the measurement of unemployment by the LFS to at least the lowest level of policy formulation.

1.2.2.2. The poor estimation of unemployment by 2001 Census

Although the release of the 1996 Census gave an improved overall profile of the demographic characteristics of the South African population, the final estimates were subject to adjustment according to an undercount derived from the PES. There were also quite significant changes in the way the employment questions were phrased and sequenced over the period 1994 to 1997 (Stats SA, 1998, p 64)., resulting in slight changes in the method of calculation of the unemployment rate over this period.

The 2001 Census questionnaire design was done taking into consideration the formulation of labour questions. However, the questions were limited in terms of probing of the respondents as to the classification of those economically active or not, and specifically a new category introduced in 1997, namely those not working, not looking for work, but available for work. There are a considerable number of persons excluded from labour market participation because they have abandoned the search for employment.

The unemployment estimates from the 2001 Census were substantially different to the estimates of September 2001 LFS (Stats SA, 2003). Some attempts to readjust the difference suggested that variation was due to LFS results being calibrated to the mid-year estimates for 2001 and to the huge adjustment arising from the undercount in the 2001 Census. The Census adjustment factors were calculated within the homogenous classes obtained using an Automatic Interaction Detection analysis with the assumption of similar undercount rates within the same classes. However, the difference of estimates of unemployment between Census and LFS is so large that there is a need to use auxiliary information from other sources to investigate if this might produce more plausible estimates. This study looked at possible improvements in the measurement of unemployment using Census information corrected by the LFS and the centroid x and y (latitude and longitude) coordinates of each target geographical area represented in the small area framework.

1.2.2.3. The change in the spatial distribution of the population in the country

The settlement pattern of the South African population has been influenced by many endogenous and exogenous factors varying from political influence to spatial demographic growth. Since the 1996 Census, the spatial population distribution has been measured based on the full demarcation of the country into small units of data collection called enumeration areas (EAs). The EA has been considered as the building block of the aggregation of the population into place-names or suburbs, local government areas, District Council and provinces.

In the 2001 Census operations, the EA groupings and definitions were redefined based on population or household concentrations. This re-demarcation showed that there are other spatial elements attracting people to a given area. For instance, infrastructure such as schools, roads, hospitals and industrial establishments are some of the elements attracting population settlement. These elements are closely correlated with employment creation in different areas.

If one generalises any value measured from an EA to a geographic coordinate point, the point pattern distribution of the central point of EA or municipality can be spatially regressed on the point patterns of infrastructure, in order to improve the estimation in the spread of the measurement of unemployment. The distribution within the municipality prompts the need for spatially distributed correction factors of unemployment, in order to implement policy intervention programmes such as job creation investment or labour intensive projects.

1.2.2.4. The lack of relevant comparative data source at low level

The need to support local government programmes in job creation requires a statistical tool for the estimation of unemployment. Since the LFS is able to give the required trend at national and provincial level, the use of auxiliary variables available also in the census data and geographical information system data could assist in getting a SAE of unemployment.

1.3. The objective of the study

The aim of this study is to bring together the information on unemployment estimated by the LFS and other information from census and geographical data, in order to get an improved spatial estimate of unemployment per target area (municipality or district council). Therefore, this study will explore methodological issues arising from a combined/integrated model for the estimation of unemployment, using LFS data together with other socio-economic variables and census variables in a regression. Both sources are relatively less reliable than the expected true estimate for local areas due to sample design limitations in the LFS and poor direct measurement of unemployment by the 2001 Census.

A variety of SAE approaches will be examined in this study, with the intention of assessing the most reliable methodology by combining the model fitting methods and other estimation methods. The aim is to produce enhanced estimates for the counts and rates of unemployment for each proclaimed municipality in South Africa (based on Local Government: Municipal Structure Act No 117 of 1998 with the subsequent amendment).

Though the study could use data collected over different periods, the objective is to get a feasible estimate of unemployment for small areas in 2001, using the September 2001 LFS, the 2001 Population Census and the Geographic Information System (GIS) identification of the 2001 EAs fitting the proclaimed boundaries. Some of the issues to be investigated are:

- to assess whether it is possible to estimate the unemployment rate within different levels of geographical hierarchy;
- whether SAE methods are able to yield estimates that are acceptable; and
- whether the labour market profile (e.g. unemployment among youth versus unemployment among other age groups) is maintained at the level of the traditional domains of estimation (National and Province).

The other issue is the variability of estimates across data sources. The census figures are adjusted for the undercount according to a dual-system-estimation (DSE) analysis used in the PES. The adjustment factors from the 2001 PES are not reliable beyond the domain of estimation of the survey design: National level, Urban/non-urban nationally and Provincial level (Stats SA, 2004). The PES universe was also limited to residents of housing units and workers' hostels, with the exclusion of other collective living quarters such as residential hotels, homes for the aged, student residences, tourist hotels and institutions (Stats SA, 2004). The master sample used for the LFS used a similar universe with identical exclusions.

1.4. Summary of the road map in this research report

This research report comprises five main sections:

(1) introduction in chapter 1,

(2) overview of the concepts and methods in chapter 2,

(3) literature survey in chapter 3,

(4) application of SAE to the estimation of unemployment in chapter 4,

(5) conclusions in chapter 5.

The introduction covered the context, relevance and needs of the estimation of unemployment statistics for small areas for policy planning. The context gives an understanding of reasons motivating this study, both methodological and practical. It is understood that the different methodologies of estimation of statistical information within the domain of estimation are limited by the sampling errors, in the case of surveys, or coverage errors in the case of the census. There are also practical reasons that motivate the need for estimation for small areas. These are referenced by existing sources of information for labour market statistics. The official LFS has limitations of accuracy within its domain of estimation. Even the use of the 2001 Census has not been able to give a better estimate of unemployment in South Africa, most probably due to the huge undercount and the depth of the screening questions for the population in or out of the work force. The other practical reason is the rapid change in the pattern of the spatial population observed in the last 10 years. There is evidence that internal migration has increased since 1994 (Stats SA, 2002).

The analysis presented in this report is supported by the relevant literature. The review attempts to cover different estimation methods, supported by a succinct discussion of practical experiences in the estimation of small area statistics. Based on different statistical methods of estimation for small areas, special importance is given to their application to the labour market. The focus in terms of practical application to the South Africa employment status profile will be the techniques of direct estimation, synthetic estimation and the unit level SAE methods. Further approaches are defined in Chapter 2, as background for the illustration of experiences of other countries recorded in the literature.

CHAPTER 2: CONCEPTS AND METHODS

This chapter gives the theoretical foundation of the report. In addition to the understanding of the SAE concepts, the chapter explorers the standard definitions of the terms used concerning labour market statistics before giving the methods and procedures of SAE.

2.1. Concepts and Definitions

2.1.1. Definition of small areas

The concept of small area estimation can be defined as a derivative of the sample survey definition of the domain of estimation. In this context, the domain is the subpopulation of interest, where the sample design should yield direct estimates with adequate precision associated with the inferred weighted population estimates. This is usually referred to as design-based estimation. A domain (area) is regarded as "small" if the domain-specific sample is not large enough to support direct estimates of adequate precision (Rao, 2003a).

In some other instances, SAE will use model-assisted estimation with the objective of deriving robust inferences, which correct the data distortion within the domain (small area). It is often necessary to use indirect estimates by fortifying the data with good auxiliary data such as a recent Census or current administrative records of the domain, and the determination of suitable models linking areas.

2.1.2. Definition of some terms used

The definitions given in this section were taken from the statistical release P0210 "Labour Force Survey September 2001" (Stats SA, 2002) as referenced in the metadata notes on pages xiii, xiv and xv.

2.1.2.1.1. Definition of unemployment

The official definition of unemployment is referred to as the *strict definition* where the characteristics of *unemployed* persons are for those people within the *economically active population* who: (a) did not work during the seven days prior to the interview, (b) wanted to work and were available to start work within two weeks of the interview, and (c) had taken active steps to look for work or to start some form of self-employment in the four weeks prior to the interview. South Africa has been reporting, in addition to the official definition, another type of definition called the *expanded definition of unemployment* that excludes criterion (c).

The target group for the measurement of unemployment is all persons aged between 15 and 65 years old. Even though the International Labour Organisation (ILO) prescription for the age universe is 15 to 64 (inclusive), the upper limit is set to 65 because this is the most common retirement age in South Africa. The ultimate aim is to be able to classify all persons aged between 15 and 65 into economically active or inactive.

For those economically active, the measure of employment and unemployment is determined. Among the not economically active, there are those who do not want to work (e.g. pensioner) and those who might still want to work but are not currently able to do so (e.g. student, raising children ….) (Figure 2.1.2.1).

Figure 2.1.2.1

In the LFS, the other important concepts are:

- The *working age population* comprises all persons aged 15–65 years.
- The *economically active population* consists of both those who are employed and those who are unemployed.
- The *employed* are those who performed work for pay, profit or family gain in the seven days prior to the survey interview, or who were absent from work during these seven days, but did have some form of work to which to return.
- The people who are *out of the labour market* or who are *not economically active* are those who are not available for work. This category includes full-time scholars and students, full-time homemakers, those who are retired, and those who are unable or unwilling to work.
- *Workers* include the self-employed, employers and employees.
- *The formal sector* includes all businesses that are registered in any way.
- *The informal sector* consists of those businesses that are not registered in any way. They are generally small in nature, and are seldom run from business premises. Instead, they are run from homes, street pavements or other informal arrangements.
- *Labour market dynamics* refer to movement into, out of, and within the labour market over a specified period of time.

2.1.2.1.2. Additional definitions used in the Census and the Labour Force Survey

A *Population Census* is the total process of collecting, processing, analysing and publishing or otherwise disseminating demographic, economic and social data pertaining to all persons in a country or in a well-defined part of a country at a specified time. In South Africa, the listing unit is sometimes referred to as dwelling unit (DU).

The *LFS* is a twice-yearly rotating panel household survey, specifically designed to measure the dynamics of employment and unemployment in the country. It also provides insight into a variety of issues related to the labour market, including the level and pattern of unemployment and the industrial and occupational structure of the economy. The target population is all persons residing in normal households and in workers' hostels. The survey does not cover institutions such as old age homes, hospitals, prisons and military barracks.

The *Master Sample* is a multi-stage stratified sample used by Stats SA's household surveys. The LFS overall sample size of Primary Sampling Units (PSU) is 3000. The explicit strata were the 53 district councils and 9 Provinces. A PSU corresponds in most cases to one or two EAs. The number of PSUs was allocated using the power allocation method and these were sampled using probability proportional to size principles. The measure of size used was the number of households in a PSU, as calculated in the Census. The sampled PSUs were listed with the DU as the listing unit and a systematic sample of dwelling units per PSU was drawn. These samples of dwelling units form clusters of ten dwelling units.

The LFS uses a *rotating panel methodology*, which gives a picture of movements into and out of the labour market over time. The rotating panel methodology involves visiting the same DU on a number of occasions (in this instance, five at most). After the panel is established, a proportion of the DUs is replaced each round (in this instance, 20%). New DUs are added to the sample to replace those that are taken out. The advantage of this type of design is that it provides the basis for monitoring changes in the work situation of members of the same households over time, while retaining the larger picture of the overall employment situation in the country. It also allows for both longitudinal and cross-sectional analysis.

The *dwelling unit* or *housing unit* is a unit of accommodation for a household, which may consist of one structure, or more than one structure, or part of a structure. (Examples of each are a house, a group of rondavels, and a flat.) It may be vacant, or occupied by one or more than one household. *A housing unit* usually has a separate entrance from outside or from a common space, as in a block of flats. A *dwelling unit* is any structure or part of a structure or group of structures occupied by one or more than one household; or which is vacant or under construction but could be lived in at the time of the survey. The DU is the major listing unit for these surveys. However, if multiple households are identified during listing, then each household is listed separately. However, the listing unit is not primarily households, as multiple households are sometimes discovered at the time of the survey. In workers hostels, (1) where rooms are occupied by individual persons/households, then each room is treated as a DU, and (2) in the case of dormitories/communal rooms, each bed is listed separately and treated as a DU. It is important to note that the dwelling unit as defined here was also the selection unit for these surveys.

A household is defined as a person, or group of persons, who occupied a common DU (or part of it) for at least four days in a week on average during the four weeks prior to the survey interview. They live together and share resources as a unit. Other explanatory phrases can be "eating from the same pot" and "cook and eat together".

Workers' hostel is a communal living quarter for workers, provided by a public organisation such as a local authority, or a private organisation such as a mining company. These were residential dormitories established for migrant workers during the apartheid era, and they continue to house people working in certain industries, such as the mining industry.

Domestic worker is a person employed to work in the household e.g. as a house cleaner, cook, gardener, driver or nanny. There are some domestic workers who live on the property of the employer, either in the same house or in separate domestic quarters. There are those who live in separate dwelling unit. Usually, domestic workers have families and responsibilities of their own elsewhere. Thus, they are considered as separate households.

2.2. Methods of Small Area Estimation

2.2.1. Statistical notation

For the purpose of the discussion of the different statistical estimation techniques, a population of size N is considered from which a sample of size n was drawn. The attribute or measurement of the characteristic of interest is represented by y, for instance the unemployed population. It is assumed that there are m small geographic subdivisions (such as province, district Council, or municipality) and l socio-demographic subgroups. As a result, y_{ihk} represents the value of the characteristic of interest y on the k^{th} unit in the h^{th} socio-demographic subgroup in the i^{th} small geographic area, where i= $1,2,\ldots m$, $h=1,2,\ldots l$ and $k=1,2,\ldots n_{ih}$.

Since $\sum_i \sum_h n_{ih} = n$, this implies that n_{ih} represents the sample size in the $(ih)^{th}$ cross classification. The notation for the overall population is $\sum_i \sum_h N_{ih} = N$ where N_{ih} represents the population size in the $(ih)^{th}$ cross classification. The i^{th} stratum population is represented by $N_i = \sum_h N_{ih}$, the sample size of the i^{th} stratum by $n_i = \sum_h n_{ih}$ and the total measure by $\sum_h \sum_k y_{ihk} = y_i$.

2.2.2. Direct domain estimation

2.2.2.1. Survey measurement estimation

Sample surveys have been used to provide reliable direct estimates of totals and means (Rao, 2003a). They use data from the sample units in the area of domain of interest such as the whole population and large area. Hence, the estimate of a characteristic Y_i, for simple random sampling, would be given by

$$\hat{Y}_i^{srs} = \frac{N}{n} \sum_h^l \sum_{k=1}^{n_{ih}} y_{ihk} \text{ if } n_{ih} \geq 1 \text{ or zero otherwise.} \tag{2.1}$$

In the case where N_i is known, a post-stratified estimator can be defined as $\hat{Y}_i^{pst} = \frac{N_i}{n_i} \sum_h^l \sum_{k=1}^{n_{ih}} y_{ihk} = N_i \overline{y}_i$ if $n_{ih} \geq 1$ and zero otherwise. (2.2)

A direct estimator for the mean for the small area *i* is given by

$$\hat{\overline{Y}}_i = \overline{y}_i = \frac{1}{n_i} \sum_h^l \sum_{k=1}^{n_{ih}} y_{ihk} \text{ if } n_i \geq 1 \text{ and zero otherwise.} \tag{2.3}$$

The variance for $\overline{y}_i$ may be estimated by $Var(\overline{y}_i) = \frac{1-f_i}{n_i} s_{yi}^2 = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) s_{yi}^2$, (2.4)

if N_i is known where $f_i = \frac{n_i}{N_i}$ and $S_{yi}^2 = \sum_h^l \sum_{k=1}^{N_i} \frac{\left(Y_{ik} - \overline{Y}_i\right)^2}{N_i - 1}$, $s_{yi}^2 = \sum_h^l \sum_{k=1}^{n_i} \frac{\left(y_{ik} - \overline{y}_i\right)^2}{n_i - 1}$, $N_i \geq 2$ since s_{yi}^2 is an unbiased estimator for the population parameter S_{yi}^2. If N_i is unknown, the variance of $\overline{y}_i$ may be estimated by $Var(\overline{y}_i) = \frac{1-f}{n} s_{yi}^2$ where $f = \frac{n}{N}$, if it can be assumed that the f_i's are very similar for all strata.

The variance of the estimator $\hat{Y}_i$ (total number) may be estimated in a similar way to that of $\overline{y}_i$ (the mean) (Cochran, 1977, page 23-24).

The direct estimator is used when the sample size for each small area is sufficiently large to yield acceptably precise estimates for that area. In the case of the LFS of September 2001, acceptably reliable estimates can be obtained at national and provincial level only. (The acceptably reliable estimates at District Council level were only possible after the introduction of the new master sample in 2003, which is representative at the District Council level).

There are some traditional issues of sample design related to the number of strata required, the construction of strata, the sample allocation and the selection of probabilities, that need to be taken into consideration. Although the ideal goal is to find an optimal design that minimises the mean square error (MSE) of the direct estimate within the available budget, in practice a compromise is adopted where a reduced reliability in some small areas or domains will be allowed.

While clustering may lead to a reduction of the survey cost, Rao (2003a) notes that the *minimisation of clustering* is sometimes considered in order to avoid the loss in the "effective" sample size. It may also assist estimation for domains where the sample size can only be determined after the data is collected, such as social-demographic domains. However, the choice of the sampling frame to use, the sampling units and number of sampling stages have a significant impact on the effective sample size. Rao (2003a) also suggests that, in cases where there are different surveys measuring the same characteristic, using different stratifications in the sampling design, a better sample size distribution at the small area level can be obtained by *stratification* using the intersection of these strata, or "small strata". In addition, he suggests using a compromise *sample allocation* may be sufficient for the reliability requirements at both the small area and large area, using only direct estimates. Here part of the sample is allocated so as to provide reliable estimates at the large area level, with the remaining sample being allocated to improve the estimates as small area level. It is important to have a kind of *integrated survey* program, so that the questions relating to the characteristics of measurement or variables are harmonised across surveys, across regions or countries.

The South African LFS is based on a two-stage complex sample, where the first stage involves the selection of PSUs using the probability proportional to size selection scheme with the number of households per PSU as a measure of size. The second stage involves the systematic selection of 10 dwelling units from each of the selected PSUs. The number of households used as a measure of size for selecting PSUs is obtained from the Census 1996 list of EAs. The overall weights are obtained from the design weights for unit non-response and for post-stratification by adjusting the weights to conform to known population distributions, improve precision and to compensate for non-coverage.

The final estimate of the characteristic of interest within the small area i under a stratified multistage sampling (with socio-demographic variables different in various areas) is given by $\hat{Y}_i = \sum_{h=1}^{l_i}\sum_{k=1}^{n_{ih}} w_{ihk} y_{ihk}$ (2.5),

where the overall weight is written $w_{ihk} = w_{hk1} * w_{hk2.1} * w_{ihk3.21}$ where

- $w_{hk1} = \frac{1}{p_{hk}} = \frac{1}{p_{hk}(PSU) * p_{hk}(DU)}$ where p_{hk} represent the probability of selection for the PSU and DU.
- $w_{hk2.1}$ is the adjustment factor for non-response where $w_{hk2.1} = \frac{1}{\mathrm{Re}sponse\_rate_{hk}}$
- $w_{ihk3.21} = \frac{w_{hk2.1}}{\sum_{hk} w_{hk2.1}} * \Gamma_i$ where Γ_i is the known marginal total in the i^{th} small area or domain of estimation to which the results are calibrated.

2.2.2.2. Population Census

The population census provides population counts for detailed geographical areas of a country as well as for different domains (sub-populations) such as age, sex, marital status, population group, level of education and other demographic variables. These counts give, in most cases, the basis for the calculation of resource allocation (budget for the equitable share of the municipal infrastructure grant).

The Census requires a huge fieldwork operation where each dwelling unit in the country is visited. In the South African Census, the method of data collection is the application of a questionnaire form, where *all* people who were present in the country on a given reference night are recorded. The 2001 Census reference night was on 9–10 October 2001. People living in households across the country, as well as those in hostels, hotels, hospitals and all other types of communal living quarters, and even the homeless, were all visited. The total count of persons can be written as

$$P = \sum_{j \in \cup} p_j \text{ , } \cup = 1, \ldots\ldots, N \tag{2.6}$$

where p_j represents the counted persons in a given household. Each person can also be associated with a particular characteristic, so that y_j is the number of persons in household j having the given characteristic

$$Y = \sum_{j \in \cup} y_j = \sum_i y_i = \sum_i \sum_h \sum_k y_{ihk} \text{ .} \tag{2.7}$$

In preparation for the total count, the entire country is always demarcated into EAs. The size of each EA is such that an enumerator can visit all DUs in the EA to administer the census questionnaire to the household.

Although the Census aims to have a complete enumeration, a 'perfect' census is impossible due to omissions and false enumeration. A second visit is conducted independently of the census, where a representative sample of EA's is re-visited. This second exercise is called a PES. The census and PES records for each household are matched and compared. The expected outcome is an estimate of the true population based on a dual system of estimation calculated using adjustment factors in order to correct the under-coverage or the over-coverage. In the 2001 South African Census, the estimate of true population was calculated as

$$\hat{P} = \frac{\hat{P}_{PES} * P_{cenus\_corrected}}{\hat{P}_{matched}} = af * P_{Census\_count} \tag{2.8}$$

and $af = \frac{1}{1-ur}$ where *af* is the adjustment factor and *ur* is the undercount rate (Census 2001: Post-enumeration Survey, 2004).

