Computation of Optimal Estimates in a Complex Survey Sample Design

by

THANYANI ALPHEUS MAREMBA

DISSERTATION

Submitted in fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

PROMOTER: Prof. M. Lesaoana
CO-PROMOTER: Dr. P. Nyamugure

2019

Declaration

I, **Thanyani Alpheus Maremba**, the undersigned, declare that the dissertation hereby submitted to the University of Limpopo for the degree Master of Science in Statistics, has not been previously submitted by me at this or any other institution. This is my work in design and execution under the academic supervision and guidance of Prof ‘Maseka Lesaoana and Mr Philimon Nyamugure.

Signature...

Date: 04 April 2019

Abstract

This research study has demonstrated the complexity involved in complex survey sample design (CSSD). Furthermore the study has proposed methods to account for each step taken in sampling and at the estimation stage using the theory of survey sampling, CSSD-based case studies and practical implementation based on census attributes. CSSD methods are designed to improve statistical efficiency, reduce costs and improve precision for sub-group analyses relative to simple random sample (SRS). They are commonly used by statistical agencies as well as development and aid organisations. CSSDs provide one of the most challenging fields for applying a statistical methodology. Researchers encounter a vast diversity of unique practical problems in the course of studying populations. These include, inter alia: non-sampling errors, specific population structures, contaminated distributions of study variables, non-satisfactory sample sizes, incorporation of the auxiliary information available on many levels, simultaneous estimation of characteristics in various sub-populations, integration of data from many waves or phases of the survey and incompletely specified sampling procedures accompanying published data.

While the study has not exhausted all the available real-life scenarios, it has outlined potential problems illustrated using examples and suggested appropriate approaches at each stage. Dealing with the attributes of CSSDs mentioned above brings about the need for formulating sophisticated statistical procedures dedicated to specific conditions of a sample survey. CSSD methodologies give birth to a wide variety of approaches, methodologies and procedures of borrowing the strength from virtually all branches of statistics. The application of various statistical methods from sample design to weighting and estimation ensures that the optimal estimates of a population and various domains are obtained from the sample data.

CSSDs are probability sampling methodologies from which inferences are drawn about the population. The methods used in the process of producing estimates include adjustment for unequal probability of selection (resulting from stratification, clustering and probability proportional to size (PPS), non-response adjustments and benchmarking to auxiliary totals. When estimates of survey totals, means and proportions are computed using various methods, results do not differ. The latter applies when estimates are calculated for planned domains that are taken into account in sample design and benchmarking. In contrast, when the measures of precision such as standard errors and coefficient of variation are produced, they yield different results depending on the extent to which the design information is incorporated during estimation.

The literature has revealed that most statistical computer packages assume SRS design in estimating variances. The replication method was used to calculate measures of precision which take into account all the sampling parameters and weighting adjustments computed in the CSSD process. The creation of replicate weights and estimation of variances were done using WesVar, a statistical computer package capable of producing statistical inference from data collected through CSSD methods.

Keywords: Complex sampling, Survey design, Probability sampling, Probability proportional to size, Stratification, Area sampling, Cluster sampling.

Dedication

This dissertation is dedicated to those who went ahead of me, my late parents Mr P.K and Mrs M.M Maremba for laying educational foundation in my life.

Acknowledgements

Professor ‘Maseka Lesaoana believed in me and guided me from start to the end of this dissertation. Together with Mr Philimon Nyamugure we formed a formidable team that saw me through this research study. I have always remained accountable for my studies to Mr Motale Phirwa who encouraged me to stay focused until completion of the project. To all, I would like to acknowledge the contributions you have made towards my studies and I will remain grateful.

Basis for this dissertation

Earlier conference papers

- Improving Sub-Provincial Estimates for South African Labour Force Survey, Pretoria 2008.
- Synthetic Estimation as a Method to Improve Sub-Provincial Estimation in Statistics South Africa, Pretoria 2010.
- Finding Optimal Calibration Weights for Household Surveys in Statistics South Africa Using Equalization Constraints. South Africa, Potchefstroom 2010.
- Small Area Estimation for South Africa Quarterly Labour Force Survey, Patna, India 2010. Co-Author.
- Sample Design to Optimise the Estimation of Small Micro and Medium Enterprise Owners and their Characteristics. Presented in SASA2015 at the University of Pretoria, December 2015.
- Methods of Analysing Complex Sample Survey Data: Importance of Taking Sample Design into Account, SASA2016, Cape Town, December 2016.
- Methods of Analysing Complex Sample Survey Data: Importance of Taking Sample Design into Account, WSC2017, Marrakesh, July 2017.

Data used

- South African Census 2011 small area level data used for sampling.
- Simulated survey data from Census 2011 used for weighting and estimation.
- Household surveys data used to obtain response rates for simulation.
- South African community survey 2016 population data used for benchmarking.

Statistical software used

SAS based

- SAS was used for all the data preparations and integration of datasets,
- Sampling procedures were implemented in Base SAS,
- StatMx software from Statistics Canada was used in calibration,
- SAS/IML was used within StatMx,
- SAS FASTCLUS was used for clustering, and
- SAS LP (linear programming) was used in weights optimisation.

SuperCROSS system

- Extraction of census attributes used in sampling, and
- Extraction of person and household data at small area layer (SAL) level.

ArcGIS

- Merging and clustering of sampling areas, and
- Creation of area maps for SALs, place-names and other spatial data.

WesVar

- Creation of replicate weights, and
- Estimation of totals and measures of precision using calibrated replicate weights.

Ethical consideration

Confidentiality

- Data used do not disclose personal identity and location, and
- Reporting from the datasets used is in the form of aggregates and percentages.

Approval and permission

The request for the permission to use the data was made to data owners and the input data earmarked were those that are available to use by public.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

Abbreviation	**Long Form**
BRR	Balanced Repeated Replication
CLFS	Canadian Labour Force Survey
CSSD	Complex Survey Sample Design
CV	Coefficient of Variation
DC	District Council
Deff	Design effect
DU	Dwelling Unit
EA	Enumeration Area
EU	European Union
GHS	General Household Survey
IHLCA	Integrated Household Living Condition
ISR	Inverse Sampling Rate
LFS	Labour Force Survey
LISA	Laboratory of Interdisciplinary Statistical Analysis
MOS	Measure of Size
Msc	Masters of Science
MSE	Mean Square Error

Abbreviation	**Long Form**
MSME	Medium, Small and Micro Enterprises
NAEP	National Assessment for Educational Progress
OECD	Organisation of Economic Co-operation and Development
OHS	October Household Survey
PPS	Probability Proportional to Size
PSU	Primary Sampling Unit
QLFS	Quarterly Labour Force Survey
S2SWR	Stratified Two-Stage With Replacement
SAL	Small Area Layer
SAST	South African State Theatre
SE	Standard Error
SESE	Survey of Employers and Self Employed
SRS	Simple Random Sample
SRSWOR	Simple Random Sample Without Replacement
SRSWR	Simple random sampling with replacement
SSU	Second Stage Unit
StatCan	Statistics Canada
StatsSA	Statistics South Africa
SYS	Systematic
UN	United Nations
UNSD	United Nations Statistics Division

Chapter 1

Introduction

1.1 Defining complex survey sampling design

Complex survey sample design (CSSD) is a probability sample design method developed using sampling procedures such as stratification, clustering, segmentation, probability proportional to size (PPS) selection methods and weighting. CSSD is designed to improve statistical efficiency, reduce costs and improve precision for sub-group analyses relative to simple random sample (SRS) (Heeringa and Liu, 1998). These methods are commonly used by statistical agencies as well as development and aid organisations. Implementing the CSSD to various countries may present comparability challenges due to differences that exist across countries.

The primary sources of population estimates and specific domain estimates in many countries around the world are from censuses, administrative records and other forms of big data, such as, cellphone records and transactions data where customer information is recorded. Not all countries have well-managed administrative records data that are sufficient to provide reliable estimates to be used as the base population. According to the United Nations (UN) recommendations, countries are expected to conduct censuses every 10 years to provide reliable data at lower geographical level. The United Nations Statistics Division (UNSD) indicated that some countries can afford to conduct intercensal large sample surveys that are usually treated as the source of base population in the middle of two census periods (UNSD, 1998). South Africa, for example, conducts the community survey (a large sample survey) to bridge the gap between censuses.

Administrative data in some countries are already used as the main sources of population data. One of the earliest countries to adopt population registers to substitute censuses, is Denmark. Statistics Finland has a well-functioning, coordinated system of statistical registers and Finland has drawn its population and housing census entirely from registers since 1990. Social statistics are produced mainly from the register-based system or from independent person or household interview sample surveys. Both administrative data and survey data must be used to meet all user needs (Törmälehto, 2008).

Surveys are used to meet the user needs that cannot be met with register-based statistics. Total (for example, consumer opinions and preferences, time-use) or partial (for example, household wealth) lack of administrative data on some topics together with timeliness and comparability within the European Statistical System usually dictate the relevance of a survey. All the main surveys at Statistics Finland are either European Union (EU) regulated or EU harmonised, although they usually have some national content as well (Törmälehto, 2008).

In countries such as South Africa and other African states vital registration are often under-reported or reported late after occurrence of events such as births, deaths, marriages and divorces. Udjo (2017) estimated the completeness of 2011 death registration in South Africa using post-enumeration survey weighted 2011 Census figures as denominator. The completeness of deaths registration, according to Udjo (2017), was estimated at 80.9 percent for both males and females. Reasons for under-reporting vary from late registrations to no vital registration records completely. On a more positive side, the study by Fox et al. (2010) concluded that mortality was substantially under-estimated among patients lost from a South African HIV treatment programme, despite limited active tracing. They found that linking programme data to vital registration systems could provide more accurate assessments of programme effectiveness and target lost patients most at risk for mortality. There is clearly a place for use of administrative records in providing information needed to address health and socio-economic problems, either directly or as a complement of other surveys.

Even with the availability of reliable administrative sources, census or mini-census data, the demand for data at national-level is so high that data are required more frequently for subjects that may not be fully covered in census (Kalton, 1983). Subsequently, in the 1930s, the sample surveys emerged as the alternative source of data (Kalton, 1983). Advantages of sample surveys, as compared to censuses, were that they are less expensive, results can be made available quicker and more frequently, they tend to yield estimates that are more reliable due to the fact that the sampling error is measurable, and furthermore, in-depth questions about the subject can be included in these kinds of specialised surveys.

Most of the sample surveys carried out by major organisations are probability samples. Probability sampling surveys are samples where each population unit has a known probability of inclusion in the sample. Weights generated in probability sampling are numbers of population units that are represented by a sample unit. Variance estimation depends on sampling design, while with missing data, models are needed to produce estimates (Heeringa and Liu, 1998). Sample surveys are widely used to provide estimates of population quantities such as totals and means. Many countries have set-up centralised statistical agencies that are responsible for the collection of statistical information about the state of the nation. This important statistical information includes national characteristics such as the demography, agriculture, labour force, health and living conditions, and trade. Government agencies increasingly use these results to formulate policies and allocate government funds. Särndal et al. (2003) attribute much of the developments in sample survey methodology to government statistical offices.

The use of sample surveys has also been extended to the academia and private business sectors, with the latter increasingly using the results of surveys to influence business decisions, which according to Rao (2003), is attributed to their heavy reliance on local conditions. The academic sector uses sample surveys in many research areas. For example, the statistical collaborators at the Laboratory of Interdisciplinary Statistical Analysis (LISA) at Virginia Tech have worked with sample survey data collected by graduate students and faculty from a large range of disciplines, including education, engineering, public and international affairs, and agriculture (Vance and Pruitt, 2016).

Due to increased reliance on sample survey data, the quality of the data collected from these surveys and the estimates produced thereof have increased in popularity. Sources of errors that are associated with sample surveys are classified as sampling and non-sampling errors. However, for the purpose of this study, the emphasis is on the sampling errors, how they are measured and how to minimise them through sample design, while at the same time producing optimal estimates.

The theory of survey sampling has advanced over the years, having gone beyond the basic SRS. The sample designs used in practice incorporate geographical differences through clustering, taking into account the population size of different areas, as well as other known characteristics such as socio-economic attributes. In this study, these sample design approaches that involve multi-stage sampling, clustering, stratification, application of different sampling methods in different clusters, and PPS and that also take into account unequal probabilities of selection, are referred to as "complex survey sample designs".

The history of CSSD in South Africa is just over 20 years when the October Household Survey (OHS) was conducted in 1993 which did not cover South Africa entirely as it was conducted by the Central Statistical Services (CSS, 1993). In 1994 the first nation-wide OHS was conducted which included the former homelands or bantustans formed during the apartheid government that ruled prior 1994, those were, Transkei, Bophuthatshwana, Venda and Ciskei (CSS, 1994). Several series of the OHS were conducted annually up to and including 1999. OHS surveys were then replaced by separate household-based surveys namely, bi-annual Labour Force Survey (LFS), conducted by Statistics South Africa (StatsSA, 2001), General Households Survey (GHS), (StatsSA, 2003), Income and Expenditure Survey (IES), (StatsSA, 2002) and later with other surveys such as Victims of Crime Survey (VOCS).

In the past ten-year period the number of household-based surveys in South Africa grew to cater for both domestic and international needs. The new surveys addressed subjects that included: domestic tourism, transport, crime, time-use, and so forth. After every democratic census, that is, 1996, 2001 and 2011, the master sample was drawn to support the surveys between census periods. The CSSD methods are used by Statistics South Africa to select a sample of households for each household-based survey. Sampling for business surveys is based on stratified simple random sampling where *Neyman* allocation method is used to allocate samples to strata (StatsSA, 2017). However, the scope of this dissertation is limited to application of CSSD to household-based surveys.

The subject of CSSD is important to study in developing countries such as South Africa, in order to meet the needs for data required to achieve the sustainable development goals. The sample design methods need to be advanced through modelling, in order to produce data for lower geography levels at a low cost. The methods of estimating variances are applied in the South African Quarterly Labour Force Survey (QLFS), which uses replication method (StatsSA, 2008). Promoting the reporting of measures of precision with every survey results is a good practice which builds confidence on the use of survey data. The study of CSSD methods is also important to inform and educate data users to consider the procedure involved in CSSD (as compared to SRS), during the analysis of survey data obtained from CSSD.

1.2 The research problem

Sample surveys provide one of the most challenging fields for applying the statistical methodology. Researchers are confronted with a vast diversity of unique practical problems encountered in the course of studying populations using sample surveys. The problems include, but are not limited to: non-sampling errors, specific population structures, contaminated distributions of study variables, non-satisfactory sample sizes, incorporation of the auxiliary information available on many levels, simultaneous estimation of characteristics in various sub-populations, integration of data from many waves or phases of the survey and incompletely specified sampling procedures (Wywiał and Żądło, 2012).

Creating reliable estimates for small areas, where sample size is substantially lower than for the whole country (as an example, for counties in Poland, *Nomenclature of Territorial Units for Statistics 4*) is a great challenge for statisticians (Kubacki and Jędrzejczak, 2012). Rao (2003) proposed several methods to provide estimates for small area domains that were not catered for in the design. Dealing with such conditions brings about the need for formulating sophisticated statistical procedures dedicated to specific conditions of a sample survey.

A variety of sampling problems gives birth to a multitude of approaches, methodologies and procedures of borrowing the strength from virtually all branches of statistics (Wywiał and Żądło, 2012). Examples of statistical methods from other branches of statistics include cluster analysis, regression, linear programming and estimation. The application of various statistical methods from sample design to weighting and estimation ensures that the optimal estimates of a population are obtained from the sample data.

Some of the common challenges that affect complex sample weighting and estimation are already well researched. The challenges of CSSD include incorporation of the auxiliary information available for many levels, estimation of domains with smaller sample sizes, potential bias created by non-response in sample surveys, handling the impact of outdated sampling frames and the adjustment for changes that occur between sampling and data collection (Wywiał and Żądło, 2012). While several methods to account for challenges of CSSD are employed in the research study, the emphasis is on implementation of integrated weighting. The integrated weighting method is discussed in Section 2.3.4 under bechmarking methods.

Sample design in South Africa similar to other countries, faces challenges from province to province. Examples of some of the challenges are on sample allocation for the Northern Cape and the Western Cape provinces. The two provinces have unique characteristics. Northern Cape is a large and sparsely populated province. The population of the Northern Cape province is the lowest in South Africa while it is the biggest in terms of land. If proportional allocation was to be used to draw the sample, the resulting sample sizes would be smallest and could be insufficient to cater for certain domain such as estimation of some variables by demographic variables within a province. The sample for South African labour force survey used square-root allocation method which augment the sample of provinces such as the Northern Cape (StatsSA, 2008).

On the other hand, the Western Cape province presented unique challenges different to those in the Northern Cape and other provinces in the country. The province is predominantly urban and farming with no traditional areas. Geographically the Western Cape is smaller than the Northern Cape. Based on the information from previous surveys, the province had the least response rates, on average. Lower response rates in a sample survey leads to sample reduction, and subsequently increase the variances. The initial sample sizes allocated to provinces using square-root allocation are further adjusted based on the average response rates per province. As a result, the sample sizes for provinces with high average response rates were adjusted (StatsSA, 2008).

Learning from Puckcharern (2013) the shortcomings of the proportional allocation method also surfaced in the sample design of Thailand FinScope Survey in 2013. Although the sample specifications for the FinScope study indicated that the sample should be drawn using the probability proportional to size (PPS), the allocation was not necessarily suppose to be proportional. In some countries, proportional allocations were used while in other countries allocation was disproportional. The unique situation in Thailand was the differences between population densities in urban and rural areas. The disproportional allocation led to larger sample sizes being allocated to urban areas and smaller samples to rural areas. The justification was that the population in the urban areas are more concentrated in small areas, while they are also heterogeneous. In contrast, the population in the rural areas is less concentrated and they are more homogeneous. It is common that most persons and households in the rural areas are involved in similar activities such as farming, mining or more (Puckcharern, 2013).

Every stage of a sample design has a challenge, and specific measures are employed at each stage. One of the most common problems is the accuracy of the sampling frame. Most countries design the sampling frames using census or administrative data or both. The sample designs are often done with census data, while the situation on the ground had already changed. Countries such as South Africa, where there are many informal dwellings in the high density areas, even locating the sampled structures becomes a challenge. In the context of the complex design, each and every method and its modification in the form of survey weights should be reflected during the analysis. After all the necessary weight adjustments are applied, then the final weight should be factored in the analysis.

1.3 The purpose of the study

1.3.1 Aim of the study

Several methods of estimating variances and other measures of precision are available. Sampling specialists and researchers have to closely observe the sample design of a particular sample survey in order to apply the methodology that will provide suitable measures of precision. The aim of the study is to describe the best practices in complex survey sampling designs and to apply benchmarking methods that allow production of optimal estimates and accurate measures of precision. As such, analysts and researchers will be able to understand the thermodynamics of survey sampling methodology and its effect in the survey estimates.

1.3.2 Objectives of the study

The specific objectives of the research study are to:

i. evaluate sampling approaches that are widely used in complex survey sampling process

ii. design an optimal sample suitable to produce demographic, socio-economic indicators and other common indicators at national and provincial-levels using small area layer (SAL) data from the South African Census 2011 as the sampling frame

iii. implement integrated weighting and calculate measures of precision from calibrated replicate weights of CSSD

iv. propose the approaches to complex survey sampling design suitable for several sampling scenarios, situations and conditions.

1.4 Overview of study approach

In order to achieve the objectives of the study, the complete Census 2011 data were used to implement CSSD in the South African context. Census and administrative records provide a variety of auxiliary information that are correlated to the measure of size used during sampling. The known variables were used in various stages of sample design through to estimation. Census geography, demographic and socio-economic attributes that are correlated to measure of size were used in stratification to increase homogeneity within strata.

Furthermore, appropriate weighting and estimation methods that cater for each design stage were employed in order to achieve optimal estimates. The creation of final weights is an iterative process. In addition, the process employs weight optimisation models using linear programming. A summary of the weight optimisation using linear programming is illustrated in Section 4.4.2. The sampling process starts from the creation of sampling frame through to estimation, while applying methods that ensure optimal allocation of sample in each phase of multi-stage sampling. The best practices in sampling theory were applied during the design and the lessons learned from various selected case studies were incorporated, those are, Canada, Thailand, Myanmar and South Africa.

1.5 Data used

1.5.1 Sampling frame data

The main data source used to formulate the sampling frame was the small area layer (SAL) data with selected South African Census 2011 variables for both persons and households (see Appendix 1). Data used for sample frame creation are publicly available at SAL level through the software called SuperCROSS. The SuperCROSS is a dissemination tool which allows users to create tabulation on all Census 2011 published variables. The cross-tabulation can be done on the entire Census 2011 geography hierarchy from national level to SAL.

For this study, cross-tabulations were done at SAL level for each variable used in sampling process. The final file had the SALs as records from 1 to 85,101 and the other individual variables such as total males and females appear as columns and the cells represent the aggregated totals. After all the individual tables were created at SAL level, the files were merged using the SAL variable.

The SAS Code used in this process is in “SAS Code 1”.

The sampling frame was enriched by adding the geography hierarchy which is also publicly available with the corresponding spatial data at SAL level. The SAL frame was merged with the geography hierarchy file using the SAL code variable. The variables included in the initial file are listed in Appendix 4. The frame was further enhanced by deriving variables for sampling, primary stratification, pooling of small SALs and splitting large SALs. The details on further enrichment, are covered in Chapter 3 of this study.

1.5.2 Survey data

The underlying data were obtained from the South African Census 2011 community profiles for both households and persons. In the absence of data collection based on the sample drawn, survey data were simulated for the purpose of implementing sample weighting and estimation methodologies for the complex survey sample design (CSSD). Statistics South Africa published Census 2011 unit record data at SAL level. Data are made available through SuperCROSS software which is a family of SuperSTAR Suite applications developed by the Australian based *Space Time Research*. The software allows downloading of records. However, records were without addresses and other identifiers.

The resulting files do not exactly give details of one individual or one household, but the records represent a group of persons or households with given characteristics. Without identifiers of the exact persons and household characteristics, it was not possible to obtain the direct match between persons and households. Although maximum matches could not be achieved with certainty, few available variables were used to match persons to households. There was information about the household-head which was also available in person-level file. The common variables included demographics, geography, level of education, school attendance and language.

In the final households simulated dataset, there were 524,308 records. The average response rate was calculated for each province using the South African Quarterly Labour Force Survey (QLFS) and the series used were: Quarter 1, Year 2016; Quarter 2, Year 2016; Quarter 3, Year 2016; and the South African General Household Survey (GHS) 2015. Detailed implementation of the survey data simulation is given in SAS Code 5.

1.6 Delimitations

An important requirement of drawing a probability sample is the availability of the sampling frame. A frame can be created from census data or from administrative records. Once the sample elements are defined, for example, dwellings, households or persons; a probability sample can be drawn and data collection is carried out. The frame for this study is based on census data with auxiliary variables considered for sampling. Conversely, no data collection was carried out, hence, this section describes delimitations and approaches used to substitute data collection in order to complete the estimation process.

Sampling methodology for this research study is based on the small area layer (SAL) which is the smallest geographical area made available for public use. Sampling methodology in South African household surveys is based on primary sampling units (PSUs), comprising of a group of enumeration areas (EAs) where the household size is used to group the data. Over 85 percent of SALs are made of one EA, while the remaining SALs are combined, and depending on the PSU definition, they are more likely to coincide. On that basis, SALs are justifiable as the primary sampling units for this research study and they contain a reasonable size of households to be covered by fieldworkers within a shorter period of data collection.

Although the sample was drawn using real data and can be practically implemented in the field, there was no data collection process for a nation-wide sample. Instead, data were simulated from separate household and person datasets from the South African Census 2011. The key step of simulation was to link persons to households. There were limited identifiers usable in linking households to persons and procedures implemented used assumptions to reconstruct households.

Certain procedures such as clustering of large SALs and subsampling of clusters to be listed in case of extreme growth, were discussed, but not implemented since the procedure requires field visit. Instead, the list-subsampling was implemented, which involves sample reduction through modification of SAL inverse sampling rate (ISR). Henry and Valliant (2012) gave an example of a situation in area probability samples, new construction developments with a large number of housing units that were not originally listed may be discovered. These are usually subsampled to reduce interviewer workloads, but this subsampling may create a subset of units with extremely large base weights.

There are benefits of forming explicit strata after splitting very large SALs, mainly to ensure that all the split SALs will fall within the same explicit strata. In this design, explicit strata were formed before splitting, on the basis that all SALs will be selected using PPS, where each SAL has a chance of being selected, although probability depends on the size.

Chapter 2

Evaluation of complex survey sampling methodology

2.1 Sampling frame

2.1.1 What is a sampling frame?

A sampling frame is defined essentially as comprising the materials from which a sample is selected. A sampling frame may be a list of small areas. It may also be a list of structures, households or persons. The census can be used to construct either type of frame, or both; indeed, most countries do use their census for such purposes (UNSD, 1998). Sampling frames are also obtained from administrative databases, for example, registration offices and their registers or complete lists of schools, universities and communities provided by the States

Bureaus of Statistics (Steinhauer, 2014).

Throughout the study, an example of selecting seats in a theatre will be referred to from time-to-time. A sample of theatre seats is assumed to be used to study the audience who attend shows in a specific venue at a given date and time. The example is illustrated in Figure 2.1.

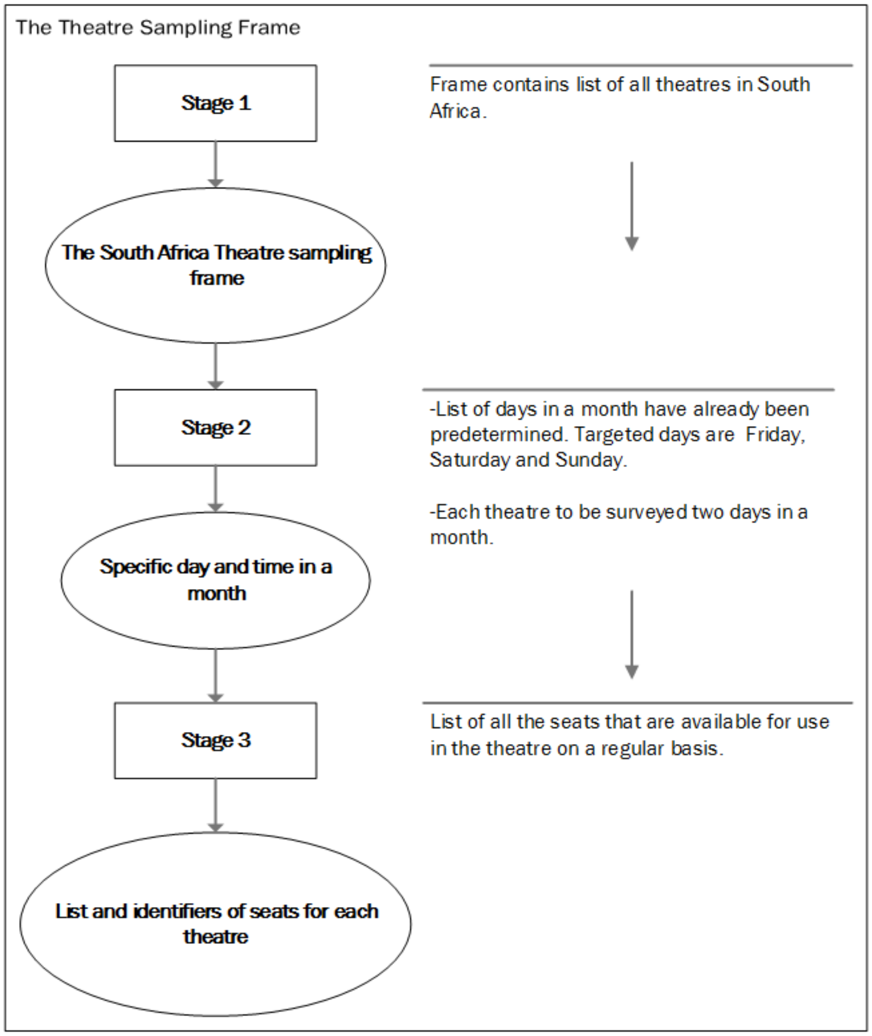
The Theatre Sampling Frame
Stage 1
Frame contains list of all theatres in South Africa.
The South Africa Theatre sampling frame
Stage 2
-List of days in a month have already been predetermined. Targeted days are Friday, Saturday and Sunday.
-Each theatre to be surveyed two days in a month.
Specific day and time in a month
Stage 3
List of all the seats that are available for use in the theatre on a regular basis.
List and identifiers of seats for each theatre

Figure 2.1: Theatre seating as an example of a sampling frame.

To complete the sampling frame for Stage 1, the list of theatres can be obtained from the administrative records owned by the relevant authorities. Accompanying the theatre list could be characteristics such as the number and list of venues within the theatre, total number of seats in the theatre, average weekly attendance per venue, number of secure parking bays, and so forth.

Stages in the sampling frame are not necessarily geographical areas, persons, items or objects. Information required for Stage 2 is two days in a month, where, only Friday, Saturday and Sunday are targeted. Within each selected theatre venue, two days in a month will be selected per theatre, where it will be determined beforehand if there are shows happening, and then the sample for the final stage is administered.

The third and final stage is a frame to be used to select seats where the survey will be administered. The list is created based on the seats that are made available for the particular show. The sitting plan or a sketch with identifiers is used to list all the available seats in the entire venue, before checking whether or not seats are made available for the event.

Table 2.1: An example of a three-stage theatre sampling frame.

Stage 1			Stage 2		Stage 3	
Name	**Venue**	**No. of seats**	**Day No.**	**Day**	**Seat No.**	**Seat ID**
SAST	Opera	1300	1	Friday	1	A1
SAST	Opera	1300	1	Friday	2	A2
SAST	Opera	1300	1	Friday	3	A3
SAST	Opera	1300	1	Friday	4	A4
SAST	Opera	1300	1	Friday	5	A5
SAST	Opera	1300	1	Friday	6	A6
SAST	Opera	1300	1	Friday	7	A7
SAST	Opera	1300	1	Friday	8	A8
SAST	Opera	1300	1	Friday	9	A9
SAST	Opera	1300	1	Friday	10	A10
SAST	Opera	1300	1	Friday	11	A11
SAST	Opera	1300	1	Friday	12	A12

Table 2.1 illustrates an example of a three-stage sampling for selecting theatres

only using one day selected in the South African State Theatre - Opera. The first in the list is the South African State Theatre (SAST). Within the theatre there are six venues and Opera theatre is the first one in the list with the largest sitting of 1,300, and it is at the level of Stage 1 of sampling where one of the theatres is sampled. Given month m in a year y the targeted days d are $Friday = \text{“}Fr\text{”}$, $Saturday = \text{“}Sa\text{”}$ and $Sunday = \text{“}Su\text{”}$, then $day\ 1 = \text{“}Fr\text{”}$, $day\ 2 = \text{“}Sa\text{”}$, $day\ 3 = \text{“}Su\text{”}$, $day\ 4 = \text{“}Fr\text{”}$, $day\ 5 = \text{“}Sa\text{”}$, $day\ 6 = \text{“}Su\text{”}$ and so forth. In columns 4 and 5, there is day number 1 within a month, which happens to be Friday. After selection of the venue in the first stage, then days in a month at that particular venue are selected in Stage 2. Lastly, in Stage 3, seats within the venue are selected using the appropriate sampling methodology to be discussed in subsequent chapters.

The frame in Table 2.1 can also be interpreted as follows. Assume SAST is in a sample and it has 1,300 seats, the list of seats 1 to 1,300 will be created to be used in Stage 3. Stage 2 in between requires that two days in a week be sampled, therefore, the final frame records for SAST will be 2,600. A representative sample of seats will be selected within each theatre and day of the week.

2.1.2 Desirable features of a sampling frame: Pros and Cons

The sampling frame must provide a high-level of coverage of the survey population. It must provide a means of uniquely identifying the sampled elements and a means to locate them. Accurate and up-to-date location information is needed. It is helpful if the frame is available in one place and is arranged in a form suitable for sampling, that is, it identifies strata and clusters, and is or can readily be arranged in groups corresponding to them. It is convenient if the final stage units are serially numbered. Computerised databases for list frame sampling are the norm today.

A complete sampling frame of areas and identifiable sampling units is necessary to draw a probability sample. Most population surveys encounter non-coverage as a result of using incomplete or outdated administrative files as sampling frames, resulting in the omission of a part of the population of interest (Mohadjer and Choudhry, 2002). Natural disasters and conflicts can also lead to major changes in the base sampling frame. The Household Living Conditions Survey survey design IHLCA (2011) shows an example of the negative effect of cyclone *Nargis* on eligible sample population.

Most population surveys are subject to some non-coverage. Surveys of low-income populations are no exception (Mohadjer and Choudhry, 2002). One source of non-coverage is the use of incomplete or outdated administrative files as sampling frames, resulting in the omission of a part of the population of interest. Similarly, non-coverage occurs when telephone interviewing is the only vehicle for data collection, because those without telephones have no chance of being selected, and are hence excluded from the survey.

Non-coverage of PSUs occurs, for example, when some regions of a country are excluded from a survey on purpose, because they are inaccessible, owing to war, natural disaster or other causes. Also, remote areas with very few households or persons are sometimes removed from the sampling frames for household surveys because they represent a small proportion of the population, and so have very little effect on the population figures. Non-coverage is a more serious problem at the household and person levels (UNSD, 1998).

In many survey applications, the omitted part of the population differs in many ways from the part that is included in the sampling frame. For example, if the objective of the study is to obtain information about the post-reform status of all low-income families, then families eligible for temporary assistance for needy families who did not become welfare recipients were included in the welfare records. These families are not covered in the administrative file used for sampling, and therefore, are not covered in the sample (Mohadjer and Choudhry, 2002).

Dealing with changes in the areas and population from the last census or creation of sampling frame, is one major challenge in probability sampling. Natural disasters and conflicts can lead to major changes in the base sampling frame. When the sample was designed for the integrated IHLCA in Myanmar, the frame which was created in 2004 had drastically changed in a 5-year period. Some of the townships that were selected for the first IHLCA-I survey were heavily affected by cyclone Nargis (IHLCA, 2011). Changes in administrative regions also need to be taken into account during the creation of the frame where the base sample frame is from previous censuses.

Another critical challenge is to draw a sample to study rare populations. A rare population is generally defined as a small proportion of a total population that possesses one or more specific characteristics. Examples include billionaires; people with a certain illness, such as gall bladder cancer; or employees in a highly technical occupation. Although the literature offers no precise definition of rare or small in this context, researchers have proposed proportions of 0.10 or less to identify rare populations. When this proportion is large, standard sampling techniques can usually be used efficiently. Kalton (2009) proposed screening, multiple frame sampling, disproportionate sampling, multiplicity sampling and snowballing as the available methods to draw a sample from a rare population.

Furthermore, sampling frames are often non-existent or incomplete for most rare populations. Although researchers can use convenience sampling (for example, snowball samples) to study rare populations, most efforts in this area have focused on probability sampling of rare populations. The costs and benefits of the various approaches can be difficult to define a priori and depend on the type and size of the rare population (Porter, 2008). Other examples of rare population include: drug users, people living with certain disease/illness (those with known status and those without), as well as small business owners.

Even in situations where listing is carried out to form a frame or listing obtained from existing administrative records, there are shortcomings, including those discussed above, that should be addressed before sampling. General solutions to sampling frame problems may include ignoring or disregarding the problem; redefining the survey population to fit the frame; and/or correcting the entire population list. However, if the problem is blanks and they are treated after sampling, eliminating them means that the realised sample will be smaller and of variable size (UNSD, 1998).

2.2 Sampling methods

2.2.1 Background

Sample surveys have long been recognised as cost-effective means of obtaining information on wide-ranging topics of interest at frequent intervals over time. They are widely used in practice to provide estimates, not only for the population of interest, but also for a variety of sub-populations (domains). Domains may be defined by geographical areas or socio-demographic groups or other sub-populations. Examples of a geographical domain (area) include a state/province, county, municipality, school district, unemployment insurance region, metropolitan area and health service area. Also, a socio-demographic domain may refer to a specific age-sex-race group within a large geographical area. An example of “other domains” is the set of business firms belonging to a census division by industry group (Rao, 2003).

As a historic example on survey sampling, in 1895 Anders Kiaer, founding Director of Statistics Norway, proposed sampling as a way of learning about a population without having to study every unit in the population (Tillé, 2011). There are two types of sampling: non-probability and probability sampling. The one chosen depends primarily on whether reliable inferences are to be made about the population. Non-probability sampling uses a subjective method of selecting units from a population. It provides a fast, easy and inexpensive way of selecting a sample. However, in order to make inferences about the population from the sample, the data analyst must assume that the sample is representative of the population. This is often a risky assumption to make in the case of non-probability sampling as pointed out in Statistics Canada (StatCan) survey methodology book (StatCan, 2003).