The dual system of estimation assumes that there was a closed population with insignificant migratory movement in the period between the census night and the PES visit. There is an independency between the Census and PES with respect to the fieldwork teams; there are almost no erroneous inclusions in either system such as duplication, fabrications, boundary misallocation or the inclusion of units outside the target universe. In addition, it assumes that all matching cases had been resolved, considering that neither the sample nor the full census count is better than the other. The estimation of the true population using the dual system of estimation is affected by some errors, as with any other sample that includes sampling errors (i.e. variance and bias) and non-sampling errors (i.e. non-response bias, correlation bias and matching bias).

The PES *variance* is obtained by a deduction of the sampling procedure where the estimates of variance are represented by the standard error, the absolute error and the confidence intervals from each dual system estimate. The PES *sampling bias* is controlled by maintaining adequate sample size in each cell of estimation. The PES *non-response bias* is a result of the refusals, non-contacts and unusable questionnaires. The operational management of the fieldwork is paramount in reducing types of error. The PES *correlation bias* is observed in situations where the same event results in difficulty in enumerating a group of persons in both the PES and Census. This arises if there is a lack of operational independence between the PES and Census under the same organization or division structure. The other bias concern, the PES *matching bias,* refers to errors during the matching process, such as erroneous matches and erroneous non-matches.

2.2.3. *Indirect Estimation*

The indirect estimator for the small area borrows strength from sample units outside the domain and/or time period of interest using a statistical model. This study will investigate which indirect methods yield the most plausible estimates based on the available data: the Population Census, the LFS and the geographical coordinates.

2.2.3.1. Synthetic Estimation procedure

A simple indirect estimator would assume that there is the same proportion of the characteristic of interest at the small area level as at the large area (e.g. national or provincial) level. This proportion can be written as: $p_i = p$ where *i* represents the different small areas. Given an unbiased estimate from a sample for the large area, this estimator can be used to derive estimates for sub-areas on the assumption that the small areas have the same characteristics as the larger area. This method is referred to as synthetic estimation. The method assumes the availability of estimates from, e.g. a survey for a large subset of the population. In this case, the reliable estimates of the LFS were published in September 2001 for the whole country, provinces and demographic groups.

If it is assumed that the small area of interest is the small geographical area (i.e. municipality) and the reliable direct estimate $\hat{Y}_{.h}$ is available for each subgroup *h* at the large area level from the survey, the corresponding estimates can be deduced as $\hat{Y}_i = \sum_h p_{ih} \hat{Y}_{.h}$ where $p_{ih} = \frac{N_{ih}}{N_i}$ represents the weights of the subgroup *h* in the small area *i* as observed in census and $\sum_h p_{ih} = 1$. The variance of the synthetic estimate is based on inverse of the sample size for the large area, and therefore it is smaller than that of the post-stratified estimator. However, the synthetic estimates are biased if the assumption of homogeneity for a given socio-demographic subgroup for both the large and small areas is not be valid (Francisco, 2003). Moreover, the structure of the population may have changed between the Census and the survey even though the October 2001 Census and September 2001 LFS are close together in time.

2.2.3.2. Composite Estimation

The composite estimator is a weighted average of a direct estimator and an indirect estimator (Ghosh and Rao, 1994). It is expected that the composite estimator will balance the potential bias of the synthetic estimator against the instability of the direct estimator (Francisco, 2003). The composite estimator can be written as $\hat{Y}_i^c = w_i * \hat{Y}_i + (1 - w_i) * \hat{Y}_i^s$ (2.9) where $\hat{Y}_i^c$ is the composite estimator for small area *i*, $\hat{Y}_i$ is the direct estimator, $\hat{Y}_i^s$ is the synthetic estimator and w_i is a selected weight with $0 \le w_i \le 1$.

There are different ways of determining the weights. The optimal weights suggested by Ghosh and Rao (1994) are obtained by minimizing the MSE of $\hat{Y}_i^c$ with respect to w_i under the assumption that $\text{cov}(\hat{Y}_i, \hat{Y}_i^s) = 0$. The result is $w_i = \dfrac{MSE(\hat{Y}_i^s)}{MSE(\hat{Y}_i^s) + V(\hat{Y}_i)} = \dfrac{\left(\hat{Y}_i^s - \hat{Y}_i\right)^2 - V(\hat{Y}_i)}{\left(\hat{Y}_i^s - \hat{Y}_i\right)^2}$. (2.10)

The composite estimator is presented here for illustration purposes as a method that has been used in the literature, but will not be investigated in this study.

2.2.3.3. Structure Preserving Estimation

The Structure Preserving Regression Estimation (SPREE) is a generalization of the synthetic estimator, based on reliable direct estimates. The parameter of interest is a count, such as the number of unemployed persons in a small area. It assumes that the totals of N_{iab} (as provided by the census) are known, where i denotes the small area of interest, a denotes the categories of the variable of interest (y) and b denotes the categories of the categories of an auxiliary variable x correlated with y. SPREE uses the iterative proportional fitting (IPF) method to adjust the cell counts M_{iab} of the survey, to those of the N_{iab}, using reliable survey totals of the auxiliary variable $N_{i.b}$ The SPREE method is similar to calibration estimation using the Horvitz-Thomson estimator, leading to $\hat{Y} = \sum_k w_k y_k$ where w_k represents the calibrated weight subject to the constraint $\sum_{k \in s} w_k x_{ki} = X_i$ representing the total within a specific domain or area *i* of x_{ki} auxiliary data or reference data.

The above estimator (SPREE) is also presented as an illustration of possible methods that have been applied to the measurement of unemployment in the literature.

2.2.3.4. Synthetic Regression

Under the assumption of the same population characteristics for both the large and small areas, a regression-synthetic estimator can be derived using a vector of domain-specific auxiliary variables in the form of known totals (Rao, 2003a). Suppose that the vector of auxiliary characteristic information X_h of the known population totals is available (as in Census 2001) and that the similar auxiliary characteristics x_h for $h \in s$ are also measured in the sample ***s*** of the universe (as in the September 2001 LFS). This implies that the

measured characteristic data (y_k, x_k) for each element $k \in s$ is part of the observed universe. The generalised regression estimator that makes efficient use of the auxiliary information can be written as
$\hat{Y}_{GR} = \hat{Y} + (X - \hat{X})^T \hat{\mathrm{B}}$. (2.11)

The synthetic regression estimator is written as: $\hat{Y}_{igrs} = X_i^T \hat{\mathrm{B}}$ (2.12)

where $\hat{B}$ is a solution of the sample weighted least squares equations: $\left(X^T W X\right) \hat{\mathrm{B}} = X^T W Y$ where $w_{ih} = \dfrac{1}{p_{ih}}$ are elements of the design sampling weights in W. The residual measure of the bias of $\hat{Y}_{igrs}$ is approximately equal to $X_i^T \mathrm{B}_i - Y_i$ where $\mathrm{B}_i = (X_i{}^T W X_i)^{-1} X_i{}^T W Y_i$ is the domain specific regression of coefficients over the universe which will give small bias relative to $Y_i = X_i^T \mathrm{B}_i$ when the β_j coefficients are obtained from areas in the universe similar to those of the sample area and large when β_j coefficients are from dissimilar areas. In order to deal with the possible large bias, the synthetic estimator is commonly subtracted from the bias giving a generalised regression estimator in the form $\hat{Y}_{igrs} = X_i^T \hat{\mathrm{B}} = \hat{\bar{Y}}_i + (\bar{X}_i - \hat{\bar{X}}_i)^T B$ where $(\hat{\bar{Y}}_i, \hat{\bar{X}}_i)$ are the Horvitz-Thompson estimators of $(\bar{Y}_i, \bar{X}_i)$. Even though the above estimation may perform well, the bias of the variance needs to be corrected by the factor $\dfrac{N_i}{w_{ih}}$ (Pfeffermann, 2010).

The above synthetic regression is also presented as an illustration of possible methods that have been applied to the measurement of unemployment in the literature.

2.2.3.5. Area level and unit level models

The area level and unit level models are models that take the area variability into account (Rao, 2003a) using the mixed model framework.

Mixed models are differentiated into two categories: (i) area level models where information on the response variable is available at the small area (municipal) level and (ii) the unit level models where information on the response variable is available at the unit (demographic profile within municipality) level.

The literature indicates that area level models have a wider scope than unit level models because area level auxiliary information is more readily available than unit level auxiliary data (Rao, 2003a, pp 163). However, this study uses the unit level model because almost all records in this study are exhaustively available from Census 2001, covering all required municipalities. There are no other auxiliary data to support the area level model at municipality level. Hence, this report uses the unit level model applied to the 2001 Census and LFS data because they have similar demographic characteristics at the unit level of data records.

2.2.3.5.1. Area level model

It is assumed there exists a suitable link function $Y_i = f(X_i) = x_i^{'}\beta + v_i$, *i=1,…, m* (2.13)

where β is the vector of regression parameters and the v_i's are independently and identically distributed (iid) normal variables with mean 0 and variance σ_v^2 (Rao, 2003a). From the previous relation, the estimate $\hat{Y}$ of Y is deduced from the model: $Y_i = f(X_i) + \varepsilon_i$, *i=1,…, m* (2.14)

where $\hat{Y}_i = f(\hat{X}_i)$ and the sampling errors $\varepsilon_i \sim N(O, \Psi_i)$ with known Ψ_i. The direct estimator $\hat{Y}_i$ is unbiased since $\mathrm{E}\left(\hat{Y}_i\right) = Y_i$.

The area level linear mixed model of Fay and Herriot (1979) is defined as a combination of the above expressions: $\underline{\hat{Y}_i = x_i^{'}\hat{\beta} + \hat{v}_{i\,i}}$, *i=1,…, m* and $\hat{Y}_i$~*iid* $N(x_i^{'}\beta, \sigma_v^2\Psi_i)$, *i=1,…, m.*

The general model would be $\underline{\hat{Y}_i = x_i^{'}\hat{\beta} + b_i\hat{v}_i}$ (2.15)

where b_i are known positive constants. This is a special case of a linear mixed model.

The area level model is also presented for illustration of possible methods that have been applied to the measurement of unemployment in the literature. In this study, there is no available reliable auxiliary data that can be used as a source of reference for the target small area domain of estimation (municipality).

2.2.3.5.2. Unit level model

Since the auxiliary data are available from Census 2001 for each element in each small area (almost all unspecified cases of the demographic variables in Census data have been imputed using dynamic or static imputation), we can assume that the variable y_{ih} (employment status) is related to x_{ih} (demographic variables) through a multivariate nested error regression model. The proposal is to build a unit level model with the link function between the Census unit records and the LFS records using similar covariates. The proposed unit level model formulated for each person-record within a specified area would be $y_{ihk} = \beta_0 + x_{ih1}^{'}\beta_1 + ... + x_{ihb}^{'}\beta_b + v_i + e_{ihk}$ (2.16) where $v_i \sim_{\text{iid}} N(O,\sigma_v^2)$ and $e_{ihk} \sim_{\text{iid}} N(O,\sigma_e^2)$ (Model A), where v_i are area specific effects and e_{ihk} are the random error vectors.

In this study, the proposed unit level model would be in two stages where the unemployment status is estimated in order to correct the misclassification, based on the demographic profile, after which the area level covariates (i.e. coordinates) are included to arrive at a two-level small area model (Rao, 2003b). The two-level model integrates the use of unit level and area level covariates into a single model: $y_{ihk} = \alpha_0 + x'_{ih1} z_{ih1} \alpha_1 + ... + x'_{ihb} z_{ihb} \alpha_b + x'_i v_i + e_{ihk}$. (Model B). (2.17)

This is an indication that part of the covariates have been modelled using β_j into α_j to arrive at two-level model (Moura and Holt, 1999).

2.3. Summary on methodology of Small Area Estimation

This chapter has defined the methodology that can be followed for determining small area estimates. In the context of this study, the chapter laid out the concepts and definition of small areas as well as concepts of unemployment and its derivatives with additional socio-economic definitions related to the population census and the LFS.

The methods of SAE that are considered for this study are defined based on the underlying statistical theory as referenced in the literature. The SAE methods include direct domain estimation through the survey estimation and population census counts. The direct methods are heavily dependent on the survey or census questionnaires applied in different households through direct interviews.

This chapter has also indicated the indirect methods that can be used: for instance, synthetic estimation can be considered assuming that the estimates from the small area (i.e. municipality) have the same characteristics as the large area (i.e. the province). It was explained that the synthetic estimates might be biased due to the assumption of homogeneity between large and small areas not being valid. Hence, sometimes, a composite estimate may be constructed as a weighted average of the direct estimate and the indirect estimate to balance the potential bias of the synthetic estimator.

Another indirect approach that has been described is symptomatic regression using the census data as auxiliary information to derive estimates using a regression model constructed from the survey direct estimates. Beyond the implicit models derived from the data, the illustration has shown that the area level models can be used. However, this study will use the unit level small area model to derive the estimates of unemployment at municipality level, as discussed above.

CHAPTER 3: LITERATURE SURVEY

3.1. Standard approaches to Small Area Estimation

"The problem of SAE is how to produce reliable estimates of characteristics of interest such as means, counts, quintiles, etc for small areas of domains for which only small samples or no samples are available." (Pfeffermann, 2010, page 2). When the sample data are available, the sample direct measurement takes into account weights or any other adjustment factors depending on the design. However, when there is no sample data available for a specific area, a statistical model is defined which borrows information from neighbouring areas. In most cases, the point estimator is given priority, but a related problem is the assessment of the estimation error. (Pfefferman, 2010)

A variety of methods have been documented in the literature (Satorra and Ventura, 2006). Broadly speaking, the SAE methods can be divided into "design-based" methods and "model-based" methods (Pfeffermann, 2010). Pfeffermann (2010, page 3) states that "Design-based methods often use a model for the construction of the estimators (known as 'model-assisted'), but the bias, variance and other properties of estimators are evaluated under a randomisation (design-based) distribution." of all the samples that can be selected within a fixed population. Pfeffermann (2010) also states that the model-based methods are conditioned on the selected sample and inference is with respect to the underlying model. "The common feature to design-based and model-based SAE is the use of covariate (auxiliary) information as obtained from surveys and/or administrative records, such as censuses or registers." (Pfeffermann, 2010, page 2). Some covariates are in the form of area means whilst other covariates are for every unit in the population.

As discussed in Chapter 2, the methods for SAE include direct, synthetic and indirect estimators. Direct estimators use only data from the small area being examined. Indirect, composite and model-based estimators are better since they are unbiased for the entire domain because they use observations from related variables or neighbouring areas. The composite estimator is a linear combination of direct and indirect estimators.

In the case of direct estimation, sample surveys are often used to estimate not only quantities related to the population (the domain) but also quantities for sub-domains. An estimator is said to be direct if it is based solely on the sub-sample for the sub-domain concerned.

In case of indirect estimation for a given small area, the indirect estimator borrows strength by using values of the variable of interest from related areas and/or times to increase the effective sample size. An implicit or explicit model is used to link the different areas and/or times, often through a source of auxiliary information, such as a census or an administrative register.

The traditional indirect estimator is based on an implicit model whereas the direct estimator is based on an explicit small area model (e.g. synthetic estimator, composite estimator). The inference from the model-based estimator uses the distribution implied by the assumed model, hence the model selection and its validation are critical in the estimation process. There are also aggregate- (or area-) level models that relate small area data summaries to area specific covariates, as well as unit-level models that relate the values of the units of the study variable to unit level (and in some models also area-level) covariates.

3.1.1 Design-based methods

Design-based methods assume a target population of fixed values (closed population). The only random variation present is due to the sampling scheme, meaning that a replication of the survey would yield a different sample (even though drawn from the same population), with the same division of the domain into areas (Satorra and Ventura, 2006).

3.1.2 Model-based approach: mixed effects regression

In the model-based approach, the data come from a realization of a stochastic process characterized by some parameters (means, ratios, …), whose values are the target of the analysis.

Small area models may be regarded as special cases of a general linear mixed model involving fixed and random effects (e.g., Prasad and Rao, 1990; Jiang and Lahiri, 2006). Means or totals for the areas can be expressed as linear combinations of fixed and random effects. Best linear unbiased predictors (BLUP) of such quantities can be obtained by minimizing the MSE in the class of linear unbiased estimators and do not require the assumption of normality of the random effects. An early application of this approach is by Ericksen (1973 and 1974), who uses the term criterion variable and symptomatic information for the dependent variable and the covariates, respectively.

3.2. Considerations for the practical application of small area methods

In practice, the area specific estimator for typical sample surveys has a small coefficient of variation (CV) for large areas, while the non-sampling errors (i.e. measurement and coverage errors) contribute much more than the sampling errors to total MSE as a measure of quality of estimator. It is in the small areas where the sample is not large enough to provide a direct estimator with acceptable quality in terms of MSE (or CV) that the area-specific estimator may not provide adequate precision. This is the reason why indirect estimators based on linking models are used (Rao, 2003b).

The use of indirect estimators borrows information from related areas through explicit (or implicit) linking models, using census and administrative data associated with the small areas. The following are the advantages of indirect estimators based on explicit linking models over the traditional indirect estimators based on implicit models:

(i) Explicit model-based methods make specific allowance for local variation through complex error structures in the model that links the small areas.
(ii) Models can be validated using the sample data.
(iii) Methods can handle complex cases such as cross-sectional and time series data, binary or count data, spatially correlated data and multivariate data.
(iv) Area-specific measures of variability associated with the estimates may be obtained, unlike overall measures commonly used with the traditional indirect estimator (Rao, 2003b).

Rao (2003b) covered several important new developments related to model-based SAE: basic SAE, basic area level model under Empirical Best Linear Unbiased Predictor (EBLUP), including unmatched sampling and linking area level models, use of sampling weights in unit level models (pseudo-EBLUP method for the basic unit level model), jack-knife methods for MSE estimation, and MSE estimation for area level models when the sampling variances are estimated. However, the following are practical issues related to SAE:

3.2.1 Design issues

Rao (2003b) presented practical design issues that have an impact on SAE. He suggested that a proper resolution of the design issue could lead to the enhancement of the reliability of direct as well as indirect estimates for both planned and unplanned domains (areas). He gave certain guidelines that might be useful in minimising the need for indirect estimators namely: using list frames such as address registers, as a replacement of some incomplete or non representative clusters; using many small strata from which samples are drawn; using compromise sample allocations to satisfy reliability requirements at a small area level as well as large area level; integrating various surveys by harmonising questions across surveys of the same population and using multiple frame surveys as well as using "rolling samples" as a method of cumulating data over time.

3.2.2 Model selection and validation

Even though the methodological developments and applications of model-based estimation are remarkable, caveats should be included with respect to the model assumptions. More effort should be given to the compilation of the auxiliary variables and predictors of study variables in determining the suitable linking models (Rao, 2003b).

Model diagnostics are always used to find suitable model(s) that fit the data. They include residual analysis to detect departures from assumed models, selection of auxiliary variables and case-deletion diagnostics to detect influential observations (Rao, 2003b).

3.2.3 Area level vs. unit level models

As stated in section 2.2.3.5, area level models are more widely used than unit level models in view of the availability of area level auxiliary information, which may give a consistent model design-weighted direct estimator. Good approximations to the sampling variances as well as methods that incorporate the variability associated with estimated sampling variances in MSE estimation need to be calculated (Rao, 2003b).

3.2.4 Non-sampling errors

It is important to notice that non-sampling errors can have a substantial effect on SAE. Hence, suitable designs must be developed as well as methods of estimation that take into account non-sampling errors. Measurement errors in the responses can lead to biased estimates. In the context of direct estimation, Fuller (1995) proposed methods at the design stage that can lead to bias-adjusted estimators.

3.3. Other considerations related to model specification

This discussion is pooled from several contributions (Longford, 1999 and 2004) on the methodology for SAE. Following Fay and Herriot (1979) and Battese, Harter and Fuller (1988), random coefficient (two-level) models have become the models of choice for SAE. The general method follows the widely accepted approach of finding a suitable model for the data and then using it for all subsequent inferences. Draper (1995) and Chatfield (1995) pointed out that model uncertainty makes a substantial contribution to the MSE of an estimator when the data is used for model selection.

Longford (1999) introduced multivariate shrinkage to estimate local-area rates of unemployment and economic inactivity using the United Kingdom (UK) LFS. The method exploits the similarity in the rates of claiming unemployment benefits and the unemployment rates as defined by the ILO. This is done without any distributional assumptions, relying only on the high correlation between the two rates. The estimation is integrated with a multiple-imputation procedure for missing employment status of the subjects in the data base (item non-response).

There is also a review of the assumption of associating small areas with random effects (Longford, 2005a). In the design-based perspective, these 'effects' are fixed, because a hypothetical replication of the survey would yield exactly the same population, with the same division to areas; the sampling process is the only source of variation. The problematic nature of the assumption of random effects can be demonstrated when estimators are applied to the samples drawn from an artificial population. It shows that the estimation for the areas that have means close to the national mean is more precise and for those that have means distant from the national mean is less precise than that indicated by the random-effects model (Rao, 2003b).

Longford (2005b) states that the following are the conditions for an auxiliary variable to be useful: high correlation of its area-level population means with the targets, small sampling variance, and small correlation with the population means of the other auxiliary variables. This differs from the conditions for a good model fit (reduction of the variance components in the random-effects model).

Although most large-scale (national) surveys have a wide inferential spread, their design does not take into consideration the possible use of SAE models. In order to take this into account the general approach is to design the sample survey to allow inferences about the small areas (sub-domains). The design-based perspective yields the 'correct result' if the assumption made for the random coefficient model is appropriate. The estimation approach allows borrowing strength and avoids the inability of design-based methods to benefit from the auxiliary information (Longford, 2007).

The literature review suggests that, depending on the available information, there are many approaches that can be followed in SAE, such as methods without any covariates and methods that make use of auxiliary variables (covariates). Extensive research has focused on design issues and on comparing the efficiency of alternative small area estimators. Covariates can be handled by mixed-effects regression. The limitation of the approach to SAE in official statistics is that the choice of the covariates introduces some uncertainty into the data, different to the culture of official statistics. Further, there is the assumption of the mixed-effects regression stipulating that covariates do not have measurement error (Longford 2003 and 2005a).

3.4. Practical examples of the application of Small Area Estimation methods

There are practical references in the literature where SAE techniques have been applied in the estimation of unemployment around the world. Differences between census-based estimates of unemployment and survey-based estimates of unemployment have been experienced in other places such as the UK, where the model-based estimates derived from the LFS (applied to the 1991 census) showed significant differences from 2001 census-based population estimates.

3.4.1 Canadian experience

The Canadian experience used a number of primary sources of data: the Canadian census of population and housing, the Canadian LFS and the Federal Government Unemployment Insurance (UI). Other small area labour market data were used as auxiliary information. Since the main goal was to minimise the model estimation error, they tested three estimation techniques: synthetic estimation, SPREE and regression estimation. The Canadian experiment showed that there are highly correlated variables between data sources, despite basic differences in definition; and this led to the modelling techniques such as SPREE and regression having low model errors. The measure of model error used was the percent Absolute Relative Difference (ARD) calculated as the relative difference between the model-based estimate and the count of unemployed as estimated in the 1981 Census at the Census Division level. The SPREE method had lower model error than the synthetic method since it had the advantage of additional cross-classification structures. The SPREE and regression models gave similar results with small relative errors for small Census Divisions (Earwaker, Brien, Gosselin, 1984).

3.4.2 Turkish experience

The Turkish experience considered the revision of concepts for international conformity. The Household Labour Force Survey (HLFS) was based on a multistage stratified sampling design consisting of two sampling sets: a sample for periodic applications and a sample for regional estimation purposes. The sub-region estimates required the use of SAE methods. It was critical to determine the most effective auxiliary information variables for stratification and estimation using several data sources such as HLFS, administrative registers, population census, extended population surveys, Household Income and Consumption Expenditure Surveys and other auxiliary small area data sets. The Turkish statistical office experimented with direct estimation (expansion estimator and ratio estimator) and model-based estimation (regression estimator and synthetic estimator). It was found that the regression synthetic estimator within post-strata was reliable for some domains whenever the linear association measure between Y and its covariate X was high enough. At the same time, the bivariate ratio synthetic estimator gave good results in other areas if a second variate could be found which is highly explanatory of the variability in Y at time of the post-stratification stage. The problem was observed for some sample dependent estimators in composite estimation around the linear combination of coefficients between several small areas for cross-class groups (Gebislioolu *et al*, 1996).

3.4.3 United Kingdom: Office for National Statistics (ONS)

The ONS in the UK, in conjunction with Southampton University, conducted research into the estimation of unemployment based on the ILO definition. A number of different approaches were explored utilising the high correlation between the LFS estimates of unemployment and the number of job seeker allowance benefit claimants. The main approach was to develop regression models linking the unemployment estimates to the claimant count information by local authority district (ONS, 2001b). A validation strategy was developed: internal validation using model diagnostics and external validation with other sources of data for small areas and with independent experts. As the SAE in the UK is an ongoing project, there is considerable future work to be covered such as multivariate estimation, time series methods, consistency between direct and model estimates, and using spatial information (ONS, 2001b). In South Africa there are not as many auxiliary sources of administrative data at local area level, and there are no records concerning job seekers' allowance benefit claimants per local authority district and unitary authority.

3.4.4 Practical example from the United States of America (USA)

The application of small area techniques was done in Ohio to uninsured-estimates by county (Suciu, Hoshaw-Woodard, Elliott, and Doss, 2001). The authors presented a review of options and issues, looking at issues such as the accuracy of and difficulty in deriving confidence intervals. The focus was on prediction models or indirect estimators grouped into three types: domain indirect, time indirect and domain/time indirect. The domain indirect estimators were concerned with auxiliary criterion measures from different domains but at the same time period (i.e., borrowing strength from neighboring counties using data collected in the same survey). The time indirect estimators referred to auxiliary criterion measures from the same domain but from different time periods (e.g. using county data collected from annual surveys). The domain/time indirect estimator used a combination of both above described indirect estimators. Each type was classified into classes of estimators such as synthetic estimators, regression models, base unit estimators, composite small area estimators and Bayesian estimators. All of these small area estimators were applied to the data concerning the proportion of uninsured residents in different counties in Ohio, borrowing strength from surrounding small areas in the estimation. Even though all of the methods presented were applied to estimate county-specific uninsured rates using data from the Ohio Family Health Survey (OFHS), point estimates could only be obtained for the number and proportion of uninsured adults and children in Ohio. Only the Hierarchical Bayesian Model approach was able to easily predict intervals for the point estimates as measure of variability.

3.4.5 Italian experience

The most relevant experience for this study was done in Italy, where both the issues of misclassification of employment status as well as that of sub-regional level estimates were presented. Although there are international guidelines to classify individuals into the three labour force states: employment, unemployment, and inactivity, the resulting statistics are known to be sensitive to slight operational variations of the definitions (Battistin, Restore and Trivellato, 2007). Comparing the definition followed by EuroStat and the definition followed by the Italian Statistical Office, there were considerable differences particularly at the boundary between unemployment and inactivity. The results show that individuals who are not at work, reporting that they are seeking work and are immediately available for work but with no recent steps undertaken, are similar to individuals who are not in the labour force. EuroStat classified those individuals as inactive. The result was robust to changes in the business cycle, to geographical area effects and to different levels of participation of married women in the labour market, as well as to changes in the survey questionnaire. Even though the additional variables omitted from the criterion to establish similarities across groups may have not been robust, those variables were considered as sufficiently important in the labour force classification to be useful (Battistin, Restore and Trivellato, 2007).

The other relevant experience from Italy was an attempt to improve the performance of the estimation of the unemployment rate at sub-regional level, from the Italy Statistics (ISTAT) LFS, with reference to domains cutting across survey strata that can be aggregated to municipalities (D'Al'O, di Consiglio, Falorsi, Pratesi,. Ranalli, Salvati, Solari, 2008). The unemployment rate was estimated using a linear mixed model with spatially correlated area effects and covariates such as sex, age and area level unemployment from the previous census. They investigated the use of non-parametric additive models to prevent misspecifications of the functional relationship between the target variable and some covariates. They also used model-based direct estimation as well as SAE based on three logistic models (no interaction between identity variables, with interaction between identity variables and with an uncorrelated random area effect). The different methods investigated were put through simulated experiments using the 2001 Census data. The main finding was that the spatial structure of the data helps to improve the accuracy of the estimates as measured by the empirical MSE for all the estimators considered.

3.4.6 An example applied in the Philippines

The existing methodology on SAE has been applied to unemployment rates for the Philippines (Francisco, 2003). The report gives an overview of the following existing small area techniques: direct, indirect and model-based. Illustrative examples with comparisons in terms of the efficiency of small area estimates are given for each of the estimation techniques. This includes the use of case studies to illustrate the potential use of the empirical Bayes (EB) technique in estimating the unemployment rates at regional and age group level in the Philippines. Since the direct estimates were unreliable for the smaller regions, the use of indirect model-dependent estimates was advantageous in giving EB estimates of the unemployment rate that were more efficient than the direct estimates. The distribution of the EB estimates was not significantly different from that of the direct estimators.

3.4.7 SAE estimation of unemployment in Iran

A synthetic estimator, a composite estimator and the EB estimator were compared in the context of estimating unemployment rates for the provinces of Iran (Khoshgooyanfard and Monazzah, 2006). The three estimators were compared using their MSEs based on the complete data (enumeration) for 1996. A cost-effective strategy was proposed to estimate the inter-censal unemployment rate at the provincial level. The findings indicated that the composite and EB estimator performed well and similarly to one another. The study assumed that the true values are known from a Census. The strategy for small area estimates was to allocate the sample, at time of sample design, into all small areas or domain of interest. The synthetic estimator used only one variable (age) for partitioning. The EB model was improved on by using supplementary variables related to economic activity (the response variable) in the model.

3.5. Summary of the literature survey

This chapter confirms that the use of SAE has increased over the years, since the methods are broadly applied. The different methods discussed in the literature, are based on the need to alleviate some limitations of the available information. For instance, the deficiencies in official statistics shows that administrative records are not often available both at small and large area levels. Surveys may yield satisfactory estimates for large areas but not for small areas. Periodic censuses do not meet all demands for effective policies and planning. The importance of SAE is evident in providing estimates to be used, for example, for fund allocation in needy areas such as education and health programmes. The standard approaches of SAE presented were based on their separation into design-based methods and model-based methods. Considerations for the practical application of SAE were presented looking at design issues, model selection and validation, comparison of area level models and unit level models as well as the impact of non-sampling error on the SAE.

This chapter took into consideration the issues related to measures of precision and the handling of item non-response using a multiple-imputation procedure. The requirements of applying the SAE methods successfully are the high correlation of auxiliary variables with the target means, small sampling variances and low correlations between auxiliary variables. There are proposals in the literature regarding the design of the sample survey taking into account the targeted small areas to allow inference in those sub-domains (small areas). What is important in the estimation approach is to borrow strength to allow the design-based methods to benefit from the auxiliary information. The limitation of the approach to SAE in official statistics is that the choice of the covariates introduces some uncertainty into the data, contrary to the culture of official statistics.

This chapter has also given illustrations of the experiences in different countries. The focus has been around the estimation of unemployment related estimates. The Canadian analysis combined the unemployment status data from the Census, the LFS and the Federal Government records of UI. The Turkish example used two samples from the HLFS: a sample for periodic applications and a sample for regional estimation purpose. The ONS in the UK used the LFS estimates of unemployment and the records of job seekers allowance benefit claimants (claimant count). The example from the USA estimated the proportion of uninsured residents in different counties in Ohio using data from the OFHS. The experience in Italy of the misclassification of economic activities is relevant to the situation in South Africa, where the discouraged worker seekers are sometimes considered as not economically active. The example used in the Philippines gave a comparison of SAE methods, looking at their efficiency in estimating the unemployment rate. Similarly, the Iranian team compared the SAE methods, looking at improved estimation of unemployment rates.

CHAPTER 4: APPLICATION OF SMALL AREA ESTIMATION TO EMPLOYMENT STATUS

4.1. Introduction

This chapter will first outline the methodology that will be applied using the available data as explained in Section 4.2. The approach will be to compare two methods falling under the implicit approach (i.e. direct and synthetic estimation) in Section 4.3, and the model-based approach (which is a combination of discriminant analysis and multinomial logistic regression) in Section 4.4. Section 4.5 will give an evaluation of the results while section 4.6 will summarise the chapter.

Due to the need for of using different data sources, different computer packages were used in this study namely: PAWS statistics 18 (formally known as SPSS) for most of the SAE analysis, SAS 9.1 for the LFS data preparation, SuperCross 4.5.5 for Census tabulation, ArcGIS 9.3 for the extraction of x,y coordinates and Microsoft office suite 2003 for the report writing.

4.2. Available sources of information

4.2.1. Labour Force Survey

By default, the best available estimate of unemployment in South Africa is the LFS. Stats SA has made the LFS an established official statistical series since 2000. There have been other surveys having the labour force questions as added modules such as the October Household Survey, targeting households, and the SEE targeting employers. As stated before, the choice of data is limited to the LFS of September 2001 to avoid introducing time fluctuation bias with the linking data set. In addition, the September 2001 LFS was reweighted to 2001 Census frame rather the original 1996 Census frame.

4.2.2. Population Census

The most recent Census data is the 2001 Population census. The choice of this data set is driven by the fact that it has coverage of all municipalities as demarcated. The problem with the 2001 Census is its large undercount. However, the effect of the undercount might be considered minimal when deriving the estimates of unemployment at municipal level because the expected estimate is a ratio (unemployment rate). The other problem is that the census data do not give a good measure of unemployment.

4.2.3. Geographic information database

It is important to acknowledge that the 2001 Census is based on the full demarcation of the EAs. These EAs are linked in a hierarchical structure to all old municipalities as well as the new municipalities, and to the provinces. The 2001 EAs allow a spatial link with the required small areas of analysis. They are also linked to the EAs used in the 1996 census, which constitute the base frame of selection of PSUs for the LFS.

4.3. Estimation based on the implicit approach

4.3.1. Direct estimation

4.3.1.1. Results from the Labour Force Survey (September 2001)

Table 4.3.1.1 gives the results of the direct estimation of the unemployment rate per province, using calibration weights obtained by using the integrated weighting method to adjust the estimates from the LFS of September 2001 to the Census 2001 population as benchmark. The universe is all people aged 15 years to 65 years old considered as economically active. Table 4.3.1.1 below shows that there are around 12.3 million persons aged 15-65 years old who are excluded from the calculation.

Table 4.3.1.1: People aged 15-65 from 2001 September LFS by Province and economic status

Province (Domain of estimation) (a)	Total persons aged 15-65 (b)=(c)+(d)	Persons not economically active (c)	Persons economically active (d)=(e)+(f)	Employed (e)	Unemployed (f)	**Unemployment rate** (g) =100*(f)/(d)
Western Cape	2,864,663	964,266	1,900,397	1,563,949	336,448	**17.7**
Eastern Cape	3,870,039	2,070,034	1,800,005	1,235,585	564,420	**31.36**
Northern Cape	561,113	235,549	325,564	244,200	81,364	**24.99**
Free State	1,945,712	746,886	1,198,826	875,030	323,796	**27.01**
KwaZulu-Natal	5,786,792	2,757,777	3,029,014	2,005,835	1,023,179	**33.78**
North-West	2,333,903	1,132,675	1,201,228	858,062	343,166	**28.57**
Gauteng	5,861,155	1,801,903	4,059,252	2,824,531	1,234,721	**30.42**
Mpumalanga	1,870,521	844,042	1,026,478	726,888	299,590	**29.19**
Limpopo	3,023,464	1,728,310	1,295,154	846,544	448,610	**34.64**
Total	**28,117,361**	**12,281,442**	**15,835,919**	**11,180,625**	**4,655,294**	**29.4**

The target population for analysis is the employed (e) and unemployed (f) which are used to deduce the unemployment rate. The LFS of September 2001 reported 4.7 million persons as unemployed whereas only 11.2 million were employed, which resulted in an unemployment rate of 29.4. The next reliable domain of estimation (based on the design of the sample) is the breakdown at provincial level and further into urban and non-urban areas.

4.3.1.2. Results from Population Census (October 2001)

Using responses from Census 2001, persons between 15 years and 65 years old reported their employment status using a reduced version of the LFS questions. Table 4.3.1.2 shows that the target population of those 15 - 65 years old in the October 2001 Census is the almost same (28.4 versus 28.1 million) observed in the September 2001 LFS.

Table 4.3.1.2: People aged 15-65 from 2001 Census by Province and economic status

Province (a)	Total persons aged 15-65 (b)=(c)+(d)	Persons not economically active (c)	Persons economically active (d)=(e)+(f)	Employed (e)	Unemployed (f)	**Unemployment rate** (g) =100*(f)/(d)
Western Cape	3,074,287	1,057,570	2,016,717	1,489,722	526,995	**26.13**
Eastern Cape	3,697,692	2,035,372	1,662,320	754,338	907,982	**54.62**
Northern Cape	529,842	216,519	313,323	208,745	104,578	**33.38**
Free State	1,752,695	715,422	1,037,273	591,002	446,271	**43.02**
KwaZulu-Natal	5,755,280	2,629,798	3,125,482	1,602,270	1,523,212	**48.74**
North West	2,352,701	1,020,777	1,331,924	748,889	583,035	**43.77**
Gauteng	6,432,054	1,877,665	4,554,389	2,894,777	1,659,612	**36.44**
Mpumalanga	1,908,020	838,420	1,069,600	630,175	439,425	**41.08**
Limpopo	2,924,559	1,627,751	1,296,808	663,847	632,961	**48.81**
Total	**28,427,126**	**12,019,294**	**16,407,832**	**9,583,762**	**6,824,070**	**41.59**

The difference between October 2001 Census and September 2001 LFS is noticeable for those who are reported as employed and unemployed. Since Census 2001 used less probing questions, the Census estimates of employed and unemployed are not considered reliable. If the key variable of employed and unemployed cannot be measured at provincial level, it would not be advisable to continue to use direct counts from the Population Census at lower levels.

The difference between the measures by the LFS in Table 4.3.1.1 and the Census in Table 4.3.1.2 is due to the misclassification of people between employed, unemployed and not economically active. However, the overall target population (economically active) seems to be measured similarly by the LFS and Census.

4.3.1.3. The problem of misclassification

The measurement of labour force employment status relies on the conventional definitions of employed and unemployed. This is based on the ILO guidelines that have become the international standard, even though they may be adapted for operational rules in different countries (Battistin, Restore and Trivellato, 2007). According to general ILO guidelines, a person who is above a specific age (usually 15 years old) is classified as:

a) *Employed*, if during the *reference period* she or he worked at least *a little* (or was not at work for any reason, but will return to a job that she or he has an attachment to);

b) *Unemployed,* if

   i. During the reference period she or he did not work at all,

   ii. She or he is looking for a job and recently took the specific steps to seek work and

   iii. She or he is immediately available to work, and

c) *Inactive*, (i.e. out of labour force).

For the Stats SA LFS, the above definition has been adapted as follows:

- The *employed* are those who performed work for pay, profit or family gain in the seven days prior to the survey interview, or who were absent from work during these seven days, but did have some form of work to which to return.
- The *unemployed* persons are those persons who:

  (a) did not work during the seven days prior to the interview,

  (b) wanted to work and were available to start work within two weeks of the interview, and

  (c) had taken active steps to look for work or to start some form of self-employment in the four weeks prior to the interview.
- The persons who are *not economically active* are those who are not available for work (i.e. full-time scholars, students, full-time homemakers, retired, and unable or unwilling to work).

The questions used in the LFS of September 2001 and Census 2001 are given in Annexure II and III. As can be seen, the 2001 Census has a reduced number of questions on labour force status. This has resulted in shifting some people who did not take steps to look for work while they consider themselves unavailable to have been classified as unemployed instead of not economically active. The order of those questions and the lack of skips on the Census questionnaire between active steps (yes or no) and availability (period) for work, resulted in including in the unemployed group some of discouraged unemployed (active) while they are actually no longer interested in looking for work (inactive).

The labour force questions in the Census questionnaire are also affected by the interpretation of the terms such as *reference period*, *a little*, r*ecently*, *specific steps* and *immediately available* (Battistin, Restore and Trivellato, 2007), required to get the correct classification of labour force status.

4.3.2. Synthetic estimation

The need for the realignment of Census data using the LFS can be done using synthetic estimation, based on the profile of the observed population in the LFS being applied to the Census data. The most reliable profile is given by the LFS of September 2001 with good representation by province and urban and non-urban. Each table cell value is calculated as

$$\hat{Y}_{ih} = p_{ih} Y_c \tag{4.1}$$

where $p_{ih} = p_{s_h} p_{c_{ih}} = \frac{N_{s_h}}{N_s} \frac{N_{c_{ih}}}{N_{c_i}}$ represents the proportion of subgroup *h* in the small area *i* as product of the proportion p_{sh} in the LFS at provincial/settlement type level and the proportion p_{cih} in the Census. Y_c is the Census estimate of the target universe (15 - 65 years). The approach uses the profile derived from the LFS as a product with the profile observed in the Census for subgroup *h* in the small area *i*. The profile derived from the survey is given in Table 4.3.2.1 below.

Table 4.3.2.1: Proportion (in percentage) of persons aged 15-65 in LFS 2001 by Province and economic status

Province	Urban Employed	Urban Unemployed	Urban Not economically active	Non-urban Employed	Non-urban Unemployed	Non-urban Not economically active	Total
Western Cape	4.77	1.13	3.09	0.79	0.06	0.34	10.19
Eastern Cape	2.12	1.03	2.15	2.27	0.98	5.21	13.76
Northern Cape	0.46	0.25	0.61	0.40	0.04	0.23	2.00
Free State	1.95	0.83	1.90	1.16	0.32	0.76	6.92
Kwa-Zulu Natal	4.83	1.97	3.97	2.31	1.67	5.84	20.58
North-West	1.22	0.50	1.28	1.84	0.72	2.75	8.30
Gauteng	9.66	4.27	6.24	0.38	0.12	0.17	20.85
Mpumalanga	1.12	0.52	1.02	1.47	0.55	1.99	6.65
Limpopo	0.62	0.19	0.50	2.39	1.41	5.65	10.75
Total	26.76	10.67	20.76	13.01	5.88	22.92	100.00

The estimate of the unemployment rate for each municipality is calculated as:

$$r(Y_i) = \frac{\hat{Y}_{2i}}{\hat{Y}_{1i} + \hat{Y}_{2i}} * 100 \tag{4.2}$$

where $\hat{Y}_i = \left(\hat{Y}_{1i}, \hat{Y}_{2i}, \hat{Y}_{3i}\right)$ is the set of inputs into the employment measures, namely the number of persons employed, unemployed and not economically active, respectively. The resulting unemployment rate presents exactly the same profile as reflected in Table 4.3.1.1. at national and provincial level. However, the above approach cannot be used to reflect measurement of unemployment rates at municipal level because the LFS does not have a similar profile on municipality level.