The basic statistical theory of probability sampling developed rapidly in the first half of the twentieth century and underpinned the growth of surveys. The essence is that units must be selected at random with known and non-zero selection probabilities. This enables unbiased estimation of population parameters and estimation of the precision (standard errors) of estimates (Groves et al., 2004). Design features such as stratified sampling and multi-stage (clustered) sampling are commonly used within a probability sampling framework. Some surveys, particularly in the commercial sector, use non-probability methods such as quota sampling (Tillé, 2011).

There are several different ways in which a probability sample can be selected. The design chosen depends on a number of factors such as the available survey frame, how different the population units are from each other (that is, their variability) and how costly it is to survey members of the population. For a given population, a balance of sampling error with cost and timeliness is achieved through the choice of design and sample size. The purpose of this chapter is to present different probability sample designs and factors to consider when determining the appropriate method for a specific survey. The background provided would form the basis of choosing the appropriate sample selection methods (StatCan, 2003).

2.2.2 Non-probability sampling

Non-probability sampling is a method of selecting units from a population using a subjective (that is, non-random) method. Since non-probability sampling does not require a complete survey frame, it is a fast, easy and an inexpensive way of obtaining data. The problem with non-probability sampling is that it is unclear whether it is possible to generalise the results from the sample to the population. The reason for this is that the selection of units from the population for a non-probability sample can result in large biases (StatCan, 2003).

Non-probability samples use procedures for selection which are not based on chance. With this type of sample, there is no way to accurately estimate the chance of any element being selected. The quality of a non-probability sample depends on the knowledge, judgement, and expertise of the researcher. At the same time, non-probability samples are quite convenient and economical (Israel, 1992). Non-probability samples include haphazard, convenience, quota, and purposive samples. Haphazard samples are those in which no conscious planning or consistent procedures are employed to select sample units (Cochran, 1965).

In a common non-probability sample design, the interviewer subjectively decides who should be sampled. Since the interviewer is most likely to select the most accessible or friendly members of the population, a large portion of the population has no chance of ever being selected, and this portion of the population is likely to differ in a systematic manner from those selected members.

Not only can this bias the results of the survey, but it can also falsely reduce the apparent variability of the population due to a tendency to select 'typical' units and eliminate extreme values. By contrast, probability sampling avoids such bias by randomly selecting units. Section 2.2.3 discusses probability sampling in greater detail.

Due to selection bias and (usually) the absence of a frame, an individual's inclusion probability cannot be calculated for non-probability samples, so there is no way of producing reliable estimates or estimates of their sampling error. In order to make inferences about the population, it is necessary to assume that the sample is representative of the population. This usually requires assuming that the characteristics of the population follow some model, or are evenly or randomly distributed over the population. This is often dangerous due to the difficulty of assessing whether or not these assumptions hold.

2.2.3 Probability sampling

Probability sampling is a method of sampling that allows inferences to be made about the population based on observations from a sample. In order to be able to make inferences, the sample should not be subject to selection bias. Probability sampling avoids this bias by randomly selecting units from the population (using a computer or table of random numbers). It is important to note that random does not mean arbitrary. In particular, the interviewers do not arbitrarily choose respondents, if that was the case, sampling would be subject to their personal biases. Random means that selection is unbiased, it is based on chance. With probability sampling, it is never left up to the discretion of the interviewer to subjectively decide who should be sampled (StatCan, 2003).

There are two main criteria for probability sampling: one is that the units be randomly selected, the other is that all units in the survey population have a known non-zero inclusion probability in the sample and that these probabilities can be calculated. It is not necessary for all units to have the same inclusion probability, indeed, in most complex surveys, the inclusion probability varies from unit to unit.

There are many different types of probability sample designs. The most basic is simple random sampling (SRS), and the designs increase in complexity to encompass systematic sampling, probability proportional to size (PPS) sampling to be discussed in the sections that follow, cluster sampling, stratified sampling, multi-stage sampling, multi-phase sampling and replicated sampling. Each of these sampling techniques is useful in different situations. If the objective of the study is simply to provide overall population estimates of a homogeneous population, then stratification would not be necessary and SRS may be the most appropriate sampling method. If the cost of survey collection is high and the resources are available, cluster sampling is often used. If sub-population estimates are also desired (such as estimates by province, age-group, or size of business), stratified sampling is usually performed. Cluster sampling is also used if no up-to-date sampling frame of the observation units exists. For example, if no up-to-date list of Grade 10 learners exists, schools are drawn first. Hence, the schools are the clusters.

Most of the more complex designs use auxiliary information on the survey frame to improve sampling. If the frame has been created from a previous census or from administrative data, there may be a wealth of supplementary information that can be used for sampling. For example, for a farm survey, the statistical agency may have the size of every farm in hectares from the last agricultural census. For a survey of people, information (for example, age, sex, ethnic origin, and so forth) may be available for everyone from the last population census. For a business survey, the statistical agency may have administrative information such as the industry (for example, retail, wholesale, manufacturing), the type of business (for example, food store), the number of employees, and so forth. Statistics South Africa calls this frame a business sampling frame. In order for the auxiliary information to improve sampling, there must be a correlation between the auxiliary data and the survey variables.

The main advantage of probability sampling is that since each unit is randomly selected and each unit's inclusion probability can be calculated, reliable estimates and an estimate of the sampling error of each estimate can be produced. Therefore, inferences can be made about the population. In fact, with a probability design, a relatively small sample can often be used to draw inferences about a large population.

The main disadvantages of probability sampling are that it is more difficult, takes longer and is usually more expensive than non-probability sampling. In general, the expense of creating and maintaining a good quality frame is substantial. Because probability samples tend to be more spread out geographically across the population than non-probability samples, their sample sizes are generally much larger and data collection is often more costly and difficult to manage. However, for a statistical agency, the ability to make inferences from a probability sample usually far outweighs these disadvantages.

Simple random sampling

Simple random sample (SRS) provides a natural starting point for a discussion of probability sampling methods, not because it is widely used, but because it is the simplest method and underlies many of more complex methods (Kalton, 1983). SRS is a one-step selection method that ensures that every possible sample of size n has an equal chance of being selected (refer Table 2.2 for the notations for the selection of n from N elements). As a consequence, each unit in the sample has the same inclusion probability. This probability, π, is equal to n/N, where N is the number of units in the population (StatCan, 2003).

Sampling may be done with or without replacement. Sampling with replacement allows for a unit to be selected more than once. Sampling without replacement means that once a unit has been selected, it cannot be reselected. Simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR) are practically identical if the sample size is a very small fraction of the population size. This is because the possibility of the same unit appearing more than once in the sample is small. Generally, sampling without replacement yields more precise results and is operationally more convenient. For the purpose of this study, sampling is assumed to be without replacement, unless otherwise specified.

Consider a population of five people and suppose that a sample of three is selected using SRSWOR. Label the people in the population 1,2,3,4 and 5 and denote the population as the set 1,2,3,4 and 5. There are ten possible samples of three people that can be selected from this population: (1,2,3); (1,2,4); (1,2,5); (1,3,4); (1,3,5); (1,4,5); (2,3,4); (2,3,5); (2,4,5) and (3,4,5). Each of these samples has an equal chance of being selected and each individual is selected in 6 out of the 10 possible samples. Thus, each individual has an inclusion probability of $n/N = 3/5$.

To select a simple random sample, the statistical agency usually constructs a complete frame (either a list or area frame) before sampling. On a list frame, the units are generally numbered 1 to N, although the method of assigning a unique number to each unit is not important. Next, n units from the list are chosen at random using a random number table or computer-generated random numbers, and the corresponding units make up the sample.

As a means of illustrating the technique of SRSWOR, consider a survey of audience attending a show in a theatre. Assume that a suitable list of theatre seats was available or was created from the theatre seating plan or by visiting and listing the theatre seats physically. This list serves as the survey or sampling frame. Now, suppose that the population list contains $N = 128$ theatre seats of which a sample of size $n = 10$ is required from this particular theatre. The next step is to decide how to select the 10 seats that represent potential occupants during the show (see Figure 2.2).

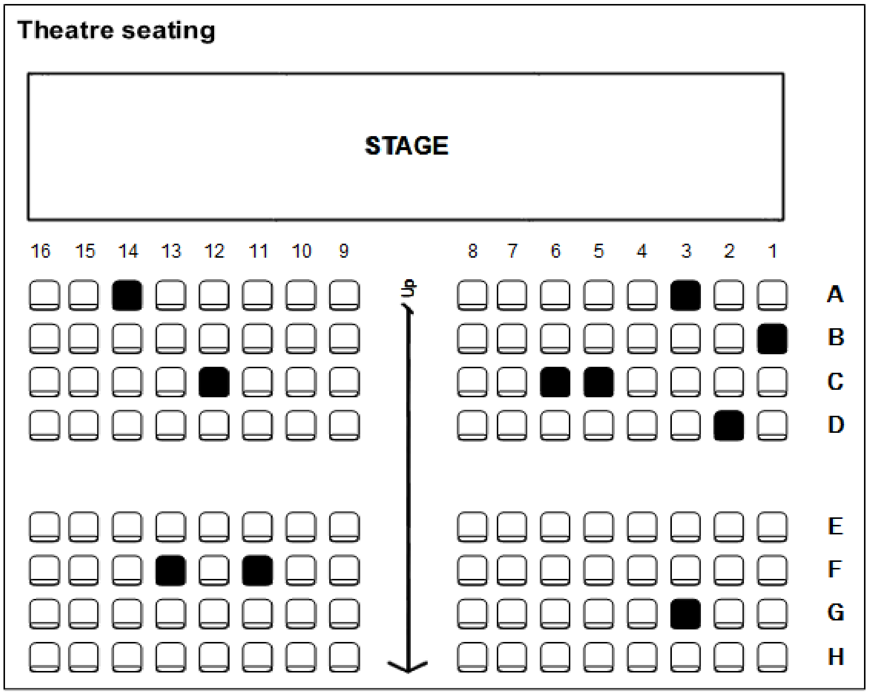

Figure 2.2: Example of sampling seats in a theatre using SRS method.

Sample selection can be done using a table of random numbers or computer generated random numbers. The first step involves selecting a three-digit number (three since this is the number of digits in 001 to 128). Sampling began by selecting a number anywhere in the table and then proceeding in any direction. The first 10 three-digit numbers that do not exceed 128 were selected.

Assuming that the random numbers selected using the table of random numbers or automatically generated random numbers are 3, 14, 17, 37, 38, 44, 50, 91, 93 and 99. Selection continued until 10 different numbers were obtained.

The result is a sample that consists of seats with the corresponding numbers in the listing of the population, seat (1) is A1 and seat (128) is H16 (see Figure 2.2). (Since the method under discussion is SRSWOR, any number that appears more than once is ignored). Although a random number table was used to illustrate the manual selection of a SRS, practically speaking, a computer program would be ideal for randomly selecting units. Notations for estimating parameters under SRS are as follows:

Table 2.2: SRS methods for N elements.

Statistic	**Population**	**Sample**
N elements	$Y_1, Y_2, ..., Y_N$	$y_1, y_2, ..., y_n$
Total	$Y = \sum Y_i$	$y = \sum y_i$
Mean	$\bar{Y} = \frac{\sum Y_i}{N} = \frac{Y}{N}$	$\bar{y} = \frac{\sum y_i}{n} = \frac{y}{n}$
Variance	$S^2 = \frac{1}{N-1} \sum (Y_i - \bar{Y})^2$	$s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$

Systematic sampling

Systematic (SYS) samples are widely used and easy to implement. A systematic sample selects the first element randomly and then every k^{th} element on the list afterwards (Israel, 1992). SYS sampling is sometimes chosen when the statistical agency wishes to use SRS but no list is available, or when the list is roughly in random order, for which SYS sampling is simpler to conduct than SRS. When a list frame is used and the population size, N, is a multiple of the sample size, n, every k^{th} unit is selected, where the interval k is equal to N/n. The random start, r, is a single random number between 1 and k, inclusively. The units selected are then: $r, r + k, r + 2k, ..., r + (n - 1)k$. Like SRS, each unit has an inclusion probability, π, equal to n/N, but unlike SRS, not every combination of n units has an equal chance of being selected. SYS sampling can only select samples in which the units are separated by k. Thus, under this method, only k possible samples can be drawn from the population (StatCan, 2003).

To illustrate SYS sampling, let us return to the theatre seats example where the population contained $N = 128$ units and a sample of size $n = 10$ units was to be drawn. The sampling interval would be $k = N/n = 128/10 = 12.8$ (rounded to 13). Next, a random number between 1 and $k = 13$, for instance 4, is chosen. The population units selected for the sample are then numbered: 4, 17, 30, 43, 56, 69, 82, 95, 108 and 121 (see Figure 2.3). With a sampling interval of 13 and a population of size 128, there are only 13 possible SYS samples, while for a SRS of size 10, there are over 25 million possible samples.

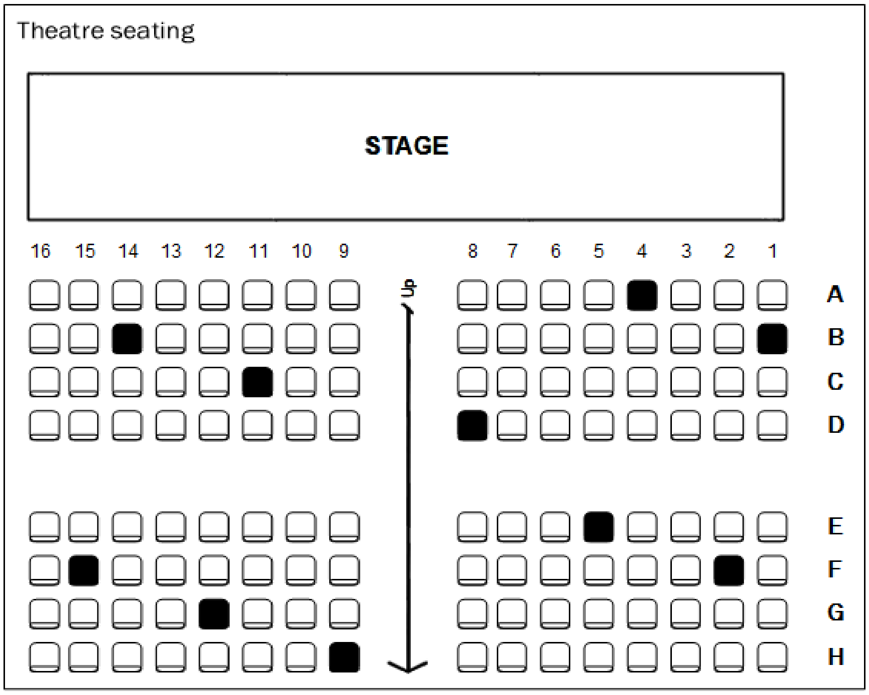

Figure 2.3: Example of sampling seats in a theatre using SYS sampling method.

One advantage of SYS sampling is that it can be used when there is no list of the population units available in advance. In this case, a conceptual frame can be constructed by sampling every k^{th} seat until the end of the population is reached, that is, the total number of seats in the theatre. One problem with

SYS sampling is that the sample size, n, is not known in advance, until after the sample has been selected.

Another problem arises when the sampling interval, k, matches some periodicity in the population. For example, suppose that a survey of clinic visits by TB patients is to be conducted in a given region and only one day of the week can be sampled, in other words, k is every 7th day. The survey's estimated clinic visits were dramatically different if the sampled days are all Sundays as opposed to all Tuesdays. Of course, if the sampling period is every 5th day, then every day of the week could be surveyed. Unfortunately, in most cases, periodicity is not known in advance.

If N cannot be evenly divided by n, the sampling interval for SYS sampling is not a whole number as is the case within the theatre example. In this case, k could be set equal to the nearest whole number, but then the sample size would vary from sample to sample. For example, suppose that $N = 128$ and $n = 10$, then $k = 128/10 = 12.8$. If k is assumed to be 13, and if $r = 4$, the sample contains those units numbered: 4, 17, 30, 43, 56, 69, 82, 95, 108 and 121. If the random start is $r = 1$ and every 13^{th} unit is selected, then the sample consists of units: 1, 14, 27, 40, 53, 66, 79, 92, 105 and 118, and in this case both samples are equal to 10. Let us now assume that the start $r = 12$, which is one less than the interval 13. The resulting systematic sample is: 12, 25, 38, 51, 64, 77, 90, 103, 116 (note that 129 is excluded because it exceeds $N = 128$, the total population, which yield the sample size of 9, and not 10 as required).

Another approach is to set each of the values $r, r+k, r+2k, \ldots, r+(n-1)k$ to the nearest whole number. With this approach, the realised sample size is fixed. For example, suppose again that $N = 128$ and $n = 10$, so that $k = 128/10 = 12.8$. If $r = 12$, the sample consists of units 12(12), 25(24.8), 38(37.6), 50(50.4), 63(63.2), 76(76), 89(88.8), 102(101.6), 114(114.4) and 127(127.2).

Alternatively, if N cannot be evenly divided by n, then to avoid a varying sample size, circular systematic sampling could be performed. With this method, the population units are thought to exist on a circle and modular counting is used. The value of k is set equal to the whole number nearest to N/n, but now the random start, r, can be between 1 and N, rather than 1 and k (that is, the first unit can be anywhere on the list). The selected units, as before, are: $r, r+k, r+2k, ..., r+(n-1)k$. If the j^{th} unit is such that $r+(j-1)k > N$, then the selected unit is $r+(j-1)k-N$. That is, when the end of the list is reached, sampling continues at the beginning of the list. The advantage of the circular method is that each unit has an equal chance of being in the sample. For example, using the previous example, suppose that $N = 128$ and $n = 10$ and $k = 13$. A random start, r, between 1 and 128 is selected, say $r = 56$. Then the selected population units are: 56, 69, 82, 95, 108, 121, 6, 19, 32 and 45.

Probability proportional to size sampling

When information on a measure of size MOS exists for every element in the population and this size measure stores valuable information about the "importance" of element i to be included in the sample, this information can be used in the sample design. Sample designs that make explicit use of such size measures are called probability proportional to size (PPS) sample designs. The inclusion probability of element i of a PPS sample of size n is

$$\pi_i^{(PPS)} = \frac{MOS_i}{\sum\limits_{i\epsilon U} MOS_i}, \tag{2.1}$$

where

$\pi =$ inclusion probability of i^{th} unit, e.g., PSU
$U =$ sample elements contained in the frame, i.e., PSUs

MOS = measure of size for each sample unit i.

Sample designs with PPS are often used in business surveys when it is important to include the largest firms in an industry in the sample since they contribute a large amount to the industry's production of goods or services. However, PPS can also be combined with cluster sample designs or general multi-stage sample designs, which are introduced in the next sections.

A good example of a PPS size of the sampling unit in area sampling is a variable "area". Farm surveys often use PPS, where the size measure is the size of the farm in hectares. Admittedly, the size of a farm can grow (or shrink) if the farmer buys or sells land, but for the most part, farm size is constant from year-to-year. In addition, typical questions for farm surveys, such as income, crop production, livestock holdings and expenses are often correlated with land holdings. Other size measures for business surveys include the number of employees, annual sales and the number of locations, although these variables are more likely to change from year-to-year. In PPS sampling, the size of the unit determines the inclusion probability. Using farms as an example, means that a farm with an area of 200 hectares has twice the probability of being selected, compared to a farm with 100 hectares.

To further illustrate, assume that there is a population of six theatres and that the client is interested in estimating the total revenue from events of this theatre population by sampling one theatre. A sample of size one is used for the purpose of illustration. In practice, a statistical agency rarely selects only one unit. Suppose that there is a stable size measure for each theatre (the seating capacity of the theatre), and to illustrate the efficiency gains over SRS, assume that each theatre's revenue is known. Obviously, in real life, if the revenue were known, there would be no need to conduct the survey. The theatre with the most seats (largest capacity) has the higher probability of being selected as compared to the rest of the theatres in the frame.

Very often PPS sample selection method can be confused with proportional allocation method. Regardless of the allocation method applied to strata, the sampling units contained can still be selected using PPS. The process of selecting clusters/SALs using PPS is discussed in Section 3.5.1.

Cluster sampling

In a cluster sample, the population is divided into non-overlapping subpopulations, usually based on geographical or political boundaries. For a simple cluster sample, a random sample of subpopulations (clusters) is obtained and, within each selected cluster, each subject is sampled (Hoshaw-Woodard, 2001). When a list of the entire population is non-existent, or hard to obtain or the cost of surveying dispersed individuals is prohibitive, cluster sampling can facilitate the data collection process. Kish (1965) defined cluster sampling as a method of selecting sampling units in which the unit contains a cluster of elements. Some types of clusters are employees of business firms, children in schools, dwellings in city blocks, and residents in counties or states. The last two are geographical clusters or areas (Israel, 1992).

Cluster sampling is the process of randomly selecting complete groups (clusters) of population units from the survey frame. It is usually a less statistically efficient sampling strategy than SRS. In survey work, much time and effort are spent in following a method of simple random sampling until an individual is actually identified and enrolled in the sample. Savings can be achieved if, instead of performing this 96 times for 96 individuals as required by the above calculations, the sampling process can be performed fewer times, each time taking several persons. This is called cluster sampling. Because there is usually a tendency for individuals found within a cluster to share characteristics, however, use of cluster sampling can be expected to decrease the precision of the sample result (Henderson and Sundaresan, 1982).

The cluster sampling method is performed for several reasons. First, sampling clusters can greatly reduce the cost data of collection, particularly if the population is spread out and personal interviews are conducted. Second, it is not always practical to sample individual units from the population. Sometimes sampling groups of the population units is much easier (for example, entire households). Finally, it allows the production of estimates for the clusters themselves (for example, average revenue per household) (StatCan, 2003).

Each ultimate sampling element belongs to exactly one PSU and each PSU contains one or more ultimate sampling units. A clustered population consists of M PSUs, which are of size $N_i, i = 1, ..., M$. Assume that a complete frame of PSUs exists from which a sample of m PSUs is drawn. The set of possible samples of m of the M clusters is denoted by S and a specific sample of m PSUs is denoted by s. The cluster sample design is defined as $p(s)$. The inclusion probabilities of each of the M clusters is denoted by π_i. The value of π_i depends on the characteristics of the sample design. After s has been obtained, y is surveyed for each of the

$$n = \sum_{i=1}^{m} N_i \tag{2.2}$$

ultimate sample elements (ignoring contact and non-response issues for the time being).

There are two reasons that motivate the use of cluster sampling by various countries. First, the absence or poor quality of listings of households or addresses makes it necessary to first select a sample of geographical units, and then to construct lists of households or addresses only within those selected units. The samples of households can then be selected from those lists. Second, the use of multi-stage designs controls the cost of data collection (UNSD, 1998).

Cluster sampling is a two-step process. First, the population is grouped into clusters (this may consist of natural clustering, for example, households, schools). The second step is to select a sample of clusters and interview all units within the selected clusters. The survey frame may dictate the method of sampling. Until now, the focus has been on sampling individual units of the population from a list-frame. If the units of the population are naturally grouped together, it is often easier to create a frame of these groups and sample them rather than try to create a list frame of all individual units in the population. For example, the client may be interested in sampling teachers, but only have available a list of schools.

In the case of household or farm surveys, as an example, many countries do not have complete and up-to-date lists of the people, households or farms for any large geographical area, but they do have maps of the areas. In this case an area frame could be created, with the geographical areas divided into regions (clusters), and then sample the regions and interview everyone within the sampled regions. Different sample designs can be used to select the clusters, such as SRS, SYS sampling or PPS. A common design uses PPS where sampling is proportional to the size of the cluster.

Probability proportional to size (PPS) sample selection method discussed in the previous section, is also commonly used to select a sample of clusters such as PSUs. PPS sampling is a technique that employs auxiliary data to yield dramatic increases in the precision of survey estimates, particularly if the measures of size are accurate and the variables of interest are correlated with the size of the unit. It is the methodology of choice for sampling PSUs for most household surveys. PPS sampling yields unequal probabilities of selection for PSUs. Essentially, the measure of size of the PSU determines its probability of selection. However, when combined with an appropriate sub-sampling fraction for selecting households within selected PSUs, it can lead to an overall self-weighting sample of households in which all households have the same probability of selection regardless of the PSUs in which they are located. Its principal attraction is that it can lead to approximately equal sample sizes of SSUs per PSU and also increase the statistical efficiency (UNSD, 1998).

Stratified sampling

With stratified (STR) sampling, the population is divided into homogeneous, mutually exclusive groups called strata, and independent samples are then selected from each stratum. The population may also be divided into non-overlapping subpopulations defined on the basis of some known characteristic that is believed to be related to the variable of interest. For example, the population may be stratified with respect to sex, race, geographical region, and so forth (Hoshaw-Woodard, 2001).

Any of the sample designs mentioned in this chapter can be used to sample within strata, from the simpler methods such as SRS or SYS sampling to the more complex methods such as PPS, cluster, multi-stage or multi-phase sampling (discussed later in this chapter). For example, with cluster sampling, it is very common to first stratify, then draw the cluster sample. This is called stratified cluster sampling (StatCan, 2003). Stratification can be explicit or both explicit and implicit.

In STR sampling as described in UNSD (1998), the population of interest can be divided into H non-overlapping sub-populations or strata of size $N_h(h = 1, ..., H)$ according to a stratification variable Z. The stratification variable is either discrete or has to be recoded into a discrete variable with as many unique values as the desired number of strata. The values of Z are denoted by Z_h. The total sample size n is then allocated to the strata, so that $n = \sum\limits_{h=1}^{H} n_h$, where n_h is the sample size drawn from stratum of size N_h.

For example, in a single-stage stratified sample, the mean is given by:

$$\bar{y}_{st} = \sum_h \frac{N_h}{N} \sum_h \frac{y_{hi}}{n_h} = \sum_h W_h \bar{y}_h, \tag{2.3}$$

where n_h is the size of the sample selected from the N_h units in stratum h, $N = \sum N_h$ is the population size, $W_h = N_h/N$ is the proportion of the population in stratum h, y_{hi} is the value for sampled unit i in stratum h, and $\bar{y}_h = \sum_i y_{hi}/n_h$ is the sample mean in stratum h. In practice, $\bar{y}_{st}$ is computed as a weighted estimate, where each sampled unit is assigned a base weight that is the inverse of its selection probability (ignoring for the moment, the sample and population weighting adjustments). Here each unit in stratum h has a selection probability of n_h/N_h and hence a base weight of $w_{hi} = w_h = N_h/n_h$. Thus, $\bar{y}_{st}$ may be expressed as

$$\bar{y}_{st} = \frac{\sum_h \sum_i w_{hi} y_{hi}}{\sum_h \sum_i w_{hi}} = \frac{\sum_h \sum_i w_{hi} y_{hi}}{\sum_h n_h y_h}. \tag{2.4}$$

Assuming that the finite population correction can be ignored, the variance of the stratified mean is given by

$$V(\bar{y}_{st}) = \sum \frac{W_h^2 S_h^2}{n_h}, \tag{2.5}$$

where $S_h^2 = \sum i(Y_{hi} - Y_h)2/(N_h - 1)$ is the population unit variance within stratum h.

Explicit stratification

Stratification is commonly applied at each stage of sampling. However, its benefits are particularly strong in sampling PSUs/SALs. It is, therefore, important to stratify the PSUs/SALs efficiently before selecting them. Stratification partitions the units in the population into mutually exclusive and collectively exhaustive sub-groups or strata. Separate samples are then selected from each stratum. A primary purpose of stratification is to improve the precision of the survey estimates. In this case, the formation of the strata should be such that units in the same stratum are as homogeneous as possible and units in different strata are as heterogeneous as possible with respect to the characteristics of interest to the survey. Other benefits of stratification include, (i) administrative convenience and flexibility, and (ii) guaranteed representation of important domains and special sub-populations.

Implicit stratification

Within each explicit stratum, a technique known as implicit stratification is often used in selecting PSUs/SALs. Prior to sample selection, PSUs/SALs within an explicit strata are sorted with respect to one or more variables that are deemed to have a high correlation with the variable of interest, and that are available for every PSU/SAL in the stratum. A systematic sample of PSUs/SALs is then selected. Implicit stratification guarantees that the sample of PSUs will be spread across the categories of the stratification variables. When there are many variables available, implicit stratification can take the form of clustering where an explicit strata can be divided into several homogeneous clusters that are taken as the final strata.

For many household surveys in developing and transition countries, implicit stratification is based on geographical ordering of units within explicit strata. Implicit stratification variables, sometimes used for PSU selection, include residential area (low-income, moderate-income, high-income), expenditure category (usually in quintiles), ethnic group and area of residence in urban areas;

and area under cultivation, amount of poultry or cattle owned, proportion of nonagricultural workers, and so forth, in rural areas. For socio-economic surveys, implicit stratification variables include the proportion of households classified as poor, the proportion of adults with secondary or higher education, and distance from the centre of a large city. Variables used for implicit stratification are usually obtained from census and administrative record data.

In the design of the master sample used to draw a sample for South African Quarterly Labour Force Survey, implicit stratification was implemented (StatsSA, 2008). To implement implicit stratification, socio-economic variables were obtained for all the enumeration areas from the most recent census data. The variables included demographics (age, gender, race), household-based variables (access to services and goods) and economic variables (industry, income, occupation and employment). The number of auxiliary variables where such that basic method of sorting to determine implicit strata would not be sufficient, as a result, specialised clustering method was employed, that is, k-means clustering. The EAs were clustered based on the socio-economic variables that resulted to new strata formed.

A population can be stratified by any of the variables that are available for all units on the frame prior to the survey being conducted. For instance, this information may simply be the address of the unit, allowing stratification by province, or there may be income data on the frame, allowing stratification by income group. Commonly used stratification variables include: age, sex, geography (for example, province), income, revenues, household size, size of business, type of business, number of employees, and so forth. Stratification can be explicit, using natural groupings; or implicit, using auxiliary variables that are correlated to the MOS or domains of interest.

In most household-based surveys, countries use natural geography clusters that are derived from census. They are the lowest geography units for data collection or even for dissemination. In some countries the census enumeration area is usually the basis of forming such geography areas. In South Africa, the SALs are formed from EAs and they are publicly available and are also suitable for sampling. After strata are formed, dwelling units are selected in PSUs. Assume a two-stage sampling where PSUs are to be selected using PPS method, then select households using systematic sampling method. The formula for selecting PSUs in a stratum will be as follows:

$$P_{PSU} = \frac{n_i}{\sum n_i} \times m, \tag{2.6}$$

where n_i is total households per PSU within a stratum, $\sum n_i$ is total households in the stratum and m is number of PSUs to be selected within a stratum.

The formula for selecting households in a PSU will be as follows:

$$P_{HH} = \frac{b}{n_i}, \tag{2.7}$$

where

b is the households sample size to be drawn per PSU.
n_i is the number of households within the sampled from each i_{th} PSU
HH is the number of households within the sampled PSU

Importance of stratification

In order to improve the statistical efficiency of a sampling strategy with respect to SRS, there must be strong homogeneity within a stratum (that is, units within a stratum should be similar with respect to the variable of interest) and the strata themselves must be as different as possible (with respect to the same variable of interest). In general, this is achieved if the stratification variables are correlated with the survey variables of interest. The reason why stratification can increase the precision of the estimates relative to SRS is explained by Cochran (1977). If each stratum is homogeneous in that the measurements vary little from one unit to another, then a precise estimate of any stratum mean can be obtained from a small sample in that stratum. Such estimates can be combined in a precise estimate for the whole population.

Stratification is particularly important in the case of skewed populations (that is, when the distribution of values of a variable is not symmetric, but leans to the right or to the left). For example, business and farm surveys often have highly skewed populations – the few large business establishments and farms often have large values for variables of interest (e.g., revenues, expenditures, number of employees). In such cases, a few population units can exert a large influence on estimates (if they happen to be selected in the sample, they can greatly increase the estimate, and if they are not selected, the estimate is much lower). In other words, these units can increase the sampling variability of the estimate. Therefore, such units should be placed in a separate stratum to ensure that they do not represent other potentially much smaller units in the population.

Usually the stratification variables are chosen based on their correlation with the most important survey variables. This means that for those less important survey variables that are uncorrelated to the stratification variables, estimates for a stratified sample can be less efficient than SRS. One of the objectives for stratification is to ensure adequate sample sizes for known domains of interest. When designing a survey, often the overall goal is to estimate a total. How many people were unemployed last month? What were the total retail sales last month?

In addition to overall totals, the client often requires estimates for sub-groups of the population, called domains. For example, the client may wish to know how many men were unemployed and compare this with the number of women who were unemployed. Similarly, the client may want to know the sales last month for clothing stores, or for all retail stores in a certain province. Creating estimates for sub-groups is called domain estimation. If domain estimates are required, the ability to calculate them with a large enough sample in each domain should be incorporated into the sample design. If the information is available on the frame, the easiest way to do this is to ensure that strata exactly correspond to the domains of interest.

Moreover, stratification is often used for operational or administrative convenience. It can enable the statistical agency to control the distribution of fieldwork among its regional offices. For example, if data collection is conducted by province, then stratification by province is appropriate, in which case the provincial regional office can be given their portion of the sample. Once the population has been divided into strata, the statistical agency needs to determine how many units should be sampled from each stratum. This step is referred to as allocation of the sample.

Inclusion probabilities usually vary from stratum to stratum; depending on how the sample is allocated to each stratum. To calculate the inclusion probabilities for most sample designs, the size of the sample and the size of the population in each stratum must be considered. To illustrate, consider a population with $N = 1000$ units stratified into two groups: one stratum has $N_1 = 250$ units and the other has $N_2 = 750$ units.

Suppose that SRS is used to select $n_i = 50$ units from the first stratum and $n_2 = 50$ units from the second, then the probability π_2 that a unit in the second stratum is selected is $\pi_2 = 50/750 = 1/15$. Units thus have different probabilities of inclusion, which implies that a unit in the first stratum is more likely to be selected than the one in the second stratum.

Stratification examples

To stratify businesses, a size variable based on the number of employees, for example, is often used. If the size variable has three values, say, small, medium and large, then the statistical efficiency is improved if the large businesses have similar sales, the medium businesses have similar sales, and the small businesses have similar sales. Similarly, for a sample design using area frames, the proper representation of large cities can be ensured by placing them in a separate stratum, and sampling each stratum separately. In some instances, those strata made of very large sampling elements are treated as “take all strata” where all elements contained are sampled with certainty.

In the previous example, it was reasonable to stratify by the number of employees, since this is a measure of the size of the company and is likely to be highly related to sales. However, if a survey is interested in the age of its employees, it makes no sense to stratify by the number of employees since there is no correlation. Also, stratification that is statistically efficient for one survey variable may not work well for others.

Returning to the theatre example, suppose that the selected theatre contained 128 seats that are all available for seating during the event. The event organisers partitioned the seats into four groups and the prices vary from the front most to the back seats. The front rows (A - B) were offered at VIP pricing followed by the next groups (C - D), (E - F) and (G - H) with their own prices (see Figure 2.4).

The partitioned groups were treated as strata due to the existing natural groupings, therefore, the systematic sample was to be drawn from each stratum. Strata had equal sizes and the total sample of 16 seats was to be distributed equally to each stratum.

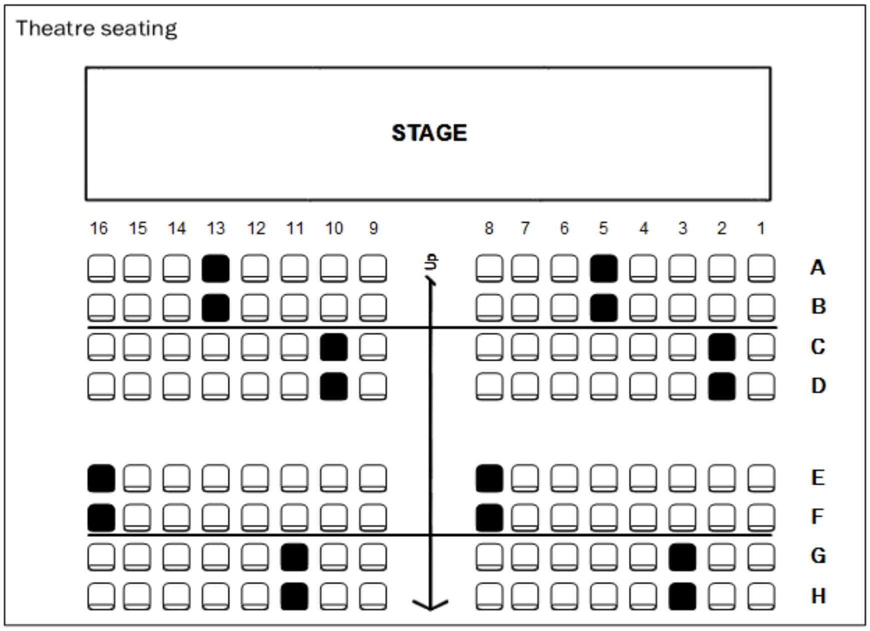

Figure 2.4: Example of sampling seats in a theatre using stratified sampling method.

Multi-stage sampling

Most surveys in developing and transition countries are based on stratified multi-stage cluster designs. Thus far, the discussion has centred around one-stage sample designs. Multi-stage sampling is the process of selecting a sample in two or more successive stages. The units selected at the first stage are called primary sampling units (PSUs), units selected at the second stage are called second stage units (SSUs), and so forth. The units at each stage are different in structure and are hierarchical (for example, people live in dwellings, dwellings make up a city block, city blocks make up a city, and so forth). In multi-stage sampling, the SSUs are often the individual units of the population (StatCan, 2003).

A common multi-stage sample design involves two-stage or more cluster sampling using an area frame at the first stage to select regions (the PSUs) and then a systematic sample of secondary sampling units (SSUs), i.e., dwellings within a region at the second stage, and so forth. With the one-stage cluster sampling presented earlier, every unit within a sampled cluster is included in the sample. In two-stage sampling, only some of the units within each selected PSU are sub-sampled.

Multi-stage sampling is commonly used with area frames to overcome the inefficiencies of one-stage cluster sampling, which is, in fact rarely used. If the neighbouring units within a cluster are similar, then it is more statistically efficient to sample a few SSUs from many PSUs than to sample many SSUs from fewer PSUs.

Multi-stage samples can have any number of stages, but since the complexity of the design (and estimation) increases with the number of stages, designs are often restricted to two or three stages. It should be noted that the frame for the first stage is generally quite stable. For example, an area frame covering large geographical areas does not change rapidly over time. Second (and subsequent) stage frames required to sample units at subsequent stages are usually less stable. Often these frames are list frames created in the field during data collection. For example, for the geographical areas sampled at stage one, a list frame could be created for all those dwellings or households within the sampled areas. Note that listing only sampled areas requires much less effort than trying to list the whole population.

In two-stage sampling, for example, ultimate sampling units are nested directly within superordinate clusters. Under two-stage sampling m of M, clusters are selected at the first stage. The set of possible samples of m primary sampling units is denoted by $S(1)$. A specific sample of m primary sampling units is denoted by $s(1)$; inclusion probabilities for each of the M PSUs are denoted by $\pi_i, i = 1, ..., M$. At the second stage, n_i SSUs of the i^{th} PSU of size N_i are selected within each selected PSU. Thus,

$$n = \sum_{i \epsilon s^{(1)}}^{m} n_i. \tag{2.8}$$

Elements of the i_{th} cluster are denoted by $1, ..., j, ..., n_i$. The set of possible samples of n_i from N_i SSUs in the i_{th} PSU is denoted by $S_i^{(2)}$ and a specific sample by $s_i^{(2)}$.

The sum of all $S_i^{(2)}$ is S and the sum of all $s_i^{(2)}$ is s.

The inclusion probability of the j^{th} element given the i^{th} PSU selected is denoted by $\pi_{j|i}$. The magnitude of $\pi_{j|i}$ depends on the sample design that is used to select the elements within the PSU. A convenient notation system for the ultimate sample elements selected into the sample is required. Suppose there are

$$n = \sum_{i=1}^{m} n_i \tag{2.9}$$

elements selected in total. Let us refer to the j^{th} element of the i^{th} PSU as the k^{th} element, $k = 1, ..., n$. Table 2.3 illustrates this notational scheme.

Following this notation, the overall inclusion probability of the j^{th} element in the i^{th} PSU is denoted by π_{ij} or simply by π_k, and it is the product of the inclusion probabilities in the two stages, which is expressed as

$$\pi_{ij} = \pi_k = \pi_i * \pi_{j|i}, \tag{2.10}$$

and the design weight is expressed as

$$w_{ij} = w_k = \frac{1}{\pi_k} = \frac{1}{\pi_i.\pi_{j|i}}. \tag{2.11}$$

Table 2.3: Notation scheme.

PSU		**SSU**		**Ultimate sample element**
$(i = 1, ..., M)$	$i\epsilon s^{(1)}$	$(j = 1, ..., N_i)$	$j\epsilon s^{(1)}$	$(k = 1, ..., n)$
1	yes	1	yes	1
1	yes	2	no	1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1	yes	N_1	no	1
2	no	1	no	1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2	no	N_2	no	1
i	yes	j	yes	k
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
i	yes	N_j	no	1
M	yes	1	no	k
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
M	yes	N_M	yes	n

Complex sampling

Complex survey sample design (CSSD) is defined by Heeringa and Liu (1998) as a probability sample design method developed using sampling procedures such as stratification, clustering, segmentation, probability proportional to size (PPS) selection methods and weighting. CSSD is designed to improve statistical efficiency, reduce costs and improve precision for sub-group analyses relative to simple random sample (SRS). The probability sampling methods that are discussed in previous sections have their own challenges to draw samples and make inference. Most modern surveys employ more than one sampling methods and also derive weights for domains that were catered and not catered for during the design, that qualifies them to be complex sample designs.

In order to illustrate the use of complex sampling, examples from Canada, South Africa and Thailand are given. The Canadian Labour Force Survey (CLFS) sample is an example of a multi-stage stratified sample. The country is divided into over 1,100 strata. Each stratum consists of a group of EAs (StatCan, 2003). On the other hand the South African Labour Force Survey, when it was redesigned in 2008, was based on PSUs made of combining smaller EAs (StatsSA, 2008). The survey sample had 363 strata formed. The Thailand FinScope study was also a CSSD based on EAs. The illustration went a step further and included the sampling of one person within the household. EAs are geographical areas defined by the census of population so that the area that they cover can be canvassed and enumerated by one census representative (they are created by keeping in mind the size of territory and the density of the population). The first stage of sampling is a stratified sample of clusters (EAs or groups of EAs) from within these strata. At the second stage, the clusters are mapped, all dwellings in them are listed, and the census representative selects a systematic sample of dwellings from each list. All persons within a selected dwelling are then interviewed for the survey.

The Thailand FinScope 2013 survey was a multi-stage sample where the EAs for the first stage of sampling were EAs in urban and rural areas (Puckcharern, 2013). The PPS method was applied to draw samples from each stratum. The MOS used was the total number of households within an EA.

The selection probability of the i^{th} block/village was given by:

$$P_{hi} = \frac{\alpha_h M_{hi}}{M_h}, \tag{2.12}$$

where

M_{hi} is the number of households from i^{th} EA in h^{th} stratum

M_h is the number of households in the h^{th} stratum

α_h is the number of EAs selected

$i = 1, 2, \ldots, m_h$

$h = 1, 2, \ldots, 9.$

The systematic sampling was used for PPS sampling.

The list of households will be used as the sampling frame for the selection of households in the second stage sampling. In order to prepare the sampling frame, the enumerators will list all dwellings and record the number of all households located in the sampled EAs. The listing information includes:

- identification number of building,
- identification number of household,
- address,
- name of the head of household,
- the total number of household members, and
- the socio-economic condition of a household.

The systematic random sampling is then applied to select equal number of households within EAs. The reason for applying 10 sampled households in each sampled EA is that EA sizes in urban and rural are 150 and 300 households on average, respectively. There is more heterogeneity among households in urban areas than in rural areas.

The selection probability at second stage of sampling is given by:

$$P_{hij} = \frac{n_{hi}}{N_{hi}}, \tag{2.13}$$

where

n_{hi} is the number of households selected from i^{th} EA in h^{th} stratum
N_{hi} is the number of households counted from the listing of the i^{th} EA in the h^{th} stratum.

From previously selected households, all eligible people will be listed, say n_{hi}. One eligible person was randomly selected in the third stage with a probability of $1/n_{hij}$, where n_{hij} is the number of eligible persons selected from the j^{th} household, i^{th} EA in the h^{th} stratum.

The final sample distribution of FinScope Thailand study is provided in Appendix 3.

Finally, it should be noted that although the examples provided thus far use an area frame at the first stage, this is by no means a requirement for multi-stage sampling. An example of a multi-stage sample using a different kind of frame is a travel survey conducted at an airport. The PSU could be time (for example, days in a month), while the second stage unit could be actual travellers. For a more complex travel survey, the second stage unit could be arriving passenger planes, while the third stage unit could be the actual seats on the plane. The theatre example at second stage of sampling has three days a week as sample elements.

Multi-phase sampling

Despite the similarities in name, multi-phase sampling is quite different from multi-stage sampling. Although multi-phase sampling also involves taking two or more samples, all samples are drawn from the same frame and the units have the same structure at each phase. A multi-phase sample collects basic information from a large sample of units, and then for a sub-sample of these units, it collects more detailed information. The most common form of multi-phase sampling is two-phase sampling (or double sampling), but three or more phases are also possible. However, as with multi-stage sampling, the more phases, the more complex the sample design and estimation (StatCan, 2003).

Multi-phase sampling is useful when the frame lacks auxiliary information that could be used to stratify the population or screen out part of the population. For example, suppose information is needed about cattle farmers, but the survey frame only lists farms, with no auxiliary information. A simple survey could be conducted whose only question is: “Is part or all of your farm devoted to cattle farming?” With only one question, this survey should have a low cost per interview (especially if done by telephone) and consequently the agency should be able to draw a large sample.

Once the first sample has been drawn, a second, but smaller sample can be drawn from amongst the cattle farmers and more detailed questions asked of these farmers. Using this method, the statistical agency avoids the expense of surveying units that are not in scope (that is, who are not cattle farmers). Multi-phase sampling can also be used to collect more detailed information from a sub-sample when there is insufficient budget to collect information from the whole sample, or when doing so would create excessive response burden.

The multi-phase sampling is also used by Statistics South Africa in a survey of employers and self-employed (SESE) (StatsSA, 2013). The first phase in SESE is identification of employers and self-employed using LFS through screening. Screening is a process where a filtering question is asked to identify candidates of the target population, for example, in the labour force survey the self-employed, business owners and employers are identified. The persons who meet the above criteria become part of the sample for the survey of employers and self-employed. SESE uses cell-based non-response adjustments that are applied to the original person weight from Phase I, that is, the quarterly labour force final weights.

According to Henry and Valliant (2012), these adjustments involve categorising the sample dataset into cells using covariates available for both respondents and non-respondents that are believed to be highly correlated with response propensity and key survey variables. Assuming that non-response is constant within each cell ("missing at random;" age/race/sex are often used in household surveys), the reciprocal of the cell-based response rate is used to increase the weights of all units within the cell. The final weight after Phase II is a product of the Phase I weight and the non-response adjustment based on the defined adjustment classes (StatsSA, 2013). The final weight is therefore calculated as $w_{Phase\ II} = w_{Phase\ I} * \mathrm{NR}_{adj}$ where $\mathrm{NR}_{adj} = \sum_{j=1}^{J} er_j / \sum_{j=1}^{J} en_j$ and $er =$ total eligible person within adjustment classes and $en =$ eligible respondents within each adjustment class. Other non-response weighting adjustments are discussed by Brick and Montaquila (2009) and are described in Section 2.2.3 (Brick and Montaquila, 2009). FinScope survey of micro, small and medium enterprises (MSME) makes use of multi-phase sampling methodology to study the population of small businesses (FinMark, 2010).

Multi-phase sampling can also be used when there are very different costs of collection for different questions on a survey. Consider a health survey that asks some basic questions about diet, smoking, exercise and alcohol consumption. In addition, suppose the survey requires that respondents be subject to some direct measurements, such as running on a treadmill and having their blood pressure and cholesterol levels measured. It is relatively inexpensive to ask a few questions, but the medical tests require the time of a trained health practitioner and the use of an equipped laboratory, therefore conducting survey without screening can be relatively expensive. Similar health survey could be done as a two-phase sample, with the basic questions being asked at the first phase and only the smaller, second phase sample receiving the direct measurements.

Replicated sampling

Replicated sampling involves the selection of a number of independent samples from a population rather than a single sample. Instead of one overall sample, a number of smaller samples, of roughly equal size, called replicates, are independently selected, each based upon the same sample design. Replicated sampling might be used in situations where preliminary results are needed quickly. Such preliminary results might be based upon the processing and analysis of a single replicate (StatCan, 2003).

The main reason for replicated sampling is to facilitate the calculation of the sampling variance of survey estimates (sampling variance is a measure of sampling error). While it is generally possible to calculate the sampling variance based on probability samples, such calculations can be exceedingly difficult depending on the complexity of the sample design. The problem is that some mathematical expressions for sampling variance are difficult to derive and tedious and costly to program. In particular, in the case of systematic sampling, variance estimates cannot be calculated directly, unless assumptions are made about the arrangement of units in the list.

Measures of sampling error are determined by examining the extent to which sample estimates, based upon all possible samples of the same size and design, differ from one another. Replicated sampling simulates this concept. Instead of drawing all possible samples (which is not practical), a reasonable number of smaller samples are selected using identical methods. For example, instead of selecting one sample of size 10,000, ten independent samples of size 1,000 could be drawn. The estimates from each of these ten samples can be compared and estimates of sampling variance derived. The reliability of the sampling variance estimates increases with the number of replicates selected.

There are a number of other procedures that use replication to estimate the sampling variance for complex sample designs. These include balanced repeated replication (BRR), Jackknife (Shao and Tu, 1995) and Bootstrap. These techniques, which all extend the basic idea of replicated sampling, differ from one another in terms of the accuracy with which they measure the sampling variance of different types of survey estimates and their operational complexity, as well as the situations to which they best apply.

There are drawbacks to this approach. One disadvantage of this scheme is that estimates of sampling variance, in general, tend to be less precise than if they were based directly on the statistical expressions that incorporate sample design features such as multi-stage, stratification, and so forth.

Statistical efficiency

SRS is used as a benchmark for evaluating the efficiency of other sampling strategies. In order to understand the concept of efficient sampling, some definitions are presented (StatCan, 2003). A parameter is a population characteristic that the client or data user is interested in estimating, for example, the population average, proportion or total. An estimator is a formula by which an estimate of the parameter is calculated from the sample, and an estimate is the value of the estimator using the data from the realised sample. The sampling strategy is the combination of the sample design and estimator used. For example, the parameter of interest might be the population average, $\bar{Y}$, which is calculated as follows:

$$\bar{Y} = \sum_{(i \in U)} \frac{Y_i}{N}, \tag{2.14}$$

where Y_i is the value of the variable Y for the i^{th} unit, U is the set of units in the population, and N is the number of units in the population.

For an SRS with 100 percent response rate, the usual but not the only estimator for the population average, is:

$$\bar{Y} = \sum_{(i \in S_r)} \frac{Y_i}{n}, \tag{2.15}$$

where S_r is the set of respondents in the sample and there are n units in the sample. The value of $\bar{Y}$ in Equation 2.15 for a particular sample is called the estimate.

Estimates calculated from different samples differ from one another. The sampling distribution of an estimator is the distribution of all the different values that the estimator can have for all possible samples from the same design and population. This distribution thus depends on the sampling strategy. Estimators have certain desirable properties. One is that the estimator be unbiased or approximately unbiased. An estimator is unbiased if the average estimate over all possible samples is equal to the true value of the parameter.

Another desirable property of an estimator is that the sampling distribution be concentrated as closely as possible about the average (that is, the sampling error be small). The sampling error of an estimator is measured by its sampling variance, which is calculated as the average squared deviation about its mean calculated across all possible samples generated from the sample design. An estimator with small sampling variance is said to be precise. Precision increases as the sampling variance decreases. Note that an estimator can be precise but biased. Accuracy is a measure of both the bias and precision of the estimator. An accurate estimator has good precision and is nearly unbiased. Accuracy is thus calculated by the mean square error defined as follows:

$$MSE = var + bias. \tag{2.16}$$

One sampling strategy is more efficient than another if the sampling variance of the estimator for the sampling strategy is smaller than that of another sampling strategy. So as not to confuse this type of efficiency with other types, for example, cost efficiency which was referred to as statistical efficiency, statistical efficiency is an important consideration when comparing different possible designs. If one design can provide improved or equivalent precision using a smaller sample size, the design can provide considerable cost savings. Formally, this is measured by calculating the design effect (Deff) to be discussed in Section 5.4.3.

2.2.4 Modified probability sampling

Modified probability sampling is a combination of probability and non-probability sampling. The first stages are usually based on probability sampling. The last stage is a non-probability sample, usually a quota sample. For example, geographical areas may be selected using a probability design, and then within each region, a quota sample of individuals may be drawn (StatCan, 2003).

2.3 Design weights and adjustments

2.3.1 Background

The probability methods of sample allocation and selection result to unequal inclusion probabilities. In a multi-stage probability sample design inclusion probabilities are calculated at each stage. For example, the first stage may involve selection of clusters or PSUs. The inclusion probability will be calculated for each unit. To compensate for unequal inclusion probabilities, weights for each unit are calculated. The weights are simply the inverse of the inclusion probabilities. The same procedure of calculating design weights is followed in subsequent sampling stages. Incomplete frame contributes to over/under-coverage and the compensation for those frame deficiencies is performed through various methods of benchmarking. Another form of sample survey problems that require adjustments is non-response which is compensated through non-response adjustment.

Almost all large scale sample surveys suffer from the problem of missing data, which result with the smaller sample size than expected. A far more serious effect of non-response is that estimates of population characteristics may be biased (Dihidar, 2014). The study of low-income population referred to in Mohadjer and Choudhry (2002) focuses on unit non-response, which occurs when a sampled unit (person or family/household) fails to participate in the survey. In some cases, collected data may be lost during data transmission stages (Mohadjer and Choudhry, 2002).

Missing data problem may occur even if an investigator tries to have all questions fully responded to in a survey (item non-response), or if the respondent is not available at home to answer the questionnaire (unit non-response). This would lead to less accurate, but still valid estimates of population characteristics, which can be taken care of by taking the larger sample size initially.

This situation occurs if, due to non-response, some groups in the population are over- or under-represented, and these groups behave differently with respect to the characteristics to be investigated. Consequently, wrong conclusions would be drawn from the survey data. The amount of bias created due to missing values often increases with the rate of occurrence of non-response. Above all, the large number of missing values in the dataset can also lead to computational difficulties (Dihidar, 2014).

Although both unit and item non-responses exist in surveys, for the purpose of this study, the focus was mainly on unit non-response. A unit non-response occurs when a sampled unit (person or family) fails to participate in the survey. Unit non-responses can occur, for example, because the sampled dwelling/-household/person cannot be located, refuses to participate, is too ill to participate, cannot participate because of language or hearing problems, or is away from the area for the period of the survey fieldwork (Mohadjer and Choudhry, 2002).

A unit non-response in surveys occurs for various reasons, including the failure to locate sampled persons and the refusal of sampled persons to be interviewed. In welfare studies that collect outcome data from administrative files, a non-response can occur because of inability to match the sampled case to the administrative file that includes the outcome data. Statistics derived from survey data may be biased if the missed persons are different, with respect to the variable of interest to the survey, from those who participated in the survey (Mohadjer and Choudhry, 2002).

2.3.2 Design weights

Once a “clean” data file has been obtained, the last processing step is the calculation of weighting or expansion factors (the “weights”) to be added to each record. The weights depend on the sample design. They indicate the number of population units represented by each sample unit. The weights will reflect the selection probabilities based on the sample design, adjustments for non-response provided that the respondents represent the non-respondents and adjustments based on auxiliary data, taking into account their relationship to the variables under study (discussed in the subsequent sections).

In the case of the South African Labour Force Survey (StatsSA, 2008) the design/base weight for each sampled household is equal to the reciprocal of the probability of selection, which is simply the inverse of the sampling rate. The sampling rate has been assigned at the province-level, that is, all design strata within a province have been sampled at the same rate. Thus, the initial base weight (or design weight) assigned to each household in a province is simply the inverse sampling rate (ISR) for the province.

A compromise allocation was used aiming at striking a balance between producing reliable provincial estimates and reliable national estimates. A number of procedures were available to achieve this compromise. The simplest and most commonly used allocation is the “square root” allocation. Under this allocation, the sample would be allocated to the provinces proportional to the square root of the population of each province. Under “square root” allocation, the sample is reallocated from very large provinces to the smaller ones, compared to what would have been obtained under proportional allocation.

A more general compromise allocation method is the “power allocation” method discussed by Bankier (1988), under which the sample is allocated proportional to x^{λ}, where x is the measure of size and the parameter λ can take values between 0 and 1. The value of $\lambda = 0.5$ corresponds to “square root allocation”. The two extreme values of λ give “equal allocation” and “proportional allocation”. In other words, $\lambda = 0$ corresponds to “equal allocation” and $\lambda = 1$ corresponds to “proportional allocation”. Kish (1988) also discussed a number of compromise allocations, including the “square root” allocation. Since the target is both national estimates and the estimates for each province, the “square root” compromise allocation was used to allocate the sample across the provinces.

It should be noted that the provincial ISRs are the initial stage of allocation and therefore the final ISRs for sampling strata were further adjusted. The ISR is simply the inverse of the inclusion probability and the latter is calculated as follows:

Apply the sampling rate b/M_i to the i^{th} PSU to obtain a sample size of

$$b_{psu} = \frac{b}{M_i} \times N_i, \tag{2.17}$$

where

b_{psu} is a households sample size
b sample size based on the MOS from design before listing of households
M_i is the total of MOS based on the design information
N_i is the total households within a PSU after listing and household count was done.

Moreover, the design weight, based on the sample design, depends only on the selection probability of each sample unit, called probability p. The design weight was $w = 1/p$ which implied that each unit in the sample represents w units in the population, including itself. If every unit in the population has the same selection probability p, then every unit in the sample has the same design weight, w. For more complex designs, the calculation of the design weights is more complicated. For example, for population of $N = 100$, $n = 25$ is to be selected using SRSWOR, $p = 25/100 = 1/4$, hence design weight $w = 4$ (Choudhry, 2009).

In a typical household survey, the initial design weights are common for all persons in the same household. These initial weights are usually adjusted for non-response to form a set of new weights, and sub-weights (Wu et al., 1997). Reid and Hall (2001) introduced weight equalisation approach which depends on the level at which the design weights are equal. In the case of multi-stage, sample elements within the same cluster are sampled together and are also adjusted for non-response in the similar manner. In this case, base weights are adjusted at the following levels: Stratum, PSU, segments and clusters, until household design weights produced. Section 2.3.4 on integrated weighting discusses further how the final weight is achieved through weight equalisation.

2.3.3 Non-response adjustment

Most surveys suffer from non-response, which occurs when all or some of the information requested from sampled units is unavailable for some reason. As presented in Section 2.3.1, there are two main types of non-response: item and total non-response. Item (or partial) non-response occurs when information is available for only some items, such as when the respondent answers only part of the questionnaire. In this case, the most common approach is to impute the missing values. Different approaches for imputing item or partial non-response will not be addressed in this study. The reader is referred to StatCan (2003).

This section focuses on total non-response, the case when all or almost all data for a sampled unit are missing. This can occur when the sample unit refuses to participate, no contact is made, the unit cannot be located or the information obtained is unusable. The easiest way to deal with such a non-response is to ignore it.

When non-response is present, a weight adjustment can partially compensate for the loss of data. This weight adjustment increases the weights of the sampled cases for which data were collected. The first step in adjusting for non-response is the construction of weighting classes. As discussed in the following text, within each weighting class, the base weights are inflated by the inverse of the response rate so that the sum of the adjusted base weights for respondents is equal to the sum of the base weights for the total eligible sample selected in the weighting class (Mohadjer et al., 1996).

In some exceptional circumstances, proportions or averages estimated without adjusting for total non-response are the same as those produced using a non-response adjustment. However, not compensating for the non-responding units leads to the under-estimation of totals (for example, the size of a population, total income or total acreage of crops).

The most common way of dealing with total non-response is to adjust the design weights based on the assumption that the responding units represent both responding and non-responding units. This is reasonable under the assumption that, for the characteristics measured in the survey, the non-respondents are like the respondents. The design weights of the non-respondents are then redistributed amongst the respondents. This is often done using a non-response adjustment factor that is multiplied by the design weight to produce a non-response adjusted weight.

A more technical overview of dealing with non-response is given by Holt and Elliot (1991). One way is to adjust the number of participants to the initial sample size. That is weight is multiplied by an adjustment factor δ such that:

$$w_i = d_i.\delta, \quad \text{with} \quad \delta = \frac{n_r + n_n}{n_r}. \tag{2.18}$$

Here n_r denotes the number of participants and n_n is the number of non-participants. This approach implicitly assumes that unit non-response occurs completely random (Steinhauer, 2014).

2.3.4 Benchmarking

It is often the case that the achieved sample does not represent the target population as closely as intended in terms of certain sub-groups being under-/

over-represented. This occurrence (such as too few young males or small households), is quite common in practice and could lead to biased results, if ignored. Therefore, it should be identified and controlled. Some of the approaches to handling coverage errors due to under-/ over-representation, are:

(1) Improved field procedures, and/or,

(2) Compensating for over-coverage and/or under-representation through the adjustment of design weights (Luus, 2016).

The final stage of weight construction is where this design weight adjustment occurs. This stage makes use of auxiliary information obtained from census data or other population data sources, to further adjust the non-response adjusted weights of the sampled units such that the weighted estimates of the population totals conform to the actual known population totals of such variables (Neethling and Galpin, 2006). The following weight adjustment methods exist under this approach and were considered in this study:

(1) Post-stratification,

(2) Calibration weighting,

(3) Integrated weighting, and

(4) Raking.

Difficulty in analysing complex surveys arises due to survey weighting adjustments for known or expected differences between a sample and its population. These differences arise from sampling design, under-coverage, non-response, and limited sample size in sub-populations of interest. Weights are constructed based on design and benchmarking calibration variables that are predictors of inclusion probability, which is defined as the probability that unit i was included in the sample, where inclusion refers to both selection into the sample and response given selection. However, weights have problems. As pointed out by Gelman (2007) and the associated discussions, current approaches for construction of weights are not systematic, with much judgement required on which variables to include in weighting, which interactions to consider, and how weights should be trimmed.

Post-stratification

Post-stratification is a popular estimation procedure in which the weights of the respondents are adjusted so that the sums of the adjusted weights are equal to known population totals. For example, suppose that the age and sex of all respondents are obtained in a survey and that the population age-by-sex distribution is known (for example, from a census or administrative file). Post-stratification adjusts the weights of the respondents so that the respondent distribution by age and sex (when weighted by the post-stratified weights) is the same as the population distribution.

Since post-stratification involves adjusting the weights of the respondents, the post-stratified estimates are different from the estimates produced without post-stratification. The variances of the estimates are also different. Post-stratification is often used to reduce the variance of the estimates or to correct for survey under-coverage of some types of units. Cochran (1977) describes post-stratification in more detail.

Since post-stratification is used frequently and it affects the estimates and their precision, WesVar software has a function to post-stratify weights. One must provide an auxiliary information with post-strata cell identifiers and control totals. For the above example, this would be identifiers of the age-by-sex cell and the population total in that cell. In addition, the values of the cell identifiers must match exactly, the values for a user-specified variable on the WesVar data file. For example, if cell identifier 3 is for males aged 24 to 35 years on the file with the control totals, it must have the same meaning for the cell identifier variable on the WesVar file. If this correspondence is not exact, the post-stratification cannot be done (Westat, 2006).

Henry and Valliant (2012) stated that post-stratification survey weights are adjusted such that they add up to external population counts by available domains. This widely-used approach allows us to correct the imbalance that can occur between the sample design and sample completion, that is, if the sample respondent distribution within the external categories differs from the population (which can occur if, for example, subgroups respond or are covered by the frame at different rates), as well as reduce potential bias in the sample-based estimates. Denoting the poststrata by $d = 1, \ldots, D$, the post-stratification estimator for a total involves adjusting the base-weighted domain totals ($\hat{T}_d$) by the ratio of known (N_d) to estimated ($\hat{N}_d$) domain sizes:

$$\hat{T}_{PS} = \sum_{d=1}^{D} N_d \hat{T}_d / \hat{N}_d. \tag{2.19}$$

Raking

Raking (also known as iterative proportional fitting) is done in place of post-stratification. Unlike post-stratification, raking is performed iteratively to two or more different distributions of a population total (for example, gender and age). It is typically used in situations in which the interior cells of a cross-tabulation are either unknown, or some sample sizes in the cells are too small for efficient estimation. In raking, the marginal population totals, $N_{i.}$ and $N_{.j}$ are known (those are, age and gender population counts). However, the interior cells of the cross-tabulation N_{ij} (the age by gender cells) are estimated from the sample by $\hat{N}_{ij}$, where these are the sum of weights in the cells (Allen et al., 1999).

The raking algorithm proceeds by proportionally scaling the N_{ij} such that the following relations are satisfied:

$$\sum_{j} \hat{N}_{ij} = N_{i.} \tag{2.20}$$

and

$$\sum_{i} \hat{N}_{ij} = N_{.j.}. \tag{2.21}$$

Wallace and Rust (1996) compared post-stratification and raking using National Assessment for Educational Progress (NAEP). The findings revealed that post-stratification reduced variances of NAEP by 50 percent over what would be achieved using non-post-stratified weights. It was postulated that raking, as opposed to post-stratification, might reduce variances further. The reason for the improvement was that raking would control the distribution of the final sample weights with respect to greater variety of variables related to educational achievement than could be achieved through stratification.

Calibration

One of the goals of weighting survey data is to reduce variances in the estimation procedure by using auxiliary information that is known with a high degree of accuracy (Mohadjer et al., 1996). Incorporating auxiliary totals to survey estimates, characteristics in the survey that are correlated with the auxiliary totals can be estimated with precision.

Calibration weighting offers a way to incorporate auxiliary information into survey estimates so that, in general, characteristics that are correlated with the auxiliary variables are estimated with greater precision (Reid and Hall, 2001). The information required for calibration is a set of population control totals $\sum_U x_k$, where U is the finite population universe and the x_k are vectors of auxiliary information that are known individually only for elements in the respondent sample.

Calibration uses this information by constructing weights such that $\sum_s w_k x_k = \sum_U x_k$, where s represents the respondent sample and w_k is the calibrated weight for element k.

Typically, there are many possible choices of weights that satisfy this benchmarking constraint. Calibration, by its classical definition, produces the one that is closest to the design weights, with closeness determined by a suitable distance function (see Deville et al. (1993), for details). So selecting a calibration estimator reduces to the selection of a distance function.

Unavailability of auxiliary totals for some stages of sample design household and family level weighting is one of the common problems. In a typical household survey, the initial design weights are common for all persons in the same household. These initial weights are usually adjusted for non-response to form a set of new weights, called sub-weights. The sub-weights are then calibrated to known auxiliary controls to improve efficiency as well as to adjust for coverage bias.

The calibrated weights need not be common for persons in the same household because auxiliary controls may include person-level characteristics such as population counts for age and sex groups. When an estimate of characteristics of household (for example, household income) is needed, each household must have a single weight. Several methods have been proposed for producing this single weight, including generalised regression methods (Wu et al., 1997). This objective of producing single weight is further explored in a special case of calibration called integrated weighting.

Integrated Weighting

Integrated weighting is also referred to as weight equalisation, which implies that in producing final household weight, persons' attributes are included and all persons in the household will have one weight, and the same household weight is also used to estimate household total. Reid and Hall (2001) stated that when two sample elements have the same design weight and the same values of their corresponding auxiliary vectors, their calibrated weights will be the same. This property leads to the method of equalising weights that is followed in this study.

When two sample elements have the same design weight and the same values of their corresponding auxiliary vectors, their calibrated weights will be the same. This property leads to the method of equalising weights. The objective of calibration has been stated in the previous section as $\sum_s w_k x_k = \sum_U x_k$ and the additional constraint is $w_k = w_l$.

Now there is a situation where there are two or more sampled units, such as grouped persons within the household. Each person will be classified in a category determined by calibration class for which there are corresponding vectors of auxiliary data.

New vectors formed

$$\sum_u z_k = \sum_U x_k, \tag{2.22}$$

$$\sum_s w_k z_k = \sum_s w_k x_k, \tag{2.23}$$

$$z_k = z_l, \tag{2.24}$$

wherever $w_k = w_l$ is wanted, only allowing the constraint if elements k and l are in the same ultimate sampling unit or household.

Elements that are analysed should be coming from the same PSU/SAL for the following reasons. Firstly, their design weights will be equal, and secondly, the z-vectors must be known for all elements in the sample. So, if z_k is a function of x_k and the x-vectors of the elements, those other elements should be guaranteed to be in sample whenever k is in the sample. Constraining these other elements to be in the same PSU/SAL as element k would satisfy the requirement (Reid and Hall, 2001).

The construction is simple if all elements in a group p are needed to have the same weight after calibration, define z_k, for k in p, to be the average of the x-vectors in p. Group p is the container of sample elements in which weights must be equal. That is, $z_k = \sum_p x_k/n_p$, where n_p is the number of elements in p. There are two main reasons to require that elements be in the same USU/household/family if their weights are constrained to be equal. The first reason is that their design weights will be equal, which was one of the assumptions. This is a necessary condition for the approach to lead to equal calibration weights. Secondly, the z-vectors must be known for all elements in sample. So, if z_k is a function of x_k and the x-vectors of other elements, those other elements should be guaranteed to be in sample whenever k is in sample. Constraining these other elements to be in the same USU as element k would satisfy this requirement. Equation 2.22 ensures that the elements have the same weight after calibration and all the 2.23 and 2.24 equations are satisfied.

2.4 Estimation

2.4.1 Estimation of population totals

Surveys are frequently required to produce estimates for sub-populations, sometimes for a single sub-population and sometimes for several sub-populations in addition to the total population. Some domains may represent a rare population which was not catered for during the design, for example, a population of small business owners and self-employed whose population is not known. When membership of a rare sub-population (or domain) can be determined from the sampling frame, selecting the required domain sample size is relatively straightforward. Statistical estimation methodologies are employed to optimise the estimation of such domains.

Further than estimation of population totals the CSSD should be designed in such a way that they enable estimation of domains. A domain (area) is regarded as large (or major) if the domain-specific sample is large enough to yield “direct estimates” of adequate precision. A domain is regarded as “small” if the domain-specific sample is not large enough to support direct estimates of adequate precision. Some other terms used to denote a domain with small sample size include: “local area”, “sub-domain”, “small sub-group”, “sub-province” and “minor domain”. In some applications, many domains of interest (such as counties) may have zero sample size (Rao, 2003). Rao (2003) uses the term *“small area”* to denote any domain for which direct estimates of adequate precision cannot be produced. Although small area estimation is identified as one of the problems in complex sampling surveys that are designed for larger areas, the methods to produce small area estimates will not be fully covered in this study.

Following the theoretical framework is often assumed in estimating population parameters. A survey population U consists of N distinct elements identified through the labels $j = 1, ..., N$. The characteristic of interest y_j associated with element j is exactly known by observing the element j. A sample is a subset, s, of U and associated y-values, that is, $\{(i, y_i), i\epsilon s\}$, selected according to a specified sampling design which assigns a known probability $p(s)$ to s such that $p(s) > 0$ for all $s\epsilon S$, the set of possible samples s, and $\sum_{s\epsilon S} p(s) = 1$. Consider general parameters of interest

$$H = \sum_{j\epsilon U} h(y_j) \text{ and } \bar{H} = N^{-1}H \tag{2.25}$$

for specified function h. The choice $h(y) = y$ gives the population total $H = Y$ and the population mean $\bar{H} = \bar{Y}$, while the choice $h(y) = \Delta(t - y)$ with $\Delta(a) = 0$ otherwise, gives the distribution function:

$$\bar{H} = F(t) = N^{-1} \sum_{j\epsilon U} \Delta(t - y_j) \tag{2.26}$$

for each t, where t is a pivotal quantity, which means its distribution does not depend on y.