Table 4.3.2.2: Distribution of municipalities by class of unemployment rate

Class of unemployment rate in percentages	Number of municipalities as a result of Synthetic estimation of the unemployment rate	Number of municipalities as a result of direct measurement of the unemployment rate from Census	Percentage of municipalities as a result of Synthetic estimation of the unemployment rate	Percentage of municipalities as a result of direct measuresment of the unemployment rate from Census
0 – 10	23	12	8.78	4.58
10 – 20	52	23	19.85	8.78
20 – 30	79	30	30.15	11.45
30 – 40	61	53	23.28	20.23
40 – 50	27	61	10.31	23.28
50 and plus	15	78	5.73	29.77
No rate calculated	5	5	1.91	1.91
Total	**262**	**262**	**100.00**	**100.00**

Table 4.3.2.2 shows that the use of synthetic estimation has shifted the distribution of the unemployment rate per municipality towards 20% or 30% as opposed to the Census erroneous measurement, where the unemployment rate is above 30% in the majority of cases. It must be stressed that the results for the synthetic estimation are given for illustration purposes only since they are based on the assumption of equal profiles across the small areas (municipalities), which is unrealistic.

4.4. Model-based estimation

4.4.1. Model formulation

The model-based estimation will use a model constructed from two estimation stages: discriminant analysis and multinomial logistic regression. The first stage aims at reducing the misclassification of discouraged work seekers from unemployed to not economically active. The second stage aims at using the outcome from the first stage, in combination with the geographic variables and unit level characteristics, in order to spread the survey estimates of the labour force across the small areas of estimation. The reason of a two stage approach is to resolve two problems: (1) the correction of misclassification and (2) the inclusion of the spatial spread of the estimation. This kind of estimation is referred to in the literature as a two-level small area model (defined as Model B in Section 2.2.3.5.2.). The theoretical definition of the model B to be used in this study is $\hat{\theta}_{ji} = g(X_i, f(X_i), Z_j)$ where (4.3)

$\hat{\theta}_{ji}$ represents the ultimate expected estimate of the classification of employment status related to the i^{th} person in the j^{th} domain of estimation, after taking the geographic information into account

$g(.)$ represents the multinomial logistic regression function

$f(.)$ represents the discriminant function

X_i represents demographic covariates used in both Census and LFS for each i^{th} person

Z_j represents the spatial coordinates related to the j^{th} domain of estimation

$f(X_i) = Y_i$ represents the estimate obtained by applying the discriminant function estimated from the demographic variables and employment status from the LFS, to the demographic variables related to the i^{th} person in the census.

Based on the above definition, the employment rate is calculated at each stage, using the standard definition of unemployment based on the employment status. More specifically, the calculation of employment rate is performed based on the estimated counts of employed people and unemployed people in each small area.

For the i^{th} person in the j^{th} small area or domain of estimation, the original employment status X_{ji} is used as one of the demographic covariates, and is decomposed into X_{1ji}, X_{2ji} and X_{3ji} representing respectively the employed; unemployed and not economically active. Similarly, the results of the discriminant analysis for employment status Y_{ji} can be structured as Y_{1ji}, Y_{2ji} and Y_{3ji} representing respectively predicted employed; predicted unemployed and predicted not economically active. Finally, the true employment status θ_{ji} would be structured as θ_{1ji}, θ_{2ji} and θ_{3ji} representing, respectively final predicted employed; final predicted unemployed and final predicted not economically active.

The Census unemployment rate within the j^{th} domain of estimation is

$$r(X_{2j}) = \frac{\sum_i w_i X_{2ji}}{\left(\sum_i w_i X_{1ji}\right) + \left(\sum_i w_i X_{2ji}\right)} = \frac{\sum_i w_i X_{2ji}}{\left(\sum_i w_i X_{.ji}\right) - \left(\sum_i w_i X_{3ji}\right)} \tag{4.4}$$

with w_i representing the adjustment factors in the Census or the sample weights in LFS for each i^{th} person.

Subsequently, the unemployment rate within j^{th} domain of estimation is

$$r(Y_{2j}) = \frac{\sum_i w_i Y_{2ji}}{\left(\sum_i w_i Y_{1ji}\right) + \left(\sum_i w_i Y_{2ji}\right)} = \frac{\sum_i w_i Y_{2ji}}{\left(\sum_i w_i Y_{\cdot ji}\right) - \left(\sum_i w_i Y_{3ji}\right)} \tag{4.5}$$

derived from the discriminant analysis.

Therefore, the SAE of unemployment rate within the j^{th} domain of estimation is

$$r(\hat{\theta}_{2j}) = \frac{\sum_i w_i \hat{\theta}_{2ji}}{\left(\sum_i w_i \hat{\theta}_{1ji}\right) + \left(\sum_i w_i \hat{\theta}_{2j_i}\right)} = \frac{\sum_i w_i \hat{\theta}_{2ji}}{\left(\sum_i w_i \hat{\theta}_{.ji}\right) - \left(\sum_i w_i \hat{\theta}_{3j_i}\right)} \quad (4.6)$$

where the $\hat{\theta}$'s are the estimates from the multinomial logistic regression.

In order to achieve the estimation procedure, the preparation of the data requires the merging of Census data and LFS data linking similar variables in both data sets.

4.4.2. Discriminant Analysis model

The goal of discriminant analysis is to classify cases into one of several mutually exclusive groups based on their values for a set of predictor variables. This uses a classification rule developed using cases for which group membership is known (the LFS), using demographic variables common to both data sets.

This rule is then applied to the census data to obtain a new reclassification of the status of each individual in Census 2001 into employed, unemployed and not economically active. However, the calculation of unemployment rate is obtained using only the employed and unemployed (excluding the not economically active) (See Section 4.4.1.)

The set of independent variables that are investigated are gender, age group, population group, education level, income and marital status. The choice of variables is based on the fact that they are in both data sources (Census and LFS), they have little correlation between themselves and they have a known impact on employment status.

Linear discriminant analysis will be used on the LFS data to create discriminant functions. The number of functions depends on the numbers of groups. In this case of employment status, every case is expected to belong to one of the 3 predicted groups (predicted employed; predicted unemployed and predicted not economically active) which correspond to 2 functions. The discriminant functions are defined as:

$$d_{ik} = b_{0k} + b_{1k}x_{i1} + \ldots + b_{pk}x_{ip} \quad (4.7)$$

where d_{ik} is the value of the kth discriminant function for the ith case;

- p is the number of predictors
- b_{jk} is the value of jth coefficient of the kth function

- x_{ij} is the class of the j^{th} demographic variable for the i^{th} case.

The discriminant analysis is done under the assumptions of homoscedasticity, which means that there is an assumption of the same correlation structure between the predictors across groups. This is reasonable as the covariates are not highly correlated with each other.

The covariates have been recoded into a set of indicator variables representing whether the person has the characteristic (1) or not (0), which gives a standard measure of comparison between the different variables on a scale between 0 and 1. The number of indicator variables per covariate is the number of categories minus 1. The recoding of those covariates into indicator variables for each unit in the population was done as follows:

- Gender: 1 for male and 0 otherwise; implying that female is the base category
- Age_{ij} representing the age of unit *i* in the small area domain *j* was grouped into class of age as 15 – 24, 25-34 and 35-44: 1 if the individual belong to the specified age class and 0 otherwise; implying that 45-64 is the base category
- Marital status: 1 as ever married and 0 otherwise;
- Settlement type: 1 for urban settlement and 0 otherwise;
- Population group: Black, Coloured and Asian Indian; implying that the base category is White.
- Education level: 1 for matric or above and 0 otherwise;

Table 4.4.2.1 gives the proportions of the dichotomised variables in the various categories, and their standard deviations (applying the LFS weights and the Census adjustment factors of undercount). The proportions of the indicator variables in the LFS show almost the same profile as in Census in most of the cases with the exception of employment status. The group statistic tables (Table 4.4.2.1 and Table 4.4.2.2) give an indication of variables that may be important in differentiating groups via the group proportions and standard deviations.

Table 4.4.2.1: Group statistics: LFS and Census

Indicator variables	LFS		Census	
	Proportion	Std. Deviation	Proportion	Std. Deviation
Economically Active population	0.56	0.4960	0.58	0.4938
Employed	0.40	0.4894	0.34	0.4728
Unemployed	0.17	0.3717	0.24	0.4276
Urban settlement type	0.58	0.4933	0.62	0.4847
Male	0.48	0.4998	0.47	0.4993
Ever Married	0.47	0.4994	0.52	0.4997
Black	0.76	0.4246	0.77	0.4191
Coloured	0.09	0.2863	0.09	0.2889
Asian Indian	0.03	0.1700	0.03	0.1663
Matric and above education level	0.36	0.4795	0.29	0.4548
AGE1524	0.33	0.4699	0.33	0.4688
AGE2534	0.26	0.4406	0.26	0.4362
AGE3544	0.19	0.3875	0.19	0.4002

Table 4.4.2.2 shows the estimated proportions for each of the demographic variables, in each of the employment categories in the LFS,. This provides an indication of differentiation of the employment status in the various demographic groups. The proportions employed are highest for age group 35 to 44, male, ever married, those living in urban settlements, Coloured, Asian or Indian, and those with Matric and above. The proportions unemployed are highest for age group 25-34 and blacks The proportion of inactive persons is highest for those 15-24 years old..

Table 4.4.2.2.: LFS Group statistics by employment status

	1=Employed		2=Unemployed		3=Inactive	
Variable of analysis	Proportion	Std. Deviation	Proportion	Std. Deviation	Proportion	Std. Deviation
AGE1524	0.12	0.3222	0.32	0.4680	0.52	0.4994
AGE2534	0.33	0.4690	0.41	0.4919	0.15	0.3576
AGE3544	0.29	0.4514	0.17	0.3746	0.10	0.2970
Male	0.58	0.4942	0.48	0.4996	0.40	0.4903
Ever Married	0.67	0.4710	0.32	0.4670	0.36	0.4793
Urban	0.67	0.4692	0.64	0.4786	0.48	0.4994
Black	0.66	0.4745	0.88	0.3287	0.82	0.3857
Coloured	0.11	0.3183	0.07	0.2614	0.07	0.2620
Asian Indian	0.04	0.1920	0.02	0.1445	0.03	0.1568
Matric and above education level	0.46	0.4982	0.43	0.4947	0.24	0.4280

The attempt to separate the classification using discriminating variables is supported by the low inter-correlations between variables of analysis as shown by the pooled within groups matrix (Table 4.4.2.3). This is in line with the assumption of discriminant analysis of having independent variables not to be highly correlated.

Table 4.4.2.3: Correlations pooled within groups matrices

	AGE1524	AGE2534	AGE3544	Male	Ever Married	Urban	Black	Coloured	Asian Indian	Matric and above education level
AGE1524	1.00	-0.39	-0.27	0.10	-0.52	-0.03	0.04	0.00	-0.01	0.05
AGE2534	-0.39	1.00	-0.34	-0.03	-0.08	0.02	0.02	-0.01	-0.01	0.17
AGE3544	-0.27	-0.34	1.00	-0.04	0.21	0.00	0.02	0.00	-0.01	-0.07
Male	0.10	-0.03	-0.04	1.00	-0.13	0.00	0.02	-0.02	0.00	-0.02
Ever Married	-0.52	-0.08	0.21	-0.13	1.00	0.01	-0.15	0.04	0.05	-0.08
Urban	-0.03	0.02	0.00	0.00	0.01	1.00	-0.33	0.14	0.14	0.22
Black	0.04	0.02	0.02	0.02	-0.15	-0.33	1.00	-0.57	-0.31	-0.23
Coloured	0.00	-0.01	0.00	-0.02	0.04	0.14	-0.57	1.00	-0.06	-0.05
Asian Indian	-0.01	-0.01	-0.01	0.00	0.05	0.14	-0.31	-0.06	1.00	0.09
Matric and above education level	0.05	0.17	-0.07	-0.02	-0.08	0.22	-0.23	-0.05	0.09	1.00

The Box's M statistic for testing the equality of covariance matrices gives a large value which is significant at $p < 0.0001$. However, since this analysis is based on the Census data set that a very large number of records (28.1 million), any statistical test for differences can be expected to be significant. Table 4.4.2.4 shows the canonical discriminant functions based on different covariance matrices for separate groups.

Table 4.4.2.4.: Group covariance of canonical discriminant functions

Employment status	Function	1	2
1=Employed	1	0.926	-0.040
	2	-0.040	1.219
2=Unemployed	1	1.000	0.226
	2	0.226	0.953
3=Inactive	1	1.068	-0.049
	2	-0.049	0.818

Looking at the information on each of the discriminant functions (equations) produced; the first two canonical discriminant functions with their associated eigenvalues are shown in the Table 4.4.2.5.

Table 4.4.2.5: Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.343	83.2	83.2	0.505
2	0.069	16.8	100.0	0.255

The canonical correlation, representing the correlation between the predictors and the discriminant function is 0.505 for the first function and 0.255 for the second function. The proportion of discriminating ability of the three classifications of employment status shows that the first function accounts for 83.2% of the total, whereas the second function accounts for 16.8%.

The significance of the discriminant functions is indicated by the Wilks' lambda indicator (Table 4.4.2.6), which is the proportion of the total variability in the discriminant scores not explained by differences among groups: 70% is unexplained between the first and second function whereas 94% is unexplained by the second function. The significant function ($p<0.0001$) for the Chi-square test indicates that there is a highly significant difference between the groups' means, indicating that the two functions have a significant discriminating power.

Table 4.4.2.6: Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1 through 2	0.70	10,170,028.02	20	0.0001
2	0.94	1,886,717.41	9	0.0001

Table 4.4.2.7 provides the standardized canonical discriminant function coefficients. The strongest predictors are age group 15 to 24 for the first function and being black for the second function. Being coloured and Indian is less useful as a predictor for the first function as is being male for the second function. The same pattern holds for those variables when looking at the structure matrix (Table 4.4.2.8). However, the structure matrix reveals further useful variables, looking at structure coefficients or discriminant loadings above 0.30. Table 4.4.2.8 shows that categories Age 15 - 24, Age 25 – 34, Age 35 – 44, being Male, being black or being coloured are important contributors to the first discriminant function. For the second function, the important variables are the population group categories (being black, coloured or Asian-Indian).

Table 4.4.2.7: Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
AGE1524	-.461	.012
AGE2534	.231	.541
AGE3544	.318	.199
Male	.387	-.066
Ever Married	.279	-.454
Urban	.192	.339
Black	-.108	.786
Coloured	.036	.334
Asian Indian	-.044	.138
Matric and above education level	.339	.235

Table 4.4.2.8: Structure Matrix

	Function	
	1	2
AGE1524	-0.733	0.010
AGE2534	0.386	-0.084
AGE3544	0.362	0.235
Male	0.320	0.206
Ever Married	0.275	-0.025
Urban	0.109	-0.101
Black	0.505	-0.569
Coloured	0.325	0.567
Asian Indian	-0.297	0.471
Matric and above education level	0.059	-0.087

The analysis gave the classification functions that were used to assign cases to groups or classification of employed, unemployed or inactive. Table 4.4.2.9 gives the canonical discriminant function coefficients that can be used to create the discriminant function equation. These coefficients indicate the partial contribution of each variable to the discriminant function controlling for all other variables in the equation.

Table 4.4.2.9: Canonical Discriminant Function Coefficients

Variables	Function	
	1	2
AGE1524	-1.067	0.028
AGE2534	0.539	1.263
AGE3544	0.840	0.525
Male	0.785	-0.134
Ever Married	0.589	-0.958
Urban	0.396	0.700
Black	-0.260	1.894
Coloured	0.125	1.168
Asian Indian	-0.261	0.811
Matric and above education level	0.724	0.501
(Constant)	-0.900	-2.082

The two discriminant functions below show how the discriminant function coefficients provide information on the relative importance of each variable in the expressions:

$D_1 = (-1.067 \times Age1524_1) + (0.539 \times Age2534_1) + (0.840 \times Age4565_1) + (0.785 \times Male_1) + (0.589 \times Ever\ married_1) + (0.396 \times Urban\ Settlement_1) + (-0.260 \times Black_1) + (0.125 \times Coloured_1) + (-0.261 \times Indian_1) + (0.724 \times Matric\ and\ above_1) - 0.900;$

$D_2 = (0.028 \times Age1524_2) + (1.263 \times Age2534_2) + (0.525 \times Age4565_2) + (-0.134 \times Male_2) + (-0.958 \times Ever\ married_2) + (0.7 \times Urban\ Settlement_2) + (1.894 \times Black_2) + (1.168 \times Coloured_2) + (0.811 \times Indian_2) + (0.501 \times Matric\ and\ above_2) -2.082.$

The outcome of the analysis managed to classify the different cases into discriminant categories where the group means of the predictor variables are described in terms of the employment profile. The groups means are called centroids as illustrated in Table 4.4.2.10. Cases with discriminant scores near to a centroid are predicted as belonging to that group or category.

Table 4.4.2.10: Functions at Group Centroids

Employment status	Function	
	1	2
1=Employed	0.667	-0.123
2=Unemployed	0.024	0.591
3=Inactive	-0.616	-0.112

The final aspect of the assessment of a discriminant analysis is the Confusion Table showing the observed and predicted categories (Table 4.4.2.11). The Confusion Table shows that the prediction accuracy of the model for the analysis sample is only 60.7%. No cross validation is possible since the classification has used separate group covariance matrices for the discriminant functions.

Application of the discriminant functions to the census data is achieved by combining the two data sets, and specifying that the Census records be treated as the holdout sample. It can be seen from Table 4.4.2.11 that the hold-sample, representing records extracted from Census, achieved a hit ratio of 50.9% of unselected original grouped cases correctly classified. The hit ratio is the percentage of correctly classified cases. The goodness of results shows that the hit ratio for the holdout sample exceeded the maximum chance value (42%) and proportional chance value (35%) calculated from Table 4.4.2.12. using the Census unselected cases. The maximum chance value is the maximum of the proportion of persons among the predicted employment status (1=employed, 2=unemployed, 3=inactive). The proportional chance value is $p_1^2 + p_2^2 + p_3^2$ where p_i is the proportion of persons in the employment status (1=employed, 2=unemployed, 3=inactive).

Table 4.4.2.11: Classification Results

Data sources		Employment status	Predicted Group Membership			Total
			1=Employed	2=Unemployed	3=Inactive	
Cases Selected (LFS)	Count	1=Employed	7,190,191	1,471,252	2,519,183	11,180,625
		2=Unemployed	1,377,035	1,333,213	1,945,046	4,655,294
		3=Inactive	2,829,546	908,466	8,543,430	12,281,442
	%	1=Employed	64.3	13.2	22.5	100.0
		2=Unemployed	29.6	28.6	41.8	100.0
		3=Inactive	23.0	7.4	69.6	100.0
Cases Not Selected (CENSUS)	Count	1=Employed	5,297,998	1,523,418	2,730,838	9,552,254
		2=Unemployed	2,966,339	1,083,794	2,767,050	6,817,182
		3=Inactive	3,030,376	884,069	8,025,622	11,940,067
	%	1=Employed	55.5	15.9	28.6	100.0
		2=Unemployed	43.5	15.9	40.6	100.0
		3=Inactive	25.4	7.4	67.2	100.0

Even though the discriminant model has some level of predictive power, the results of the classification could be improved by using the contribution from additional area levels in order to achieve improved classification within the different domains of estimation. Table 4.4.2.12 illustrates the distribution of the classification from the original data (LFS 2001 and Census 2001) and the predicted Census.

Table 4.4.2.12: Outcome of the reclassification of Census 2011

Employment status	LFS	Census	Census predicted (discriminant)
1=Employed	11,180,625	9,552,254	11,294,713
2=Unemployed	4,655,294	6,817,182	3,491,281
3=Inactive	12,281,442	11,940,067	13,523,510
Total	28,117,361	28,309,503	28,309,504
Unemployment rate	***29.40%***	***41.60%***	***23.61%***

The discriminant model has pooled those discouraged work seekers that are poorly identified between Unemployed and Inactive and reclassified some as employed or inactive. This has increased the estimate of the labour force participation and improved the prediction of unemployment. The next section will look at additional improvements by introducing other geographical factors to the small area estimation.

4.4.3. Multinomial Logistic model

A model-based regression estimator can be derived using domain-specific auxiliary variables of the same population observed in both Census 2001 and September LFS 2001. The problem is to be able to classify the population into the employment status categories. For each unit level variable, the estimate gives the classification into employed, unemployed and inactive. The proposed approach is a multinomial logistic model that predicts more than the two categories used in a logistic regression. Since there is no natural ordering to the value of dependent variables, the multinomial logistic regression is the alternative approach.

The proposed multinomial logit is fitted using a full factorial model where the parameter of estimation is obtained through the maximum likelihood algorithm. For a dependent variable with K categories, consider the existence of K unobserved continuous variables, Z_1, ... Z_K, each of which can be thought of as the "propensity toward" a category. The relationship between the Z's and the probability of a particular outcome is described in formula below:

$$\pi_{ik} = \frac{e^{z_{ik}}}{e^{z_{i1}} + e^{z_{i2}} + \ldots + e^{z_{iK}}} \qquad (4.8)$$

where π_{ik} is the probability that the i^{th} case falls in category k and z_{ik} is the value of the k^{th} unobserved continuous variable for the i^{th} case. z_k is also assumed to be linearly related to the predictors. $z_{ik} = b_{k0} + b_{k1}x_{i1} + \ldots + b_{kJ}x_{kJ}$ where (4.9)

- x_{ij} is the j^{th} predictor for the i^{th} case;
- b_{kj} is the j^{th} coefficient for the k^{th} unobserved variable
- j is the number of predictors

If z_k were observable, a linear regression could be fitted. However, since z_k is unobserved, the predictors are related to the probability of interest by substituting for z_k

$$\pi_i = \frac{e^{b_{k0} + b_{k1}x_{i1} + \ldots + b_{kJ}x_{iJ}}}{e^{b_{10} + b_{11}x_{11} + \ldots + b_{1J}x_{1J}} + .. + e^{b_{K0} + b_{K1}x_{i1} + \ldots + b_{KJ}x_{iJ}}} \qquad (4.10)$$

The K^{th} category is called the reference category, because all parameters in the model are interpreted in reference to it. The selection of the reference category determines the "standard" category to which others would naturally be compared.