Rao (1994) outlined three different approaches to inference of H or $\bar{H}$, those are, (1) design-based also called probability, (2) model-dependent approach and hybrid or model assisted report. Probability sampling approach refers to repeated sampling from survey population U involving all samples $s\epsilon S$ and associated probabilities $p(s)$. It provides valid inferences irrespective of the population y-values in the sense that pivotals $t_1 = (\hat{H} - H)/s(\hat{H})$ and $t_2 = (\hat{\bar{H}} - H)/s(\hat{\bar{H}})$ are approximately distributedN(0,1), at least for large samples, where $(\hat{H}, \hat{\bar{H}})$ and $(s^2(\hat{H}), s^2(\hat{\bar{H}}))$ are design-consistent estimators of $(H, \bar{H})$ and $Var(\hat{H}), Var(\hat{\bar{H}}))$, respectively.

Survey estimates for any estimation domain d can be computed using the set of final weights for the respondents in the domain of interest. For example, the domain could be a geographical domain, such as, a region or a province; or it could be a characteristic domain, such as, persons employed in a particular industry group, and so forth. Moreover, a domain can also be defined as the intersection of domains, such as, persons employed in a certain industry in a given province. Let us say that the interest is in estimating the number of persons employed in agriculture in the province of Western Cape, and then the domain of interest would be the persons employed in agriculture in Western Cape. The symbol $(hijk)$ will be used to denote the respondent k in the sampled household j from the sampled PSU_j in design stratum h. An indicator variable defined as ${}_d\delta_{hijk}$ to indicate whether the respondent $(hijk)$ belongs to the estimation domain d, is as follows.

$$
{}_d\delta_{hijk} = \begin{cases} 1 & \text{if the respondent } (hijk) \text{ belongs to the estimation domain } d, \\ 0 & \text{otherwise.} \end{cases}
$$

The desired estimate of total number of persons in the domain d can then be obtained as ${}_d\hat{N} = \sum\limits_{hijk} \times {}_d\delta_{hijk}$, where w_{hijk} is the final weight of the respondent $(hijk)$ and the sum is over all respondents. Estimates of ratios and averages for the domain can be obtained as the ratios of the estimated domain totals.

In general estimation of totals for stratified samples can be calculated as follows:

$$
t_{str}(x) = \sum_{h=1}^{L} \sum_{i=1}^{nh} t_{hi}, \tag{2.27}
$$

where

$$t_{hi} = \sum_{j \in PSU_{hi}} w_{hij} x_{hij} \tag{2.28}$$

is the estimate of the total based on the i^{th} PSU/SAL in the h^{th} stratum.

2.4.2 Variance estimation

Variance estimation in complex surveys serves two goals. First, applied researchers need standard errors to test their hypotheses of substantive interest and construct (Wald) tests and confidence intervals. Second, sample designers use variance estimates to gauge performance of existing designs and choose design parameters for future surveys of similar populations (Kolenikov et al., 2010). There are several variance estimation methods commonly used with complex survey data. This section of the study starts by discussing the direct variance estimation and then resort to resampling methods.

For a very complex survey design, exact accounting for all its features is extremely cumbersome. At the data analysis stage, approximations are often made to yield a usable estimation formula. The most common approximate design is stratified two-stage sampling with replacement (S2SWR). In this design, the population is divided into strata. From each stratum, a sample of PSUs is taken with replacement, and from each PSU, samples of ultimate units are taken.

In the S2SWR design, the variance of $t_{str}(x)$ of the above estimate under stratified sample can be directly estimated by

$$v_{str}\{t_{str}(x)\} = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (t_{hi} - \bar{t}_h)^2, \bar{t}_h = \frac{1}{n_h} \sum_{i=1}^{nh} t_{hi} \tag{2.29}$$

The finite population corrections are ignored because sampling is assumed to be with replacement in the S2SWR design. When complex survey designs are used to collect data, special techniques are needed to obtain meaningful and accurate analyses. Replicate variance estimation is a robust and flexible approach that can reflect the complex sampling and estimation procedures used in practice.

In producing estimates for sample surveys, it is a generally recommended practice that due concern be given to quantifying the extent of sampling error (Rust, 1985). There are formulae provided for unbiased estimators of sampling variance of simple (linear) estimators of, say, totals and means from a population of known size. This is provided for a variety of the more straightforward sample designs. However, such formulae are usually inadequate in survey practice because non-linear estimators are often used for estimating parameters of interest.

Replication can be used with a wide range of sample designs, including multi-stage, stratified, unequal probability samples as well as complex sample designs. Replication variance estimates can reflect the effects of many types of estimation techniques, including non-response adjustment, post-stratification, raking, and ratio estimation. Rust (1985) stated that if a given sample design is used to select a sample, and then repeated r times, the final sample is made up of r independent replicates, each with identical design. The resulting sample is called a simple replicated sample. Each replicate sample then provides an estimate of the parameter of interest. The variability among the r replicate estimates then gives a measure of variance of overall sample estimator, which is the simple average of the r replicate estimators.

Rao and Shao (1999) stated that balanced half-samples or balanced repeated replication (BRR), has long been used as one of the methods to estimate the variances of non-linear estimators from stratified sampling designs. Compared with the Taylor linearisation method for variance estimation, the BRR requires longer central processing unit time to run, but has the following advantages. Firstly, it requires no theoretical derivation of a variance formula for each problem, which can be difficult and messy. Secondly, programming case in complex situations, meaning that computer programs can be used to process BRR. Lastly, Using a unified recipe for a variety of problems, for example, variance estimations for functions of averages and for sample quantiles.

BRR also has a wider application scope than Jackknife, another popular replication method, since the latter is known to have problems in estimating variances of non-smooth estimators such as sample quantiles. BRR works by forming balanced replicates, half-samples or subsets of the whole dataset, re-computing the survey estimate on each replicate and taking the mean squared derivation of the replicate estimates as the variance estimate.

BRR was introduced by McCarthy (1969) for the class of designs in which $n_h = 2$ for all strata. In each replicate, one of the two PSUs is omitted, and the other one is retained and replicated twice to ensure that the totals are on the right scale. Because exactly half of the PSUs are used, the replicates are also referred to as half-samples. The replicate weights are:

$$w_{hij}^{(r)} = \begin{cases} 2w_{hij,} & PSU_{hi} \text{ is retained} \\ 0, & PSU_{hi} \text{ is omitted} \end{cases}$$

These weights are used to compute $\bar{\theta}_{BRR}^{(r)}$. The BRR variance estimator is as follows:

$$v_{BRR} = \frac{1}{R} \sum_{r=1}^{R} \left\{ \hat{\theta}_{BRR}^{(r)} - \hat{\theta} \right\}^2 \qquad (2.30)$$

The number of all possible half-samples is 2^L, where L is PSU with a stratum. If all half-samples are used, $v_{BRR} = v_L = v_r$ in the linear case. One of the well-known limitation is that BRR requires 2 PSU/SAL per strata. This can cause a problem if the strata are already formed and they do not contain 2 PSU/SAL per strata. The limitation can often be overcome through grouping PSU/SAL using some characteristics in the frame. The example is that of the sampling frame used to draw a sample for the South African Quarterly Labour Force Survey (QLFS) (StatsSA, 2008). The South African QLFS sampling frame was designed in such a way that households within PSUs would be rotated after four times in the sample. In the frame there were four rotation groups and each PSU was assigned a rotation group numbers. Using the rotation groups, a variable that represents two PSUs per stratum was formed by combining odd number rotation groups (1 and 3) as well as even number rotation groups.

Many procedures available in standard statistical software packages are not appropriate for analysis of data from complex survey design, because the analyses are based on the assumption that the sample has been drawn with SRS. Choudhry and Valliant (2002) further discussed WesVar software that computes estimates and replication variance estimates by properly reflecting complex sampling and estimation procedures.

Nearly all surveys use complex sample designs to collect data that are frequently used for statistical analyses beyond the estimation of simple descriptive parameters of the target population. Many procedures available in popular statistical software packages are not appropriate for this purpose because the analyses are based on the assumption that the sample has been drawn with SRS. Therefore, the results of the analyses conducted using these software packages would not be valid when the sample design incorporates multi-stage sampling, stratification, or clustering (methods defined in this study as the features of CSSD).

In our study Fay's method will be used to estimate variances. Fay's method is a variant of the BRR, but has better properties in certain situations (Fay and Dippo, 1989). Standard BRR can run into problems when computing an estimate for a small domain or estimating a ratio with very few sample cases for estimating the denominator. The basic idea of Fay's method is to correct this problem by modifying the sample weights less than in BRR, where half the sample is zero-weighted while the other half is double weighted in each replicate. Using Fay's method, one-half sample is weighted down by a factor K $(0 < K < 1)$ and the remaining half is weighted up by a factor 0.5. A perturbation factor of around 70 percent is generally recommended with Fay's method, which is achieved by using a value equal to 0.5 in the FAY_K box in WesVar.

The variance estimate under Fay's method is computed as:

$$v(\hat{\theta}) = \frac{1.0}{R(1.0 - K^2)} \sum_{r=1}^{R} (\hat{\theta}_{(r)} - \hat{\theta}), \tag{2.31}$$

where

θ is an arbitrary parameter of interest,

$\hat{\theta}$ is the estimate of θ based on the full sample,

$\hat{\theta}_{(r)}$ is the estimate of θ based on the r^{th} replicate, and

$v(\hat{\theta})$ is the estimated variance of $\hat{\theta}$.

R is the number of replicate samples, and K is the Fay's K-Factor, where $0 < K < 1$. It should be noted that the replicate estimates are obtained using the replicate final weights that are obtained by applying the non-response adjustment, and then benchmarking to the independent population counts.

The standard error of an estimate is defined as the square-root of the variance of the estimate. The estimate of $\hat{\theta}$ is denoted by an arbitrary population parameter θ, and $v(\hat{\theta})$ is the corresponding variance estimate. Then the standard error of the estimate $\hat{\theta}$ is given by $se(\hat{\theta}) = \sqrt{v(\hat{\theta})}$. The standard error can be used to express the precision of an estimate by computing the 95 percent confidence interval, or the coefficient of variance (CV) of the estimate. Those measures of precision are discussed in the section that follow.

The Bootstrap method is also used in variance estimation process for smooth and non-smooth functions, such as quantiles. Inference in parametric statistical procedures is based on sampling distributions of parameter estimates and test statistics. These distributions can often be derived by transformations of the underlying random variables or by asymptotic arguments. The Bootstrap provides an alternative paradigm: it mimics the original sampling procedure to obtain approximations to the sampling distributions of the statistics of interest. The Bootstrap samples are taken from a distribution that is close, in some suitable sense, to the unknown population distribution. A typical choice is the empirical distribution of the data (Kolenikov et al., 2010).

Let the sample data $x_1, ..., x_n$ be independent and identically distributed (i.i.d.) from distribution F characterised by parameter $\theta = T(F)$. The empirical distribution function of the data is F_n, and the associated parameter estimate is $\hat{\theta} = T(F_n)$. The Bootstrap takes a simple random sample with replacement $(x_1^*, ..., x_m^*)$ of size m from $x_1, ..., x_n$. The empirical distribution function of the Bootstrap sample is F_m^*, and the associated parameter estimate is $\hat{\theta}_m^* = T(F_m^*)$. The Bootstrap distribution of $\hat{\theta}_m^*$ is obtained by taking different Bootstrap samples and computing $\hat{\theta}_m^*$ for each of them.

The plug-in principle of the Bootstrap states that relation of the Bootstrap values $\hat{\theta}_m^*$ to $\hat{\theta}_n$ is approximately the same as that of $\hat{\theta}_n$ to the unknown parameter θ. Typically, but not necessarily, $m = n$. If this is the case, the Bootstrap estimates of the moments and the distribution function of $\hat{\theta}_n$ are

$$\begin{aligned}
Bias(\hat{\theta}_n) = E(\hat{\theta}_n - \theta) &\doteq E^*(\hat{\theta}_n^* - \hat{\theta}_n), \\
V(\hat{\theta}_n) = E\left[\left\{\hat{\theta}_n - E(\hat{\theta}_n)\right\}^2\right] &\doteq E^*\left[\left\{\hat{\theta}_n^* - E(\hat{\theta}_n)\right\}^2\right], \\
MSE(\hat{\theta}_n) = E\left\{(\hat{\theta}_n - \theta)^2\right\} &\doteq E^*\left\{(\hat{\theta}_n^* - \hat{\theta}_n)^2\right\}, \\
cdf_{\theta n}(x) = Prob(\hat{\theta}_n - \theta < x) &\doteq Prob^*(\hat{\theta}_n^* - \hat{\theta}_n < x),
\end{aligned} \qquad (2.32)$$

where the starred quantities are taken with respect to the Bootstrap distribution. The particular strength of the Bootstrap is the last equation of (2.32). The Bootstrap accounts for asymmetry of the sampling distributions and gives better one-sided confidence interval coverage than the confidence intervals based on asymptotic normality (Efron and Tibshirani, 1994).

Chapter 3

Sampling methodology applied to the South African Census 2011 data

3.1 Sampling frame creation

3.1.1 Background

This chapter covers the implementation of sampling methodologies described in Chapter 2, particularly using probability sampling. The two main types of sampling are probability sampling and non-probability sampling as described in Chapter 2. Non-probability sampling is of limited use for surveys conducted by statistical agencies, since the biased selection of units does not readily permit inferences to be made about the surveyed population. However, it is fast, easy and can be useful for exploratory studies or during the development phase of a survey (for example, to test the questionnaire).

The emphasis and application of this study is on probability sampling which should be used when inferences about the population are to be made based on the survey results. In a probability sample, every unit on the frame has a known non-zero probability of being selected and the units are selected randomly. As a result, selection is unbiased and it is possible to calculate the probabilities of inclusion and the sampling variance of estimates, and also make inferences about the population. The main disadvantages of probability sampling are that it requires more time, it is costlier than non-probability sampling and requires a high quality sampling frame.

The simplest probability sample designs are simple random sampling (SRS) and systematic (SYS) sampling, which result in equal probabilities of inclusion. More complex designs that can result in unequal probabilities of inclusion and most of which require auxiliary information include: stratified, probability proportional to size (PPS), cluster, multi-stage and multi-phase sampling. Unequal probability designs are typically used to improve the statistical efficiency of the sampling strategy or reduce the cost of sampling. Sometimes their use is dictated by the sampling frame.

The above considerations should be taken into account when deciding between the various possible designs. We first determine what designs are feasible given the survey frame, units on the survey frame, domains of interest, response burden, the method of data collection, budget, and so forth.

Some other important issues to consider are:

- Does the survey frame have auxiliary data that could be used to improve the efficiency of sampling?

- Does the survey frame lack auxiliary information that could be used to screen out units or that would be useful for stratification? Is data collection very expensive or burdensome? (Should a two-phase sampling be performed?)
- Is the population naturally clustered or are the units on the survey frame are clusters themselves? Is the population spread out geographically and personal interviews need to be conducted? (Should single-stage or multi-stage cluster sampling be performed?)
- Examine several special applications of sample designs that can be made depending on the specific needs of the survey.
- Determine the size of sample required to satisfy a given level of precision, and how to compare the efficiency of different sample designs by comparing design effects (Deff).

3.1.2 The primary sampling units for area sampling

The primary sampling units (PSU) for the current sample design is small area layer (SAL) defined by Statistics South Africa after Census 2011 (see the map in Figure 3.1). Throughout the sampling process, SAL was used to refer to the ultimate geography level of selection or the PSU. These areas are made of a combination of Census 2011 enumeration areas (EAs) to create geography areas that are large enough to preserve confidentiality of information provided by the respondents and also small enough to be PSUs. These sampling units are clearly identifiable geographical areas that are represented by geo-referenced polygons linked to the entire South African census geography.

Figure 3.1 contains a map of SALs in the main-place *Mankweng*, situated on the eastern side of the Polokwane city centre, Polokwane municipality, Capricorn district, Limpopo province, Republic of South Africa. Limpopo province is one of the nine provinces of South Africa. Any of the numbered SALs on the map has a chance to be selected through the probability sample design process.

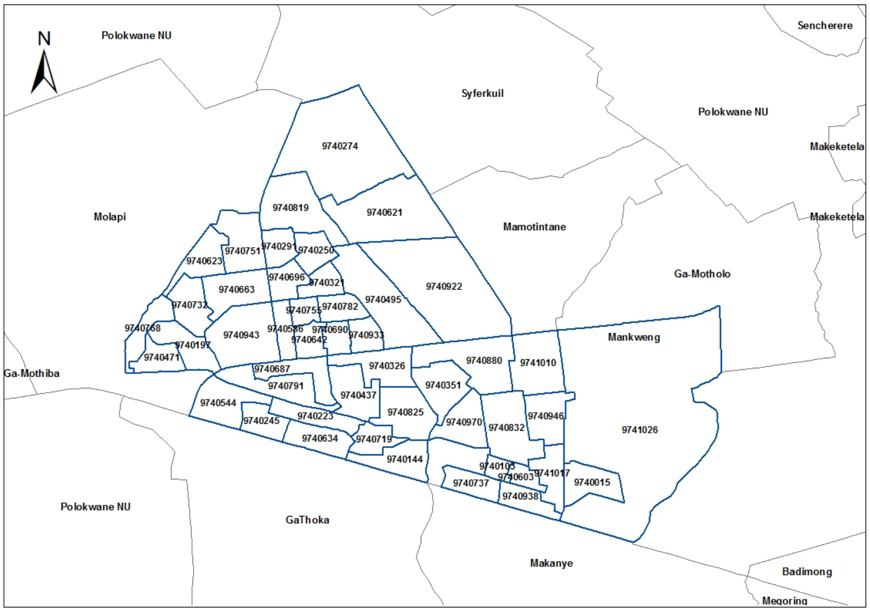

Figure 3.1: SALs under the place name Mankweng.

The total number of SALs published were 85,101 and they were spread across all the nine South African provinces. Among the list of SALs, a total of 3,840 were not considered for sampling. They belonged to EA types that were not primarily residential, although there may be people living inside those SALs. Excluded EAs are made up of parks and recreation, industrial and educational institutions, vacant areas and commercial EAs.

In Figure 3.1, SAL number "9741026" on the far right is an educational institution, the University of Limpopo. Based on the inclusion rules, this SAL was not part of the sample selection since the sampling frame was only focused on conventional households and converted hostels. There are cases where EAs or SALs were not classified as hostel or institution, however, hostels or institutions are identified during data collection. When institutions are discovered during data collection, enumeration still proceeds since the probability sampling in most instances, and in this study, in particular, is without replacement.

3.1.3 Enriching the frame

Identifying SALs to be modified

The first SAL sampling frame after the exclusion of non-targeted or out-of-scope EA types, while retaining very small SALs, contained 81,261 SALs (see Appendix 1). The new dataset had only SALs that contained eligible households from which the sample could be drawn. The remaining SALs were also grouped based on the household size and the **`SALgrp`** (that is, SAL group was formed). The following SAS Code was used to group the SALs based on total households. The code is an extract from a program (SAS Code 1).

```
If total_hh lt 50 then SALgrp=0;
Else If total_hh lt 100 then SALgrp=1;
Else If total_hh lt 150 then SALgrp=2;
..................................;
Else If total_hh lt 1000 then SALgrp=9;
Else SALgrp=10;
```

The SALs with less than 100 households, contained in groups 0 and 1, were collapsed. SALs in groups 2 to 7 were sampled as they were. It was expected that the number of households in an SAL would be between 100 and 400. Those that fall in group 8 were split conceptually into two groups, while those in groups 9 and 10 were split into three groups conceptually.

There were 81,261 observations read from the dataset **`msc.msc_frame006`** (“Msc" is a **`SAS`** project created for the Master of Science(Msc) project implementation). The dataset for pooled SALs (**`msc.msc_frame_collapse`**) had 10,897 observations. The SALs that met the criteria to be included in the frame without pooling or splitting (**`msc.msc_frame007`**) had 67,569 observations. The dataset for SALs to be split (**`msc.msc_frame_split`**) had 2,795 observations.

Out of 81,261 SALs, 10,897 were collapsed within the same geography type and sub-place to reach the minimum of 100 households in an SAL and resulted in 8,557 pseudo SALs. A total of 67,569 SALs were within the required range of 100 to 400 households, while 2,795 SALs were split conceptually. There were 8,557 observations read from the dataset **`msc_sal_poolx`**. The dataset **`msc_sal_poolx1`** has 818 observations. The dataset **`exclude_sp`** had 7,739 observations.

The newly formed SALs were checked if they met the minimum required household size, and 7,739 still did not meet the minimum criteria and were excluded from sampling frame. The remaining 818 were further assessed against the minimum criteria and the limit of the formation of pseudo-SALs was set to a maximum of 6 SALs. This was to ensure that not more than 6 SALs formed a final pooled SAL for sampling.

There were 8,713 observations in the dataset **`msc_final_pooled`** and the dataset **`msc.msc_final_pooled1`** had 972 observations.

The final frame consisted of 75,314 SALs after the splitting of big SAL and the pooling of smaller SALs (see Appendix 2).

Explicit stratification using natural clusters

There were two options for creating explicit strata: strata formed before pooling and splitting; and strata formed after pooling. The latter had an advantage that all the new split-SALs would fall within the same explicit strata, however, in this study, the former option was implemented. Explicit stratification was implemented using geography hierarchy variables to group SALs that are situated in similar geography areas. The following geography variables from the frame were used:

- **`Pr_Code_2011`** (provincial code of 2011 boundaries)
- **`Metro`** (distinguishes metro and non-metro variables within each province)
- **`EA_Gtype`** (EA geography type).

A total of 39 explicit strata were formed through concatenating (1) province, (2) metro and non-metro grouping of SAL, as well as the (3) geography type (see Appendix 1).

Combining SALs

SALs that formed part of a list to be collapsed were 10,897. The initial criteria of pooling is that the small areas should fall in a sub-place and a geography type. Those SALs were collapsed or pooled using the following process:

Step 1

Summarise total households by sub-place and geography type:

```
Proc summary data=msc_SAL_pool01 sum nway;
Class SP_Code EA_GTYPE_C;
    Var Total_HH;
    Output out=msc_SAL_poolx sum=Tot;
Run;
```

Step 2

Within each sub-place and a geography type, order the SAL numbers following the spatial order numbering already implied in the SAL number (that is, from South-West to North-East). Take the first six SALs to form a new SAL and all the SALs will be identified based on the first SAL number.

Step 3

Create the household totals for each SAL following the similar order of the above SALs. The sum of households for the SALs within a sub-place and geography type becomes the households' total for the new pooled/combined SALs.

Step 4

Obtain other census characteristics from the initial frame for each SAL. Aggregate the census attributes of each pooled SAL into one SAL number used to identify the rest of the linked SALs.

Step 5

Check if there are SALs with household totals below 100 and only retain those with totals of more than 100 to add to the final sampling frame.

Figure 3.2 is an example of the resulting group of SALs that were pooled to create one SAL/PSU. These SALs were all part of the same sub-place *Melodie AH* in the North West province, South Africa, and they are adjacent to each other. The group of EAs in the frame are identified by the SAL code "6610021". During data collection, enumerators should have all the associated pooled SALs (those are: "6610021", "6610024", "6610028", "6610029" and "6610031").

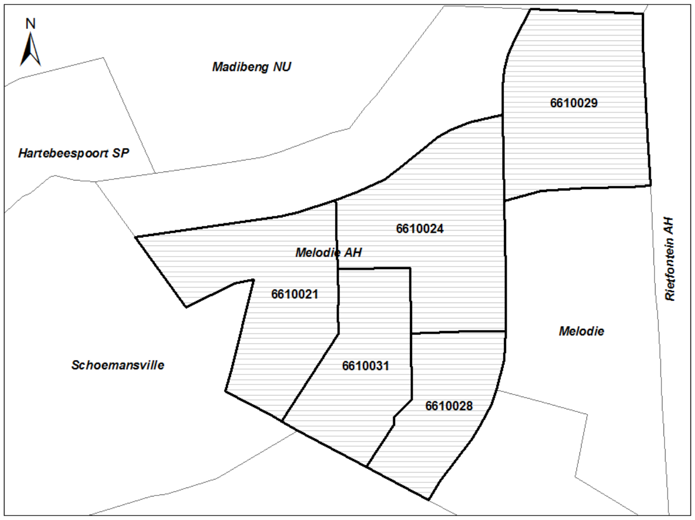

Figure 3.2: Example of pooled SALs to form SAL for sampling.

Splitting SALs

A total of 2,795 SALs formed part of the list to be split. The rule is that all SALs falling within SAL Flags 8, 9 and 10 were subjected to splitting. Flag 8 was split into two SALs and Flags 9 and 10 were split into three SALs each. The following steps were carried-out to split SALs.

Step 1

Create two split files from the SAL Flag 8 by dividing all the census attributes (contained in the frame dataset) by two, followed by splitting Flags 9 and 10 into three SALs and follow the same process as in two splits, dividing census attributes by three. The first split of two will be numbered 1 and the second 2, while, the first split of three will be numbered 1, the second 2 and the third 3.

Step 2

Create the new **`SAL_Number_S`** (small area layer number after splitting) by adding the split number after the original SAL number to create a unique SAL number which incorporates and distinguishes split SALs. All the SALs in the frame that were not split were assigned 0 as a split number and the 0 was added after the original SAL number.

Step 3

Combine the split file with the entire frame to create a frame for sampling. The new unique identifier is **`SAL_CODE_S`**, which incorporates splits and also takes pooled SALs into account.

Table 3.1: Numbering of SALs after splitting.

SAL Code	**Split No.**	**SAL Code Split**
7994290	1	79942901
7994290	2	79942902
7994290	3	79942903

Table 3.1 is an example of a split SAL in Gauteng province, South Africa. The SAL “7994290” had over 1,000 households, and it was conceptually split into three different SALs which all have an equal chance of being selected independently. It should be noted that there were no physical boundaries resulting from splitting of the SALs.

Clustering for implicit stratification

At this point the 39 strata are already formed. It was thought that using auxiliary variables at SAL level to further stratify SALs the following benefits can be realised. Firstly, secondary stratification would create more homogeneous strata and secondly, result in improved precision of the estimates. The following person and household variables were used for implicit or secondary stratification. These variables are also provided in Appendices 4 and 5.

Person

Population

- Proportion 15 to 64

Gender

- Proportion Male
- Proportion Female

Population group

- Proportion Black
- Proportion Coloured
- Proportion Indian
- Proportion White

Employment status

- Proportion Employed
- Proportion Unemployed

Industry

- Proportion Agricultural; hunting; forestry and fishing
- Proportion Industry: Mining and quarrying
- Proportion Manufacturing
- Proportion Electricity, gas and water supply
- Proportion Construction
- Proportion Wholesale and retail trade
- Proportion Transport, storage and communication

Household

Toilet facility

- Proportion None
- Proportion Flush toilet (connected to sewerage system)
- Proportion Flush toilet (with septic tank)
- Proportion Chemical toilet
- Proportion Pit latrine with ventilation (VIP)
- Proportion Pit latrine without ventilation
- Proportion Bucket latrine

Piped water

- Proportion Piped (tap) water inside the dwelling
- Proportion Piped (tap) water inside the yard
- Proportion Piped (tap) water on community stand: distance less than 200m from dwelling
- Proportion Piped (tap) water to community stand: distance less than 200m and 500m from dwelling
- Proportion Piped (tap) water to community stand: distance less than 500m and 1000m from dwelling
- Proportion Piped (tap) water on community stand: distance greater than 1000m (1 km) from dwelling
- Proportion Piped Water: No access to piped (tap) water

The secondary stratification was carried out independently within each primary/explicit stratum. The explicit strata are formed by (3) area types, (9) provinces and (6) and (1) non-metro category. The resulting strata may still be distinguished by other socio-economic attribute. For example, not all SALs that are classified as urban have the same status. As a result, clustering approach discussed in Huang (1997) was considered in order to identify more homogeneous clusters to be used to form the final strata. Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters, according to some defined criteria. The attributes that are listed above per SAL were used for clustering. The SAS procedure **`Proc Fastclus`** was chosen to perform the clustering of SALs within the primary strata.

In applying **`Proc Fastclus`** it is crucial to decide whether variables should be standardised in some way, since variables with large variances tend to have more effect on the resulting clusters than those with small variances. If all variables are measured in the same units, standardisation might not be necessary. Otherwise, some form of standardisation is strongly recommended (SAS, 2014). Although the variables are already presented as proportions, the resulting proportions themselves may vary from variable to variable. For example, the “households services” variable may have different variation compared to person “industry” variable.

Before the SAS procedure for clustering, **`Proc Fastclus`** was executed, standardisation of the clustering variables was performed. All the variables were standardised within the primary stratum or explicit strata **`ex_strata`** such that the mean is 0 and standard deviation is one 1. The clustering variables were modified in order to fit the standard normal distribution where $\mu = 0$ and $\sigma = 1$. The SAS procedure **`Standard`** was executed for standardising the data.

The final design stratum (Stratum) will be a 5-digit ID with the following fields: province, metro code (71 through 76) or non-metro, geographical area type and secondary stratum or a cluster within the primary stratum.

The example in Table 3.2 shows nine clusters created from explicit stratum "111", which is an urban cluster in Western Cape metro-municipality. The aim was for a sample size to be in the range of 40 to 200 households for the secondary strata. The secondary strata with sample size less than 40 were collapsed with the "nearest neighbour" or the nearest small cluster, and those with more than 200 sample sizes were further stratified.

Table 3.2: Example of clusters formed from explicit stratum "111".

Explicit Strat	Cluster	Stratum	Tot. Households	Sample
111	.	1 287	1 075 632	836
111	1	1 287	83 378	65
111	2	1 287	148 698	116
111	3	1 287	168 120	131
111	4	1 287	53 071	41
111	5	1 287	168 682	131
111	6	1 287	17 202	13
111	7	1 287	108 837	85
111	8	1 287	172 096	134
111	9	1 287	155 548	121

In the example (see Table 3.2), primary or explicit stratum "111", which was a metro area of geography type urban-formal in province 1, that was, Western Cape province, had a sample of 836 households. The stratum had been divided into nine clusters. The cluster 6 had only 13 households and was collapsed with the nearest neighbour, that is, cluster 7. Thus, the primary stratum "111" was divided into 8 secondary strata.

After the completion of secondary stratification, the new variable "Stratum" was created and the variable contained final strata, where a two-digit code was added to a three-digit **ex_strata**, hence the final strata had five digit codes. The total number of 158 strata was formed and was distributed across all South African provinces (see Appendix 7).

Content of the sampling frame

To start this section, variables that formed part of the final sampling frame which was used for sampling are listed in Table 3.3.

Table 3.3: Variables in the sampling frame.

Number	Variable	Label
1	Pr_Code	Province Code
2	Stratum	Stratum
3	SAL_Code	Small Area Layer Code
4	SAL_Code_S	SAL Code After Splitting
5	SAL_Group	Small Area Layer Group
6	Prov_Sample	Provincial Sample
7	ISR	Inverse Sampling Rate
8	TOT_SAL	Total SAL per Stratum
9	MOS	Measure of Size
10	STR_MOS	Stratum Measure of Size

The variables in Table 3.3 were to be used further in the sampling design process. Sample allocation to provinces was represented in two ways, firstly as a provincial sample size, and secondly as provincial inverse sampling rate (ISR), which is the inverse of the selection probability. The two parameters were calculated assuming the national sample size of 15,000 households using Census 2011 household totals. Sample size determination was not particularly covered in this study. The assumption of sample size was based on other studies that were observed and employed CSSD. The possibility of incorporating population growth as well as the average response rates was explored; however, it was not implemented in order to manage the complexity of the study.

The measure of size (MOS) was derived from the household totals for each SAL. The stratum MOS was the total number of households calculated, first within each explicit stratum and later to the implicit strata. The explicit stratum was created by combining province code, metro/non-metro code, and geography type.

After allocating the sample to the provinces, the SALs were further stratified by geography (primary stratification), and by population attributes using the Census 2011 data (secondary stratification). This chapter further provides specifications to implement the stratification and the corresponding sample allocation.

In order to implement the sample design, the ISR, rather than sample sizes at different levels of stratification, were used. Finally, the inverse sampling rates (or ISRs) were translated into sample sizes. The ISRs used for the sample allocation were based on the Census 2011 counts and are calculated in Section 3.2.3. The consideration was made to re-compute the ISRs before sampling SALs to account for the growth between the years 2011 and 2017, the Census year and the survey implementation year, respectively. However, the procedure was not implemented in order to purge the complexity of the study. The effect of growth realised after simulation was taken into account through clustering and segmentation of large SALs and in the weighting system, the details of which are discussed in Chapter 4.

Table 3.4: Provincial sample allocation percentage.

Province	**Percent Sample**
Western Cape	12.2
Eastern Cape	11.3
Northern Cape	5.2
Free State	8
Kwazulu Natal	15.4
North West	9.2
Gauteng	18.1
Mpumalanga	9.8
Limpopo	10.9

The final provincial sample allocation based on percentages was presented in Tables 3.4. Gauteng province got the largest allocation of the sample (18.1 percent) followed by Kwazulu Natal with 15.4 percent and lastly Northern Cape province with 5.2 percent. It was not surprising that Northern Cape had the lowest population sample (as per allocation) due to its small total population size. Similarly Gauteng has the largest population followed by Kwazulu Natal. Even lower sample sizes could have been realised if a proportional allocation was followed. The subsequent sections will reveal how the final sample allocations were realised using square-root allocation method which takes sample from large provinces to smaller provinces (StatCan, 2003).

The square-root allocation is a compromise between the proportional and the equal allocation where the sample is allocated proportionally to the square-root of the stratum size. Square-root allocation has been used for the master samples in Vietnam and South Africa. Kish (1988) has proposed an alternative compromise based on an allocation proportional to $n\sqrt{W_h^2 + H^{-2}}$ where n is the overall sample size, W_h is the relative size of stratum h and H is the number of strata. For very small strata, the second term H^{-2} dominates the first W_h^2, thereby ensuring that allocations to the small strata are not too small (UNSD, 1998). Section 3.2 expand on the detailed implementation of the square-root allocation.

3.2 CSSD sample allocation and stratification

3.2.1 Background

The stratification (discussed in Section 3.1) in the sample design for this study was done in two stages. The first level (explicit stratification) of stratification was based on the geography, i.e., provinces, area type and metro and non-municipalities. The next level (implicit stratification) of stratification was based on the census characteristics of the individuals and the households. This section describes the steps followed to allocate the sample to the provinces as well as sample selection, within each strata. After the provincial sample allocation was implemented in `SAS Code 2`, the allocation to further stages was implemented in `SAS Code 3`.

3.2.2 Methodology overview

The sample design was based on the Census 2011 using original 2011 municipal boundaries. After the frame construction, the first step in the sample design was to allocate the total sample to the provinces. There was a large variation in the sizes of the provinces, ranging from 5.2 percent in the Northern Cape to 18.1 percent in Gauteng. Since survey estimates were to be produced both at the national-level and for each of the provinces, allocation of the sample to the provinces was done using a form of disproportional allocation called square-root allocation method. The square-root allocation shifts the sample from the larger provinces to the smaller ones, compared to what would have been under the proportional-allocation (StatCan, 2003).

Note that the proportional-allocation would be suitable for national estimates, but it could not have been possible to produce reliable estimates for the smaller provinces. Using proportional allocation, the provinces that are small will receive a smaller sample size which is proportional to their measure of size. If the objective is to produce estimates at provincial-level, those smaller provinces will not have sufficient sample sizes for estimation.

Tambay and Catlin (1995) also stated that proportional-allocation is not "optimal", either to minimise costs for a given reliability or to maximise reliability for a given cost (collection costs are greater in rural and remote strata), when a survey is focused on the measurement of a single set of related characteristics, such as income. Proportional allocation also does not consider sub-provincial estimation requirements: the sample size is often inadequate to produce reliable estimates in certain regions.

Allocation to strata using $\sqrt{N}$, proportional-allocation can improve the precision of the stratum estimates. In this case, the allocation parameter, is calculated as:

$$a_h = \frac{\sqrt{N_h}}{\sum_{h=1}^{L} \sqrt{N_h}}, \tag{3.1}$$

where $N_h =$ the measure of size in a stratum and,
$h = 1, 2, 3, ..., L$ are individual strata.

In other words, the allocation parameter a_h is equal to the ratio of the square-root of the population size in the stratum to the sum of the square-root of the population size of all strata (StatCan, 2003).

3.2.3 Sample allocation to provinces

The variable **`Pers_15_64`** (as is referred to in the SAS program) is the number of persons aged between 15 and 64, that is, the economically active population, and it was used to allocate total sample to provinces. In this instance the population of persons aged between 15 and 64 was extracted from the South African Census 2011. The reason for using 15 to 64 is that most surveys that seek to measure economic activity target mainly the economically active population. The economically active population was only used for sample allocation to provinces, while the households were used as the measure of size for sample allocation and selection during the subsequent sampling stages.