In case of this study, the categories of dependent variable are "employed", "unemployed" and "inactive" with "employed" used as the reference category. The dependent variable was set to use the results from the discriminant analysis for the Census records and the data set with the original estiamtes from the LFS. The independent variables for each unit level (person) are indicator variables for the case belonging to age group 15 to 24 (1=Yes or 0=No), 25 to 34 (1=Yes or 0=No), 35 to 44 (1=Yes or 0=No); being male (1=Yes or 0=No), Ever married (1=Yes, 0=No), being black (1=Yes or 0=No), being coloured (1=Yes or 0=No), being Indian (1=Yes or 0=No), having matric or above level of education (1=Yes or 0=No), municipality X, Y coordinates, province X,Y coordinates, and the outcome of discriminant classification as employed (1=Yes or 0=No), unemployed (1=Yes or 0=No) and inactive (1=Yes or 0=No).

Table 4.4.3.1: Profile of Employment status for LFS and Census

Employment status	LFS (%)	Census (%)	Census predicted by discriminant (%)	LFS predicted by discriminant (%)
1=Employed	39.76	33.74	39.90	40.53
2=Unemployed	16.56	24.08	12.33	13.21
3=Inactive	43.68	42.18	47.77	46.26
Total	100.00	100.00	100.00	100.00

Table 4.4.3.1 shows the original and predicted distribution of the employment status. The proportion of unemployed has been aligned from Census to LFS profile using the discriminant analysis. The next concern is to look at subdivisions of the above national profile into the small domains of estimation such as municipality.

The goal is to build an association between employment status and different demographic variables and to estimate the employment status within the different levels of domains representing the small areas of estimation. The building of the multinomial logistic model needs to link the employment status with the variables of location area (X, Y coordinates). This allows pooling strength from those spatial variables (coordinates) to assist in improving the estimation for municipalities for the Census data. This is done by using the predicted class of employment status from the discriminant analysis as a dependent variable (only for Census with the original measure for LFS). This means that the original estimates of employment status is included as an independent variable for the census data.

Table 4.4.3.2: Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	94,588,782.8	94,588,814.5	94,588,778.8			
Final	22,219,404.8	22,219,880.3	22,219,344.8	72,369,434.0	28	.000

Table 4.4.3.2 gives the likelihood ratio test of the final model against one in which all the parameter coefficients are 0 (Null). The chi-square statistic is the difference between the -2 log-likelihoods of the Null and Final models. Since the significance level of the test is less than 0.05, the final model is outperforming the Null. The final model gives improved predictions compared to the Intercept Only (Null model).

The Akaike information criterion (AIC) is a measure of the relative goodness of fit of a statistical model. It describes the tradeoff between bias and variance in model construction, or between accuracy and complexity of the model. The AIC value gives a ranking of several candidate models providing a means for comparison for model selection. If all candidate models have the same number of parameters, then using AIC is similar to using the 2 log-likelihood which is valid only for nested models.

The Bayesian information criterion (BIC) is also a criterion for model selection among a finite set of models. It is closely related to both -2 log-likelihood and AIC except that the BIC introduces a penalty term for the number of parameters in the model larger in BIC than in AIC.

The model fitting information in Table 4.4.3.2 shows similar values of AIC, BIC and the -2 log-likelihood which means the final model is preferred to the intercept only model.

Caution should be used since it is difficult to check if the model adequately fits the data due to large number of data records from Census. The tests of significance of the Pearson and deviance statistics with p value <0.05 would usually imply that the goodness of fit is not good. Rather than rejecting the fit of the model to the data, the model is be considered useful for analysis purposes since one can still draw inferences from the model. Given the large number of records, no model can be obtained that will not be rejected by goodness of fit test.

The variables selected for constructing the model are considered as those predictors that contribute significantly to the model. The likelihood ratio statistics in the tables below test each variable's contribution to the model. For each effect, the -2 log-likelihood is computed for the reduced model; that is, a model without the effect. The chi-square statistic is the difference between the -2 log-likelihoods of the reduced model (Table 4.4.3.3) and the Final model reported in the model fitting information (Table 4.4.3.2). Since the significance of the tests are small (less than 0.05), all effects contribute to the model.

Table 4.4.3.3: Likelihood Ratio Tests

Effect	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	22,255,399.0	22,255,842.8	22,255,343.0	35,998.2	2	.000
Employed (original)	46,785,769.3	46,786,213.1	46,785,713.3	24,566,368.5	2	.000
Unemployed (original)	33,626,267.1	33,626,710.9	33,626,211.1	11,406,866.3	2	.000
Munic X	22,244,999.0	22,245,442.7	22,244,943.0	25,598.1	2	.000
Munic Y	22,229,965.6	22,230,409.4	22,229,909.6	10,564.8	2	.000
Province X	22,226,938.2	22,227,381.9	22,226,882.2	7,537.3	2	.000
Province Y	22,228,616.6	22,229,060.4	22,228,560.6	9,215.8	2	.000
AGE 15-24	24,764,836.0	24,765,279.7	24,764,780.0	2,545,435.1	2	.000
AGE 25-34	23,345,667.1	23,346,110.9	23,345,611.1	1,126,266.3	2	.000
AGE 35-44	22,950,524.3	22,950,968.0	22,950,468.3	731,123.5	2	.000
Male (1=Yes or 0=No)	25,513,434.7	25,513,878.5	25,513,378.7	3,294,033.9	2	.000
Ever Married (1=Yes or 0=No)	26,054,766.8	26,055,210.6	26,054,710.8	3,835,366.0	2	.000
Black (1=Yes or 0=No)	22,518,593.0	22,519,036.7	22,518,537.0	299,192.2	2	.000
Coloured (1=Yes or 0=No)	22,257,519.5	22,257,963.3	22,257,463.5	38,118.7	2	.000
Matric or above (1=Yes or 0=No)	25,044,159.6	25,044,603.4	25,044,103.6	2,824,758.8	2	.000

In the case of multinomial logistic regression model, a Pseudo R-Squared Statistic is computed to measure the variability in the dependent variable that is explained by the model. The Pseudo R-Squared Statistic summarizes the proportion of the variance in the dependent variable associated with the predictor (independent) variables. For categorical variables, the Pseudo R-Squared Statistic is an approximation based on Cox and Snell's R^2 (Cox and Snell, 1989), Nagelkerke's R^2 (Nagelkerke, 1991) and McFadden's R^2 (McFadden, 1974).

The Cox and Snell's R^2 is based on the log likelihood for the model compared to the log likelihood for a baseline model. The Nagelkerke's R^2 is an adjusted version of the Cox and Snell R^2 that adjusts the scale of the statistic to cover the full range from 0 to 1. Lastly, the McFadden's R^2 is another version, based on the log-likelihood kernels for the intercept-only model and the full estimated model (Table 4.4.3.4). The model explains more than 70% of variability in the dependent variable using the first two statistics.

Table 4.4.3.4:Pseudo R-Square

Cox and Snell	.723
Nagelkerke	.833
McFadden	.634

Table 4.4.3.5 shows the parameter estimates as a summary of the effect of each predictor. The indicators to look at are the squared ratio of the coefficient to its standard error, which equals the Wald statistic. If the significance level of the Wald statistic is small (less than 0.05) then the parameter is different from 0. The parameters with significant negative coefficients decrease the likelihood of that response category with respect to the reference category. Parameters with positive coefficients increase the likelihood of that response category.

Since the reference category is the "Employed persons", the parameter estimates are determined for "Unemployed person" as well as for "Inactive person". The choice of reference category is due to the poor classification of employment status between economically active and not economically, where there might be problem of correctly classifying the discouraged workers who may be either unemployed or not economically active.

The first part of the Table 4.4.3.5 shows the outcome for the "Employed" category compared to the second part "Unemployed" category. Looking at odds ratios in column under Exp(B), persons aged between 15 and 24 are more likely to be classified as Inactive. Reciprocally, those who where originally classified unemployed, as well as those persons aged between 25 and 34 including black persons are more likely to be classified as unemployed.

Table 4.4.3.5: Parameter Estimates

Employment status (Classified Discriminant analysis)		B	Std. Error	Wald	Df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
3.=Inactive	Intercept	2.037	.011	32,138.7	1	.000			
	Employed (original)	-7.477	.005	2,729,358.7	1	.000	.001	.001	.001
	Unemployed (original)	-1.697	.001	1,507,313.3	1	.000	.183	.183	.184
	Munic X	.059	.001	13,747.6	1	.000	1.061	1.060	1.062
	Munic Y	.069	.001	9,052.0	1	.000	1.072	1.070	1.073
	Province X	-.035	.001	3,770.2	1	.000	.965	.964	.967
	Province Y	-.057	.001	5,082.0	1	.000	.944	.943	.946
	AGE 15-24	2.326	.002	2,203,540.0	1	.000	10.238	10.207	10.270
	AGE 25-34	-.005	.002	10.2	1	.001	.995	.991	.998
	AGE 35-44	-1.060	.002	357,101.1	1	.000	.346	.345	.348
	Male (1=Yes or 0=No)	-1.730	.001	2,243,662.5	1	.000	.177	.177	.178
	Ever Married (1=Yes or 0=No)	-.289	.001	65,100.0	1	.000	.749	.747	.750
	Black (1=Yes or 0=No)	.274	.002	25,345.8	1	.000	1.315	1.310	1.319
	Coloured (1=Yes or 0=No)	-.283	.002	12,850.5	1	.000	.754	.750	.757
	Matric or above (1=Yes or 0=No)	-2.005	.001	2,425,234.1	1	.000	.135	.134	.135
2.=Unemployed	Intercept	.346	.014	638.6	1	.000			
	Employed (original)	-4.898	.003	2,072,112.0	1	.000	.007	.007	.008
	Unemployed (original)	2.545	.002	2,805,292.5	1	.000	12.745	12.707	12.783
	Munic X	-.028	.001	2,141.0	1	.000	.972	.971	.973
	Munic Y	.007	.001	69.2	1	.000	1.007	1.006	1.009
	Province X	.019	.001	735.2	1	.000	1.019	1.017	1.020
	Province Y	.023	.001	575.6	1	.000	1.024	1.022	1.025
	AGE 15-24	.775	.002	140,154.7	1	.000	2.170	2.161	2.179
	AGE 25-34	1.766	.002	821,751.6	1	.000	5.847	5.825	5.870
	AGE 35-44	.643	.002	101,313.1	1	.000	1.902	1.895	1.910
	Male (1=Yes or 0=No)	-1.859	.001	1,884,918.7	1	.000	.156	.155	.156
	Ever Married (1=Yes or 0=No)	-2.397	.001	2,811,233.6	1	.000	.091	.091	.091
	Black (1=Yes or 0=No)	1.374	.003	266,083.1	1	.000	3.951	3.930	3.972
	Coloured (1=Yes or 0=No)	.415	.004	13,686.2	1	.000	1.514	1.504	1.525
	Matric or above (1=Yes or 0=No)	-.931	.001	446,497.6	1	.000	.394	.393	.395

The classification in Table 4.4.3.6 shows the outcome of reclassification into the predicted response category. This will allow us to verify whether the model has improved the classification of the employment status.

Table 4.4.3.6: Multinomial Logistics regression classification Results

Data sources	Employment status	Predicted categories			Total
		1=Employed	2=Unemployed	3=Inactive	
LFS	1=Employed	7,055,592	1,652,982	2,471,680	11,180,254
	2=Unemployed	1,351,257	1,497,892	1,908,369	4,757,519
	3=Inactive	2,776,577	1,020,680	8,382,331	12,179,588
	Total	11,183,426	4,171,555	12,762,380	28,117,361
	%	39.77	14.84	45.39	100.00
Census	1=Employed	5,071,157	1,974,820	2,619,594	9,665,571
	2=Unemployed	2,839,331	1,404,931	2,654,331	6,898,593
	3=Inactive	2,900,626	1,146,026	7,698,688	11,745,340
	Total	10,811,115	4,525,777	12,972,612	28,309,504
	%	38.19	15.99	45.82	100.00

The result shows that there has been change in reclassification of Census data using the profile of the population distribution from the LFS to the 'true' employment status profile. The outcome shows that the LFS and Census have similar estimated profile distributions while maintaining the profile of the direct estimates of employment status of the LFS (Table 4.4.3.7).

Table 4.4.3.7: Employment status profile comparison for LFS and Census in 2001

Employment status	Census (counts) predicted by MLR	LFS (counts) predicted by MLR	Census predicted MLR (%)	LFS predicted MLR (%)	Census predicted by discriminant (%)	LFS predicted by discriminant (%)	Direct measure LFS (%)	Direct measure Census (%)
1=Employed	10,811,115	11,183,426	**38.19**	**39.77**	39.9	40.53	39.76	33.74
2=Unemployed	4,525,777	4,171,555	**15.99**	**14.84**	12.33	13.21	16.56	24.08
3=Inactive	12,972,612	12,762,380	**45.82**	**45.39**	47.77	46.26	43.68	42.18
Total	28,309,504	28,117,361	**100**	**100**	**100**	**100**	**100**	**100**

The new result allows tabulation of the persons in Census data with a new profile allocation within the different domain of the small area estimation. The model-based estimation, in comparison with the direct measurement of the labour force, presents small shifts in the profile of employment across the provinces. There is little remarkable difference when comparing the original LFS data and the model-based estimates. Larger changes in terms of the unemployment rate are the increase of the unemployment rate in the Western Cape and Gauteng. The model-based estimation from the survey data has lowered the unemployment rates in the Eastern Cape, KwaZulu-Natal and North West (Table 4.4.3.8 and Table 4.3.1.1.1).

As expected, a change is observed in the Census data (see Table 4.3.1.2) when comparing the model-based estimates to the original count. The only increase in unemployment rate is observed in the Western Cape and Gauteng. The remaining provinces are estimated to have an unemployment rate less than what the Census has erroneously counted before. Nationally, the estimation of unemployment rate is recorded as having reduced from 41.6% in Census 2001 to 29.5%. There is also a reduction of unemployment rate in tribal rural provinces (Eastern Cape, KwaZulu-Natal and Limpopo) (Table 4.4.3.9).

Table 4.4.3.8: Model-based estimation of People aged between 15 - 65 by Employment status for LFS Sept. 2001

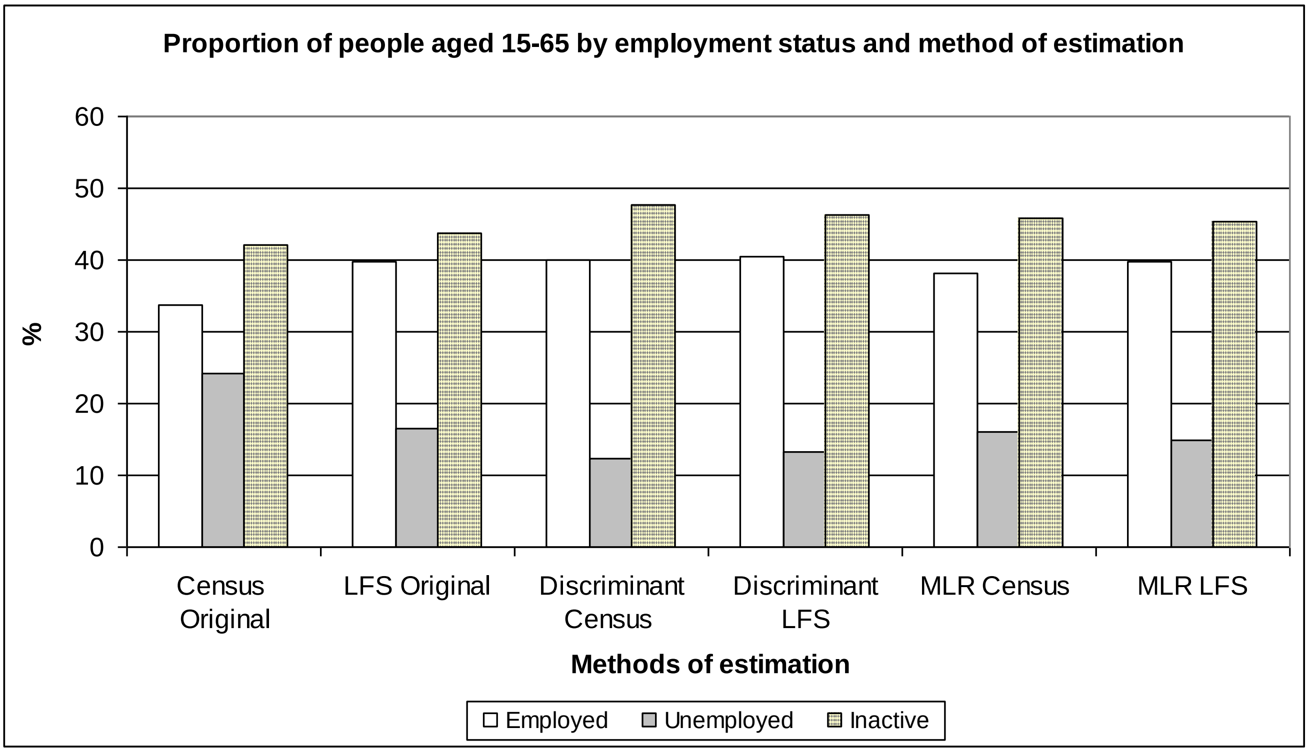
Province (Domain of estimation) (a)	Total persons aged 15-65 (b)=(c)+(d)	Persons not economically active (c)	Persons economically active (d)=(e)+(f)	Employed (e)	Unemployed (f)	Unemployment rate (g) =100*(f)/(d)
Western Cape	2,864,663	1,852,509	1,012,154	721,769	290,385	28.7%
Eastern Cape	3,870,039	1,454,461	2,415,578	1,908,392	507,186	21.0%
Northern Cape	561,113	278,672	282,441	215,292	67,149	23.8%
Free State	1,945,710	980,797	964,913	703,848	261,065	27.1%
KwaZulu-Natal	5,786,791	2,279,402	3,507,389	2,569,910	937,479	26.7%
North-West	2,333,903	967,683	1,366,220	1,031,900	334,320	24.5%
Gauteng	5,861,154	3,151,670	2,709,484	1,632,882	1,076,602	39.7%
Mpumalanga	1,870,520	798,940	1,071,580	788,464	283,116	26.4%
Limpopo	3,023,462	998,242	2,025,220	1,610,968	414,252	20.5%
Total (South Africa)	**28,117,355**	**12,762,376**	**15,354,979**	**11,183,425**	**4,171,554**	**27.2%**

Table 4.4.3.9: Model-based estimation of People aged between 15 - 65 by Employment status for 2001 Census

Province (Domain of estimation) (a)	Total persons aged 15-65 (b)=(c)+(d)	Persons not economically active (c)	Persons economically active (d)=(e)+(f)	Employed (e)	Unemployed (f)	Unemployment rate (g) =100*(f)/(d)
Western Cape	3,054,351	2,073,384	980,967	650,032	330,935	33.7%
Eastern Cape	3,683,216	1,198,233	2,484,983	1,870,739	614,244	24.7%
Northern Cape	525,957	284,938	241,019	180,072	60,947	25.3%
Free State	1,741,726	760,087	981,639	650,624	331,015	33.7%
KwaZulu-Natal	5,741,128	2,292,087	3,449,041	2,559,291	889,750	25.8%
North-West	2,333,510	1,005,509	1,328,001	940,903	387,098	29.1%
Gauteng	6,408,709	3,664,651	2,744,058	1,645,247	1,098,811	40.0%
Mpumalanga	1,902,485	796,777	1,105,708	798,492	307,216	27.8%
Limpopo	2,918,425	896,950	2,021,475	1,515,714	505,761	25.0%
Total (South Africa)	**28,309,507**	**12,972,616**	**15,336,891**	**10,811,114**	**4,525,777**	**29.5%**

By running the table from Census data on the model estimated employment status by municipalities, the detailed results of the unemployment rate estimation per municipality are obtained as presented in the annexure (Table A1). The province and municipalities have been considered as domain factors determining the sub-populations of the small area domain. Figure 4.4.3.1 shows the overall classification of the South African population aged between 15 and 65 into employed, unemployed and inactive using the different data methods. The Census direct misclassification of economic activities is readjusted to the LFS classification first using the discriminant analysis and later using the multinomial logistics regression. The change in profile is noticeable for Census data whereas a little variation is observed from the LFS control sample data.