The calculation of the provincial sample sizes was done using the following formula:

$$a_p = n_p \times \frac{\sqrt{N_p}}{\sum_{p=1}^{K} \sqrt{N_p}} + 0.5, \tag{3.2}$$

where

$N_p =$ the measure of size in a province and,
$p = 1, 2, 3, ..., K$ are individual provinces.

The implementation of provincial allocation is as follows:

Step 1: Quality assurance

The frame file contains the variable **Pers_15_64** at the SAL-level. The SAS procedure **Proc Univariate**, was used to check for any outliers.

Step 2: Province-level totals

The SAS procedure **Proc Summary**, was used to compute the total population aged between 15 and 64 at the province-level. Note that the 2011 provincial boundaries were used for the sample design.

The following SAS Code was used to compute the total population at the province-level. The option **Missing** in the **Proc Summary** was used to check for any missing values for the province.

```
Proc Summary Data=msc.msc_frame009  Missing Nway;
Class pr_code_2011;
    Var AG02 Pers_15_64;
    Output Out=PR_Total Sum=Pr_Pers_15_64 N=;
    Format pr_code_2011;
Run;
```

The variable **Pr_Pers_15_64** in the temporary SAS output dataset **Pr_Total** provided the total population at the province-level.

Step 3: Calculating proportions and sample sizes

The SAS dataset **Pr_Total** created in Step 2 above gives the total population aged between 15 and 64 at the province-level (one record per province). The square-root of the variable **Pr_Pers_15_64** was compute as follows:

```
Rt_Pr_Pers= Sqrt (Pr_Pers_15_64);
```

The sum of the variable **Rt_Pr_Pers** at the national-level was computed.

The name given to the variable was **Sum_Rt_Pr_Pers**. The general formula for calculating provincial proportions is:

$$a_p = \frac{\sqrt{N_p}}{\sum_{p=1}^{K} \sqrt{N_p}}, \tag{3.3}$$

where

$N_p =$ the measure of size in a province and,
$p = 1, 2, 3....K$ are individual provinces.

Then, **Pr_Proportion=Rt_Pr_Pers/Sum_Rt_Pr_Pers;**

The total national sample size required was assumed to be approximately 15,000 households.

Set Total_Sample = 15000;

Computation of the provincial sample sizes was implemented as:

```
Pr_Samp=INT (Total_Samp * Pr_Proportion + 0.5);
```

Step 4: Creating the output file

A permanent SAS dataset at the province-level (one record per province) was created as follows using variables in Table 3.5:

`Pr_Code_2011`, **`Pr_Pers_15_64`**, **`Pr_Proportion`** and **`Pr_Samp`**.

The total provincial samples were allocated to the subsequent levels of geographical stratification using the output file from Step 4.

Table 3.5: Provincial sample allocation.

Province	Person(15 to 64)	Proportions	Sample Size
Western Cape	3 827 647	0.122	1 834
Eastern Cape	3 255 706	0.113	1 691
Northern Cape	686 155	0.052	776
Free State	1 642 582	0.08	1 201
Kwazulu Natal	6 045 443	0.154	2 305
North West	2 158 004	0.092	1 377
Gauteng	8 374 451	0.181	2 712
Mpumalanga	2 462 391	0.098	1 471
Limpopo	3 035 309	0.109	1 633

3.3 Sample allocation to strata and SALs

The distribution of SALs and households within each strata are reported in Appendix 7. The latter formed the basis of sample allocation to both strata and SALs.

The starting point is to calculate the strata proportions to be used as input to sample allocation.

$$a_h = \frac{\sqrt{N_h}}{\sum_{h=1}^{L}\sqrt{N_h}}, \tag{3.4}$$

where

$N_h =$ the measure of size in a strata and,

$h = 1, 2, 3, ..., L$ are individual strata.

The stratum household sample was calculated as follows:

$$n_h = N_h \times \frac{\sqrt{N_h}}{\sum_{h=1}^{L}\sqrt{N_h}} + 0.5 \quad or \quad n_h = N_h \times a_h + 0.5, \tag{3.5}$$

where

$N_h =$ the measure of size in a strata and,

$h = 1, 2, 3, ..., L$ are individual strata.

The stratum SAL sample was then calculated as follows:

$$n_{SAL} = n_p \times \frac{\sqrt{N_h}}{\sum_{h=1}^{L}\sqrt{N_h}} + 0.5 \quad or \quad n_{SAL} = n_p \times a_h + 0.5, \tag{3.6}$$

where

$N_h =$ the measure of size in a strata,

$n_p =$ provincial sample allocation and,

$h = 1, 2, 3, ..., L$ are individual strata.

The implementation of the above equations result in stratum proportions, using the proportions to derive households and SAL sample allocation.

3.4 Creation of sampling parameters

The ISR is an important sampling parameter which is used as the basis of sample selection in all stages. It is derived from the inclusion probability. The inclusion probability at stratum level is calculated as follows:

$$i_h = \frac{n_h}{N_h}, \tag{3.7}$$

where

$n_h =$ the sample of SAL in a stratum,
$N_h =$ the population of SAL in a stratum and,
$i_h =$ are inclusion probabilities per stratum.

The inclusion probability at SAL level is calculated as follows:

$$i_{SAL} = \frac{n_{SAL}}{N_{SAL}}, \tag{3.8}$$

where

$n_{SAL} =$ the sample of households in a SAL,
$N_{SAL} =$ the population of households in a SAL and,
$i_{SAL} =$ are inclusion probabilities of households in a per SAL.

Stratum ISR is calculated as:

$$Stratum_{ISR} = \frac{1}{N_{i_h}}, \tag{3.9}$$

where

$i_h =$ are inclusion probabilities of households in a stratum.

Table 3.6 shows the contents of sampling parameter file which is also used in sampling and weighting.

Table 3.6: Contents of the sampling parameter file.

No.	Variable	Description
1	pr_code	Province code
2	stratum	Stratum
3	Sal_Code_S	Small area layer
4	Sal_isr	SAL inverse sampling rate
5	Sal_start1	SAL random start
6	Sal_flag	SAL flag
7	sample_flag	Sample flag
8	n_starts_a	Number of starts available
9	n_starts_o	Number of starts over-used
10	n_starts_u	Number of starts used
11	Sal_adj	SAL Adjustment
12	VarUnit	VarUnit

Sampling process produces a variety of parameters. These parameters are stored in the dwelling/household sample file and the household sampling frame. SAS Code 6 provides detailed process for the creation of sampling parameters. The sampling parameters contain information which can be used to draw samples for more than one survey using one sampling frame. Information such as ISR is further used to calculate base weights.

3.5 CSSD sample selection

3.5.1 Selection of SALs sample

Computation of selection probabilities and selection of SALs

The selection probabilities of SALs with randomised PPS systematic sampling methods are discussed in this section. For the sake of simplicity, one stratum is used to describe the procedure. It should be understood that one sampling procedure was applied independently within each design stratum.

Let N be the total number of SALs in a stratum, and the number of SALs to be selected from the stratum is denoted by n. Also, let x_i denote the size measure of the SAL i within the stratum, where $i = 1, 2, ..., N$. Then, the selection of the sample of size n SALs with randomised PPS systematic sampling method can be described as follows:

Step 1: Randomising the SALs within the stratum

The list of SALs within the stratum is randomised by generating uniform random numbers between 0 and 1, and then sorting the SALs in ascending or descending order of these random numbers. Once the SALs have been randomised, permanent sequence numbers for the SALs are generated.

Step 2: Defining normalised measures of size for the SALs

The measure of size (MOS) of SAL i within the design stratum is denoted as x_i. Then, the MOS for the stratum is given by $X = \sum_{i=1}^{N} x_i$. The normalised size measure p_i of SAL i is $p_i = x_i/X$; $i = 1, 2, ..., N$, where N is the total number of SALs in the design stratum. Then, p_i is the relative size of the SAL i in the stratum, and $\sum_{i=1}^{N} p_i = 1$ for all strata. It should be noted that the value of

$n \times p_i$, which is the inclusion probability of SAL, must be less than 1.

Step 3: Obtaining ISRs

At this point the provincial sample sizes and their corresponding ISRs are already calculated using proportional allocation. The stratum ISR is the same as the corresponding provincial ISR because of the proportional allocation within the province. It should also be noted that the provinces are not explicit strata since the explicit strata further group SALs within each province by geography type and metro/non-metro. It should also be noted that the proportional allocation within the province also results in a self-weighting design. In situations where the elements are selected with equal probability within strata, this type of sample design results in equal probability of selection method "*epsem*" sampling, and therefore, "self-weighted" estimates of population parameters (Ross, 1978). According to the organisation of economic cooperation and development (OECD) definition (OECD, 2007), if the raising factors of all the sample units are equal, the common raising factor is called the raising factor of the sample, and the sample itself is called self-weighting. In this instance the strata inherit the provincial weight or a rasing factor.

Let R be the stratum ISR, then, the SAL-ISRs were obtained as follows. First, N real numbers $Z_i = n \times p_i \times R$; $i = 1, 2, ..., N$ were defined. It is easy to verify that $\sum_{i=1}^{N} Z_i = n \times R$. Next, round the N real numbers Z_i, $i = 1, 2, ..., N$ to integer values R_i, $i = 1, 2, ..., N$ such that each R_i is as close as possible to the corresponding Z_i value and the R_i values add up to $n \times R$ within the stratum. In other words, the sum of the absolute differences between the R_i and the corresponding Z_i values is minimised subject to the constraint that the R_i values add up to $n \times R$ within the stratum.

Scholars Drew et al. (1978) provided a simple algorithm to obtain the integer R_i values as follows: Let "d" be the difference between the value $n \times R$ and the sum $S = \sum_{i=1}^{N} [Z_i]$, where $[Z_i]$ is the integer function, then R_i values can be obtained by rounding up the "d", Z_i values with the largest fraction parts, and by rounding down the remaining $(N - d)$ of them. It should be noted that the integer sizes R_i; $i = 1, 2, ..., N$ are also the SAL-ISRs for systematic sampling of households.

To further illustrate the SAL-ISR consider a dataset containing 10 SALs from which a sample of two SALs will be drawn using PPS, that is, $N = 10$ and $n = 2$. The measure of size x_i in each SAL is households and they are contained in the dataset. Dividing each individual measure of size by the sum of all measure of sizes creates a normalised measure of size, that is, $P_i = x_i / \sum x_i$. The individual selection probabilities of SALs are calculated as $\pi_i = np_i$.

Assume that the sampling rate is 2 percent. The ISR $R = 50$. It follows that the individual inverse probability of selection is calculated as

$$\begin{aligned} R_i &= \pi_i R \\ \sum R_i &= \sum \pi_i R \\ &= nR \end{aligned}$$

where π_i is the selection probability.

The final steps of obtaining the ISR are implemented as follows:

Table 3.7: Rounding algorithm to obtain SAL-ISRs.

Steps.	**Description**
Step1	Sort the R_i by size in descending order
Step2	Split the values into integers and fractional components
Step3	Sort the fractions by size in descending order
Step4	Sum the integers to obtain total T
Step5	Subtract total T from N to obtain the difference "d"
Step6	Use D to increment the integers corresponding to the highest fractions for the first "d" integers
Result	This yields the new inverse sampling rate L

Step 4: Obtaining cumulative ISR values

The cumulative ISRs of the SALs are denoted by C_i, $i = 1, 2, ..., N$ within the stratum. It should be noted that the SALs within the stratum have been sorted according to the sequence numbers that were assigned after the randomisation. Then, the cumulative ISRs are defined as follows:

$$\begin{aligned} &C_1 = R_1; \\ &C_j = C_{(j-1)} + R_j; j = 1, 2, ..., N. \end{aligned}$$

It should also be noted that the value C_N will be equal to $n \times R$, which is also the total number of systematic samples of households that can be selected from the stratum.

*Step 5: Generate an integer random number **r** between* 1 *and* R*, interactively and compute* n *integers* $r_1, r_2, ..., r_n$ *as follows*:

$$\begin{aligned}
&r_1 = r;\\
&r_2 = r_1 + R;\\
&r_3 = r_2 + R;\\
&.\\
&.;\\
&r_i = r_{(i-1)} + R;\\
&.\\
&.;\\
&r_n = r_{(n-1)} + R.
\end{aligned}$$

Step 6: Selecting n *SALs out of the* N *SALs in the stratum with the label (sequence numbers) numbers* $i_1, i_2, ..., i_n$ *such that*:

$$\begin{aligned}
&C_{i_1-1} < r_1 \leq C_{i_1};\\
&C_{i_2-1} < r_2 \leq C_{i_2};\\
&.\\
&.;\\
&C_{i_n-1} < r_n \leq C_{i_n}.
\end{aligned}$$

Then, the n SALs with the labels $i_1, i_2, ..., i_n$ would get selected with PPS, and the selection probability of SAL i is given by R_i/R.

The sample selection of SALs was done using the randomised PPS systematic sampling method following the above process. After randomising the SALs within strata and computing the SAL-ISRs, the SALs were selected within strata using the stratum random start. There were 1,849 sampled SALs, that represented the entire country (South Africa), from 75,314 SALs. In the section that follows, the implementation of sampling procedure is illustrated as it was carried out in SAS.

Implementation of sample selection procedure for SALs

Sample selection of SALs was done using randomised PPS systematic sampling method. After randomising the SALs with strata and computing the SAL-ISRs, the SALs were selected within strata using the stratum random start **`Str_Rand_Start`**.

The following steps were followed to select the sample of SALs with randomised PPS systematic method.

Step 1

Input the SAS file **`msc_frame012`**
and use the SAS procedure **`Proc Sort`** to sort the file by

`Pr_Code Stratum Seq_Num`.

The sort procedure would randomise the list of SALs within each stratum.

Step 2

The stratum random start was generated for each "Stratum". The SAS variable **`Str_Rand_Start`** on the file **`msc_frame012`** gives the stratum random start.

Step 3

If it is the first SAL in the stratum, then

```
Set Sel_ISR = Str_Rand_Start
```

Find the first SAL in the stratum such that

```
Sel_ISR le Cum_ISR
```

Select this SAL by setting the flag

`Selection = "*"`, `Set Sel_ISR = Sel_ISR + Str_ISR;`.

Find the next SAL in the stratum such that

`Sel_ISR le Cum_ISR`

Select this SAL by setting the flag

`Selection = "*"`, `Set Sel_ISR = Sel_ISR + Str_ISR`.

Continue the above process until there are no more SALs in the stratum. Check that the above process selected the required number of SALs from the Stratum. The number selected should be exactly `N_Sample_SALS` (number of SALs allocated per stratum).

Step 4

Repeat the above procedure for all strata to select the SALs.

Check the SAL sample

Subset the SAS dataset after sampling using the subset condition

`If Selection = "*"`.

There were 1,849 sampled SALs. The detailed SAS Code given in SAS Code 3 was executed to obtain the final sample of SALs. The permanent SAS dataset was created and named: `MSC_Sal_Sample`.

3.5.2 Selection of households sample

After selecting the SALs, the list of the ultimate sampling unit (in this case households), can be obtained in various ways. Firstly, through listing structures on the ground by visiting each SAL and identifying households from structures, and secondly, from a georeferenced dwelling frame. However, for this study simulation was carried out.

The objective of the simulation exercise was to obtain a list of households from 1 to the total number of households recorded in Census 2011 data for each SAL. As a result of simulation, growth that was usually observed between sample selection and data collection did not exist, however, this was common when actual listing and data collection were conducted.

The selection of the sample of households was done with systematic sampling procedure. This section provides specifications for selecting the sample of dwellings using the SAL-ISR and the initial random start for the SAL.

The following steps provide specifications for sampling households from the sampled SALs.

Step 1

The SAS input file **`SAL_delling_frame`** was created from the household database generated through simulation of a household list obtained after the field listing operation and data capture. This file was at the household level (one record per household). All the households within the SAL were numbered sequentially. The variables in Table 3.8 (next) were on the input SAS file.

Step 2

SAS procedure **`Proc Summary`** was used to obtain the dwelling count **`Sum_Count`** within each SAL by counting the number of household records for the SAL. Check against the number reported, that is, households (**`HH`**). If the numbers are different, then output the SAL record on an **`ERROR`** file. Print the **`ERROR`** file, and check the discrepancy.

Table 3.8: Contents of the SAL household/dwelling frame file.

No.	Variable	Description
1	**PR_CODE**	Province Code
2	**STRATUM**	Sampling Stratum
3	**SAL_Code**	Small Area Layer Code
4	**SAL_Code_S**	Small Area Layer Code after splitting
5	**SAL_ISR**	SAL Inverse Sampling Rate
6	**SAL_r_Start**	SAL Random Start
7	**N_DUS**	Number of Private Dwellings/households in the SAL
8	**DU_Num**	DU Number within the SAL (for PDs or Households)
9	**SAL_Flag**	SAL Flag for Conceptual split.

Step 3

Modify the **SAL_ISR** and **SAL_r_Start** for the SALs with conceptual split. The variable **SAL_Flag** identifies the conceptual split and the number of SALs created from the split.

```
If SAL_Flag ge 2 then Do; /*it is a conceptual split*/
    SAL_R_Start = (SAL_Flag-1)*SAL_ISR + SAL_R_Start1;
    SAL_ISR=SAL_Flag*SAL_ISR;
End;
```

Step 4

Assign the sample selection flag **Sample_Flag** as follows. Use the SAS function **MOD** to define the random start.

```
R_Start = MOD(HH_Num, SAL_ISR);
If R_Start = 0 then R_Start=SAL_ISR;
```

Note that the variable **HH_Num** contains the sequential household numbers within the SAL, and **N_HH** is number of households in the SAL. The household was sampled if the SAL sampled random start **SAL_r_Start** was the same as the **R_Start**. Define the **Sample_Flag** as:

```
    If R_Start = SAL_r_Start then Sample_Flag = 1;
    Else Sample_Flag = 0;
```

Thus, all sample households had the **`Sample_Flag`** equal to 1, and the non-sampled households had the **`Sample_Flag`** value equal to 0. The SAS Code given in (SAS Code 4) was executed to create the permanent SAS dataset of the households sample, that is, **`MSC_dwelling_Sample`** of 17,076 households. The final sample of households was larger than the initially proposed sample size. The reason for differences is due to varying measures of size within strata and the effect of the unequal probability of selection of SALs. Another practical reason is that the sample allocation and initial ISRs were calculated using previous census data. In practice, due to growth that may occur in areas between census and sample implementation, the initial ISRs would yield much higher sample sizes despite the adjustments.

Chapter 4

Weighting methodology

4.1 Background of sample weighting

Weighting samples is important to reflect not only sample design decisions made at the planning stage and use of auxiliary data to improve the efficiency of estimators, but also practical issues that arise during data collection and cleaning that necessitate weighting adjustments. Planned and unplanned adjustments to survey weights are used to account for these practical issues (Henry and Valliant, 2012).

The sample for this study was based on a stratified two-stage design with PPS sampling of SALs at the first stage, and sampling of households with systematic sampling at the second stage. The sampling methods were discussed in the previous chapters. The current chapter provides detailed description of sample weighting. At this point it is important to note there was no data collection conducted from sampled SALs. Data used as a proxy to survey data were simulated from Census 2011 responses for each SAL. Detailed processes of data simulation are in SAS Code 5, while household/dwelling frame was first used in SAS Code 4 and obtained from SuperCROSS, that is, a cross tabulation tool used by StatsSA for data dissemination.

After the data were simulated, the sampling weights were constructed. The starting point for weighting used ISRs at provincial-level (also used in provincial sample allocation). The provincial ISRs were calculated as follows:

Starting with the inclusion probabilities for provinces:

$$i_p = \frac{n_p}{N_p}, \tag{4.1}$$

where

n_p = the sample of person aged (15 to 64) in a province,
N_p = the measure of size (person aged (15 to 64)) in a province and,
i_p = are inclusion probabilities per province.

The provincial ISRs are then calculated as an inverse of the inclusion probabilities for each province.

$$ISR_p = \frac{1}{i_p} \tag{4.2}$$

The ISRs for each province were calculated and presented in Table 4.1, and were thereafter used to obtain ISRs for stratum as well as PSU/SAL level. The resulting provincial ISRs and the original base weights were adjusted in order to obtain the final weights for estimation. The calculation was the result of factors including original selection probabilities, adjustment for non-response and benchmarking to known population counts from the Community Survey 2016 (a large sample household-based survey conducted by StatsSA).

Table 4.1: ISRs for the provinces.

Province	**ISR**
Western Cape	895
Eastern Cape	860
Northern Cape	376
Free State	651
Kwazulu Natal	1 072
North West	764
Gauteng	1 465
Mpumalanga	716
Limpopo	838

4.2 Base weights and adjustments

4.2.1 Base weights

The base weight for each sampled household is equal to the reciprocal of the probability of selection, which is simply the inverse of the sampling rate. The sampling rate has been assigned at the province-level, that is, all design strata within a province have been sampled at the same rate. Thus, the initial base weight (or design weight) assigned to each household in a province was simply the ISR for the province. Table 4.1 presents the ISRs for the South African provinces.

The sampling rates would have to be modified in certain situations because of reasons related to operational feasibility and/or cost implications. There are two types of adjustments (or modifications) that were considered and discussed in Section 4.2.2. They are adjustment for moderate growth and extreme growth. This approach was used by StatsSA in the sample design for quarterly labour force survey from year 2008.

4.2.2 Adjustment for growth SALs

The sample for this study is based on the Census 2011 data. It is possible that some areas would have grown in population at a much faster rate than the rest of the country. Therefore, the sample sizes in terms of numbers of households from the growth SALs would proportionately be larger as well. In terms of growth there were two situations discussed with respect to the magnitude of growth, Situation A: moderate growth, and Situation B: extreme growth. The effect of growth where the old sample frame is used is inevitable, therefore, it is important to discuss in detail because all frames based on previous censuses are subjected to growth of sampling units in population size.

Moderate growth

If the household count in the SAL had grown by a factor less than 3 from the design count, it is categorised as moderate growth. In the case of moderate growth, it could be possible to list the SAL. The sample yield from the SAL would be large, such that it would not be operationally feasible to interview large number of households within the time constraint of the survey period. In order to cope with this situation, the SAL-ISR for the growth SAL was modified (increased) so that the actual sample of households was close to 10 households. This method is also known as list sub-sampling.

Let R_{hi} be the original ISR of the SAL based on the sample design, and R^*_{hi} be the modified ISR. Then, the adjustment factor for the growth SAL was the ratio of the two ISRs, that is,

$$(Adj\_Growth)_{hi} = \frac{(R^*_{hi})}{R_{hi}} \tag{4.3}$$

It should be noted that the adjustment factor from Equation 4.3 for the growth SAL was greater than 1.0 because R^*_{hi} is greater than R_{hi}. The adjustment was applied to the growth SAL in the design stratum h. The adjustment factor $(Adj\_Growth)_{hi}$ was set to 1 for the non-growth SAL. In practice there is the possibility that some EAs, PSUs or SALs that were part of the frame are found to have less households during listing or sometimes there are no households found during listing of data collection.

The sampling specialists do not have control over certain occurrences. As an example, the city council may decide to relocate the households in the informal settlements in order to locate the formal non-residential structures. In such cases the areas become out of scope and they contribute to the reduction in the sample, and subsequently the undercoverage. Such undercoverage is indirectly compensated through benchmarking. The most proactive way of avoiding undercoverage of such kind is to obtain the up-to-date sampling frame.

Extreme growth

If the SAL household count had grown by a factor more than 3 times the design count, then it was categorised as extreme growth. For the extreme growth situation it might not even be operationally feasible to complete the listing task within the given listing period. Moreover, the task of listing the extreme growth SAL would be too costly. Therefore, the clustering (or segmentation) approach would be used for the extreme growth SALs.

Clustering approach was used for the growth SALs because of the magnitude of listing task in situations where listing is required. Therefore, the rationale for the clustering in this situation was to reduce the workload. The clustering sampling method was used for dividing the extreme growth SALs. The extreme growth SALs were divided into a number of clusters (or segments) with identifiable boundaries.

After clustering, the selection of one cluster from the extreme growth SAL was carried out with PPS. The above method of reducing the sample yield was referred to as cluster sub-sampling. If the sample is drawn from the SAL with large growth without modifying the ISR to cater for growth, it would be difficult to cover all the sampled households during the given time for enumeration due to the large sample size.

The following procedure is followed in order to manage the extreme growth. Let D_{hij} be the household count for the segment/cluster j in the growth SAL i in the design stratum h. Then, the household count for the SAL is given by $D_{hi} = \sum_j D_{hi}$.

Let R_{hi} denote the SAL-ISR. Then, the expected number of households that would have been sampled from the SAL, if the entire SAL was listed and sampled with the ISR R_{hi}, would be given by D_{hi}/R_{hi}.

Computation of the segment ISRs R_{hij} was done, which were integers such that R_{hij} were proportional to D_{hij}, and $\sum_j R_{hij} = \lambda \times R_{hi}$, where λ is a factor by which the sample yield from the growth SAL would be dampened. It should be noted that $\lambda \times R_{hi}$ must be an integer even though λ can be a non-integer.

One segment was selected with PPS sampling method using R_{hij} as the size measure. This was done by following the two steps given below.

Step 1: Obtain cumulative ISR values

The cluster cumulative ISRs are defined as:

$$C_{hi1} = R_{hi1,}; C_{hij} = C_{hi(j-1)} + R_{hij;j=2,3,----,J,}, \tag{4.4}$$

where J is the total number of clusters created in the growth SALs. It should be noted that C_{hiJ} would be equal to $\lambda \times R_{hi.}$

Step 2: Select one segment with PPS

Generate a random integer r_{hij} between 1 and $\lambda \times R_{hi.}$ Find the cluster j^* in the growth SAL such that $C_{hi(j^*-1)} < r_hij \leq C_{hij^*}$. Then, the cluster with the label hij^* was the sampled cluster. Note that the sampled cluster is listed and the sample of households are selected with systematic sampling procedure using the cluster ISR. The adjustment for the growth SAL was equal to the dampening factor λ, that is, $(Adj\_Growth_{hi} = \lambda)$. The above adjustment factor was applied to the sampled growth SAL i in the design stratum h. It should be noted that one cluster was selected from the growth SAL with PPS sampling.

An Example

Consider a sampled SAL with a design count of 188 dwellings and an ISR equal to 20. It was found at the time of listing that the SAL had grown almost 4 times. There was lot of new development in the area with new street patterns. It would have taken too much time to list the entire SAL, and the sample yield from the SAL would have been too high to complete enumeration within the survey period. Therefore, it was decided to implement cluster sub-sampling. The SAL was divided into 8 clusters and household counts were obtained for the clusters. The cluster numbers and the corresponding household counts are given in the first two columns of Table 4.2. The SAL household count was 742. Therefore, the expected sample yield would have been 37.1 without sub-sampling.

For instance, the decision was to dampen the sample yield by a factor of 3.5.

It should be noted that the dampening factor used was less than the growth factor so that there would not be too much adverse impact on the variance due to differential weighting. Then, the cluster ISRs are computed such that they add up to 70 (that is, the original SAL-ISR of 20 multiplied by 3.5). The individual clusters' ISRs and the cumulative ISRs are given in columns 3 and 4 of Table 4.2.

Table 4.2: Cluster household counts, ISRs and cumulative ISRs.

Cluster-ID	**Household Count**	**Cluster ISR**	**Cumulative ISR**
1	136	13	13
2	103	10	23
3	111	10	33
4	68	6	39
5	61	6	45
6	44	4	49
7	107	10	59
8	112	11	70

In order to select one cluster, a random integer between 1 and 70 is generated. Suppose that the random number (integer) was 35. Therefore, the sampled cluster would be cluster number 4, and 5 random starts were used from the cluster. The cluster ISR of cluster 4 is equal to 6, therefore, random starts 2, 3, 4, 5 and 6 were used before the cluster was exhausted within the SAL. It should be noted that as many random samples (starts) from the clusters in the growth SAL were used as would have been used from the SAL without cluster subsampling. The weight adjustment factor $(Adj_{Growth})_{hi}$ was equal to 3.5, which was applied to the above sampled growth SAL.

4.2.3 Adjustment for sample stabilisation

Most government statistical agencies create a sampling frame to be used for more than one study at a different point in time. The sample sizes for the future studies will grow because of the natural growth in the population. The sample stabilisation (or random drop) would maintain the sample within sample stabilisation areas at the sizes that were established at the time of the sample design. The pre-assigned sample sizes at the time of design are also known as base sample sizes. The sample stabilisation areas are defined to be the same as those for balancing the expected sample sizes. These are defined as cross-classification of province by area-type by individual metro and non-metro SAL. The categories of area-type are the same categories used for the Census 2011, which are as follows: urban, tribal area, and farms.

The categories of the variable metro are the individual metros and the non-metro areas within each of the provinces. There were eight metros, and nine provinces, which resulted in 17 categories of metro/non-metro. The combination of province by geography type by metro or non-metro generates 49 balancing areas (with Western Cape province not having traditional areas).

After the sample of households has been selected for a given study, the aggregate sample sizes are computed for the balancing areas. The sample sizes at the balancing area levels for the quarter are compared with the corresponding base sample sizes. For those balancing areas where the sample is found to be too large compared with the base sample, the excess samples would be dropped through the following steps: Step 1, determine the access sample size; Step 2, sort the sample within the balancing area by stratum and SAL number; Step 3, take a systematic sample of size equal to the access sample from the sampled households, and Step 4, drop the cases sampled in Step 3 from the sample.

The sample stabilisation weight adjustment is computed as follows. Let A_b be the actual sample size in terms of number of dwellings for the balancing area b, and B_b be the corresponding base sample size where $A_b > B_b$. The difference $D_b = A_b - B_b$ is the access sample for the balancing area b that is dropped (deselected) at random. After dropping the excess sample of D_b households from the balancing area b, a sample of B_b households is left. The corresponding weight adjustment for the balancing area b is given by $(Adj_{Stabilisation})_b = A_b/B_b$.

The above stabilisation adjustment factor is applied to the sampled households from the stabilisation area. It is not necessary to consider sample stabilisation for about two years because the sample would not have grown much in the first two years after implementation. Regardless of the fact, the weighting system should have the flexibility to implement the adjustment for the sample stabilisation in the future. The regular household survey and sampling frame maintenance will be the basis for assessing the potential growth and suggest the need for stabilisation. Initially, the adjustment for sample size stabilisation is set to 1.

Base weight/final adjusted base weight

The base weight is defined as the product of the provincial ISR (Table 4.1) and the three adjustment factors discussed above, those are, (1) adjustment factor for sub-sampling of growth SALs, (2) adjustment for clustering, and (3) adjustment factor for sample stabilisation.

4.3 Non-response adjustments

Two types of non-response, i.e., item and unit non-response, are already discussed in the previous sections. For item non-response, imputation is assumed to be done on survey data and weight adjustment to account for the unit non-respondent, for example, refusal of households to be interviewed and no contact were done. It should be emphasised that the household is both sampling unit and the unit of observation. The sampled households not eligible for enumeration, for example, foreigners only, or no households (those are, vacant dwellings), did not contribute to the survey.

Respondents

This category consists of eligible households that completed the survey questionnaire and provided usable survey responses.

Non-respondents

These were the eligible households that did not complete the survey questionnaire, for example, refusal, no contact and temporarily absent.

In general, the non-response adjustment was applied at the SAL level. The non-response adjustment was applied at the stratum-level for only those cases where the non-response at the SAL level was too large. Let n_{hi} be the number of households sampled from SAL in the design stratum h. Also, let $(n_{hi}^{(resp)})$ be the number of respondent households out of the n_{hi} eligible households. The remaining $(n_{hi} - n_{hi}^{(resp)})$ are the non-respondent households. The non-response adjustment factor at the SAL level is defined as:

$$(Adj_{non-response})_{hi} = n_{hi}/(n_{hi}^{(resp)}) \tag{4.5}$$

The non-response adjusted weight was computed by multiplying the base weight with the non-response adjustment factor given by Equation 4.5. If the SAL level non-response rate was too high, then the non-response adjustment was applied at the stratum level. The following rule was used: SAL level non-response adjustment was applied only if the corresponding adjustment factor was less than 1.5 or some other lower threshold value for the adjustment factor.

The choice of threshold for non-response adjustment factor was a trade-off between the potential non-response bias and increase in the variance due to increase in the variability of the resulting weights. If the non-response adjustment was applied at a higher aggregate level, there was a risk of large potential non-response bias. In addition, some of the non-response adjustment factors might be too large when the non-response adjustment was applied at a very low level, which would have an adverse effect on the variance. The reason that both of those issues were important was that the total mean square error (MSE) of a survey estimate was equal to the sum of its variance and the square of the bias, that is, $MSE = Variance + (Bias)^2$. In surveys with large sample sizes, the MSE might be dominated by the bias term. When sample sizes are small, the variance might be a greater cause for concern.

4.4 Full weight calibration to produce final survey weights

4.4.1 Weight optimisation

In CSSDs and also other probability sample methodologies, the common challenge is making inference about the population. Researchers seek to accurately expand the survey households to present the population. There are several methods in literature to achieve that goal, calibration being the main focus of this study. Deville and Särndal (1992) proposed a model-assisted calibration approach that involves minimising a distance function between the base weights and final weights to obtain an optimal set of survey weights. Here "optimal" means that the final weights produce totals that match external population totals for the auxiliary variables **X** within a margin of error. Specifying alternative distance functions produces alternative estimators.

Bar-Gera et al. (2009) introduced the entropy maximisation method to estimate household survey weights to match the exogenously given distributions of the population, including both households and persons. Bar-Gera et al. (2009) also presents a *Relaxed Formulation* to deal with cases when constraints are not feasible and convergence is not achieved. The goal through this optimisation procedure is to find a weight for each household so that the distributions of characteristics in the weighted sample match the exogenously given distribution in the population, for both household characteristics as well as person characteristics.

The process to accurately handle the procedure is iterative to find weight which is as close as possible to the target distribution. Take for instance matrix A. Each column in A corresponds to a sample household. Each row within a column gives the contribution from a sample household to a certain population characteristic. Specifically, hhc_{ij} represent household characteristic values and cpt_{ij}'s represent person characteristic values. They are normally referred to as control variables. Each of the hhc_{ij} is either 1 or 0 depending on whether or not a particular household i has a certain characteristic value j. The value cpt_{ij} represents the number of persons with the person characteristic j belonging to household i.

$$A = \begin{bmatrix} hhc_{11} & . \; . \; . & hhc_{m1} \\ . & . \; . \; . & . \\ . & . \; . \; . & . \\ hhc_{1p} & . \; . \; . & hhc_{mp} \\ cpt_{11} & . \; . \; . & cpt_{m1} \\ . & . \; . \; . & . \\ . & . \; . \; . & . \\ cpt_{1q} & . \; . \; . & cpt_{mq} \end{bmatrix}$$

The hypothetical survey data are used to illustrate the construction of matrix A. If only marginal totals by each population (household and person) characteristics are separately available, then matrix A_1 is constructed.

$$A_1 = \left[\begin{array}{l} \left.\begin{array}{cccccc} 1 & 1 & 0 & 0 & 0 & Rented \\ 0 & 0 & 1 & 1 & 1 & Owned \\ 1 & 0 & 1 & 0 & 0 & Urban \\ 0 & 1 & 0 & 1 & 1 & Suburban \end{array}\right\} \text{Household} \\ \left.\begin{array}{cccccc} 0 & 1 & 2 & 1 & 1 & Male \\ 2 & 1 & 0 & 1 & 2 & Female \\ 2 & 1 & 1 & 0 & 0 & Caucasian \\ 0 & 0 & 1 & 2 & 2 & Hispanic \\ 0 & 1 & 0 & 0 & 1 & Asian \end{array}\right\} \text{Person} \end{array}\right.$$

On the other hand, matrix A_2 is constructed when frequency distribution for the composite household and person types obtained by combining more than one population characteristic, if such characteristics are available. It must be noted that the household and person composite types are created by combining household characteristics and person characteristics, respectively.

$$A_2 = \left[\begin{array}{l} \left.\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & RU \\ 0 & 1 & 0 & 0 & 0 & RS \\ 0 & 0 & 1 & 0 & 0 & OU \\ 0 & 0 & 0 & 1 & 1 & OS \end{array}\right\} \text{Household} \\ \left.\begin{array}{cccccc} 2 & 0 & 0 & 0 & 0 & FC \\ 0 & 1 & 1 & 0 & 0 & MC \\ 0 & 0 & 0 & 1 & 1 & FH \\ 0 & 0 & 1 & 1 & 1 & MH \\ 0 & 1 & 0 & 0 & 1 & FA \\ 0 & 0 & 0 & 0 & 0 & MA \end{array}\right\} \text{Person} \end{array}\right.$$

The goal is therefore to find a weight for each household so that the distributions of characteristics in the weighted sample match the exogenously given distributions in the population, for both household characteristics as well as person weight (Chen et al., 2017).