Figure 4.4.3.1



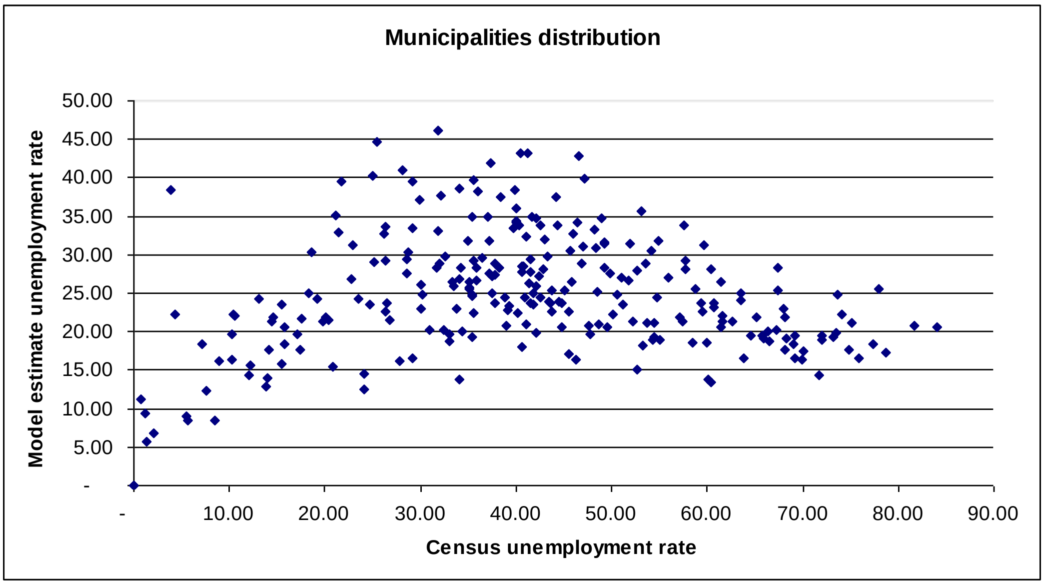
4.5. Evaluation of the results

The over-estimation of unemployment in Census has been reduced (Table 4.5.1.) from 41.6% to 29.5% by reclassifying the discouraged job seekers into the main stream of the economically active population. Although the same reduction pattern is observed for all provinces, the reduction of the unemployment rate is more pronounced in rural provinces (Eastern Cape, KwaZulu-Natal and Limpopo). However, the change in profile for municipalities has not been systematically similar across municipalities. There are some few municipalities with an increased unemployment rate whilst other municipalities show a reduced unemployment rate (Figure 4.5.1).

Table 4.5.1: Comparison of unemployment rate for different estimation approach by Province

Province	LFS Origin al	Census 2001 original	LFS 2001 Discriminant	Census 2001 Discriminant	LFS 2001 Multinomial logistic	Census 2001 Multinomial logistic
Western Cape	17.7	26.1	10.7	15.5	28.7	33.7
Eastern Cape	31.4	54.7	26.6	20.1	21.0	24.7
Northern Cape	25.0	33.4	14.2	16.4	23.8	25.3
Free State	27.0	43.2	17.2	36.2	27.1	33.7
KwaZulu-Natal	33.8	48.8	31.5	18.6	26.7	25.8
North-West	28.6	44.0	27.4	23.4	24.5	29.1
Gauteng	30.4	36.5	27.1	29.7	39.7	40.0
Mpumalanga	29.2	41.1	28.2	26.0	26.4	27.8
Limpopo	34.6	48.8	24.6	21.8	20.5	25.0
Total	29.4	41.6	24.6	23.6	27.2	29.5

Figure 4.5.1



The validation of the results is done by comparing the unemployment rate where the Labour Force Survey of September 2001 is used as the reference. However, the comparison is done on National and Provincial level as a confirmation of the estimation to be used as results (Table 4.5.2). It is assumed that the validity of the estimates at National and Provincial level can be used to support the plausible estimates on municipality level due to lack of available valid from other sources. The results of ARD (absolute relative difference) represented in Table 4.5.2 is calculated as a measure of model estimation error.

$$ARD_j = \frac{| r(X_j) - r(\hat{\theta}_j) |}{Max(| r(X_j) |, | r(\hat{\theta}_j) |)} \tag{4.11}$$

The model estimation error shows a very small percentage difference even applied to the Census or LFS..

Table 4.5.2: Percent ARD of unemployment rate between LFS 2001 (reference) and different estimation approaches by Province

Province	Census 2001 original	LFS 2001 Discriminant	Census 2001 Discriminant	LFS 2001 Multinomial logistic	Census 2001 Multinomial logistic
Western Cape	0.3225	0.3977	0.1268	0.3829	0.4752
Eastern Cape	0.4259	0.1526	0.3597	0.3304	0.2117
Northern Cape	0.2512	0.4329	0.3428	0.0487	0.0117
Free State	0.3722	0.3617	0.2545	0.0017	0.1990
KwaZulu-Natal	0.3069	0.0685	0.4506	0.2087	0.2363
North-West	0.3474	0.0410	0.1824	0.1434	0.0199
Gauteng	0.1653	0.1083	0.0234	0.2345	0.2404
Mpumalanga	0.2896	0.0350	0.1086	0.0948	0.0480
Limpopo	0.2903	0.2889	0.3692	0.4095	0.2777
Total	0.2932	0.1641	0.1968	0.0758	0.0038

4.6. Summary of the application of Small Area Estimation to employment status

This chapter gave a practical attempt to improve the measurement of unemployment within small area domains of estimation such as the Municipality. The problem has two dimensions: the poor classification of employment status by Census 2001 and the lack of good measurement of employment at the small area domain of estimation.

The choice of the data used was motivated by the strength in each data set which could be used to alleviate the weakness in the other data set. All data sets were assumed to have been concurrent because the LFS took place in September 2001 and the Census took place in October 2001. Those periods are close enough to allow the assumption of non-variability affecting the employment status. Moreover, both Census 2001 and the September LFS 2001 have a number of similar demographic variables. Although the LFS is a sample, it shares the same spatial distribution of the settlement of the population in South Africa in Census 2001 due to the sample design. In addition, the universe was limited to people in the age group 15 to 65 and the x, y coordinates of the different domains of estimation were also used as spatial covariates for province and municipality.

The inference using the LFS is based on applying the sampling weights which give reliable results within the domain of estimation of the survey. As for 2001 Census, though there is an assumption that everybody in the country has been counted, the final figure is corrected of the undercount by applying the adjustment factors for each unit observed.

The comparison between the measurements obtained from the LFS and the Census shows that there was a problem of misclassification in that the Census estimate of employment status has shifted some of discouraged job seekers to outside the economically active band. Looking at different possible approaches, the use of synthetic estimation is not recommended due to the bias resulting from the assumption of the same profiles across the small areas. An acceptable approach is to build estimation models such as discriminant analysis for the employment status categories as well as estimation using a multinomial logistic model. The use of the discriminant model assisted in reducing the effect of misclassification in the data. However, the use of municipality's coordinates as covariates ensures that the proximity of the different municipalities representing the small area domain of estimation is taken into account.

In order to improve the estimates, the result from discriminant analysis estimation was used as input into the next stage of re-estimating the employment status groups using the multinomial logistic model together as covariates with the demographic variables as well as the geographic coordinates of the target area: Municipalities and Provinces. The outcome shows that overall the profile of LFS employment status was maintained, particularly the adjustment of Census profile of employment status to resemble the LFS profile of the employment status.

The result of using the model-based approach has reduced the gap between the Census and the LFS profiles of unemployment status. This is an indication that the estimation has borrowed strength in the estimation of unemployment status from the LFS and applied this to the Census. The result shows that the estimation of unemployment status per municipality is possible.

CHAPTER 5: CONCLUSIONS

5.1. Summary of findings

This study has attempted to assess different approaches to SAE, where the domains of estimation were Provinces and local municipalities. The reference to the literature assisted in understanding the different theoretical methods of SAE. The direct estimation methods are mostly used for the surveys and the census where a questionnaire was administered in the different households. The indirect estimation methods include synthetic estimation, composite estimation, symptomatic regression, and area level or unit level models. Despite the different theoretical approaches, the practical application of SAE techniques is dependent on the available data as well as on the strength of the relation of the variables between the different sources of information. In the estimation of unemployment status across municipalities, the available data used are the LFS, the Population census and the geographic coordinates of different domain areas.

The report has attempted to show the results from the direct estimation where the issue of misclassification was clarified. A synthetic estimation that was attempted was without success, since it is based on the assumption that the labour force profile at provincial level remains the same across the municipalities. Finally, the recommended model estimates are based on a two-stage model estimation where the outcome from a discriminant model is used as input into the multinomial logistic model.

5.2. Limitation of the estimation and future work

Although the modelling approach is able to give a SAE of unemployment rate for each municipality, there are some limitations due to the nature of demarcated municipalities. Some municipalities have at best little statistical representation in the LFS because there are few people living in those municipalities (e.g. District Management Areas relating to National Parks). Even though some other municipalities gave the classification of employment status, those areas with predominantly rural or traditional profiles may have limited economic activities beyond the family business. It is sometimes difficult to classify those people involved in family businesses correctly into employed or not employed.

The lack of additional data set related to economic activities at municipal level does not allow a good validation process. It is difficult to validate the results with reference to any direct measurement related to labour market data on municipality level. The available auxiliary data with labour market variables can be reliably used at provincial level.

Future work would require building a SAE model-based on area level data where the economic factors can be added as well as time factors. In addition, the future research should look at how to update the change using the local auxiliary data even though these data might not be exhaustive.

5.3. Policy implications

A SAE such as the measurement of unemployment status or rate for each municipality in South Africa gives a basis or reference for planning and for the monitoring and evaluation of government programs. Since the local area of development and service delivery is the municipality, the estimate of unemployment at local level should be seen as a good step towards allowing targeted interventions to alleviate poverty, looking at those who have jobs and those who do not. In addition, localized interventions may be aimed at areas contributing towards production in support of overall development. Although the estimates used can be used as a reference for unemployment status as in 2001 Census, more research will be required to provide an updated estimate of current or future projected unemployment status for each municipality.

Bibliography

- Battese, G.E., Harter, R.M., Fuller, W.A. (1988), "*An Error-Components Model for Prediction of Crop Areas Using Survey and Satellite Data*", *Journal of the American Statistical Association*, 83, 28-36.
- Battistin, E, Restore, E and Trivellato, U, (2007), "*Choosing between alternative classification criteria to measure the labour force state*", Journal Royal Statistical Society, 170(1), 5 -27.
- Chatfield, C. (1995), "*Model uncertainty, data mining and statistical inference*". Journal of the Royal Statistical Society, Ser. A 158, 419–466.
- Cochran W.G., (1977), "*Sampling techniques*", Wiley, Harvard, USA, 428 pages.
- Cox D. R. and Snell E.J., (1989),"*Analysis of Binary Data*", Monographs on Statistics and Applied Probability, Chapman & Hall, London, UK.
- D'Alo M., Di Cousiglio L., Falorsi S., Pratesi M. Ranolli M.G., Solari F. (2008), "*Small area models for unemployment rate estimation at sub-provincial areas in Italy* ", Paper for Project PRIN2007, Italy.
- Draper, D. (1995). *"Assessment and propagation of model uncertainty".* Journal of the Royal Statistical Society, Ser. B 57, 45–97.
- Earwaker S., Brien N. and Gosselin J.F., (1984), "*Model-based Small area estimation of unemployment: A Canadian experience*", Statistics Canada, Canada.
- Ericksen, E. (1973),"*A method of combining sample survey data and symptomatic indicators to obtain population estimates for local areas*", Demography, 10, 147–160.
- Ericksen, E. (1974), "*A regression method for estimating population changes of local areas*", Journal of the American Statistical Association, 69, 867–875.
- Fay, R.E. and Herriot, R.A. (1979), "*Estimates of Income for Small Places: An Application of James-Stein Procedures to Census data*", American Statistical Association, USA.
- Francisco J. (2003), "*Small area estimation: An overview of existing methodologies with application to the Estimation of unemployment rates in the Philippines*", University Laval, Canada.
- Fuller, W.A. (1995). "*Estimation in the Presence of Measurement Error*". *International Statistical Review*, 63, 121-147.
- Gebizlioudlu O.L., Depep H. and Toprak O. (1996), "*Small area estimation Turkey*", Annual Research Conference, US Bureau of Census, USA.
- Ghosh and Rao (1994), "*Small Area Estimation: An Appraisal*", Statistical Science, 9(1) 55-93.
- Holt, D. (2000), "*Statistics in a Democratic Society: the Role of Methodology*", Statistics and Democracy: Proceedings of the 4th International Conference on Methodological Issues in Stockholm October 12-13, 2000.
- Jiang, J. and Lahiri, P. S. (2006b), "*Mixed model prediction and small area estimation*". 15 1–96.
- Khoshgooyanfard, A.R. and Monazzah, M.T., (2006), "*A Cost Effective Strategy for Provincial Unemployment Estimation: A Small Area Approach",* Survey Methodology, Vol. 32, No. 1, 105-114, Statistics Canada.
- Kish L., (1995), "*Survey sampling*", Wiley, Michigan, USA, 643 pages.
- Knight R., (2001), "South Africa: Economic policy and development", paper, SSHARED INTEREST (NGO), New York, USA. http://richardknight.homestead.com/files/sisaeconomy.htm

- Longford, N. T. (1999). "*Multivariate shrinkage estimation of small area means and proportions*". Journal of the Royal Statistical Society, Ser. A 162, 227–245.
- Longford N. T. (2003) "*An alternative to model selection in ordinary regression*". Statistics and Computing 13 67–80, 2003.
- Longford, N. T. (2004). "*Missing data and small-area estimation in the UK Labour Force Survey*". Journal of the Royal Statistical Society Ser. A 167, 341–373.
- Longford, N. T. (2005a). "*Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician"*. Springer-Verlag, New York. 357 pages.
- Longford, N. T. (2005b). "*On selection and composition in small area and mapping problems"*. Statistical Methods in Medical Research 14, 1–14.
- Longford, N. T. (2007). "*On standard errors of model-based small-area estimators*". Survey Methodology, Vol. 33, No 1, Statistics Canada, Catalogue No. 12-001, pp 69-79,
- Marker D., (1983), "*Organisation of small area estimators*", Westat, USA
- Marker D.A, (1999), "*Organisation of small area estimators using a generalised linear regression framework*", Journal of Official Statistics, 15, 1-24.
- Marker, D.A. (2001), "*Producing Small Area Estimates from National Surveys: Methods for Minimizing Use of Indirect Estimators.*" *Survey Methodology*, 27, pages 183-188.
- May J., Budlender D., Mokate R., Rogerson C. and Stavrou A., (1998),"*Poverty and inequality in South Africa*", Report office of Deputy President, Pretoria.
- Moura, F.A.S. and Holt, D. (1999), "*Small area estimation using Multilevel Models*", Survey Methodology, 25, 73-80.
- Office for National Statistics ONS (2001a), "*Small area estimation project*", London, UK.
- Prasad, N.G.N. and Rao, J.N.K. (1990), "*The estimation of the mean squared error of small area estimators*". *Journal of the American Statistical Association.* 85, 63-171.
- Pfeffermann, D, (2010), "*New Important Developments In Small Area Estimation*", Working Paper M10/11, Methodology, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK, 35 pages.
- Rao J.N.K, (2003a), "*Small Area Estimation",* John Wiley & Sons, Inc., New York.
- Rao J.N.K, (2003b), "*Some New Developments in Small Area Estimation*", Journal of the Iranian Statistical Society (2003), 2(2), 145-169. 313 pages.
- SAS Institute Inc. (2006), "*Base SAS 9.1.3 Procedures Guide*", Second edition, Volumes 1,2,3 and 4, Cary, NC, USA.
- Satorra A. and Ventura E, (2006), "*Small Area Estimation at IDESCAT: Current and Related Research",* IDESCAT, Barcelona, Spain, 22 pages.
- Space-Time Research Pty Limited (2006), "*SuperCross User's Guide Version 4.5.3*", Acrobat portable file, Melbourne, Australia.
- SPSS Inc (2007a), "*SPSS Advanced Statistics 17.0*", Acrobat portable file, Chicago, Illinois, USA.
- SPSS Inc (2007b), "*SPSS Statistics 17.0 User's Guide*", Acrobat portable file, Chicago, IL, USA.
- Stats SA, (1998), "*Unemployment and employment in South Africa",* Statistics South Africa, Pretoria, South Africa, 80 pages.

- Stats SA, (2002), "*Labour force survey*", Statistics South Africa, Pretoria, Statistics South Africa, Pretoria, South Africa, 98 pages.
- Stats SA, (2003), "*Census 2001: How the count was done*", Statistics South Africa, Pretoria, South Africa, 58 pages.
- Stats SA, (2004), "*Census 2001: The Post-enumeration survey - Results and methodology*", Statistics South Africa, Pretoria, South Africa, 91 pages.
- Suciu, G., Hoshaw-Woodard, S., Elliott, M., and Doss, H, (2001), "*Uninsured estimates by County: A review of options and issues*", Report, Centre for Biostatistics, The Ohio State University, USA, 30 pages.

ANNEXURE I: ESTIMATION OF UNEMPLOYMENT RATE BY MUNICIPALITIES

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate)

	Municipalities	Predicted Response Category (MLR)			Total	CENSUS: Unemployment rate (Original)	**Discriminant Model: Unemployment rate**	**MLR Model:Unemployment rate**
		1= Employed	2= Unemployed	3= Inactive				
1	101 WC011: Matzikama	7,713	1,989	23,017	32,719	15.8	7.2	**20.5**
2	102 WC012: Cederberg	5,820	1,133	18,796	25,749	10.2	7.0	**16.3**
3	103 WC013: Bergrivier	6,624	927	23,802	31,353	7.6	5.5	**12.3**
4	104 WC014: Saldanha Bay	10,295	5,047	32,296	47,638	21.5	15.1	**32.9**
5	105 WC015: Swartland	9,739	2,396	35,758	47,893	10.2	9.1	**19.7**
6	106 WC022: Witzenberg	12,728	3,567	39,082	55,377	14.6	12.7	**21.9**
7	107 WC023: Drakenstein	31,818	11,722	84,439	127,979	22.8	12.8	**26.9**
8	108 WC024: Stellenbosch	22,465	5,491	54,968	82,924	17.1	12.0	**19.6**
9	109 WC025: Breede Valley	24,664	6,683	63,071	94,418	19.7	11.0	**21.3**
10	110 WC026: Breede River/Winelands	15,467	2,858	33,177	51,502	12.2	10.5	**15.6**
11	111 WC031: Theewaterskloof	12,945	5,617	43,794	62,356	18.6	14.2	**30.3**
12	112 WC032: Overstrand	5,507	3,587	28,433	37,527	21.7	16.6	**39.4**
13	113 WC033: Cape Agulhas	4,110	884	11,926	16,920	14.2	9.2	**17.7**
14	114 WC034: Swellendam	5,063	1,141	11,638	17,842	15.8	9.5	**18.4**
15	115 WC041: Kannaland	5,073	747	8,804	14,624	13.9	5.6	**12.8**
16	116 WC042: Langeberg	8,537	1,378	18,607	28,522	14.1	6.6	**13.9**
17	117 WC043: Mossel Bay	13,705	4,217	30,461	48,383	24.7	12.9	**23.5**
18	118 WC044: George	24,373	10,626	54,620	89,619	28.7	17.1	**30.4**
19	119 WC045: Oudtshoorn	16,870	5,047	30,681	52,598	33.7	8.5	**23.0**
20	120 WC047: Plettenberg Bay	4,727	2,304	12,980	20,011	26.3	20.4	**32.8**
21	121 WC048: Knysna	9,156	3,809	21,302	34,267	28.6	17.3	**29.4**
22	122 WC051: Laingsburg	1,274	372	2,605	4,251	26.3	7.3	**22.6**
23	123 WC052: Prince Albert	2,020	727	3,589	6,336	35.0	7.8	**26.5**
24	124 WC053: Beaufort West	8,755	2,584	11,913	23,252	39.1	10.6	**22.8**
25	171 City of Cape Town	374,935	244,943	1,360,334	1,980,212	29.2	17.4	**39.5**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) - continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemplo-yment rate (Original)	**Discriminant Model: Unemplo-yment rate**	**MLR Model:Unemplo-yment rate**
		1= Employed	2= Unemplo-yed	3= Inactive				
26	191 WCDMA01: West Coast	888	175	1,449	2,512	29.2	4.6	**16.5**
27	192 WCDMA02: Breede River	745	45	3,712	4,502	1.4	6.2	**5.7**
28	193 WCDMA03: Overberg	21	6	139	166	4.3	0.9	**22.2**
29	194 WCDMA04: South Cape	2,539	542	6,240	9,321	17.4	5.4	**17.6**
30	195 WCDMA05: Central Karoo	1,458	370	1,750	3,578	32.4	9.8	**20.2**
31	201 EC101: Camdeboo	9,360	2,694	14,569	26,623	35.6	10.9	**22.4**
32	202 EC102:Blue Crane Route	8,385	3,230	11,003	22,618	40.6	20.2	**27.8**
33	203 EC103: Ikwezi	2,686	709	2,988	6,383	41.0	13.1	**20.9**
34	204 EC104: Makana	18,903	8,635	22,534	50,072	49.3	22.2	**31.4**
35	205 EC105: Ndlambe	13,667	5,437	16,806	35,910	40.7	24.4	**28.5**
36	206 EC106: Sunday's River Valley	10,593	3,500	13,446	27,539	35.5	21.5	**24.8**
37	207 EC107: Baviaans	3,968	570	5,008	9,546	24.1	7.3	**12.6**
38	208 EC108: Kouga	14,696	4,581	27,488	46,765	26.5	17.5	**23.8**
39	209 EC109: Kou-Kama	5,666	1,880	14,961	22,507	18.3	13.9	**24.9**
40	210 EC121: Mbhashe	88,252	17,494	19,266	125,012	69.2	15.8	**16.5**
41	211 EC122: Mnquma	92,936	25,063	34,452	152,451	62.7	17.6	**21.2**
42	212 EC123: Great Kei	14,227	4,050	8,419	26,696	50.1	20.0	**22.2**
43	213 EC124: Amahlathi	41,813	15,077	24,783	81,673	61.5	15.5	**26.5**
44	214 EC125: Buffalo City	170,526	94,860	208,910	474,296	53.1	23.6	**35.7**
45	215 EC126: Ngqushwa	28,074	9,661	10,935	48,670	78.0	8.9	**25.6**
46	216 EC127: Nkonkobe	37,948	14,972	24,813	77,733	67.4	13.7	**28.3**
47	217 EC128: Nxuba	6,418	2,600	6,497	15,515	53.6	20.2	**28.8**
48	218 EC131: Inxuba Yethemba	13,833	5,869	18,362	38,064	43.3	20.5	**29.8**
49	219 EC132: Tsolwana	10,501	2,815	5,400	18,716	54.4	17.9	**21.1**
50	220 EC133: Inkwanca	5,540	2,155	4,412	12,107	52.7	27.9	**28.0**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) – continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemployment rate (Original)	**Discriminant Model: Unemployment rate**	**MLR Model:Unemployment rate**
		1= Employed	2= Unemployed	3= Inactive				
51	221 EC134: Lukanji	53,441	19,727	38,512	111,680	55.9	26.4	**27.0**
52	222 EC135: Intsika Yethu	66,129	13,942	15,833	95,904	70.1	12.0	**17.4**
53	223 EC136: Emalahleni	40,336	8,698	10,978	60,012	68.1	15.8	**17.7**
54	224 EC137: Engcobo	49,593	11,236	12,429	73,258	69.1	15.6	**18.5**
55	225 EC138: Sakhisizwe	15,440	5,012	8,930	29,382	54.8	27.1	**24.5**
56	226 EC141: Elundini	42,029	13,305	14,667	70,001	63.6	23.8	**24.0**
57	227 EC142: Senqu	44,911	10,537	18,862	74,310	55.1	18.6	**19.0**
58	228 EC143: Maletswai	9,557	3,187	10,107	22,851	37.4	29.9	**25.0**
59	229 EC144: Gariep	8,176	2,769	7,762	18,707	43.8	26.6	**25.3**
60	230 EC151: Mbizana	80,289	19,805	17,787	117,881	73.5	18.6	**19.8**
61	231 EC152: Ntabankulu	47,766	10,185	8,754	66,705	74.8	19.6	**17.6**
62	232 EC153: Qaukeni	80,946	19,613	23,331	123,890	65.8	17.8	**19.5**
63	233 EC154: Port St Johns	49,117	9,614	10,016	68,747	70.0	16.3	**16.4**
64	234 EC155: Nyandeni	93,756	21,776	23,531	139,063	72.0	19.2	**18.9**
65	235 EC156: Mhlontlo	63,750	15,476	19,136	98,362	64.5	17.9	**19.5**
66	236 EC157: King Sabata Dalindyebo	129,109	36,071	58,344	223,524	57.1	23.7	**21.8**
67	237 EC05b1: Umzimkhulu	54,963	15,452	16,075	86,490	68.2	20.1	**21.9**
68	238 EC05b2: Umzimvubu	118,889	35,433	35,613	189,935	68.0	19.1	**23.0**
69	275 Port Elizabeth: Nelson Mandela	223,225	116,334	349,701	689,260	46.4	20.1	**34.3**
70	291 ECDMA10: Aberdeen Plain	1,312	221	2,751	4,284	12.1	4.2	**14.4**
71	292 ECDMA13: Mountain Zebra National Park	10	1	53	64	5.5	20.0	**9.1**
72	293 ECDMA14: Oviston Nature Reserve	1	-	4	5	**-**	-	**-**
73	294 ECDMA44: O'Conners Camp	-	-	-	-	-		-
74	301 NC061: Richtersveld	1,428	942	4,489	6,859	35.5	10.2	**39.8**
75	302 NC062: Nama Khoi	7,423	2,672	19,328	29,423	33.3	8.3	**26.5**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) – continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemployment rate (Original)	**Discriminant Model: Unemployment rate**	**MLR Model:Unemployment rate**
		1= Employed	2= Unemployed	3= Inactive				
76	303 NC064: Kamiesberg	1,225	495	4,911	6,631	32.0	5.6	**28.8**
77	304 NC065: Hantam	3,552	969	7,536	12,057	20.4	6.9	**21.4**
78	305 NC066: Karoo Hoogland	1,960	745	3,778	6,483	28.6	6.0	**27.5**
79	306 NC067: Khai-Ma	1,628	500	5,367	7,495	15.5	6.8	**23.5**
80	307 NC071: Ubuntu	3,337	1,226	5,433	9,996	34.1	9.0	**26.9**
81	308 NC072: Umsombomvu	5,664	2,601	6,110	14,375	51.9	23.3	**31.5**
82	309 NC073: Emthanjeni	8,255	2,676	11,015	21,946	41.0	14.8	**24.5**
83	310 NC074: Kareeberg	1,762	694	3,207	5,663	35.8	5.3	**28.3**
84	311 NC075: Renosterberg	1,945	1,036	2,563	5,544	48.9	13.7	**34.8**
85	312 NC076: Thembelihle	3,459	586	4,520	8,565	24.2	10.9	**14.5**
86	313 NC077: Siyathemba	3,961	1,278	5,349	10,588	38.8	10.7	**24.4**
87	314 NC078: Siyancuma	9,596	1,864	10,372	21,832	27.8	15.0	**16.3**
88	315 NC081: Mier	1,714	397	1,754	3,865	33.0	5.3	**18.8**
89	316 NC082: Kai !Garib	8,954	2,473	27,059	38,486	17.5	8.9	**21.6**
90	317 NC083: ||Khara Hais	14,697	5,098	26,015	45,810	35.2	13.3	**25.8**
91	318 NC084: !Kheis	3,418	625	5,677	9,720	20.8	6.4	**15.5**
92	319 NC085: Tsantsabane	7,935	2,818	8,712	19,465	41.3	21.3	**26.2**
93	320 NC086: NC086: Kgatelopele	3,784	1,126	5,108	10,018	30.1	21.6	**22.9**
94	321 NC091: Sol Plaatje	47,362	19,777	64,923	132,062	41.5	22.9	**29.5**
95	322 NC092: Dikgatlong	10,166	3,426	9,098	22,690	48.5	24.7	**25.2**
96	323 NC093: Magareng	5,929	2,160	5,287	13,376	51.8	25.3	**26.7**
97	324 NC01B1: Gamagara	3,444	968	6,889	11,301	20.1	20.1	**21.9**
98	81 CBLC1: Ga-Segonyana	3,595	743	4,986	9,324	45.5	12.6	**17.1**
99	87 CBLC7: Phokwane	9,733	2,335	13,455	25,523	35.5	25.9	**19.4**
100	391 NCDMA06: Namaqualand	143	18	444	605	0.7	0.3	**11.2**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) – continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemplo-yment rate (Original)	**Discriminant Model: Unemplo-yment rate**	**MLR Model:Unemplo-yment rate**
		1= Employed	2= Unemplo-yed	3= Inactive				
101	392 NCDMA07: Bo Karoo	460	130	1,678	2,268	10.5	2.3	**22.0**
102	393 NCDMA08: Benede Oranje	1,694	123	4,818	6,635	2.1	4.0	**6.8**
103	394 NCDMA09: Diamondfields	714	137	2,173	3,024	9.0	13.2	**16.1**
104	395 NCDMACB1: Kalahari	1,137	307	2,880	4,324	14.5	20.6	**21.3**
105	401 FS161: Letsemeng	9,249	3,647	13,580	26,476	31.6	29.8	**28.3**
106	402 FS162: Kopanong	13,615	5,113	16,318	35,046	37.8	29.3	**27.3**
107	403 FS163: Mohokare	8,826	3,204	10,684	22,714	35.9	32.6	**26.6**
108	404 FS171: Naledi	6,890	2,623	7,263	16,776	37.1	37.5	**27.6**
109	405 FS172: Mangaung	148,834	77,460	203,032	429,326	40.1	37.1	**34.2**
110	406 FS173: Mantsopa	13,136	5,414	15,862	34,412	35.5	35.9	**29.2**
111	407 FS181: Masilonyana	15,047	8,019	18,855	41,921	42.1	37.6	**34.8**
112	408 FS182: Tokologo	7,789	2,133	10,177	20,099	26.8	34.0	**21.5**
113	409 FS183: Tswelopele	13,433	5,011	13,818	32,262	37.5	38.3	**27.2**
114	410 FS184: Matjhabeng	88,689	66,347	122,348	277,384	46.5	38.9	**42.8**
115	411 FS185: Nala	26,652	11,912	23,728	62,292	48.3	39.5	**30.9**
116	412 FS191: Setsoto	29,639	14,217	33,027	76,883	41.1	41.1	**32.4**
117	413 FS192: Dihlabeng	29,619	14,868	37,749	82,236	39.7	38.4	**33.4**
118	414 FS193: Nketoana	17,357	5,405	15,010	37,772	37.7	38.9	**23.8**
119	415 FS194: Maluti a Phofung	97,803	50,012	71,767	219,582	57.5	28.9	**33.8**
120	416 FS195: Phumelela	13,901	3,481	13,029	30,411	34.4	29.1	**20.0**
121	417 FS201: Moqhaka	36,975	19,324	51,353	107,652	39.9	34.0	**34.3**
122	418 FS203: Ngwathe	30,642	13,757	30,258	74,657	47.1	34.2	**31.0**
123	419 FS204: Metsimaholo	25,685	13,777	39,564	79,026	37.0	37.0	**34.9**
124	420 FS205: Mafube	16,815	5,283	12,586	34,684	44.5	37.3	**23.9**
125	491 FSDMA19: Golden Gate Highlands National Park	28	8	81	117	10.5	18.2	**22.2**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) – continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemployment rate (Original)	**Discriminant Model: Unemployment rate**	**MLR Model:Unemployment rate**
		1= Employed	2= Unemployed	3= Inactive				
126	501 KZ211: Vulamehlo	28,034	6,631	10,930	45,595	65.9	10.6	**19.1**
127	502 KZ212: Umdoni	16,071	5,634	18,591	40,296	42.2	14.8	**26.0**
128	503 KZ213: Umzumbe	66,332	16,046	20,480	102,858	72.0	9.8	**19.5**
129	504 KZ214: uMuziwabantu	29,292	6,976	11,878	48,146	54.5	19.7	**19.2**
130	505 KZ215: Ezingoleni	17,953	4,668	7,298	29,919	61.5	11.6	**20.6**
131	506 KZ216: Hibiscus Coast	56,274	18,833	59,859	134,966	41.8	13.5	**25.1**
132	507 KZ221: uMshwathi	28,640	8,397	26,699	63,736	43.7	13.7	**22.7**
133	508 KZ222: uMngeni	15,160	5,990	28,158	49,308	34.2	17.1	**28.3**
134	509 KZ223: Mooi Mpofana	7,645	4,576	11,583	23,804	44.2	19.6	**37.4**
135	510 KZ224: Impendle	10,853	3,113	3,714	17,680	74.1	9.3	**22.3**
136	511 KZ225: Msunduzi	129,848	64,623	172,839	367,310	48.2	20.4	**33.2**
137	512 KZ226: Mkhambathini	15,952	4,976	14,213	35,141	43.6	13.4	**23.8**
138	513 KZ227: Richmond	16,696	5,224	15,713	37,633	43.4	16.7	**23.8**
139	514 KZ232: Emnambithi/Ladysmith	56,010	22,200	61,049	139,259	49.2	18.9	**28.4**
140	515 KZ233: Indaka	37,571	9,747	9,414	56,732	84.1	15.9	**20.6**
141	516 KZ234: Umtshezi	15,183	7,060	14,713	36,956	54.8	22.1	**31.7**
142	517 KZ235: Okhahlamba	42,653	12,476	19,740	74,869	59.5	15.9	**22.6**
143	518 KZ236: Imbabazane	38,365	12,625	15,085	66,075	73.7	14.8	**24.8**
144	519 KZ241: Endumeni	11,238	5,470	15,549	32,257	46.0	21.8	**32.7**
145	520 KZ242: Nqutu	49,450	13,001	11,957	74,408	81.6	17.5	**20.8**
146	522 KZ244: Msinga	57,908	12,119	9,978	80,005	78.7	21.7	**17.3**
147	523 KZ245: Umvoti	26,974	5,954	19,189	52,117	40.7	16.2	**18.1**
148	524 KZ252: Newcastle	83,236	36,442	88,118	207,796	54.1	21.4	**30.5**
149	525 KZ253: Utrecht	8,074	2,467	8,948	19,489	39.3	12.7	**23.4**
150	526 KZ254: Dannhauser	30,087	10,178	16,454	56,719	67.4	9.7	**25.3**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) – continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemployment rate (Original)	**Discriminant Model: Unemployment rate**	**MLR Model:Unemployment rate**
		1= Employed	2= Unemployed	3= Inactive				
151	527 KZ261: eDumbe	24,854	6,749	12,660	44,263	57.5	12.5	**21.4**
152	528 KZ262: uPhongolo	35,438	9,357	21,456	66,251	48.6	14.6	**20.9**
153	529 KZ263: Abaqulusi	57,809	17,971	33,696	109,476	59.4	15.3	**23.7**
154	530 KZ265: Nongoma	71,435	11,959	14,629	98,023	71.7	11.6	**14.3**
155	531 KZ266: Ulundi	69,168	17,591	23,785	110,544	67.2	15.1	**20.3**
156	532 KZ271: Umhlabuyalingana	49,068	11,914	11,225	72,207	69.2	21.2	**19.5**
157	533 KZ272: Jozini	67,887	10,804	15,302	93,993	60.2	19.9	**13.7**
158	534 KZ273: The Big 5 False Bay	9,463	2,324	4,988	16,775	47.7	15.3	**19.7**
159	535 KZ274: Hlabisa	60,589	14,468	16,455	91,512	73.2	10.7	**19.3**
160	536 KZ275: Mtubatuba	7,689	3,915	10,508	22,112	44.4	24.2	**33.7**
161	537 KZ281: Mbonambi	31,036	10,624	18,027	59,687	58.8	11.5	**25.5**
162	538 KZ282: uMhlathuze	70,823	28,197	86,529	185,549	40.6	18.0	**28.5**
163	539 KZ283: Ntambanana	27,719	6,282	11,894	45,895	58.5	9.5	**18.5**
164	540 KZ284: uMlalazi	71,817	15,887	34,710	122,414	53.2	12.0	**18.1**
165	541 KZ285: Mthonjaneni	14,972	3,895	9,571	28,438	49.6	13.4	**20.6**
166	542 KZ286: Nkandla	46,231	10,411	10,964	67,606	77.4	11.9	**18.4**
167	543 KZ291: eNdondakusuka	33,727	11,462	33,881	79,070	45.1	16.6	**25.4**
168	544 KZ292: KwaDukuza	36,101	12,683	56,589	105,373	33.5	23.6	**26.0**
169	545 KZ293: Ndwedwe	51,225	12,770	20,962	84,957	66.3	10.1	**20.0**
170	546 KZ294: Maphumulo	42,191	8,388	9,831	60,410	75.9	12.6	**16.6**
171	547 KZ5a1: Ingwe	35,712	8,237	11,181	55,130	66.6	14.6	**18.7**
172	548 KZ5a2: Kwa Sani	3,800	964	4,218	8,982	31.0	17.8	**20.2**
173	549 KZ5a3: Matatiele	2,586	1,532	6,360	10,478	30.0	29.2	**37.2**
174	550 KZ5a4: Greater Kokstad	9,730	7,416	20,337	37,483	41.2	27.8	**43.3**
175	551 KZ5a5: Ubuhlebezwe	31,864	9,020	14,737	55,621	61.6	12.5	**22.1**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) – continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemployment rate (Original)	**Discriminant Model: Unemployment rate**	**MLR Model:Unemployment rate**
		1= Employed	2= Unemployed	3= Inactive				
176	572 Durban: Ethekwini	699,484	328,972	1,083,379	2,111,835	43.0	21.0	**32.0**
177	591 KZDMA22: Highmoor/Kamberg Park	-	-	11	11	-	25.0	-
178	592 KZDMA23: Gaints Castle Game Reserve	29	3	424	456	1.2	9.2	**9.4**
179	593 KZDMA27: St Lucia Park	1,262	512	1,106	2,880	46.9	9.3	**28.9**
180	594 KZDMA43: Mkhomazi Wilderness Area	84	19	523	626	7.1	14.2	**18.5**
181	601 NW371: Moretele	49,461	19,387	38,402	107,250	57.7	10.9	**28.2**
182	602 NW372: Madibeng	73,562	39,406	110,115	223,083	41.7	19.8	**34.9**
183	603 NW373: Rustenburg	74,799	45,109	152,299	272,207	32.2	25.1	**37.6**
184	604 NW374: Kgetlengrivier	8,727	2,872	12,253	23,852	30.2	23.5	**24.8**
185	605 NW375: Moses Kotane	64,186	23,684	55,358	143,228	51.1	9.9	**27.0**
186	606 NW381: Setla-Kgobi	36,961	7,365	12,268	56,594	63.9	9.2	**16.6**
187	607 NW382: Tswaing	37,082	5,940	22,632	65,654	34.1	21.6	**13.8**
188	608 NW383: Mafikeng	66,357	30,653	66,841	163,851	49.3	19.6	**31.6**
189	609 NW384: Ditsobotla	41,334	13,395	38,700	93,429	42.5	23.1	**24.5**
190	610 NW385: Zeerust	42,677	11,440	27,007	81,124	53.7	12.2	**21.1**
191	611 NW391: Kagisano	32,118	6,290	14,137	52,545	46.3	13.5	**16.4**
192	612 NW392: Naledi	13,914	4,775	18,177	36,866	35.2	24.3	**25.6**
193	613 NW393: Mamusa	13,699	4,272	11,254	29,225	44.7	30.2	**23.8**
194	614 NW394: Greater Taung	59,560	16,702	24,798	101,060	65.2	10.4	**21.9**
195	615 NW395: Molopo	3,268	301	3,826	7,395	8.5	33.9	**8.4**
196	616 NW396: Lekwa-Teemane	10,763	4,208	12,391	27,362	42.8	27.1	**28.1**
197	617 NW401: Ventersdorp	12,414	3,049	11,946	27,409	33.0	23.3	**19.7**
198	618 NW402: Potchefstroom	28,357	13,169	47,962	89,488	37.1	24.2	**31.7**
199	619 NW403: City Council of Klerksdorp	80,293	45,228	123,925	249,446	40.0	30.4	**36.0**
200	620 NW404: Maquassi Hills	20,557	6,326	17,402	44,285	41.8	30.5	**23.5**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) – continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemployment rate (Original)	**Discriminant Model: Unemployment rate**	**MLR Model:Unemployment rate**
		1= Employed	2= Unemployed	3= Inactive				
201	621 NW1a1: Moshaweng	30,446	4,706	8,269	43,421	60.4	8.6	**13.4**
202	81 CBLC1: Ga-Segonyana	21,464	6,298	15,048	42,810	45.5	14.3	**22.7**
203	87 CBLC7: Phokwane	15,628	5,109	18,143	38,880	35.5	32.4	**24.6**
204	88 CBLC8: Merafong City	36,648	25,496	96,587	158,731	28.1	31.0	**41.0**
205	691 NWDMA37: Pilansberg National Park	43	4	179	226	5.6	5.5	**8.5**
206	701 GT411: Mogale City	52,628	33,078	121,954	207,660	34.1	27.7	**38.6**
207	702 GT412: Randfontein	23,890	14,842	53,136	91,868	36.1	26.9	**38.3**
208	703 GT414: Westonaria	18,797	16,113	47,865	82,775	31.9	33.0	**46.2**
209	704 GT421: Emfuleni	150,299	99,549	211,438	461,286	47.2	36.7	**39.8**
210	705 GT422: Midvaal	11,928	5,421	28,982	46,331	22.9	25.4	**31.3**
211	706 GT423: Lesedi	15,908	7,426	25,005	48,339	35.0	28.5	**31.8**
212	707 GT02b1: Nokeng tsa Taemane	8,778	4,449	23,351	36,578	26.4	24.3	**33.6**
213	773 East Rand: Ekurhuleni Metro	455,277	346,923	985,539	1,787,739	40.4	30.9	**43.3**
214	774 City of Johannesburg Metro	570,804	410,871	1,386,888	2,368,563	37.4	29.6	**41.9**
215	776 Pretoria: City of Tshwane Metro	402,172	199,055	809,602	1,410,829	31.9	26.9	**33.1**
216	82 CBLC2: Kungwini	20,599	10,318	42,482	73,399	29.2	30.3	**33.4**
217	791 GTDMA41: West Rand	926	297	2,949	4,172	13.1	20.1	**24.3**
218	801 MP301: Albert Luthuli	58,190	15,760	29,371	103,321	52.2	20.6	**21.3**
219	802 MP302: Msukaligwa	29,791	11,719	34,794	76,304	38.2	28.9	**28.2**
220	803 MP303: Mkhondo	36,644	13,200	30,981	80,825	45.8	21.5	**26.5**
221	804 MP304: Seme	22,171	7,347	16,127	45,645	50.6	24.8	**24.9**
222	805 MP305: Lekwa	23,408	9,878	32,352	65,638	36.5	28.2	**29.7**
223	806 MP306: Dipaleseng	9,944	4,352	9,775	24,071	45.7	32.3	**30.4**
224	807 MP307: Govan Mbeki Municipality	46,241	28,848	75,650	150,739	39.8	30.9	**38.4**
225	808 MP311: Delmas	13,067	6,683	16,360	36,110	42.5	31.0	**33.8**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) – continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemployment rate (Original)	**Discriminant Model: Unemployment rate**	**MLR Model:Unemployment rate**
		1= Employed	2= Unemployed	3= Inactive				
226	809 MP312: Emalahleni	58,673	35,170	96,436	190,279	38.4	30.5	**37.5**
227	810 MP313: Middelburg	29,702	15,975	50,088	95,765	35.4	28.4	**35.0**
228	811 MP314: Highlands	10,012	3,547	13,852	27,411	30.0	27.5	**26.2**
229	812 MP315: Thembisile	78,513	24,184	48,986	151,683	51.2	21.7	**23.6**
230	813 MP316: Dr JS Moroka	75,399	22,639	38,258	136,296	60.7	15.0	**23.1**
231	814 MP321: Thaba Chweu	16,302	6,690	31,681	54,673	25.1	26.9	**29.1**
232	815 MP322: Mbombela	115,708	46,877	130,784	293,369	37.7	25.7	**28.8**
233	816 MP323: Umjindi	10,340	4,280	21,056	35,676	26.3	29.3	**29.3**
234	817 MP324: Nkomazi	91,222	28,438	65,882	185,542	41.5	25.2	**23.8**
235	83 CBLC3: Greater Marble Hall	36,116	9,313	21,459	66,888	44.8	15.1	**20.5**
236	84 CBLC4: Greater Groblersdal	72,274	16,877	30,138	119,289	54.3	11.8	**18.9**
237	85 CBLC5: Greater Tubatse	88,441	24,067	31,825	144,333	61.5	13.9	**21.4**
238	86 CBLC6: Bushbuckridge	155,795	51,764	59,092	266,651	63.5	22.5	**24.9**
239	893 MPDMA32: Lowveld	72	58	252	382	25.4	34.8	**44.6**
240	996 CBDMA4: Kruger Park	230	143	2,816	3,189	3.9	34.1	**38.3**
241	901 NP331: Greater Giyani	72,272	28,209	27,564	128,045	60.4	33.8	**28.1**
242	902 NP332: Greater Letaba	69,246	17,124	34,327	120,697	42.1	22.7	**19.8**
243	903 NP333: Greater Tzaneen	101,723	37,990	82,462	222,175	42.4	21.1	**27.2**
244	904 NP334: Ba-Phalaborwa	30,636	15,612	36,503	82,751	40.4	23.8	**33.8**
245	905 NP341: Musina	5,590	3,770	17,291	26,651	25.0	38.6	**40.3**
246	906 NP342: Mutale	22,104	9,109	8,919	40,132	57.8	27.5	**29.2**
247	907 NP343: Thulamela	162,961	74,112	77,380	314,453	59.6	28.4	**31.3**
248	908 NP344: Makhado	139,091	52,968	78,893	270,952	49.8	22.4	**27.6**
249	909 NP351: Blouberg	53,806	9,539	16,238	79,583	52.6	15.9	**15.1**
250	910 NP352: Aganang	46,561	10,657	16,426	73,644	59.9	7.2	**18.6**