In this study only the person characteristics were used to construct the composite estimates. Fuller (2002) provides an excellent review of the regression related methods for survey estimation. Regression estimation was first used at Statistics Canada in 1988 for the Canadian Labour Force Survey. Regression estimation is now used to construct composite estimators for the Canadian Labour Force Survey (Gambino et al., 2001). The regression estimation is also currently being used for the StatsSA's Quarterly Labour Force Survey with population controls at the provincial-level in addition to the national-level population controls by age, gender and race (StatsSA, 2008). Using these methods, the weight achieved by taking into account person characteristics in calculating household weight is called integrated weighting, also discussed in Section 2.3.4.

4.4.2 Implementation of adjustment of survey weights to control totals

The final survey weights were constructed using regression estimation to calibrate survey estimates to the known population counts at the national-level by cross-classification of age, gender and race, and the population counts at the individual metros and non-metros within the provinces by two age groups (0-14, and 15 years and over). The computer program StatMx developed at Statistics Canada was used for constructing the final weights. Population control totals used in calibration were obtained from the South African Community Survey 2016 (see Table 4.3).

Table 4.3: Population control totals for calibration.

Age Group	Male	Female	Total
0 - 9	5 845 061	5 751 251	11 596 312
10 - 19	5 160 083	5 134 204	10 294 287
20 - 29	5 325 132	3 993 402	10 582 840
30 - 39	4 091 563	4 211 088	8 302 651
40 - 49	2 981 228	3 062 283	6 043 511
50 +	3 844 159	4 989 891	8 834 050
All	27 247 226	28 406 425	55 653 651

Population survey estimates obtained from simulated survey data before they were benchmarked to control totals (see Table 4.4).

Table 4.4: Population survey totals before calibration.

Age Group	Male	Female	Total
0 - 9	4 295 438	4 779 742	9 075 180
10 - 19	4 340 959	4 079 218	8 420 177
20 - 29	4 495 869	3 993 402	8 489 272
30 - 39	4 226 903	2 773 091	6 999 993
40 - 49	3 287 377	2 143 358	5 430 735
50 +	3 912 680	2 904 566	6 817 246
All	24 559 226	20 673 377	45 232 603

The population estimates used for benchmarking were from the Community Survey 2016. They were obtained by cross-classification of age by gender by race for broader 10-year age groups. The 10-year age groups were: 0-9, 10-19, 20-29, 30-39, 40-49 and lastly much broader 50+ age group (see Table 4.3 and 4.4). The groupings were derived in such a way that they will be usable and their corresponding survey totals will have sufficient sample sizes. In particular, the number of persons in the sample from the older age groups were very small. Therefore, the older age groups were collapsed such that the expected sample size for each cell would be at least 20 persons.

In practice the benchmarking groups correspond with common domains of estimation such as economically active population, school going and child bearing age. When specific broader age-based domains are known during the design stage they can be incorporated in the design.

The problem of estimating survey weights can indeed be formulated as a constrained optimisation problem, when one is attempting to minimise the difference between the weighted sample distributions of known population distributions across a set of control variables at both the household and person-levels. Such constraints can be embedded into a linear programming problem, with an artificially chosen linear objective function, and solved (at least in principle) by general linear programming method, such as the Simplex Algorithm. Within STATMX linear programming is used to achieve the above objective.

Chapter 5

Estimation methodology

5.1 Estimation of domain totals, means and proportions

Timely and complete social and economic data can be obtained from national sample surveys, however, it is usually only for major domains of a study. Domains can be local areas, often administrative units, such as geographical areas, for which separate estimates are planned, and which also tend to be partitions in the collection process, whether of censuses or of sample surveys. Or, on the contrary, they can be "cross-classes" of the population and of the sample, which cut across the partitions of the collection and the sample design; for example, age and sex categories. The dataset for this study contains variables, language, and education that can be viewed with other demographics.

Less commonly used than either of the above, are domains that have not been distinguished in the sample selection, but tend to concentrate unevenly in the primary sample units. Estimates for small domains (any "small" sub-class or sub-division of the original domains of a study) are generally unavailable from typical samples, and are obtained from population censuses or from administrative registers, and sometimes from special-purpose surveys. However, effective planning of social services and other activities cannot depend on these traditional data sources; the data must be more current, more complete, and more relevant than these sources provide. Estimates for local areas such as administrative units, appear as the most salient and common concern for detailed data, but cross-classes and other small domains can also be important. Several of these methods are equally pertinent to both (Purcell and Kish, 1980).

Since the sizes of the small domains influence the choice and applicability of methods, a classification of domains, based on their sizes, is presented here (Purcell and Kish, 1979). It can remind us of the practical differences between the types of domains and help us avoid the common mistake of considering "statistics for domains" as one homogeneous problem. The boundaries of this classification are stated very roughly to orders of magnitude and should not be taken too seriously; they depend on the variables and the statistics estimated, the sizes of samples and populations, the precisions and decisions involved, and so forth.

Definition of major, minor to rare domains defined by Purcell and Kish (1979).

1. Major domains, composing perhaps 1/10 of the population or more. Examples: major regions, 10-year age groups, or major categorical classes, like occupations.
2. Minor domains, comprising between 1/10 and 1/100 of the population. Examples: state populations, single years of age, two-fold classifications like occupation by education, or a single small classification like the unem-

ployed or the disabled.

3. Mini domains, comprising from 1/100 to 1/10,000 of the population. Examples: populations of counties (more than 3,000 of them in the USA), a two-fold classification like state by work force status, or a three-fold classification like age by occupation by education.
4. Rare domains, comprising less than 1/10,000 of the population. Examples: populations of small local areas, perhaps all classified by various ethnic groups.

For major domains, standard estimates, basically without bias, are generally available from probability methods of survey sampling. However, frequently for minor domains and usually for mini domains, the standard methods of survey estimation break-down, because the sample bases are ordinarily too small for any useable reliability, and new methods are needed. For rare domains, sample surveys are usually not of great use; separate and distinct methods are required (Rao, 2003).

Sample surveys are widely used to provide estimates of population quantities such as totals and means. Many countries have set-up centralised statistical agencies that are responsible for the collection of statistical information about the state of the nation. This important statistical information includes national characteristics such as the demography, agriculture, labour force, health, living conditions, and trade. Government agencies increasingly use these results to formulate policies and allocate government funds. Särndal et al. (2003) attribute much of the developments in sample survey methodology to government statistical offices.

The discussion and analyses by Reiter et al. (2005) suggest that survey design must be incorporated into multivariate analyses if conclusions from models are to be reliable. The approach taken to do so can affect conclusions substantially, as they illustrated in their comparisons of design-based and model-based analyses of the National Organizations Survey (NOS) and the National Survey of Establishments (NSE).

Chromy and Abeyasekera (2005) have discussed the role of survey weights and recognition of the sample structure in developing both descriptive and analytic statistics from survey data. Survey data analysis software that use survey weights and take account of the sample structure may be used to estimate the parameters of both linear and logistic regression models based on survey data. The estimates based on the sample are estimates of what would be obtained from fitting the models to the entire finite population. Furthermore, standard errors of the estimates can also be obtained (Chromy and Abeyasekera, 2005). Several replication methods have been discussed already and they are going to be used to correctly model and estimate variances of the CSSD.

5.2 Application of replication methods

5.2.1 Creation of replicate weights

Replicate weights are set of weights used to compute the survey estimates for each portion of the sample (replicate) when resampling methods are used for variance estimation. For example, in the Jackknife replication method, some sampled units have their weights doubled to account for the sampled units that were dropped. In Bootstrap replication methods it is an assumption that the sample of the same size is drawn for certain number of times. Replication is part of resampling methods of estimating variances. The first step for creating replicate weights is identifying and equating the highest non-certainty level to the PSU/SAL. In the first stage, after removing non-responding PSUs/SALs, not all strata had two or three PSUs/SALs and to correct that, “VarUnits” were modified in such a way that there were at least two PSUs/SALs in a stratum. The process is implemented in WesVar software (Westat, 2000).

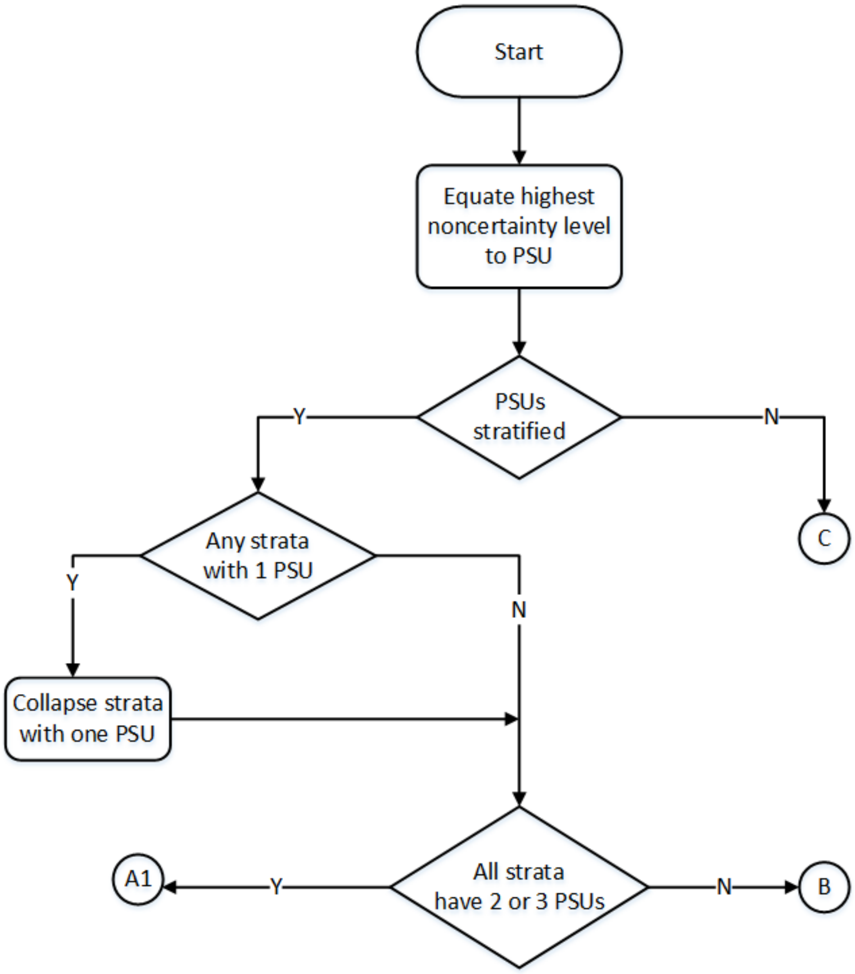
Start
Equate highest noncertainty level to PSU
PSUs stratified
Y
N
C
Any strata with 1 PSU
Y
N
Collapse strata with one PSU
All strata have 2 or 3 PSUs
A1
Y
N
B

Figure 5.1: Creation of replicates (Step one).

Figure 5.1 shows the process flow which illustrate steps that are taken to create replicate weights. The replicates are calculated for each PSU/SAL. The conditions that are required to be met are firstly, PSUs/SALs should be stratified; secondly, all strata must have 2 or 3 strata and those strata with one PSUs are collapsed. The unqualified strata with large numbers of PSUs are taken through the Step two in Figure 5.1 while those that qualified are taken through Step three for creating replicates in Figure 5.3.

VarUnits were recreated to ensure that there were at least two PSUs/SALs in a stratum. After the updating of VarStrat we move to the next process (A1) to create replicate weights.

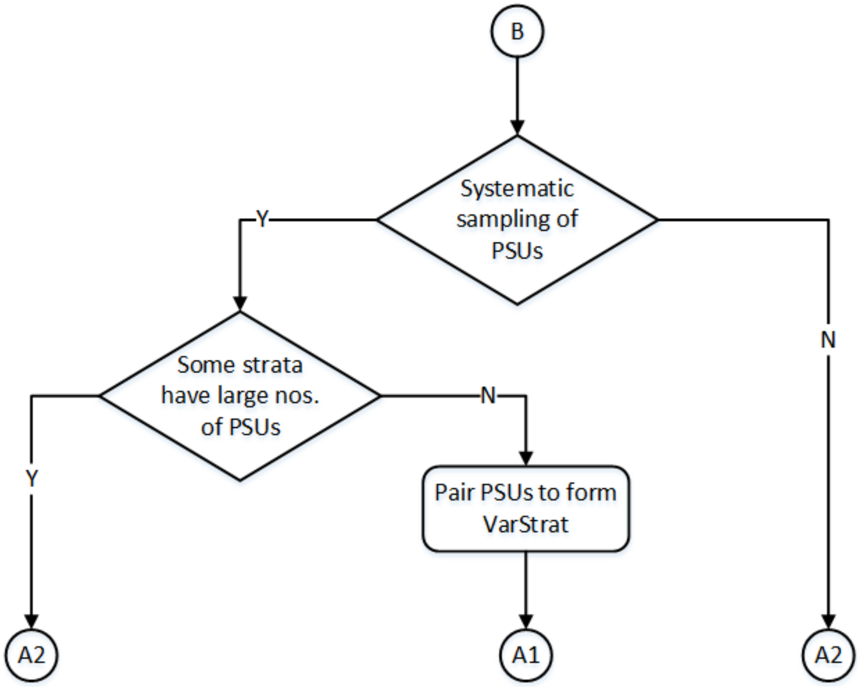
B
Systematic sampling of PSUs
Y
N
Some strata have large nos. of PSUs
Y
N
Pair PSUs to form VarStrat
A2
A1
A2

Figure 5.2: Creation of replicates (Step two).

This process was followed in the South African QLFS where many strata had more than three PSUs. In that case PSUs in each strata were combined to create two pseudo PSUs per stratum and the new sets of PSUs were called VarUnits (StatsSA, 2008). The South African example also accommodates all PSUs including those that ended up in node A2 of Figure 5.2.

When all the conditions are met, the target is to get to A1 where the replicates are created.

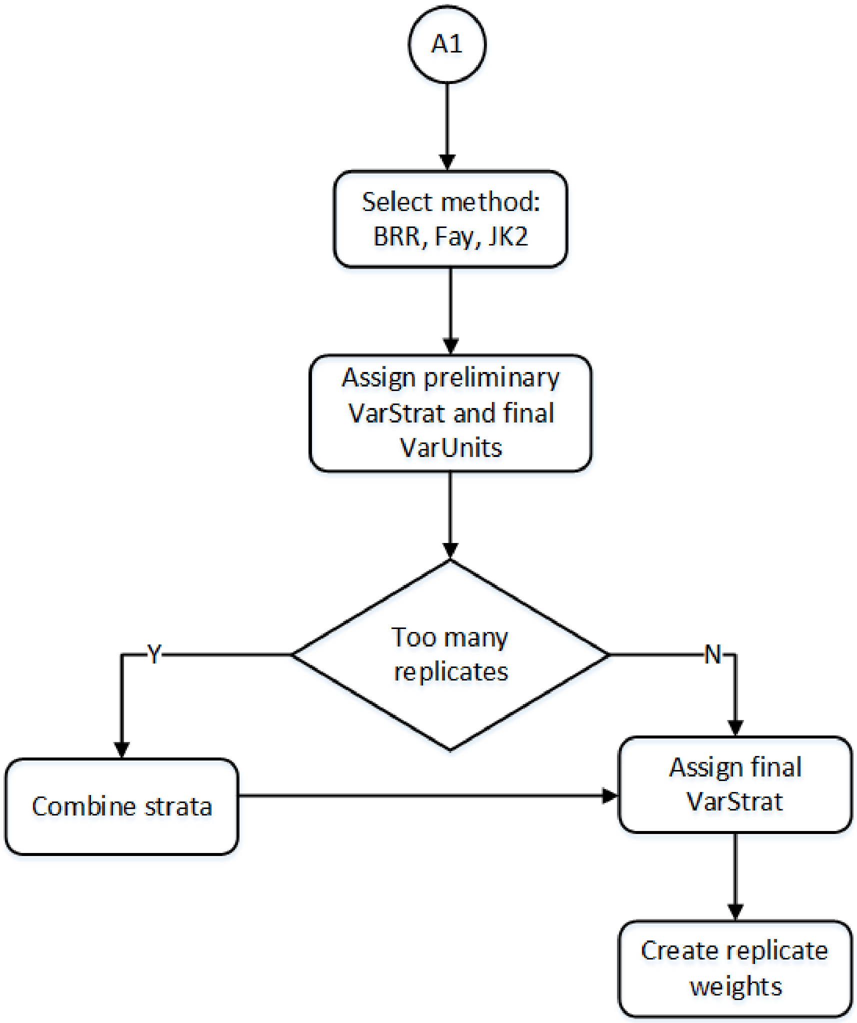

Figure 5.3: Creation of replicates (Step three).

In Figure 5.3 there is an important decision to be made regarding the replication method. Each method has implications and may require further modification of strata. Balanced repeated replication (BRR) method in particular require exactly 2 PSUs/SALs per stratum. If for example, BRR is the method of choice then the additional step of forming appropriate VarUnits is carried out. The BRR method is described in the section that follow.

5.2.2 Balanced repeated replication method

The basic idea behind replication is to select sub-samples repeatedly from the whole sample, calculate the statistic of interest for each sub-sample, and then use these sub-samples or replicate statistics to estimate the variance of the full-sample statistic. The sub-samples are called replicates and the statistics calculated from these replicates are called replicate estimates.

BRR is generally used with multi-stage stratified sample designs with two PSUs/SALs sampled per stratum. To meet the basic requirement of BRR, a variable called “varunit” was formed and it resembled 2 PSUs per strata. The variable was created by first randomising the PSUs in each stratum, creating sequential numbers, and assigning the first half to “VarUnit = 1” while the remaining half are assigned to “VarUnit = 2”. Since SALs were collapsed to form two groups (replicates), the BRR method of replication would be applicable. Each replicate half-sample is formed by selecting one of the two SALs from each stratum. Based on a Hadamard matrix, then only the selected SALs are used to estimate the parameter of interest. To construct the weights for the replicate estimate, the weights of the selected SALs are multiplied by a factor of 2.

The total number of different replicate samples that could be formed is 2^L where L is equal to the number of design strata. However, it is not necessary to form all replicates, because the variance can well be estimated using R “balance” replicates. The minimum number of replicates needed is the smallest integer divisible by 4, which is greater than or equal to L. There were 158 design strata for the design. Therefore, 160 replicates were formed. The “full” orthogonal balance with 160 replicates (the nearest multiple of 4) was actually achieved (Wolter, 1985). The replicates in Table 5.1 were created following replicates creation process in Section 5.2.1.

Table 5.1: Small area layer (SAL) replicate weights.

uqno	**fullwgt**	**rpl001**	**rpl003**	**rpl...**	**rpl159**	**rpl160**
160...009	1253.00	1879.50	626.5	......	1879.5	1879.5
160...013	1367.36	683.68	2051.04	......	2051.04	2051.04
160...002	1211.97	605.98	1817.96	......	1817.96	1817.96
161...091	1389.96	694.98	2084.94	......	2084.94	694.98
161...083	1389.96	694.98	2084.94	......	2084.94	694.981
161...068	1239.69	1859.54	619.84	......	619.84	1859.54
161...111	1118.75	1678.12	559.37	......	559.37	559.37
161...092	1118.75	1678.12	559.37	......	559.37	559.37
161...003	1342.50	671.25	2013.75	......	2013.75	2013.7
....	......	......	......	......	......	......
987...036	1047.50	1571.25	523.75	......	523.75	1571.25
987...009	1047.50	1571.25	523.75	......	523.75	1571.25
987...028	1047.50	1571.25	1571.25	......	1571.25	523.75
987...051	1047.50	523.75	523.75	......	523.75	1571.25

5.2.3 Fay's method

Fay's method is a variant of the BRR, but has better properties in certain situations (Fay and Dippo, 1989). Standard BRR can run into problems when computing an estimate for a small domain or estimating a ratio with very few sample cases for estimating the denominator. The basic idea of Fay's method is to correct this problem by modifying the sample weights less than in BRR, where half the sample is zero-weighted while the other half is double weighted in each replicate. Using Fay's method, one-half sample is weighted down by a factor K $(0 < K < 1)$ and the remaining half is weighted up by a factor 0.5. A perturbation factor of around 70 percent is generally recommended with Fay's method, which is achieved by using a value equal to 0.5 in the FAY_K box in WesVar.

The variance estimate under Fay's method is computed as:

$$v(\hat{\theta}) = \frac{1.0}{R(1.0 - K^2)} \sum_{r=1}^{R} (\hat{\theta}_{(r)} - \hat{\theta}), \tag{5.1}$$

where

θ is an arbitrary parameter of interest,

$\hat{\theta}$ is the estimate of θ based on the full sample,

$\hat{\theta}_{(r)}$ is the estimate of θ based on the r^{th} replicate, and

$v(\hat{\theta})$ is the estimated variance of $\hat{\theta}$.

R is the number of replicate samples, and K is the Fay's K-Factor, where $0 < K < 1$. It should be noted that the replicate estimates are obtained using the replicate final weights that are obtained by applying the non-response adjustment, and then benchmarking to the independent population counts.

The standard error of an estimate is defined as the square-root of the variance of the estimate. The estimate of $\hat{\theta}$ is denoted by an arbitrary population parameter θ, and $v(\hat{\theta})$ is the corresponding variance estimate. Then the standard error of the estimate $\hat{\theta}$ is given by $se(\hat{\theta}) = \sqrt{v(\hat{\theta})}$. The standard error can be used to express the precision of an estimate by computing the 95 percent confidence interval, or the coefficient of variance (CV) of the estimate. These are discussed in the following sections.

5.3 Calibration of replicate weights

The input to the replication of calibrate weights was person-level file with variables that are used to create calibration cells as well as the corresponding replicate weights. As it was discussed in Section 2.3.4 on integrated weighting, the person attributes are taken into consideration in calculating the household final weights.

Final calibrated and replicate weights are the same within one household and

Table 5.2: Person-level replicate weights.

PersonID	prov	age1	age2	race	sex	rpl001
160...901	1	2	1	2	2	1879.5
160...902	1	2	1	2	2	1879.5
160...903	1	2	1	2	2	1879.5
160...301	1	2	1	4	1	683.7
160...302	1	2	1	4	1	683.7
160...701	1	4	2	4	1	683.7
160...702	1	4	2	4	1	683.7
160...703	1	4	2	4	1	683.7
160...704	1	4	2	4	1	683.7
987...301	9	3	2	1	1	523.8
....	.	.	.	.	.	......
987...302	9	3	2	1	1	523.8
987...501	9	6	3	1	1	523.78
987...502	9	6	3	1	1	523.8
987...503	9	6	3	1	1	523.8
987...504	9	6	3	1	1	523.8

as a result, by keeping one household weight, the total will represent household totals. Table 5.2 shows first three records of one household with the same weight (replicate weight). Table 5.3 shows a subset of the calibrated replicate weights.

Table 5.3: Calibrated replicate weights.

PersonID	prov	age2	race	sex	educ	fullcal	calfac	rpcal001
160...901	1	1	2	2	07	1475.8	1.18	2231.9
160...902	1	1	2	2	06	1475.8	1.18	2231.9
160...903	1	1	2	2	06	1475.8	1.18	2231.9
160...904	1	1	2	2	06	1475.8	1.18	2231.9
160...905	1	1	2	2	07	1475.8	1.18	2231.9
160...906	1	1	2	2	08	1475.8	1.18	2231.9
160...301	1	1	4	1	06	1396.4	1.02	655.8
160...302	1	1	4	1	07	1396.4	1.02	655.8
160...701	1	2	4	1	25	625.5	0.45	281.7
260...201	2	1	1	1	06	1282.4	1.19	1938.4
260...202	2	1	1	1	05	1282.4	1.19	1938.4
260...203	2	1	1	1	05	1282.4	1.19	1938.4
260...204	2	1	1	1	05	1282.4	1.19	1938.4
260...001	2	3	1	2	05	1184.7	1.1	1735.8
260...002	2	3	1	2	11	1184.7	1.1	1735.8
766...801	7	2	1	2	12	2212.9	1.02	1109.3
766...802	7	2	1	2	12	2212.9	1.02	1109.3
766...803	7	2	1	2	12	2212.9	1.02	1109.3
978...501	9	2	4	1	23	617.2	0.59	918.9
978...502	9	2	4	1	19	617.2	0.59	918.9
978...503	9	2	4	1	12	617.2	0.59	918.9
978...504	9	2	4	1	16	617.2	0.59	918.9

5.4 Variance estimation

A number of methods are available for estimating sampling errors of estimates based on complex sample designs, among which the Taylor series linearisation method and two replication methods of Jackknife and BRR are the most widely used (Wolter, 1985). In addition, the Bootstrap variance estimation method that is based on re-sampling will also be applicable Rao and Wu (1988) and Rao et al. (1992). While taking account of the complexities of the sample design, these methods provide practically unbiased estimates of variance for most survey estimates.

Replicate base weights were constructed using WesVar estimation system. The non-response adjustment, benchmarking of full sample weights and replicate weights to known population counts were carried out outside the WesVar system in SAS. WesVar system was then used to produce the survey estimates based on the full sample calibrated weights and the corresponding variances of these estimates using the full sample and replicate final weights (Westat, 2000). The replication method can be applied to almost all sample designs, whereas the Taylor linearisation method can be applied only to those statistics that can be linearised. Although SUUDAN software can be used to estimate variances with the Taylor linearisation method, the formulae for the Taylor method would vary by sample design, weighting and estimation procedures. Moreover, the design information (strata and PSU/SAL) would have to be included on the data file.

Replication methods are flexible and can be used with a wide range of complex sample designs, including multi-stage, stratified, and unequal probability samples. The two major reasons for choosing the replication method to estimate variances were, operational convenience for researchers and the ability to reflect all components of the design and estimation in the estimates of variability. With respect to operational convenience, once replicate weights are constructed, the variance estimates can then readily be computed by a simple procedure.

Furthermore, the same procedure is applicable to most statistics desired such as means, percentages, ratios, correlation, and so forth. These estimates can also be calculated for analytic groups or sub-populations. The second reason for choosing replication is probably more important. The non-response adjustment made in developing the sampling weights affects variances. Replicate weights can be developed that reflect this aspect of weighting.

WesVar can calculate estimates of simple statistics such as totals and means, along with their standard error estimates. It is also easy to use WesVar to compute variance estimates for functions of these estimates such as ratios and differences of ratios. Moreover, WesVar can be used to estimate variances of more complex statistics, for example, quantiles, correlations, and so forth. Therefore, WesVar system was used to compute the sampling errors.

Design effect (Deff) (also covered in Section 5.4.3.) is used to determine the total effect of any complex design on the sampling variance in comparison to the alternative simple random sample design, and is defined as the ratio between sampling variance of a complex sample design and the sampling variance of SRS design (Kish, 1965). Deff can also be used to analyse the data collected through the complex sampling survey.

The stratification would reduce the variance whereas clustering would increase the variance. Moreover, differential weighting due to disproportional allocation of the sample would also have an adverse effect on the variance, resulting with an increase in the Deff. The non-response adjustment made in developing the sampling weights also affects the variances. WesVar provides the option to compute design Deff of the estimates, which is a very useful tool to monitor the efficiency of the sample design over time.

The sample design for this study is a two-stage stratified design with more than two SALs sampled from each stratum. The Jackknife replication method of variance estimation would be applicable for the design, but the number of replicates would be +/-1,849, which is the number of sampled SALs. These many replicates (that is, +/-1,849), would be computationally too intensive. Therefore, the SALs were randomly collapsed into two groups within each stratum, such that each group had the same number of SALs. It should be noted that the number of SALs sampled from each stratum is always even, and four or more SALs were sampled from each stratum. After collapsing the SALs, the BRR or the Fay's replication methods were applicable.

5.4.1 Construction of confidence intervals

The 95 percent confidence interval is the interval such that there is a 95 percent chance that the unknown population parameter θ would be within the interval. In case of replication it means that if a sample is drawn from the sample population and variances are calculated, 95 percent of the time the estimate will be within the interval. An example is when we want to measure the statistical significance of the estimate of the differences between two estimates being compared.

Assuming a large sample size, the 95 percent confidence interval is constructed as $\hat{\theta} \pm z_{0.025} \times se(\hat{\theta})$. The lower limit is $\hat{\theta} - z_{0.025} \times se(\hat{\theta})$, and the upper limit of the interval is $\hat{\theta} + z_{0.025} \times se(\hat{\theta})$. The width $z_{0.025} \times se(\hat{\theta})$ is known as half-width of the 95 percent confidence interval. The factor $z_{0.025}$ is the standardised normal value at $\alpha = 0.025$. The smaller the half-width of the confidence interval, the more precise is the survey estimate.

In practical terms when the interval contains 0, the estimate is not statistically significant. When the interval is one-sided, meaning that it is either below or above 0, then the difference is statistically significant.

5.4.2 Coefficient of variation

Alternatively, the precision of the survey estimate can also be expressed in terms of the coefficient of variation (CV) of the estimate (CV is sometimes referred to as the relative standard error). The CV of an estimate is defined as the ratio of the standard error (SE) of the estimate and the magnitude of the estimate expressed in percent, that is, $cv(\hat{\theta}) = 100 \times se(\hat{\theta})/\hat{\theta}$. CV is also a relative measure of the variability of the estimate. The smaller the CV of an estimate, the more precise is the estimate. Confidence intervals are used when differences are calculated as an example in Section 5.4.1.

Statistics Canada classifies CVs according to the following four categories and since the methodologies are similar to those of South Africa the same criteria can also be adopted for the South African studies such as the QLFS, GHS and our study:

0%	< CV <=	16.5%	A = GOOD
16.5%	< CV <=	33.3%	B = FAIR
33.3%	< CV <=	50%	C = CAUTION
	CV >	50%	D = UNREASONABLE

CVs reported together with the estimates from Appendix 8 to Appendix 13 will be evaluated based on the categories above.

5.4.3 Design effect

The Deff computed by WesVar is the ratio of the variance under the actual survey design, in this case, CSSD, to the variance under SRS with replacement. The SRSWR variance is conditional on the achieved sample size for the domain of interest. The Deff may be thought of the effect of the survey design on the variance of an estimate, as compared with that of a SRSWR (Westat, 2006). [Note that this definition of Deff differs from that of Kish (1965), who uses the variance from SRSWOR in the denominator].

WesVar calculates a Deff for proportions of class variables and also for means and totals of quantitative variables. The design effect is labelled Deff on the Wesvar output listing. If the sample size is 0 or 1 for the estimate, however, the Deff is reported as N/A. The numerator of the Deff is the WesVar estimate of variance.

For a proportion, the variance under SRS is calculated as $p(1-p)/n$, where p is the sample estimate of the population proportion using the full-sample weight and n is the sample size on which the estimate of p is based (n is the denominator of the proportion). For the mean of a continuous variable, the variance under SRSWR is estimated as:

$$v_{SRSWR}(\hat{\hat{Y}}) = \frac{\sum_{i=1}^{n} w_i(y_i - \hat{\hat{Y}})^2}{n\hat{N}}, \tag{5.2}$$

where

$v_{SRSWR}(\hat{\hat{Y}})$ is the variance under SRS with replacement, i.e., SRSWR.

The variance estimator under complex survey sample design was calculated based on Fay's method and it is given in Equation (5.1).

5.4.4 Test of significants

Users of regular surveys such as the South African QLFS data are often interested in knowing whether or not an estimate of change is statistically significant. The 95 percent confidence interval or equivalently the t-statistics can be used for this purpose. The value of $Prob > |t|$ can also be used for this purpose.

Confidence Interval

The default Confidence Interval is the 95 percent Confidence Interval. If the value "zero" lies inside the interval then the change is not statistically significant at the 95 percent level of confidence. If the lower limit of the confidence interval is greater than "zero" (> 0) then there is significant increase (Westat, 2000). On the other hand, if the upper limit of the confidence interval is less than "zero" (< 0) then there is a significant decrease, also discussed in Section 5.4.1.

T-statistics

If the absolute value of t-statistics is < 2 (that is, $-2 < t < +2$), then the change is not statistically significant at 95 percent level of confidence. If the value of t-statistics ≥ 2, then there is a significant increase. On the other hand, if the value of t-statistics ≤ -2 then there is significant decrease.

WesVar computes individual parameter estimates and their standard errors for every term in the linear, logistic, and multinomial logistic regression models. The output gives a t-statistic for each estimated parameter that can be used to test whether the parameter is significantly different from 0 (Choudhry and Valliant, 2002). The t-statistic is defined as

$$t = \frac{b_k}{\sqrt{v(b_k)}}, \tag{5.3}$$

where b_k is a parameter estimated.

Prob > |t|

The $Prob > |t|$ is the probability of observing an absolute value of t-statistics larger than the observed one when the null hypothesis is true, that is, no significant difference. If the value of $Prob > |t|$ is greater than 5.0 percent, then there is no significant change. Otherwise, the change is statistically significant. Often, the estimate of change is known as significant if the value of the probability is between 1.0 and 5.0 percent and this is indicated by a single star (*) beside the estimate of change. If the value of $Prob > |t|$ is less than or equal to 1.0 percent then the change is highly significant and this is indicated by double star (**) beside the estimate of change.

The other view of the same concept is that the p-value corresponding to an estimate of difference between two estimates being compared is the probability of observing a value larger than the particular observed value under the hypothesis that there is no statistically significant difference. If p-value <0,01, the difference is highly significant; if p-value is between 0,01 and 0,05, the difference is significant; and if p-value >0,05, the difference is not significant.

This important estimator was used in the South African QLFS for the first time to measure the changes of estimates from one quarter to the other. At one stage the unemployment rate dropped in thousands in one quarter and the question was whether the drop was statistically significant. The hypothesis was that the change reported from one quarter to the other is not statistically significant. The $Prob > |t|$ and the confidence intervals were used to test the hypothesis.

5.5 Estimation of results

5.5.1 Background

Table 5.4 gives simple results from the estimation process, where a point estimate, CV, as well as the Deff are given by gender. National, provincial and demographic totals were used as control totals during calibration, and as a result there is no variation observed and the CV is 0 and the Deff is also 0. When estimates are to be drawn for domains that were not directly used for benchmarking they are subject to variation which is also observed in the form of CV and Deff. Tables in Appendix 8 onwards report cell percentages corresponding to each value estimate as well as the CVs and Deff(s). Results for sub-national as well as other domains are presented in Section 5.5.3.

Table 5.4: National estimates by gender.

Gender	Estimate(%)	CV(%)	Deff
Male	48.96	0	0
Female	51.04	0	0
Total	100	0	.

5.5.2 Domain and point estimates

A Table Request function in WesVar was used. It allows specification of simple two-way tables (for example, B*A) and more complex multi-way tables (for example, D*C*B*A). In the two-way table, the output consists of estimates for all the variables specified: for each cell in the table, for all of the margins (that is, the cells corresponding to each level of the A variable when B is summed across all its levels), and for the grand total.

When the estimated percentages are produced for a two-way table, these are also the ordinary percentages. For example, consider the table B*A with levels as shown above, where the entries in each cell of the table are the estimated sum of weights for the cells. The percentages for cell (a, b) are estimated in a Table Request by $100\hat{N}_{ab}/\hat{N}_{.}$ for the table percentage, $100\hat{N}_{ab}/\hat{N}_{a.}$ for the row percentage, and $100\hat{N}_{ab}/\hat{N}_{.b}$ for the column percentage. These are ratio estimates in general, and the estimates of the standard errors are computed accordingly (Westat, 2006).

5.5.3 Survey estimates and measures of precision for selected variables

The concepts illustrated above are used to produce estimates for selected variables. The selected variables are demographics, provinces, language and education. The estimates are provided in the form of cell percentages, that is, an estimate shows the percentage measured at national-level taking into account all variables being measured as opposed to row and cell percentages where percentages are calculated for each province.

The estimates are provided with the corresponding measures of precision, that is, CV to measure the reliability of estimate and Deff to measure the effect of the chosen design. All the results tables also provide the Deff which explains the benefits of using a complex sample design vis-à-vis the SRS.

South African: highest-level of education estimates

Appendix 8 is a one-way table of the highest-level of education computed nationally. There are 30 categories under the variable highest-level of education of the first being Grade 0 and last being No schooling. The majority of population are part of the school going age from Grade 1/Sub A to Grade 12/Std 10/Form 5 and among the above categories majority of population have completed Grade 12/Std 10/Form 5.