Table A1: Estimation of unemployment by Municipalities (Census direct measure, Discriminant estimate and Multinomial Logistic estimate) – continued								
Municipalities		Predicted Response Category (MLR)			Total	CENSUS: Unemplo-yment rate (Original)	**Discriminant Model: Unemplo-yment rate**	**MLR Model:Unemplo-yment rate**
		1= Employed	2= Unemplo-yed	3= Inactive				
251	911 NP353: Molemole	30,682	8,070	21,121	59,873	39.0	17.2	**20.8**
252	912 NP354: Polokwane	126,798	48,932	124,974	300,704	41.5	22.6	**27.8**
253	913 NP355: Lepele-Nkumpi	67,015	20,933	31,554	119,502	60.7	17.0	**23.8**
254	914 NP361: Thabazimbi	11,034	5,949	28,884	45,867	21.2	29.0	**35.0**
255	915 NP362: Lephalale	22,467	4,225	32,107	58,799	15.5	17.0	**15.8**
256	916 NP364: Mookgopong	5,685	1,828	12,515	20,028	19.1	24.7	**24.3**
257	917 NP365: Modimolle	14,851	4,769	26,031	45,651	23.5	28.8	**24.3**
258	918 NP366: Bela-Bela	10,731	4,542	17,076	32,349	32.6	27.6	**29.7**
259	919 NP367: Mogalakwena	88,563	23,165	53,241	164,969	47.6	20.0	**20.7**
260	920 NP03A2: Makhuduthamaga	88,891	23,849	23,359	136,099	75.1	8.7	**21.2**
261	921 NP03A3: Fetakgomo	31,841	7,560	9,291	48,692	68.3	8.2	**19.2**
262	922 NP04A1: Maruleng	26,561	7,715	19,043	53,319	40.1	19.6	**22.5**

ANNEXURE II: Extract from Census 2001 questionnaire

AII.1. Census demographic questions

INFORMATION FOR PERSONS IN THE HOUSEHOLD - ASK OF EVERYONE

AGE (P-02)

What is (the person's) date of birth and age in completed years?

If date of birth not known give (the person's) age in completed years. If age not known give an estimate of age. Date of birth is recorded as DD/MM/YYYY. DD is for day / MM is for month and / YYYY is for year. For example, if the person was born on 7 September 1963, write for the day DD, for the month MM, and for the year YYYY. For babies less than one year write for age, and for person 7 years and 10 months old write for age.

SEX (P-03)

Is (the person) male or female?

M = Male

F = Female

Dot the appropriate box.

MARITAL STATUS (P-05)

What is (the person's) PRESENT marital status?

1 = Married civil/religious

2 = Married traditional/customary

3 = Polygamous marriage

4 = Living together like married partners

5 = Never married

6 = Widower/widow

7 = Separated

8 = Divorced

Write only one code per person in the box. If both civil/religious and traditional marriage, indicate civil/religious. If categories 5-8 go to (P-06).

POPULATION GROUP (P-06)

How would (the person) describe him/ herself in terms of population group?

1 = Black African

2 = Coloured

3 = Indian or Asian

4 = White

5 = Other (specify)

INCOME CATEGORY (P-22)

What is the income category that best describes the gross income of (this person) before tax?

Choose from the table below the code that corresponds to the income level.

	MONTHLY	ANNUAL
1	No income	No income
2	R 1 – R 400	R 1 – R 4 800
3	R 401 – R 800	R 4 801 – R 9 600
4	R 801 – R 1 600	R 9 601 – R 19 200
5	R 1 601 – R 3 200	R 19 201 – R 38 400
6	R 3 201 – R 6 400	R 38 401 – R 76 800
7	R 6 401 – R 12 800	R 76 801 – R 153 600
8	R 12 801 – R 25 600	R 153 601 – R 307 200
9	R 25 601 – R 51 200	R 307 201 – R 614 400
10	R 51 201 – R 102 400	R 614 401 – R 1 228 800
11	R 102 401 – R 204 800	R 1 228 801 – R 2 457 600
12	R 204 801 or more	R 2 457 601 or more

ALL AGED 5 YEARS OR MORE

LEVEL OF EDUCATION (P-17)

What is the highest level of education that (the person) has completed?

99= No schooling
00= Grade 0
01= Grade 1/Sub A
02= Grade 2/Sub B
03= Grade 3/Standard 1
04= Grade 4/Standard 2
05= Grade 5/Standard 3
06= Grade 6/Standard 4
07= Grade 7/Standard 5
08= Grade 8/Standard 6/ Form 1
09= Grade 9/Standard 7/ Form 2
10= Grade 10/Standard 8/ Form 3/NTCI
11= Grade 11/Standard 9/ Form 4/NTCII
12= Grade 12/Standard10/ Form 5/Matric./NTCIII
13= Certificate with less than Grade 12
14= Diploma with less than Grade 12
15= Certificate with Grade 12
16= Diploma with Grade 12
17= Bachelors Degree
18= Bachelors Degree and Diploma
19= Honours degree
20= Higher Degree (Masters, Doctorate)
21= Other
22= Don't know

AII.2. Census Employment related questions

ASK FOR ALL PERSONS AGED 10 YEARS AND OLDER (BORN BEFORE 10 OCTOBER 1991)

ANY WORK IN THE 7 DAYS BEFORE 10 OCTOBER

(P-18) In the SEVEN DAYS before 10 October did (the person) do any work for PAY (in cash or in kind) PROFIT or FAMILY GAIN, for one hour or more?

1 = Yes: formal registered (non-farming)
2 = Yes: informal unregistered (non-farming)
3 = Yes: farming
4 = Yes: has work but was temporarily absent
5 = No: did not have work

If YES go to P-19

<u>DID NOT HAVE ANY WORK</u>

ACTIVE STEPS (P-18b)

If NO to P-18

In the PAST FOUR WEEKS before 10 October has (the person) taken active steps to find employment?

Y = Yes

N = No

For example, (the person) went to visit factories or other employment places, placed or answered advertisements, looked for land or a building or equipment to start own business or farm.

REASON WHY NOT WORKING (P-18a)

If NO to P-18

What is the main reason why (the person) did not have work in the seven days before 10 October?

1 = Scholar or student

2 = Home-maker or housewife

3 = Pensioner or retired person/ too old to work

4 = Unable to work due to illness or disability

5 = Seasonal worker not working presently

6 = Does not choose to work

7 = Could not find work

If more than one reason, write the code of the MAIN (most important) reason.

AVAILABILITY (P-18c)

If NO to P-18 If offered work, how soon could (the person) start?

1 = Within one week

2 = More than 1 week, up to 2 weeks

3 = More than 2 weeks, up to 4 weeks

4 = Some time after 4 weeks

5 = Does not choose to work

Go to P-20

ASK FOR ALL PERSONS AGED 10 YEARS AND OLDER (BORN BEFORE 10 OCTOBER 1991) WHO HAD WORK

WORK STATUS (P-19)

If YES to P-18

How can one best describe (the person's) main activity or work status?

1 = Paid employee

2 = Paid family worker

3 = Self-employed

4 = Employer

5 = Unpaid family worker

6 = Other (specify)

BUSINESS/COMPANY NAME (P-19a)

If YES to P-18

What is the FULL name of the business/company or organisation for whom (the person) works?

If the person works for him/ herself, and the business does not have a name, write SELF in the appropriate row. If doing PAID domestic work in a private household, write DOMESTIC SERVICE. Use CAPITAL LETTERS only.

COMPANY/BUSINESS ACTIVITY (P-19b)

If YES to P-18

What does the business do (main economic activity)?

Write the MAIN INDUSTRY, economic activity, product or service of (the person's) employer or company. For example, gold mining, road construction, supermarket, police service, healthcare, hairdressing, banking. OR Write the activity of the person if self-employed. For example, subsistence farming. If doing PAID domestic work in a private household, write DOMESTIC SERVICE. Use CAPITAL LETTERS only.

OCCUPATION (P-19c)

If YES to P-18

What is the main occupation of (the person) in this workplace?

> *Occupation refers to the type of work (the person) performed in the seven days before 10 October. Use two or more words. For example, street trader, cattle farmer, primary school teacher, domestic worker, fruit vendor, truck driver, warehouse manager, filing clerk, etc. Use CAPITAL LETTERS only.*

HOURS WORKED (P-19d)

If YES to P-18

How many hours did (the person) work in the seven days before 10 October?

> *If (the person) was absent from work those seven days, but usually works, write the number of hours s/he usually works.*

PLACE OF WORK (P-19e) (P-19f)

If YES to P-18

Does (the person) work in the same sub-place in which s/he usually lives?

> Y = Yes
>
> N = No
>
> *Dot the appropriate box.*

If NO, where is this place of work?

> *If NOT the same place, write PROVINCE, MAIN PLACE (city, town, tribal area, administrative area) and SUB-PLACE (suburb, ward, village, farm, informal settlement). If another country, write the name of the country in the boxes below.*

ANNEXURE III: Extract from LFS 4 – September 2001 questionnaire

AIII.1. LFS demographic questions

FLAP This section covers particulars of each person in the household

The following information must be obtained for every person who has stayed in this household for at least four nights on average per week during the last four weeks.

Do not forget babies. *If there are more than 10 persons in the household, use a second questionnaire.*

B Has ...... stayed here for at least four nights on average per week during the last four weeks?

1 = Yes
2 = No — End of questions for this person

C Is ...... a male or a female?

1 = Male
2 = Female

D **How old is ......?**

(In completed years - In figures only) Less than 1 year = 00

E **What population group does ...... belong to?**

1 = African/Black
2 = Coloured
3 = Indian/Asian
4 = White
5 = Other, specify…………

1.1.a What is ……'s present marital status?

1 = Married or living together as husband and wife

2 = Widow/Widower

3 = Divorced or Separated

4 = Never married

1.3.a What is the highest level of education that …… has completed?

00 = No schooling	12 = Grade 11/Standard 9/Form 4
01 = Grade 0	13 = Grade 12/Standard 10/Form 5/Matric
02 = Sub A/Grade 1	14 = NTC l
03 = Sub B/Grade 2	15 = NTC II
04 = Grade 3/Standard 1	16 = NTC III
05 = Grade 4/Standard 2	17 = Diploma/certificate with less than Grade 12/Std 10
06 = Grade 5/Standard 3	18 = Diploma/certificate with Grade 12/Std 10
07 = Grade 6/Standard 4	19 = Degree
08 = Grade 7/Standard 5	20 = Postgraduate degree or diploma
09 = Grade 8/Standard 6/Form 1	21 = Other, specify in column
10 = Grade 9/Standard 7/Form 2	22 = Don't know
11 = Grade 10/Standard 8/Form 3	

Diplomas or certificates should be of at least six months study duration full time (or equivalent). If code 17-20 → Go to Q 1.3.b, If other code → Go to Q 1.4

***4.15.* What is ……'s total salary/pay at his/her main job?**

Show the categories. Make sure the respondent points at *the **correct income column** (weekly, monthly, annually) on **prompt card 3** and mark the applicable code.*

	Weekly	**Monthly**	**Annually**
01	NONE	NONE	NONE
02	R1 - R46	R1 - R200	R1 - R2 400
03	R47 - R115	R201 - R500	R2 401 - R6 000
04	R116 - R231	R501 – R1 000	R6 001 - R12 000
05	R232 - R346	R1 001 - R1 500	R12 001 - R18 000
06	R347 - R577	R1 501 - R2 500	**R18 001 - R30 000**
07	R578 - R808	R2 501 - R3 500	R30 001 - R42 000
08	R809 - R1 039	R3 501 - R4 500	R42 001 - R54 000
09	R1 040 - R1 386	R4 501 - R6 000	**R54 001 - R72 000**
10	R1 387 - R1 848	R6 001 - R8 000	R72 001 - R96 000
11	R1 849 - R2 540	R8 001 - R11 000	R96 001 - R132 000
12	R2 541 - R3 695	R11 001 - R16 000	**R132 001 - R192 000**
13	R3 696 - R6 928	R16 001 - R30 000	R192 001 - R360 000
14	R6 929 OR MORE	R30 001 OR MORE	R360 001 OR MORE
15	DON'T KNOW	DON'T KNOW	**Don't know**
16	REFUSE	REFUSE	REFUSE

AIII.2. LFS Employment related questions

<u>SECTION 3 This section covers unemployment and non-economic activities</u>

Ask for all household members aged 15 and above who did not work and were not absent from work (i.e. for those whose answer on Q 2.2 = 2)

Read out: **Now I am going to ask some questions about whether you (……) wanted and were (was) available for any of the types of work mentioned earlier**

3.1. **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 3.8
02 =Lack of skills or qualifications for available jobs
03 =Scholar or student and prefers not to work
04 =Housewife/homemaker and prefers not to work
05 =Retired and prefers not to seek formal work
06 =Illness, invalid, disabled or unable to work (handicapped)
07 =Too young or too old to work
08 =Seasonal worker, e.g. fruit picker, wool-shearer
09 =Cannot find suitable work (salary, location of work or conditions not satisfactory)
10 =Contract worker, e.g. mine worker resting according to contract
11 =Recently retrenched
12 =Other reason

3.2. If a suitable job is offered, will …… accept it?

1 = Yes
2 = No
3 = Don't know →

Go to Q 3.8

3.3. 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

3.4. During the past four weeks, has …… taken any action

a) to look for any kind of work
b) to start any kind of business

If "No" to both a) and b) → *Go to Q 3.7*

3.5. In the past four weeks, what has …… done to look for work or to start a business?

Give only one answer, the main one

1 = Waited/registered at employment agency/trade union
2 = Enquired at workplaces, farms, factories or called on other possible employers
3 = Placed/answered advertisement(s)
4 = Sought assistance from relatives or friends
5 = Looked for land, building, equipment or applied for permit to start own business or farming
6 = Sought/underwent training
7 = Waited at the street side where casual workers are found
8 = Other
9 = Don't know

3.6. How long has …… been trying to find work or start a business?

1 = Less than a month

2 = 1 month to less than 2 months

3 = 2 months to less than 3 months

4 = 3 months to less than 4 months

5 = 4 months to less than 6 months

6 = 6 months to less than 1 year

7 = 1 year to less than 3 years

8 = 3 years or more

9 = Don't know

→ *Go to Q 3.8*

3.7. *If "No" to both Q 3.4.a and b (has not been looking for work or trying to start a business in the past four weeks)*

What was the main reason why …… did not try to find work or start a business in the past four weeks?

01 = Has been temporarily laid off work

02 = Ill health/Injury/Physical disability

03 = Pregnancy

04 = Family considerations/Child care

05 = Undergoing training to help find work

06 = No jobs available in the area

07 = Lack of money to pay for transport to look for work

08 = Unable to find work requiring his/her skills

09 = Lost hope of finding any kind of work

10 = No transport available

11 = Other reason

3.8. Has …… ever worked before?

1 = Yes

2 = No

→ *Go to Q 3.12*

3.9. How long ago was it since …… last worked?

01 = 1 week - less than 1 month
02 = 1 month - less than 2 months
03 = 2 months - less than 3 months
04 = 3 months - less than 4 months
05 = 4 months - less than 5 months
06 = 5 months - less than 6 months
07 = 6 months - less than 1 year
08 = 1 year - less than 2 years
09 = 2 years - less than 3 years
10 = 3 years or more
11 = Don't know

3.10a. What kind of work did …… do in his/her last job?

Give occupation or job title. Work includes all the activities mentioned earlier. Record at least two words: Car sales person, Office cleaner, Vegetable farmer, Primary School teacher, etc. For agricultural work on own/family farm/plot, state whether for own use or for sale mostly.

3.10b. What were ……'s main tasks or duties in this job?

Examples: Selling fruit, repairing watches, keeping accounts, feeding and watering cattle, teaching children.

<u>SECTION 4 This section covers main work activity in the last seven days</u>

Ask for all persons 15 years and over who were working or absent from work in the last seven days.

Read out: **The next several questions refer to your (……'s) main job or activity. That is the one where you (he/she) <u>usually</u> work (-s) the most hours per week, even if you (he/she) were (was) absent the last seven days.**

4.1.a : *Read out*:

You said …… was doing these activities during the last seven days (or was temporarily absent). *Refer to Q 2.1*

What kind of work did …… do in his/her <u>main</u> job during the last seven days (or usually does, even if he/she was absent in the last seven days)**?**

Give occupation or job title.

Work includes all the activities mentioned earlier Record at least two words: Car sales person, Office cleaner, Vegetable farmer, Primary school teacher, etc.

For agricultural work on own/family farm/plot, state whether for <u>own</u> use or for <u>sale</u> mostly.

4.1.b ***What were ……'s <u>main</u> tasks or duties in this job?***

EXAMPLES: SELLING FRUIT, REPAIRING WATCHES, KEEPING ACCOUNTS, FEEDING AND WATERING CATTLE.

4.2.a. What is the name of ……'s place of work?

For government or large organisations, give the name of the establishment and branch or division: e.g. Education Dept – Rapele Primary School; ABC Gold Mining, Maintenance Div. Write 'Own house' or 'No fixed location', if relevant.

4.2.b. What are the main goods and services produced at ……'s place of work? What are its main functions?

Examples: Repairing cars, Selling commercial real estate, Sell food wholesale to restaurants, Retail clothing shop, Manufacture electrical appliances, Bar/ restaurant, Primary Education, Delivering newspapers to homes.

4.3. In ……'s main work was he/she

1 = Working for someone else for pay? *Payment in cash kind or accommodation.*

Category 1 includes all employees: Full-time, part-time, casual work, piecework, except private household work. → *Go to Q 4.4*

2 = Working for one or more private households as a DOMESTIC EMPLOYEE, GARDENER OR SECURITY GUARD? *Payment in cash, kind or accommodation.*

→ Go to Q 4.4

3 = Working on his/her own or on a small household farm/plot or collecting natural products from the **forest or sea**? ***→ Go to Q 4.14***

4 = Working on his/her own or with a partner, in any type of business (including commercial farms)? ***→ Go to Q 4.14***

5 = Helping without pay in a household business? ***→ Go to Q 4.14***

4.4. Does …… work for

1 = One employer

2 = More than one employer

4.5. When did …… start working with the (main) employer mentioned above (firm, institution or private individual)? Give year and month.

*State year in **four** figures, e.g. 1998* **Year**

*State month in **two** figures, e.g. 08 for August* **Month**