The CVs for all the groups are less than 6 percent showing that the estimates for the categories are reliable. As sample sizes tend to be lower for categories outside Grade 0 and Grade 12/Std 10/Form 5 even the corresponding CVs go higher, ranging from 7 to 19 percent. In contrast, the Deff for the latter tend to be lower while they also seem to be correlated to the lower sample sizes. Looking specifically at the NTC1 to NTC6, Deff(s) do not exceed 2, which implies that less than half sample size is required under SRS to yield the same precision. Population of Grade 12/Std 10/Form 5 is estimated at 16.9 percent with the lowest CV of 1.7 percent and Deff of 2.1 implying that to estimate this category under SRS, half the sample size is required to achieve the same precision achieved under CSSD.

South African: highest-level of education estimates by gender

Appendix 9 shows the breakdown of the highest-level of education by gender. While estimates for the males and females follow the pattern of national estimates, observing the estimates for Grade 12/Std 10/Form 5 among males and females shows a difference. There are 7 percent less males who have completed Grade 12/Std 10/Form 5 as compared to their females counterpart. The difference is statistically significant given the confidence interval which does not contain zero and the p-value of 0.019 which is below 0.05. Although the population size of females is greater than that of males, however, the Deff under CSSD is 2.4 times higher than could have been achieved through SRS, while the Deff for males was 1.6 times larger than if SRS was used. The CVs for both males and females are less than 3 percent, implying that the estimates are reliable.

South African: highest-level of education estimates by race

When continuing to study the population of those who completed Grade 12/Std 10/Form 5 comparing black/africans and each population group, CVs are 2.1 percent for Black/African, 6.7 percent for Coloureds, 7.8 percent for Indians and 10 percent for Whites. The differences in CVs is mainly due to smaller population sizes and subsequently the smaller sample sizes. When the sample size increases, the variances reported in the form of CVs decrease. All four estimates for the population group or race are considered to be good with all the CVs falling within category A (see Section 5.4.2). Observing the Deff for all four population groups, the Deff(s) are: 2.3 for Black/African, 1.9 for Coloureds, 1.4 for Indians and 7.5 for Whites. The Deff for whites shows that it will require a very small sample size to produce the same level of precision using SRS, that is, 1 in 7 (see Appendix 10).

South African: highest level of education estimates by age

The variable highest level of education is highly correlated to age. Age group one contains ages (0 - 19), as a result very few persons in this category will have post-matric including those with NTC1 to NTC6. Thus, post-matric results by age are not reliable, hence, the CVs are as high as 100 percent (according to CVs categories in Section 5.4.2 these estimates are unreasonable). Age group two contains ages (20 - 49) who are outside the school going age. The categories NTC1 to NTC6 seem to have very high CVs in the latter age group.

South African: language estimates

Appendix 12 shows the top three most dominant languages in South Africa, which are, IsiZulu (23.8%), IsiXhosa (16.4%) and Afrikaans (13%). The top three languages have the smallest Deff(s) and the corresponding CVs. Looking at the Deff(s) it will take very small sample sizes to achieve same precision while studying a population of languages that are least spoken country-wide.

South African: language by province

Appendix 13 shows the distribution of languages by province. Each province has at least one dominant language, those are, Western Cape (Afrikaans (6.4%), English (2.6%) and IsiXhosa (2.7%)); Eastern Cape (IsiXhosa (11.1%)); Northern Cape (Afrikaans (1.5%)); Free State (Sesotho (3.9%)); Kwazulu Natal (IsiZulu (17.6%) and English (2.4%)); North West (Setswana (5.4%)); Gauteng (IsiZulu (3.7%), Sesotho (2.5%), Sepedi (2.4%) and Afrikaans (2%)); Mpumalanga (SiSwati (2.4%)) and Limpopo (Sepedi (5.8%), Xitsonga (2.4%) and Tshivenda (1.8%)). Variances of languages spoken vary greatly from province-to-province depending on whether the language is dominant or not. Estimates for languages that are less dominant have larger CVs therefore they are less reliable.

Results summary

Better precisions could be achieved by including both languages and education level in the stratification process or alternatively control for these variables in calibration. The 30 categories for the highest level of education result in a very small sample for analysis. The shortcoming could be controlled by grouping categories into much broader categories, those could be: No schooling, Some primary, Completed primary, Some secondary, Grade 12/Std 10 and Higher.

Chapter 6

Conclusion and recommendations

6.1 Summary

CSSD methodologies were defined and illustrated using examples and literature. The study evaluated various approaches to probability sampling and the approaches to implement CSSD methodologies. South African Census 2011 data were used to implement the CSSD sampling methodology and produced a sample which can be used in practice. It was mentioned that there was no listing or data collection operations conducted for this country-wide sample. As a result, household sampling frame and survey responses data were simulated. Simulation was based on real census responses using pseudo-records that represented individuals and persons. The actual records of households and persons residing in those households were not available for public use, hence the simulation was done to achieve a proxy survey data.

Based on the theory of CSSD and the available datasets, a representative country-wide probability sample was drawn. Proxy survey data were matched with the sample to determine valid responses from the original sample supposedly sent for data collection. Part of the simulation involved using provincial average response rates from previous selected surveys to simulate response rates and distribute the response rates proportionally to SALs. Using the resulting responses at SAL-level, non-response adjustments were done, however, where the adjustment factors were too high or there was a complete non-response, then the non-response adjustments were done at stratum level. It should be noted that before non-response adjustments were applied, other adjustments for the initial base-weights were done during sampling by adjusting the ISR and they should be applied during weighting. Such adjustments include list-sub-sampling where growth had occurred between census and sample selection/implementation.

In household surveys, it is common to have weights for households as well as the weights for persons separately. It is also common that the auxiliary totals are available for persons, and not for households. Several papers proposed methods of producing one weight which can be used to estimate households and persons, and they called those methods integrated weighting. In some literature these are referred to as weight equalisation where the weights are made to be the same at household level, that is, each person in a household has the same weight, which is the household weight. The integrated weighting approach was applied in this study using generalised regression methods and implemented through the SAS-based StatMx software developed by Statistics Canada. The same procedure was used to produce full calibrated and replicate calibrated weights. The replicates were produced using Fay's method - an extension of BRR implemented in WesVar application (Westat, 2000). The Fay's method (Fay and Dippo, 1989) used in the study is a variant of BRR that was found to have better properties in certain situations. Standard BRR can run into problems when computing an estimate for a small domain or estimating a ratio if the denominator has few sample cases. Fay's method corrects this problem by retaining all sample units in each replicate while, modifying the sample weights differently than in the standard BRR (Choudhry and Valliant, 2002).

With all the complexities introduced in a process, standard statistical software are unable to incorporate design information to produce estimates. Though the estimation of totals, means and percentages leads to the same totals using various tools, the estimates of measures of precision differ. The risk of reporting reliability of estimates without taking into account the sample design information, may lead to incorrect conclusions. In this study WesVar application was further used to produce estimates, taking into account the design information and using replication methods for variance estimation.

Researchers may use various methods to calculate both first-order and second-order statistics. In the calculation of first-order statistics such as frequencies, totals and means, generally results are the same, depending on whether or not weights were applied. When making inference about the estimates obtained from complex sample survey and statistics such as variances, standard error, CV and confidence interval, it is necessary to observe the survey design.

The study used appropriate complex sampling analyses via WesVar software, which is designed to accurately account for complex sampling in analyses. Emphasis was made on using WesVar software (with replication method) to provide estimates together with measures of precision. WesVar software was utilised to appropriately model the weight, SAL, and cluster information provided in the data to account for all issues mentioned above (Westat, 2006). Wes-

Var computes estimates and replication variance estimates that do properly reflect complex sampling and estimation procedures. Replication variance estimation consists of repeatedly calculating estimates for sub-groups of the full sample and then computing the variance among these “replicate” estimates.

6.2 Recommendations

Researchers must take the time to understand the sampling methodology used and appropriately utilise weighting and Deff, which to a novice researcher can be potentially confusing and intimidating. There is mixed evidence on researcher’s utilisation of appropriate methodology (Johnson and Elliott, 1998), which highlights the need for more debate around this important issue. The goal of this study was to introduce some of the issues around using complex samples and explore the possible consequences (for example, Type I errors) of failure to appropriately model the CSSD methodology. It is well known that failure to appropriately model the complex sample can substantially bias the results of the analysis.

Most datasets from CSSD that are made available by government agencies and research organisations contain weights that have already taken account of the effects of clustering, stratification, coverage (for example, exclusions of eligible elements), non-response and other sampling errors. The standard methods used by most statistical analysis software assume that samples were drawn using SRS (Choudhry and Valliant, 2002). As a result, application of standard methods in the analysis of complex sampling surveys, could lead to drawing wrong conclusions from survey data. One of the reasons is that the effects of unequal probability of selection from complex sampling survey at this stage exist. To properly analyse data from complex sample surveys, the design features need to be included in the analysis.

The study covers the existing methods that researchers are using to analyse data from complex sampling surveys. These approaches include methods that either take into account or do not take into account (sample) design information. The population of all households is sometimes called the target population or the universe. Without the application of both probability sampling and weighting, there is no supporting statistical theory to provide a link between the sample observations and the target population parameters (Chromy and Abeyasekera, 2005).

If the sample is designed to generate equal probability sample, then the weights for estimating means, rates, or relationships among variables may be safely ignored. In general, both the population and the sample design (imposed in frames) can have some structure. While the structure does not influence the construction of first-order statistical estimates such as totals, means, ratios, or model coefficients, it does affect second-order statistics (variance estimates) that allow analysts to estimate the standard errors of the first-order statistics and to construct tests of statistical significance concerning specified hypotheses (Chromy and Abeyasekera, 2005).

Results from WesVar software for analysis of CSSD are presented in the form of CV, CI and p-value. Replication methods implemented in WesVar correctly report the estimates and also enable calculation of the measures of precision. There is a wealth of compelling data freely available to researchers, and some analysts have found evidence that researchers do not always model the sampling frame appropriately (Johnson and Elliott, 1998). In brief, most modern statistical software can take CSSD into account, either through using weights scaled for N and Deff, or through using information such as primary and secondary sampling units (often called clusters) directly in the software (Osborne, 2011). The example of the latter is the WesVar software.

References

ALLEN, N. L., CARLSON, J. E., AND ZELENAK, C. A. (1999). *The 1996 NAEP Technical Report. ERIC*.

BANKIER, M. D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, **42** (3), 174–177.

BAR-GERA, H., KONDURI, K. C., SANA, B., YE, X., AND PENDYALA, R. M. (2009). Estimating survey weights with multiple constraints using entropy optimization methods. Technical report.

BRICK, J. M. AND MONTAQUILA, J. M. (2009). Nonresponse and weighting. *In Handbook of Statistics*, volume 29. Elsevier, pp. 163–185.

CHEN, Q., ELLIOTT, M. R., HAZIZA, D., YANG, Y., GHOSH, M., LITTLE, R. J., SEDRANSK, J., THOMPSON, M., ET AL. (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, **32** (2), 227–248.

CHOUDHRY, G. H. (2009). Survey sampling: An introduction. *Statistics South Africa survey sampling training*.

CHOUDHRY, G. H. AND VALLIANT, R. (2002). WesVar: Software for complex survey data analysis. *Proceedings of Statistics Canada Symposium 2002*.

CHROMY, J. R. AND ABEYASEKERA, S. (2005). Statistical analysis of survey data. *Household sample surveys in developing and transition countries, studies in methods [text on the Internet]. New York: United Nations*.

COCHRAN, W. G. (1965). *Sampling techniques: 2nd Ed.* Wiley. New York.

COCHRAN, W. G. (1977). *Sampling Techniques: 3rd Ed.* Wiley. New York.

CSS (1993). *October Household Survey 1993*. Central Statistical Services - South Africa.

CSS (1994). *October Household Survey 1994*. Central Statistical Services - South Africa.

DEVILLE, J.-C. AND SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87** (418), 376–382.

DEVILLE, J.-C., SÄRNDAL, C.-E., AND SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, **88** (423), 1013–1020.

DIHIDAR, K. (2014). Estimating population mean with missing data in unequal probability sampling. *Sampling methods and estimation*, **369**, 369.

DREW, J. D., CHOUDHRY, G. H., AND GRAY, G. B. (1978). Some methods for updating sample survey frames and their effects on estimation. *In Proceedings of the Section on Survey Research Methods*. The Association, pp. 62–71.

EFRON, B. AND TIBSHIRANI, R. J. (1994). *An introduction to the Bootstrap*. CRC press.

FAY, R. E. AND DIPPO, C. S. (1989). Theory and application of replicate weighting for variance calculations. *In Proceedings of the Section on Survey Research Methods, American Statistical Association*, volume 12. pp. 212–217.

FINMARK (2010). *FinScope SME South Africa - Report*. FinMark Trust.

FOX, M. P., BRENNAN, A., MASKEW, M., MACPHAIL, P., AND SANNE, I. (2010). Using vital registration data to update mortality among patients lost to follow-up from ART programmes: Evidence from the Themba Lethu Clinic, South Africa. *Tropical Medicine & International Health*, **15** (4), 405–413.

FULLER, W. A. (2002). Regression estimation for survey samples (with discussion). *Survey Methodology*, **28** (1), 5–23.

GAMBINO, J., KENNEDY, B., AND SINGH, P., MANGALA (2001). Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, **27** (1), 65–74.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 153–164.

GROVES, R. M., FOWLER, F. J., COUPER, M. P., LEPKOWSKI, J., SINGER, E., AND TOURANGEAU, R. (2004). Inference and error in surveys. *Survey Methodology, Wiley Series in Survey Methodology. Groves RM, Kalton G, Rao JNK, et al (Eds). New York, Wiley*, 39–63.

HA, N. S. (2013). *Hierarchical Bayesian estimation of small area means using complex survey data*. Ph.D. thesis.

HEERINGA, S. G. AND LIU, J. (1998). Complex sample design effects and inference for mental health survey data. *International Journal of Methods in Psychiatric Research*, **7** (1), 56–65.

HENDERSON, R. H. AND SUNDARESAN, T. (1982). Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bulletin of the World Health Organization*, **60** (2), 253.

HENRY, K. AND VALLIANT, R. V. (2012). Methods for Adjusting Survey

Weights when Estimating a Total. In Proceedings of the 2012 Federal Committee on Statistical Methodology's Research Conferenc.

HOLT, D. AND ELLIOT, D. (1991). Methods of weighting for unit non-response. *The Statistician*, 333–342.

HOSHAW-WOODARD, S. (2001). Description and comparison of the methods of cluster sampling and lot quality assurance sampling to assess immunization coverage.

HUANG, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. *DMKD*, **3** (8), 34–39.

IHLCA (2011). Integrated household living conditions survey in Myanmar (2009-2010). *UNDP Myanmar Technical Report*, 29–39.

ISRAEL, G. D. (1992). *Sampling the evidence of extension program impact*. University of Florida Cooperative Extension Service, Institute of Food and Agriculture Sciences, EDIS.

JOHNSON, D. R. AND ELLIOTT, L. A. (1998). Sampling design effects: Do they affect the analyses of data from the National Survey of Families and Households? *Journal of Marriage and the Family*, 993–1001.

KALTON, G. (1983). *Introduction to survey sampling (Vol. 35)*. Sage.

KALTON, G. (2009). Methods for oversampling rare sub-populations in social surveys. *Statistics Canada*, **35** (2), 125–141.

KIM, J. K. AND SKINNER, C. J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, **100** (2), 385–398.

KISH, L. (1965). Survey sampling. Wiley. New York.

KISH, L. (1988). Multipurpose sample designs. *Survey Methodology*, **14** (1), 19–32.

KOLENIKOV, S. ET AL. (2010). Resampling variance estimation for complex survey data. *Stata Journal*, **10** (2), 165–199.

KUBACKI, J. AND JĘDRZEJCZAK, A. (2012). The comparison of generalised variance functions with other methods of precision estimation for Polish household survey. *Survey Sampling in Economic and Social Research*, **120**, 58–69.

LUUS, R. (2016). *Statistical Inference of the Multiple Regression Analysis of Complex Survey Data (Doctoral Dissertation)*. Ph.D. thesis, Stellenbosch University.

MCCARTHY, P. J. (1969). Pseudo-replication: Half samples. *Revue de l'Institut International de Statistique*, 239–264.

MOHADJER, L. AND CHOUDHRY, G. H. (2002). Adjusting for Missing Data in Low-Income Surveys. *Studies of welfare populations: Data collection and research issues*, 129–156.

MOHADJER, L., MONTAQUILA, J. M., WAKSBERG, J., BELL, B., JAMES, P., FLORES-CERVANTES, I., AND MONTES, M. (1996). National Health and Nutrition Examination Survey III: Weighting and estimation methodology. Rockville, MD: Westat.

NEETHLING, A. AND GALPIN, J. S. (2006). Weighting of household survey data: A comparison of various calibration, integrated and cosmetic estimators: Theory and methods. *South African Statistical Journal*, **40** (2), 123–150.

OECD (2007). *Glossary of statistical terms*. Organisation of Economic Cooperation and Development(OECD), Paris.

OSBORNE, J. W. (2011). Best practices in using large, complex samples: The importance of using appropriate weights and design effect compensation. *Pract Assess Res Evaluat*, **16** (12), 1–7.

PORTER, S. R. (2008). *Encyclopedia of Survey Research Methods: 2nd Ed.* SAGE Publications, Inc.

PUCKCHARERN, H. (2013). Thailand FinScope Survey 2013: The sample design. *FinScope Thailand*.

PURCELL, N. J. AND KISH, L. (1979). A biometrics invited paper. Estimation for small domains. *Biometrics*, 365–384.

PURCELL, N. J. AND KISH, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review/Revue Internationale de Statistique*, 3–18.

RAO, J. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, **10** (2), 153.

RAO, J. N. K. (2003). *Small Area Estimation*. J. Wiley & Sons.

RAO, J. N. K. AND SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 403–415.

RAO, J. N. K. AND WU, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, **83** (401), 231–241.

RAO, J. N. K., WU, C. F. J., AND YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, **18** (2), 209–217.

REID, A. AND HALL, D. W. (2001). Using equalisation constraints to find optimal calibration weights. *U.S. Bureau of the Census*.

REITER, J. P., ZANUTTO, E. L., AND HUNTER, L. W. (2005). Analytical modeling in complex surveys of work practices. *ILR Review*, **59** (1), 82–100.

ROSS, K. N. (1978). *Sample design for educational survey research*. Pergamon Press Oxford.

RUST, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, **1** (4), 381.

SÄRNDAL, C. E., SWENSSON, B., AND WRETMAN, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

SAS (2014). *SAS/STAT 13.2 User's Guide The FASTCLUS Procedure*. SAS Institute.

SHAO, J. (1993). Balanced repeated replication. *In Proceedings of the Section on Survey Research Methods, American Statistical Association*. pp. 544–549.

SHAO, J. AND TU, D. (1995). The Jackknife and Bootstrap. 1995.

STATCAN (2003). *Survey methods and practices*. Statistics Canada.

STATSSA (2001). *Labour Force Survey March 2001, Statistical release P0210*. Statistics South Africa.

STATSSA (2002). *Income and Expenditure Survey 2000, Statistical release P0111*. Statistics South Africa.

STATSSA (2003). *General Household Survey 2002*. Statistics South Africa.

STATSSA (2008). *Quarterly Labour Force Survey, Quarter 2, 2008, Statistical release*. Statistics South Africa.

STATSSA (2013). *Survey of Empoyers and Self-Employed (SESE), Statistical release*. Statistics South Africa.

STATSSA (2017). *Motor Trade Sales March, Statistical release P6343.2*. Statistics South Africa.

STEINHAUER, H. W. (2014). *Sampling techniques and weighting procedures for complex survey designs (Doctoral dissertation)*. Ph.D. thesis, University of Bamberg.

TAMBAY, J.-L. AND CATLIN, G. (1995). Sample design of the national population health survey. *Health Reports*, **7** (1), 29–38.

TILLÉ, Y. (2011). *Sampling algorithms*. Springer.

TÖRMÄLEHTO, V. M. (2008). Social statistics–integrated use of survey and administrative data at Statistics Finland. International Association for Official Statistics Conference on Reshaping Official Statistics, Shanghai.

UDJO, E. O. (2017). Can Estimating Completeness of Death Registration be used as Evidence of Inaccuracy of Population Size Estimates from a Census? The Case of the 2011 South African Population Census. *African Population Studies*, **31** (1).

UNSD (1998). Principles and recommendations for Population and Housing Censuses. *United Nations Publications*, **6** (1).

VANCE, E. AND PRUITT, T. (2016). LISA (Laboratory for Interdisciplinary Statistical Analysis). *2014-15 LISA Annual Report. VirginiaTech*.

WALLACE, L. AND RUST, K. (1996). A Comparison of raking and poststratification using 1994 NAEP Data. Westat. Rockville.

WESTAT (2000). WesVar 4.0 User's guide. Westat. Rockville.

WESTAT (2006). WesVar 4.3 User's guide. Westat. Rockville.

WOLTER, K. M. (1985). Introduction to variance estimation Springer-Verlag. *New York*, 115, 427.

WU, S., KENNEDY, B., AND SINGH, A. C. (1997). Household-level versus person-level regression weight calibration for household survey. *SSC Annual Meeting, June 1997*.

WYWIAŁ, J. AND ŻĄDŁO, T. (2012). Survey sampling in economic and social research. *Universitatis Economicae Sugillum*.

Appendices

Appendix 1: Full frame separating excluded SALs

		Included		Excluded	
No.	**Explicit Strata**	**Households**	**SAL**	**Households**	**SAL**
1	111	1 086 138	5 139	22 997	184
2	113	5 041	15	15	2
3	121	453 946	2 390	18 182	177
4	123	118 824	502	530	20
5	211	498 712	2 699	25 242	164
6	212	33 918	273	525	9
7	213	9 851	49	-	-
8	221	359 152	2 078	19 983	153
9	222	764 488	8 102	5 044	240
10	223	38 491	328	479	14
11	321	215 478	1 288	7 907	109
12	322	50 816	347	208	24
13	323	37 951	237	458	8
14	411	203 420	1 073	11 842	76
15	412	16 252	121	187	7
16	413	5 746	34	37	2
17	421	473 479	2 694	20 666	199
18	422	55 663	429	2 039	17
19	423	49 851	484	577	10
20	511	858 800	3 983	27 784	177
21	512	108 374	595	1 124	7
22	513	5 570	29	13	1
23	521	557 793	2 920	24 557	189
24	522	842 604	5 628	19 445	243
25	523	187 180	989	2 414	44

		Included		Excluded	
No.	**Explicit Strata**	**Households**	**SAL**	**Households**	**SAL**
26	621	486 853	2 504	23 283	165
27	622	484 367	2 835	2 209	26
28	623	98 935	475	2 573	24
29	711	3 373 403	13 842	142 704	709
30	712	38 648	226	218	2
31	713	32 319	146	530	12
32	721	522 930	2 568	23 726	153
33	723	30 099	188	162	4
34	821	494 980	2 676	17 580	215
35	822	492 309	3 323	3 237	73
36	823	88 859	453	5 904	73
37	921	281 058	1 342	9 594	85
38	922	1 056 640	7 703	6 015	172
39	923	91 644	554	3 625	51
00	RSA	14 610 582	81 261	454 615	3 840

Appendix 2: Frame with only included SALs

No.	Explicit Strata	Small Area Layer	Households
1	111	5 233	1 075 632
2	113	19	4 802
3	121	2 293	445 230
4	123	517	114 888
5	211	2 665	494 643
6	212	206	30 693
7	213	38	9 373
8	221	1 956	350 198
9	222	3 822	535 806
10	223	169	33 688
11	321	1 204	209 456
12	322	263	46 729
13	323	177	35 766
14	411	1 067	201 747
15	412	98	15 015
16	413	27	5 359
17	421	2 629	468 330
18	422	401	53 972
19	423	221	37 438
20	511	4 236	852 260
21	512	567	105 939
22	513	26	5 390
23	521	2 871	550 433
24	522	4 654	780 280
25	523	870	175 883
26	621	2 545	482 265
27	622	2 751	474 557
28	623	59	95 789
29	711	15 591	3 359 922
30	712	224	38 470
31	713	141	31 046
32	721	2 578	516 512
33	723	144	28 114
34	821	2 556	486 463
35	822	3 138	483 519
36	823	379	83 002
37	921	1 382	278 018
38	922	6 767	1 006 540
39	923	430	83 984

Appendix 3: Sample distribution for Thailand FinScope

		Total		Urban		Rural	
		EAs	Persons	EAs	Persons	EAs	Persons
	Central	**120**	**1,200**	**60**	**600**	**60**	**600**
11	Samut Prakan	13	130	8	80	5	50
12	Nonthaburi	9	90	6	60	3	30
13	Pathum Thani	9	90	6	60	3	30
14	Phra Nakhon Si Ayutthaya	5	50	2	20	3	30
15	Ang Thong	2	20	1	10	1	10
16	Lop Buri	5	50	2	20	3	30
17	Sing Buri	2	20	1	10	1	10
18	Chai Nat	2	20	1	10	1	10
19	Saraburi	4	40	1	10	3	30
20	Chon Buri	11	110	9	90	2	20
21	Rayong	5	50	3	30	2	20
22	Chanthaburi	3	30	2	20	1	10
23	Trat	2	20	1	10	1	10
24	Chachoengsao	4	40	1	10	3	30
25	Prachin Buri	4	40	1	10	3	30
26	Nakhon Nayok	2	20	1	10	1	10
27	Sa Kaeo	3	30	1	10	2	20
70	Ratchaburi	5	50	2	20	3	30
71	Kanchanaburi	5	50	2	20	3	30
72	Suphan Buri	5	50	1	10	4	40
73	Nakhon Pathom	6	60	2	20	4	40
74	Samut Sakhon	6	60	3	30	3	30
75	Samut Songkhram	2	20	1	10	1	10
76	Phetchaburi	3	30	1	10	2	20
77	Prachuap Khiri Khan	3	30	1	10	2	20

		Total		Urban		Rural	
		EAs	**Persons**	**EAs**	**Persons**	**EAs**	**Persons**
	Northeast	**120**	**1,200**	**60**	**600**	**60**	**600**
30	Nakhon Ratchasima	16	160	7	70	9	90
31	Buri Ram	8	80	4	40	4	40
32	Surin	6	60	2	20	4	40
33	Si Sa Ket	6	60	2	20	4	40
34	Ubon Ratchathani	10	100	4	40	6	60
35	Yasothon	2	20	1	10	1	10
36	Chaiyaphum	5	50	2	20	3	30
37	Amnat Charoen	2	20	1	10	1	10
38	Bueng Kan	2	20	1	10	1	10
39	Nong Bua Lam Phu	3	30	2	20	1	10
40	Khon Kaen	13	130	8	80	5	50
41	Udon Thani	8	80	5	50	3	30
42	Loei	4	40	2	20	2	20
43	Nong Khai	4	40	2	20	2	20
44	Maha Sarakham	5	50	2	20	3	30
45	Roi Et	7	70	4	40	3	30
46	Kalasin	7	70	5	50	2	20
47	Sakon Nakhon	6	60	3	30	3	30
48	Nakhon Phanom	3	30	1	10	2	20
49	Mukdahan	3	30	2	20	1	10
	South	**120**	**1,200**	**60**	**600**	**60**	**600**
80	Nakhon Si Thammarat	17	170	5	50	12	120
81	Krabi	4	40	1	10	3	30
82	Phangnga	3	30	1	10	2	20
83	Phuket	10	100	8	80	2	20
84	Surat Thani	14	140	8	80	6	60
85	Ranong	4	40	3	30	1	10
86	Chumphon	7	70	3	30	4	40
90	Songkhla	23	230	16	160	7	70
91	Satun	3	30	1	10	2	20
92	Trang	7	70	2	20	5	50
93	Phatthalung	8	80	5	50	3	30
94	Pattani	7	70	2	20	5	50
95	Yala	5	50	2	20	3	30
96	Narathiwat	8	80	3	30	5	50

		Total		Urban		Rural	
		EAs	**Persons**	**EAs**	**Persons**	**EAs**	**Persons**
	North	**120**	**1,200**	**60**	**600**	**60**	**600**
50	Chiang Mai	21	210	15	150	6	60
51	Lamphun	5	50	4	40	1	10
52	Lampang	9	90	6	60	3	30
53	Uttaradit	4	40	2	20	2	20
54	Phrae	4	40	2	20	2	20
55	Nan	4	40	1	10	3	30
56	Phayao	5	50	3	30	2	20
57	Chiang Rai	13	130	7	70	6	60
58	Mae Hong Son	2	20	1	10	1	10
60	Nakhon Sawan	9	90	3	30	6	60
61	Uthai Thani	3	30	1	10	2	20
62	Kamphaeng Phet	8	80	3	30	5	50
63	Tak	4	40	2	20	2	20
64	Sukhothai	6	60	2	20	4	40
65	Phitsanulok	9	90	3	30	6	60
66	Phichit	5	50	2	20	3	30
67	Phetchabun	9	90	3	30	6	60

Appendix 4: All frame auxiliary variables

No.	Variable Name	Variable Description
1	SAL_Code_S	Small Area Layer
2	ex_strata	Explicit Strata
3	metro_id	Metropolitan Area Identifier
4	metro	Metro and Non-Metro grouping
5	MN_CODE_2011	Local Municipality
6	DC_MN_C_2011	District Municipality
7	PR_CODE_2011	Province
8	EA_GTYPE_C	Geography Type
9	EA_TYPE_C	EA Type
10	AG01	Age: 0 - 4
11	AG02	Age: 15 - 34
12	AG03	Age: 35 - 64
13	AG04	Age: 65+
14	G01	Gender: Male
15	G02	Gender: Female
16	PG01	Race: Black African
17	PG02	Race: Coloured
18	PG03	Race: Indian or Asian
19	PG04	Race: White
20	PG05	Race: Other
21	ES01	Employment Status: Employed
22	ES02	Employment Status: Unemployed
23	ES03	Employment Status: Discouraged work-seeker
24	ES04	Employment Status: Other not economically active
25	ES05	Employment Status: Not applicable
26	OC01	Occupation: Legislators; senior official and managers
27	OC02	Professionals
28	OC03	Occupation: Technical and associate professionals
29	OC04	Occupation: Clerks
30	OC05	Occupation: Service workers; shop and market sales workers
31	OC06	Occupation: Skilled agricultural and fishery workers
32	OC07	Occupation: Craft and related trades workers
33	OC08	Occupation: Plant and machine operators and assemblers
34	OC09	Occupation: Elementary Occupation
35	OC10	Occupation: Domestic workers
36	OC11	Occupation: Not applicable
37	IN01	Industry: Agricultural; hunting; forestry and fishing
38	IN02	Industry: Mining and quarrying

No.	Var Name	Var Description
39	IN03	Industry: Manufacturing
40	IN04	Industry: Electricity; gas and water supply
41	IN05	Industry: Construction
42	IN06	Industry: Wholesale and retail trade
43	IN07	Industry: Transport; storage and communication
44	IN08	Industry: intermediation; insurance; real estate and business services
45	IN09	Industry: Community; social and personal services
46	IN10	Industry: Private households
47	IN11	Industry: Other
48	IN12	Industry: Not applicable
49	Total_HH	Total Households
50	TF01	Toilet Facility: None
51	TF02	Toilet Facility: Flush toilet (connected to sewerage system)
52	TF03	Toilet Facility: Flush toilet ((with septic tank)
53	TF04	Toilet Facility: Chemical toilet
54	TF05	Toilet Facility: Pit latrine with ventilation (VIP)
55	TF06	Toilet Facility: Pit latrine without ventilation
56	TF07	Toilet Facility: Bucket latrine
57	TF08	Toilet Facility: Other
58	TF09	Toilet Facility: Unspecified
59	SW01	Water Source: Regional/local water scheme (operated by a Water Service Authority or provider)
60	SW02	Toilet Facility: Borehole
61	SW03	Toilet Facility: Spring
62	SW04	Toilet Facility: Rain-water tank
63	SW05	Toilet Facility: Dam / pool / stagnant water
64	SW06	Toilet Facility: River/stream
65	SW07	Toilet Facility: Water vendor
66	SW08	Toilet Facility: Water tanker
67	SW09	Toilet Facility: Other
68	RF01	Refuse Removal: Removed by local authority at least once a week
69	RF02	Refuse Removal: Removed by local authority less often
70	RF03	Refuse Removal: Communal refuse dump

No.	Variable Name	Variable Description
71	RF04	Refuse Removal: Own refuse dump
72	RF05	Refuse Removal: No rubbish disposal
73	RF06	Refuse Removal: Other
74	RF07	Refuse Removal: Unspecified
75	PW01	Piped (tap) water inside the dwelling
76	PW02	Piped (tap) water inside the yard
77	PW03	Piped (tap) water on community stand: distance less than 200m from dwelling
78	PW04	Piped (tap) water to community stand: distance less than 200m and 500m from dwelling
79	PW05	Piped (tap) water to community stand: distance less than 500m and 1000m from dwelling
80	PW06	Piped (tap) water on community stand: distance greater than 1000m (1 km) from dwelling
81	PW07	Piped Water: No access to piped (tap) water
82	PW08	Piped Water: Unspecified
83	PV01	Property Value: Less than R50 000
84	PV02	Property Value: R50 001 - R100 000
85	PV03	Property Value: R100 001 - R200 000
86	PV04	Property Value: R200 001 - R400 000
87	PV05	Property Value: R400 001 - R800 000
88	PV06	Property Value: R800 001 - R1 600 000
89	PV07	Property Value: R1 600 001 - R3 200 001
90	PV08	Property Value: More than R3 200 001
91	PV09	Property Value: Unspecified
92	PV10	Property Value: Not applicable
93	EL01	Energy for Lighting: Electricity
94	EL02	Energy for Lighting: Gas
95	EL03	Energy for Lighting: Paraffin
96	EL04	Energy for Lighting: Candles
97	EL05	Energy for Lighting: Solar
98	EL06	Energy for Lighting: None
99	EL07	Energy for Lighting: Unspecified
100	EH01	Energy for Heating: Electricity
101	EH02	Energy for Heating: Gas
102	EH03	Energy for Heating: Paraffin
103	EH04	Energy for Heating: Wood
104	EH05	Energy for Heating: Coal

No.	Variable Name	Variable Description
105	EH06	Energy for Heating: Animal dung
106	EH07	Energy for Heating: Solar
107	EH08	Energy for Heating: Other
108	EH09	Energy for Heating: None
109	EH10	Energy for Heating: Unspecified
110	EC01	Energy for Cooking: Electricity
111	EC02	Energy for Cooking: Gas
112	EC03	Energy for Cooking: Paraffin
113	EC04	Energy for Cooking: Wood
114	EC05	Energy for Cooking: Coal
115	EC06	Energy for Cooking: Animal dung
116	EC07	Energy for Cooking: Solar
117	EC08	Energy for Cooking: Other
118	EC09	Energy for Cooking: None
119	EC10	Energy for Cooking: Unspecified
120	GT01	Geography Type: Urban
121	GT02	Geography Type: Traditional
122	GT03	Geography Type: Farms
123	Total_Pers	Total Population
124	SALgrp	Small Area Layer Groups
125	SAL2	Pooled SAL Two
126	SAL3	Pooled SAL Three
127	SAL4	Pooled SAL Four
128	SAL5	Pooled SAL Five

Appendix 5: Frame continuous variables

Variable	Label	Sum
Total_Pers	Total Population	47 977 607
Total_HH	Total Households	14 087 151
G01	Gender: Male	23 303 787
G02	Gender: Female	24 673 820
PG01	Race: Black African	37 852 644
PG02	Race: Coloured	4 394 538
PG03	Race: Indian or Asian	1 207 517
PG04	Race: White	4 256 127
PG05	Race: Other	261 532
AG01	Age: 0 - 4	13 962 025
AG02	Age: 15 - 34	18 049 293
AG03	Age: 35 - 64	13 438 395
AG04	Age: 65+	2 527 462
ES01	Employment Status: Employed	12 455 889
ES02	Employment Status: Unemployed	5 318 574
ES03	Employment Status: Discouraged work-seeker	1 688 772
ES04	Employment Status: Other not economically active	12 022 954
ES05	Employment Status: Not applicable	16 489 852
OC01	Occupation: Legislators; senior official and managers	1 062 177
OC02	Professionals	931 318
OC03	Occupation: Technical and associate professionals	1 226 012
OC04	Occupation: Clerks	1 543 034
OC05	Occupation: Service workers; shop and market sales workers	2 064 971
OC06	Occupation: Skilled agricultural and fishery workers	114 738
OC07	Occupation: Craft and related trades workers	1 562 647
OC08	Occupation: Plant and machine operators and assemblers	847 588
OC09	Occupation: Elementary Occupation	2 112 208
OC10	Occupation: Domestic workers	1 226 485
OC11	Occupation: Not applicable	35 271 479
IN01	Industry: Agricultural; hunting; forestry and fishing	650 003
IN02	Industry: Mining and quarrying	371 845
IN03	Industry: Manufacturing	1 251 684
IN04	Industry: Electricity; gas and water supply	104 762
IN05	Industry: Construction	1 029 063

Variable	Label	Sum
IN06	Industry: Wholesale and retail trade	2 193 831
IN07	Industry: Transport; storage and communication	774 100
IN08	Industry: intermediation; insurance; real estate and business service	1 935 337
IN09	Industry: Community; social and personal services	2 908 848
IN10	Industry: Private households	1 467 132
IN11	Industry: Other	3 268
IN12	Industry: Not applicable	35 271 479
TF01	Toilet Facility: None	662 649
TF02	Toilet Facility: Flush toilet (connected to sewerage system)	8 152 929
TF03	Toilet Facility: Flush toilet (with septic tank)	431 884
TF04	Toilet Facility: Chemical toilet	345 487
TF05	Toilet Facility: Pit latrine with ventilation (VIP)	1 169 998
TF06	Toilet Facility: Pit latrine without ventilation	2 659 251
TF07	Toilet Facility: Bucket latrine	303 305
TF08	Toilet Facility: Other	278 478
TF09	Toilet Facility: Unspecified	71 014
SW01	Water Source: Regional/local water scheme (operated by a Water Service Authority or provider)	11 342 797
SW02	Source of Water: Borehole	829 909
SW03	Source of Water: Spring	153 366
SW04	Source of Water: Rain-water tank	119 145
SW05	Source of Water: Dam / pool / stagnant water	209 802
SW06	Source of Water: River/stream	519 313
SW07	Source of Water: Water vendor	171 107
SW08	Source of Water: Water tanker	360 223
SW09	Source of Water: Other	370 139
RF01	Refuse Removal: Removed by local authority at least once a week	8 915 072
RF02	Refuse Removal: Removed by local authority less often	215 542
RF03	Refuse Removal: Communal refuse dump	270 669
RF04	Refuse Removal: Own refuse dump	3 790 076
RF05	Refuse Removal: No rubbish disposal	696 364
RF06	Refuse Removal: Other	119 561
RF07	Refuse Removal: Unspecified	71 014
PW01	Piped (tap) water inside the dwelling	6 509 170
PW02	Piped (tap) water inside the yard	3 936 284

Variable	Label	Sum
PW03	Piped (tap) water on community stand: distance less than 200m from dwelling	1 619 047
PW04	Piped (tap) water to community stand: distance less than 200m and 500m from dwelling	499 989
PW05	Piped (tap) water to community stand: distance less than 500m and 1000m from dwelling	221 223
PW06	Piped (tap) water on community stand: distance greater than 1000m (1 km) from dwelling	121 607
PW07	Piped Water: No access to piped (tap) water	1 099 068
PW08	Piped Water: Unspecified	71 014
PV01	Property Value: Less than R50 000	6 182 706
PV02	Property Value: R50 001 - R100 000	2 043 635
PV03	Property Value: R100 001 - R200 000	1 022 467
PV04	Property Value: R200 001 - R400 000	1 212 437
PV05	Property Value: R400 001 - R800 000	1 357 638
PV06	Property Value: R800 001 - R1 600 000	1 021 676
PV07	Property Value: R1 600 001 - R3 200 001	477 707
PV08	Property Value: More than R3 200 001	193 375
PV09	Property Value: Unspecified	71 014
PV10	Property Value: Not applicable	488 767
EL01	Energy for Lighting: Electricity	11 945 474
EL02	Energy for Lighting: Gas	31 605
EL03	Energy for Lighting: Paraffin	405 628
EL04	Energy for Lighting: Candles	1 535 260
EL05	Energy for Lighting: Solar	45 570
EL06	Energy for Lighting: None	43 216
EL07	Energy for Lighting: Unspecified	71 014
EH01	Energy for Heating: Electricity	8 370 054
EH02	Energy for Heating: Gas	340 564
EH03	Energy for Heating: Paraffin	1 204 918
EH04	Energy for Heating: Wood	1 961 365
EH05	Energy for Heating: Coal	284 687
EH06	Energy for Heating: Animal dung	38 998
EH07	Energy for Heating: Solar	35 461
EH08	Energy for Heating: Other	1 910
EH09	Energy for Heating: None	1 765 059
EH10	Energy for Heating: Unspecified	71 014

Variable	Label	Sum
EC01	Energy for Cooking: Electricity	10 483 042
EC02	Energy for Cooking: Gas	479 238
EC03	Energy for Cooking: Paraffin	1 238 878
EC04	Energy for Cooking: Wood	594 132
EC05	Energy for Cooking: Coal	99 574
EC06	Energy for Cooking: Animal dung	33 724
EC07	Energy for Cooking: Solar	20 470
EC08	Energy for Cooking: Other	26 057
EC09	Energy for Cooking: None	29 949
EC10	Energy for Cooking: Unspecified	71 014
GT01	Geography Type: Urban	9 772 654
GT02	Geography Type: Traditional	571 167
GT03	Geography Type: Farms	743 330

Appendix 6: Distribution of SALs per explicit strata

Explicit Strata	No. of SALs	No. of households in a SAL
111	5 233	1 075 632
113	19	4802
121	2 293	445 230
123	517	114 888
211	2 665	494 643
212	206	30 693
213	38	9 373
221	1 956	350 198
222	3 822	535 806
223	169	33 688
321	1 204	209 456
322	263	46 729
323	177	35 766
411	1 067	201 747
412	98	15 015
413	27	5 359
421	2 629	468 330
422	401	53 972
423	221	37 438
511	4 236	852 260
512	567	105 939
513	26	5 390
521	2 871	550 433
522	4 654	780 280
523	870	175 883
621	2 545	482 265
622	2 751	474 557
623	459	95 789
711	15 591	3 359 922
712	224	38 470
713	141	31 046
721	2 578	516 512
723	144	28 114
821	2 556	486 463
822	3 138	483 519
823	379	83 002
921	1 382	278 018
922	6 767	1 006 540
923	430	83 984

Appendix 7: Distribution of SALs per final strata

Final Strata	No. of SALs	No. of Households
11101	795	140 715
11102	426	113 871
11103	310	45 862
11104	833	164 710
11105	666	145 059
11107	637	155 318
11108	159	47 475
11109	619	98 391
11111	788	164 289
11301	19	4 800
12101	497	99 375
12102	25	7 542
12103	160	42 737
12105	981	171 149
12106	248	37 517
12107	382	86 932
12302	331	65 431
12303	186	49 469
21101	822	148 971
21102	318	44 494
21103	611	95 782
21105	605	129 967
21107	309	75 441
21201	206	30 693
21301	38	9 374
22101	868	126 220
22102	892	172 289
22103	196	51 694
22201	728	90 465
22202	52	13 206
22204	1 425	202 185
22205	510	84 671
22207	835	115 390
22208	272	29 892
22301	104	15 192
22302	65	18 498
32102	93	16 263
32103	227	39 462
32104	126	16 528

Final Strata	No. of SALs	No. of Households
32105	180	28 040
32107	329	55 352
32109	249	53 812
32202	91	21 275
32203	25	2 958
32205	97	13 887
32206	50	8 609
32301	31	11 844
32302	146	23 926
41101	865	149 131
41102	202	52 621
41201	98	15 015
41301	27	5 360
42101	287	36 737
42102	392	57 016
42103	480	97 964
42104	556	90 577
42106	275	70 509
42107	639	115 537
42201	324	40 784
42202	77	13 187
42301	101	11 971
42303	99	17 694
42305	21	7 773
51101	393	102 537
51102	747	152 791
51103	72	25 529
51104	727	130 031
51105	806	154 888
51106	638	144 544
51107	568	95 302
51109	285	46 689
51202	156	28 144
51203	65	14 717
51204	116	20 415
51205	43	5 939
51207	135	26 152
51209	52	10 575
51301	26	5 390
52101	563	92 136

Final Strata	No. of SALs	No. of Households
52102	374	55 606
52104	1 285	249 635
52105	189	49 467
52107	414	95 785
52108	46	7 824
52201	288	72 431
52202	880	130 309
52203	795	104 730
52205	958	167 546
52206	775	143 394
52207	526	109 963
52208	432	51 924
52301	98	23 185
52302	24	9 920
52304	196	44 355
52305	138	24 213
52306	160	30 160
52307	34	9 458
52309	220	34 601
62101	314	48 856
62102	212	55 188
62103	533	102 234
62104	117	16 038
62106	449	99 125
62107	507	91 277
62108	413	69 557
62201	241	64 426
62202	407	52 672
62203	544	106 485
62205	839	145 818
62207	720	105 181
62301	151	27 882
62303	167	41 357
62304	33	12 697
62306	108	13 860
71101	4 468	1 016 973
71102	546	197 939
71103	5 132	988 641
71104	2 279	639 577
71105	3 166	517 080
71201	25	5 716

Final Strata	No. of SALs	No. of Households
71202	42	6 058
71203	79	12 609
71205	78	14 087
71301	141	31 047
72102	708	147 536
72103	644	136 146
72104	76	27 364
72105	623	119 282
72106	527	86 208
72301	144	28 114
82101	454	111 675
82102	64	19 859
82103	505	69 686
82104	781	133 334
82106	752	151 931
82201	1 461	229 320
82202	1 205	157 241
82203	472	96 959
82301	50	20 125
82302	147	33 422
82303	182	29 457
92101	298	43 566
92102	587	109 378
92103	390	90 692
92104	107	34 400
92201	420	80 809
92203	1 361	177 951
92204	926	141 555
92205	41	11 925
92206	971	140 510
92207	800	130 200
92209	824	115 089
92210	853	108 543
92211	571	99 960
92302	180	24 340
92303	105	19 692
92304	101	23 895
92305	44	16 071

Appendix 8: South African - highest level of education estimates

South African - highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade 0	4.0	4.8	3.5
1	grade 1/sub A	3.2	5.6	3.8
2	grade 2/sub B	3.3	5.5	3.8
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	3.6	5.1	3.5
4	grade 4/std 2	3.5	3.8	1.9
5	grade 5/std 3/ABET 2	3.8	4.1	2.4
6	grade 6/std 4	4.1	4.1	2.5
7	grade 7/std 5/ABET 3	4.7	3.5	2.1
8	grade 8/std 6/form 1	6.9	3.2	2.7
9	grade 9/std 7/form 2/ABET 4	5.3	3.3	2.2
10	grade 10/std 8/form 3	7.6	2.9	2.6
11	grade 11/std 9/form 4	6.9	3.2	2.7
12	grade 12/std 10/form 5	16.9	1.7	2.1
13	NTC I/N1/NIC/(V) Level 2	0.1	18.2	1.7
14	NTC II/N2/NIC/(V) Level 3	0.2	16.6	1.8
15	NTC III/N3/NIC/(V) Level 4	0.2	14.7	1.6
16	N4/NTC 4	0.1	15.7	1.3
17	N5/NTC 5	0.1	18.3	1.5
18	N6/NTC 6	0.2	15.8	1.5
19	certificate with less than grade 12/std 10	0.1	31.8	4.4
20	diploma with less than grade 12/std 10	0.1	17.5	1.6
21	certificate with grade 12/std 10	0.9	9.7	2.9
22	diploma with grade 12/std 10	1.1	7.0	2.0
23	Higher Diploma	1.0	10.8	4.3
24	Post Higher Diploma (Masters; Doctoral diploma)	0.2	18.0	2.0
25	bachelors degree	0.9	7.1	1.6
26	bachelors degree and Post graduate diploma	0.3	13.1	1.7
27	honours degree	0.3	13.5	2.1
28	higher degree (masters; doctorate)	0.2	19.6	3.1
29	other	0.1	17.9	1.5
98	no schooling	20.1	2.3	4.8

Appendix 9: South African - highest level of education estimates by gender

Male - Highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade 0	2.1	5.9	2.7
1	grade 1/sub A	1.8	6.4	2.7
2	grade 2/sub B	1.8	6.4	2.7
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	1.6	6.6	2.6
4	grade 4/std 2	1.8	5.5	2.1
5	grade 5/std 3/ABET 2	2.1	6.3	3.1
6	grade 6/std 4	2.1	7.0	3.9
7	grade 7/std 5/ABET 3	2.2	6.0	2.9
8	grade 8/std 6/form 1	3.3	3.7	1.8
9	grade 9/std 7/form 2/ABET 4	2.6	4.5	2.0
10	grade 10/std 8/form 3	3.9	4.2	2.6
11	grade 11/std 9/form 4	3.1	3.6	1.5
12	grade 12/std 10/form 5	8.1	2.2	1.6
13	NTC I/N1/NIC/(V) Level 2	0.1	24.4	1.6
14	NTC II/N2/NIC/(V) Level 3	0.1	22.1	2.1
15	NTC III/N3/NIC/(V) Level 4	0.2	18.1	1.7
16	N4/NTC 4	0.1	17.4	1.2
17	N5/NTC 5	0.1	26.2	1.4
18	N6/NTC 6	0.1	20.6	1.7
19	certificate with less than grade 12/std 10	0.1	31.0	2.1
20	diploma with less than grade 12/std 10	0.1	24.8	1.8
21	certificate with grade 12/std 10	0.4	10.4	1.5
22	diploma with grade 12/std 10	0.6	9.0	1.7
23	Higher Diploma	0.5	9.6	1.6
24	Post Higher Diploma (Masters; Doctoral diploma)	0.1	21.2	1.5
25	bachelors degree	0.5	11.3	2.3
26	bachelors degree and Post graduate diploma	0.2	14.1	1.3
27	honours degree	0.1	19.4	1.9
28	higher degree (masters; doctorate)	0.2	20.6	2.9
29	other	0.1	21.4	1.3
98	no schooling	9.3	2.8	2.8

Female - Highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade	1.9	6.2	2.7
1	grade 1/sub A	1.5	8.8	4.2
2	grade 2/sub B	1.6	9.7	5.4
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	2.0	8.3	5.0
4	grade 4/std 2	1.7	5.0	1.6
5	grade 5/std 3/ABET 2	1.7	5.0	1.6
6	grade 6/std 4	2.0	5.4	2.1
7	grade 7/std 5/ABET 3	2.5	4.4	1.8
8	grade 8/std 6/form 1	3.5	4.9	3.2
9	grade 9/std 7/form 2/ABET 4	2.7	4.7	2.2
10	grade 10/std 8/form 3	3.8	3.7	2.0
11	grade 11/std 9/form 4	3.9	4.2	2.6
12	grade 12/std 10/form 5	8.8	2.6	2.4
13	NTC I/N1/NIC/(V) Level 2	0.1	20.1	1.0
14	NTC II/N2/NIC/(V) Level 3	0.1	20.9	1.0
15	NTC III/N3/NIC/(V) Level 4	0.1	21.6	1.0
16	N4/NTC 4	0.0	31.1	1.1
17	N5/NTC 5	0.1	26.8	1.8
18	N6/NTC 6	0.1	26.1	1.4
19	certificate with less than grade 12/std 10	0.1	52.1	5.9
20	diploma with less than grade 12/std 10	0.1	25.6	1.5
21	certificate with grade 12/std 10	0.5	13.0	2.9
22	diploma with grade 12/std 10	0.6	8.7	1.5
23	Higher Diploma	0.5	16.5	5.4
24	Post Higher Diploma (Masters; Doctoral diploma)	0.1	27.8	2.2
25	bachelors degree	0.4	11.5	1.8
26	bachelors degree and Post graduate diploma	0.1	27.3	2.4
27	honours degree	0.2	17.9	2.1
28	higher degree (masters; doctorate)	0.0	41.6	2.2
29	other	0.1	38.1	2.7
98	no schooling	10.8	3.8	6.2

Appendix 10: South African - highest level of education estimates by race

Black/African - Highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade 0	3.4	5.3	3.5
1	grade 1/sub A	2.9	5.9	3.7
2	grade 2/sub B	2.9	6.0	3.9
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	3.1	5.1	3.0
4	grade 4/std 2	3.1	3.9	1.8
5	grade 5/std 3/ABET 2	3.3	4.4	2.3
6	grade 6/std 4	3.4	5.0	3.2
7	grade 7/std 5/ABET 3	3.9	3.8	2.1
8	grade 8/std 6/form 1	5.4	3.0	1.8
9	grade 9/std 7/form 2/ABET 4	4.4	3.5	2.1
10	grade 10/std 8/form 3	5.9	3.4	2.5
11	grade 11/std 9/form 4	6.0	3.5	2.8
12	grade 12/std 10/form 5	13.0	2.1	2.3
13	NTC I/N1/NIC/(V) Level 2	0.1	20.3	1.7
14	NTC II/N2/NIC/(V) Level 3	0.1	20.4	1.5
15	NTC III/N3/NIC/(V) Level 4	0.1	16.7	1.0
16	N4/NTC 4	0.1	19.5	1.3
17	N5/NTC 5	0.1	17.8	1.0
18	N6/NTC 6	0.1	18.1	1.2
19	certificate with less than grade 12/std 10	0.1	31.4	2.6
20	diploma with less than grade 12/std 10	0.1	17.3	1.2
21	certificate with grade 12/std 10	0.7	11.4	3.1
22	diploma with grade 12/std 10	0.8	7.6	1.6
23	Higher Diploma	0.6	10.3	2.3
24	Post Higher Diploma (Masters; Doctoral diploma)	0.1	23.2	1.5
25	bachelors degree	0.5	9.1	1.4
26	bachelors degree and Post graduate diploma	0.1	16.6	1.3
27	honours degree	0.2	16.1	1.5
28	higher degree (masters; doctorate)	0.1	22.1	1.3
29	other	0.1	22.8	1.7
98	no schooling	17.2	2.6	5.1

Coloured - Highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade 0	0.3	16.2	3.0
1	grade 1/sub A	0.3	15.2	2.2
2	grade 2/sub B	0.3	12.5	1.4
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	0.3	15.6	2.2
4	grade 4/std 2	0.3	14.9	2.5
5	grade 5/std 3/ABET 2	0.4	12.9	2.4
6	grade 6/std 4	0.5	10.6	2.0
7	grade 7/std 5/ABET 3	0.6	12.0	3.1
8	grade 8/std 6/form 1	0.9	12.9	5.5
9	grade 9/std 7/form 2/ABET 4	0.7	10.4	2.8
10	grade 10/std 8/form 3	1.0	7.6	2.2
11	grade 11/std 9/form 4	0.5	12.1	2.9
12	grade 12/std 10/form 5	1.2	6.7	1.9
13	NTC I/N1/NIC/(V) Level 2	0.0	94.2	1.0
14	NTC II/N2/NIC/(V) Level 3	0.0	47.1	1.2
15	NTC III/N3/NIC/(V) Level 4	0.0	41.4	1.5
16	N4/NTC 4	0.0	59.5	1.4
17	N5/NTC 5	0.0	71.2	1.1
18	N6/NTC 6	0.0	58.5	1.0
19	certificate with less than grade 12/std 10	0.0	81.3	2.8
20	diploma with less than grade 12/std 10	0.0	79.5	1.5
21	certificate with grade 12/std 10	0.0	36.1	1.3
22	diploma with grade 12/std 10	0.1	30.9	4.1
23	Higher Diploma	0.0	28.4	1.0
24	Post Higher Diploma (Masters; Doctoral diploma)	0.0	64.1	1.8
25	bachelors degree	0.0	29.6	1.3
26	bachelors degree and Post graduate diploma	0.0	68.7	0.9
27	honours degree	0.0	49.0	1.6
28	higher degree (masters; doctorate)	0.0	88.1	5.0
29	other	0.0	79.0	1.5
98	no schooling	1.7	8.1	4.1

Indian - Highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade 0	0.1	38.1	3.8
1	grade 1/sub A	0.0	49.4	2.4
2	grade 2/sub B	0.0	37.3	2.1
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	0.2	42.8	13.7
4	grade 4/std 2	0.0	52.6	4.5
5	grade 5/std 3/ABET 2	0.1	46.2	3.5
6	grade 6/std 4	0.0	45.1	3.3
7	grade 7/std 5/ABET 3	0.1	31.9	2.9
8	grade 8/std 6/form 1	0.2	25.4	3.5
9	grade 9/std 7/form 2/ABET 4	0.1	32.3	2.5
10	grade 10/std 8/form 3	0.2	25.9	3.6
11	grade 11/std 9/form 4	0.1	23.4	2.5
12	grade 12/std 10/form 5	0.6	7.8	1.4
13	NTC I/N1/NIC/(V) Level 2	0.0	87.2	1.4
14	NTC II/N2/NIC/(V) Level 3	0.0	99.8	0.5
15	NTC III/N3/NIC/(V) Level 4	0.0	62.6	1.1
16	N4/NTC 4	0.0	.	N/A
17	N5/NTC 5	0.0	.	N/A
18	N6/NTC 6	0.0	69.2	0.8
19	certificate with less than grade 12/std 10	0.0	102.0	0.8
20	diploma with less than grade 12/std 10	0.0	.	N/A
21	certificate with grade 12/std 10	0.0	52.6	1.8
22	diploma with grade 12/std 10	0.0	36.9	1.6
23	Higher Diploma	0.1	35.8	2.5
24	Post Higher Diploma (Masters; Doctoral diploma)	0.0	48.7	3.0
25	bachelors degree	0.1	37.6	2.8
26	bachelors degree and Post graduate diploma	0.0	35.0	1.2
27	honours degree	0.0	42.1	1.0
28	higher degree (masters; doctorate)	0.0	50.3	2.0
29	other	0.0	94.3	0.5
98	no schooling	0.4	20.1	6.0

Whites - Highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade 0	0.2	19.3	2.7
1	grade 1/sub A	0.1	27.7	1.6
2	grade 2/sub B	0.1	24.8	2.1
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	0.1	34.2	3.2
4	grade 4/std 2	0.1	36.6	2.8
5	grade 5/std 3/ABET 2	0.1	21.4	1.4
6	grade 6/std 4	0.1	26.7	2.3
7	grade 7/std 5/ABET 3	0.1	23.4	2.8
8	grade 8/std 6/form 1	0.4	25.1	8.7
9	grade 9/std 7/form 2/ABET 4	0.1	19.2	1.7
10	grade 10/std 8/form 3	0.6	13.3	4.1
11	grade 11/std 9/form 4	0.2	15.9	2.1
12	grade 12/std 10/form 5	2.0	10.0	7.5
13	NTC I/N1/NIC/(V) Level 2	0.0	50.1	1.2
14	NTC II/N2/NIC/(V) Level 3	0.1	30.5	2.0
15	NTC III/N3/NIC/(V) Level 4	0.1	26.0	1.7
16	N4/NTC 4	0.0	33.5	1.5
17	N5/NTC 5	0.0	53.3	3.7
18	N6/NTC 6	0.1	35.4	2.7
19	certificate with less than grade 12/std 10	0.0	82.9	8.3
20	diploma with less than grade 12/std 10	0.0	44.2	2.3
21	certificate with grade 12/std 10	0.1	27.6	3.8
22	diploma with grade 12/std 10	0.2	15.0	1.7
23	Higher Diploma	0.3	24.5	7.2
24	Post Higher Diploma (Masters; Doctoral diploma)	0.0	36.9	2.1
25	bachelors degree	0.3	11.5	1.6
26	bachelors degree and Post graduate diploma	0.1	28.8	3.4
27	honours degree	0.1	23.6	2.5
28	higher degree (masters; doctorate)	0.1	22.8	2.1
29	other	0.0	33.5	1.3
98	no schooling	0.8	7.9	1.7

Appendix 11: South African - highest level of education estimates by age

Age Group One (0 - 19) - Highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade 0	3.7	5.0	3.4
1	grade 1/sub A	2.7	6.0	3.6
2	grade 2/sub B	2.4	7.8	5.5
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	2.4	6.1	3.3
4	grade 4/std 2	2.1	5.1	2.0
5	grade 5/std 3/ABET 2	2.2	4.7	1.8
6	grade 6/std 4	2.3	4.4	1.6
7	grade 7/std 5/ABET 3	2.4	5.1	2.3
8	grade 8/std 6/form 1	2.7	4.5	2.0
9	grade 9/std 7/form 2/ABET 4	2.2	5.5	2.4
10	grade 10/std 8/form 3	2.3	7.1	4.3
11	grade 11/std 9/form 4	1.5	6.6	2.5
12	grade 12/std 10/form 5	1.2	6.2	1.7
13	NTC I/N1/NIC/(V) Level 2	0.0	100.0	0.7
14	NTC II/N2/NIC/(V) Level 3	0.0	60.2	1.8
15	NTC III/N3/NIC/(V) Level 4	0.0	68.6	1.3
16	N4/NTC 4	0.0	99.9	1.0
17	N5/NTC 5	0.0	78.8	3.0
18	N6/NTC 6	0.0	.	N/A
19	certificate with less than grade 12/std 10	0.0	99.8	0.9
20	diploma with less than grade 12/std 10	0.0	.	N/A
21	certificate with grade 12/std 10 0.0 57.5 1.1			
22	diploma with grade 12/std 10	0.0	.	N/A
23	Higher Diploma	0.0	70.4	1.2
24	Post Higher Diploma (Masters; Doctoral diploma)	0.0	69.1	3.3
25	bachelors degree	0.0	.	N/A
26	bachelors degree and Post graduate diploma	0.1	24.5	1.2
27	honours degree	0.0	.	N/A
28	higher degree (masters; doctorate)	0.0	88.7	1.5
29	other	0.0	78.0	1.8
98	no schooling	12.6	2.8	4.2

Age Group Two (20 - 49) - Highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade 0	0.1	24.3	3.1
1	grade 1/sub A	0.4	15.5	3.2
2	grade 2/sub B	0.5	11.2	2.1
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	0.6	10.2	2.2
4	grade 4/std 2	0.9	8.6	2.3
5	grade 5/std 3/ABET 2	0.9	7.1	1.7
6	grade 6/std 4	1.1	8.5	2.9
7	grade 7/std 5/ABET 3	1.6	6.0	2.2
8	grade 8/std 6/form 1	2.6	4.6	2.1
9	grade 9/std 7/form 2/ABET 4	2.7	4.4	2.0
10	grade 10/std 8/form 3	4.5	2.5	1.1
11	grade 11/std 9/form 4	5.2	3.2	2.1
12	grade 12/std 10/form 5	14.1	1.9	2.1
13	NTC I/N1/NIC/(V) Level 2	0.1	17.8	1.3
14	NTC II/N2/NIC/(V) Level 3	0.1	17.1	1.4
15	NTC III/N3/NIC/(V) Level 4	0.1	14.8	1.1
16	N4/NTC 4	0.1	16.0	1.2
17	N5/NTC 5	0.1	17.8	1.0
18	N6/NTC 6	0.1	15.2	1.2
19	certificate with less than grade 12/std 10	0.1	27.8	2.0
20	diploma with less than grade 12/std 10	0.1	17.2	1.2
21	certificate with grade 12/std 10	0.8	10.4	3.0
22	diploma with grade 12/std 10	1.0	7.3	1.9
23	Higher Diploma	0.8	8.8	2.2
24	Post Higher Diploma (Masters; Doctoral diploma)	0.1	20.3	1.6
25	bachelors degree	0.7	8.1	1.6
26	bachelors degree and Post graduate diploma	0.2	17.7	1.9
27	honours degree	0.3	14.9	2.1
28	higher degree (masters; doctorate)	0.1	24.5	2.6
29	other	0.1	21.6	1.4
98	no schooling	3.2	6.0	4.4

Age Group Three (50 Plus) - Highest level of education estimates

Code	Description	Estimate (%)	CV (%)	Deff
0	grade 0	0.1	26.2	3.2
1	grade 1/sub A	0.2	12.0	1.0
2	grade 2/sub B	0.4	14.2	3.1
3	grade 3/std 1/ABET 1/Kha Ri Gude; SANLI	0.7	15.7	5.9
4	grade 4/std 2	0.6	8.8	1.8
5	grade 5/std 3/ABET 2	0.7	13.7	4.5
6	grade 6/std 4	0.7	13.8	4.6
7	grade 7/std 5/ABET 3	0.7	8.1	1.8
8	grade 8/std 6/form 1	1.6	10.1	6.0
9	grade 9/std 7/form 2/ABET 4	0.5	13.1	3.0
10	grade 10/std 8/form 3	0.9	9.6	3.0
11	grade 11/std 9/form 4	0.2	15.6	2.1
12	grade 12/std 10/form 5	1.5	11.2	6.9
13	NTC I/N1/NIC/(V) Level 2	0.0	41.1	1.1
14	NTC II/N2/NIC/(V) Level 3	0.0	41.3	1.8
15	NTC III/N3/NIC/(V) Level 4	0.1	27.6	1.5
16	N4/NTC 4	0.0	66.6	2.1
17	N5/NTC 5	0.0	60.4	3.2
18	N6/NTC 6	0.0	41.4	1.8
19	certificate with less than grade 12/std 10	0.1	61.3	6.3
20	diploma with less than grade 12/std 10	0.0	40.2	2.1
21	certificate with grade 12/std 10	0.1	34.8	3.5
22	diploma with grade 12/std 10	0.1	16.4	1.3
23	Higher Diploma	0.2	29.1	7.0
24	Post Higher Diploma (Masters; Doctoral diploma)	0.0	33.4	1.7
25	bachelors degree	0.2	17.7	2.3
26	bachelors degree and Post graduate diploma	0.1	27.1	1.5
27	honours degree	0.1	21.9	1.0
28	higher degree (masters; doctorate)	0.1	23.0	1.9
29	other	0.0	29.3	1.2
98	no schooling	4.3	5.0	4.1

Appendix 12: South African language estimates

South African languages estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	13.0	2.7	3.8
2	English	7.4	4.6	6.1
3	IsiNdebele	1.7	13.8	12.2
4	IsiXhosa	16.4	2.6	4.8
5	IsiZulu	23.8	1.9	3.9
6	Sepedi	9.2	5.2	9.9
7	Sesotho	7.8	4.2	5.4
8	Setswana	8.7	3.8	4.9
9	Sign language	0.5	12.6	2.8
10	SiSwati	2.8	13.5	19.1
11	Tshivenda	2.1	10.9	9.3
12	Xitsonga	5.3	7.4	11.1
13	Other	1.1	10.4	4.3
98	Unspecified	0.4	20.9	5.7

Appendix 13: Language estimates by province

Western Cape language estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	6.4	4.8	5.7
2	English	2.6	11.0	11.6
3	IsiNdebele	0.0	38.5	0.9
4	IsiXhosa	2.7	11.1	12.2
5	IsiZulu	0.0	40.6	1.3
6	Sepedi	0.0	48.6	2.5
7	Sesotho	0.1	15.1	0.7
8	Setswana	0.0	48.0	1.6
9	Sign language	0.1	44.4	7.7
10	SiSwati	0.0	102.2	1.2
11	Tshivenda	0.0	58.8	1.1
12	Xitsonga	0.0	31.2	0.4
13	Other	0.2	19.7	2.7
98	Unspecified	0.2	40.6	9.4

Eastern Cape language estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	1.3	14.9	10.5
2	English	0.7	22.3	13.1
3	IsiNdebele	0.0	30.1	1.1
4	IsiXhosa	11.1	2.8	3.4
5	IsiZulu	0.1	27.6	1.6
6	Sepedi	0.0	32.9	0.5
7	Sesotho	0.2	20.7	3.6
8	Setswana	0.0	45.4	0.9
9	Sign language	0.1	27.5	2.0
10	SiSwati	0.0	70.9	0.9
11	Tshivenda	0.0	68.1	4.8
12	Xitsonga	0.0	42.8	1.1
13	Other	0.1	21.3	1.2
98	Unspecified	0.1	58.2	5.8

Northern Cape language estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	1.5	11.3	7.1
2	English	0.0	30.0	0.7
3	IsiNdebele	0.0	42.9	1.1
4	IsiXhosa	0.1	30.4	2.2
5	IsiZulu	0.0	45.4	1.2
6	Sepedi	0.0	56.7	0.8
7	Sesotho	0.1	31.0	1.7
8	Setswana	0.6	21.9	10.6
9	Sign language	0.0	43.9	0.4
10	SiSwati	0.0	226.9	0.2
11	Tshivenda	0.0	0.6	0.0
12	Xitsonga	0.0	.	N/A
13	Other	0.0	30.0	0.7
98	Unspecified	0.0	70.7	2.7

Free State language estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	0.6	28.6	18.0
2	English	0.1	18.4	1.2
3	IsiNdebele	0.0	26.0	0.6
4	IsiXhosa	0.4	17.8	5.0
5	IsiZulu	0.2	34.1	8.3
6	Sepedi	0.0	55.5	0.6
7	Sesotho	3.9	5.6	4.7
8	Setswana	0.2	19.6	2.2
9	Sign language	0.1	16.2	0.7
10	SiSwati	0.0	72.3	0.8
11	Tshivenda	0.0	70.4	0.5
12	Xitsonga	0.0	80.4	4.0
13	Other	0.0	40.4	1.1
98	Unspecified	0.0	118.8	0.4

Kwazulu Natal language estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	0.3	13.3	1.9
2	English	2.4	10.1	9.0
3	IsiNdebele	0.2	19.1	2.8
4	IsiXhosa	0.6	19.9	9.0
5	IsiZulu	17.6	1.6	2.1
6	Sepedi	0.0	34.7	1.7
7	Sesotho	0.1	24.7	1.6
8	Setswana	0.1	24.6	2.7
9	Sign language	0.1	19.6	0.8
10	SiSwati	0.0	66.9	5.0
11	Tshivenda	0.0	70.6	1.2
12	Xitsonga	0.0	98.9	2.2
13	Other	0.2	28.7	4.4
98	Unspecified	0.0	36.7	1.4

North West language estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	0.3	25.6	6.4
2	English	0.1	28.4	3.8
3	IsiNdebele	0.1	22.7	2.2
4	IsiXhosa	0.4	15.5	3.3
5	IsiZulu	0.2	14.7	1.3
6	Sepedi	0.1	23.5	1.6
7	Sesotho	0.4	16.2	4.2
8	Setswana	5.4	2.5	1.3
9	Sign language	0.0	22.2	0.6
10	SiSwati	0.0	27.5	0.4
11	Tshivenda	0.0	32.4	0.7
12	Xitsonga	0.2	28.8	5.0
13	Other	0.1	25.2	1.9
98	Unspecified	0.0	38.6	1.5

Gauteng language estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	2.0	8.0	4.6
2	English	1.1	6.7	1.8
3	IsiNdebele	0.6	28.9	17.3
4	IsiXhosa	1.0	19.0	13.7
5	IsiZulu	3.7	8.4	9.7
6	Sepedi	2.4	5.8	3.0
7	Sesotho	2.5	8.9	7.5
8	Setswana	1.9	11.6	9.4
9	Sign language	0.1	24.7	2.2
10	SiSwati	0.3	10.1	1.0
11	Tshivenda	0.3	6.3	0.4
12	Xitsonga	1.5	7.7	3.4
13	Other	0.3	21.1	4.9
98	Unspecified	0.0	68.6	3.3

Mpumalanga language estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	0.4	20.4	5.9
2	English	0.3	30.5	10.6
3	IsiNdebele	0.6	27.1	15.8
4	IsiXhosa	0.1	29.6	2.9
5	IsiZulu	1.9	9.2	5.9
6	Sepedi	0.8	17.3	8.2
7	Sesotho	0.3	13.7	2.2
8	Setswana	0.4	26.7	9.6
9	Sign language	0.0	38.6	0.8
10	SiSwati	2.4	15.5	21.8
11	Tshivenda	0.0	27.5	1.2
12	Xitsonga	1.2	20.1	17.1
13	Other	0.1	39.0	5.8
98	Unspecified	0.0	34.6	1.0

Limpopo language estimates

Code	Description	Estimate (%)	CV (%)	Deff
1	Afrikaans	0.3	30.6	9.6
2	English	0.2	17.5	1.8
3	IsiNdebele	0.2	21.3	2.5
4	IsiXhosa	0.0	32.8	1.0
5	IsiZulu	0.2	22.8	3.6
6	Sepedi	5.8	7.4	12.4
7	Sesotho	0.2	25.9	4.1
8	Setswana	0.1	56.5	12.8
9	Sign language	0.0	49.5	2.0
10	SiSwati	0.0	34.9	1.4
11	Tshivenda	1.8	13.0	11.1
12	Xitsonga	2.4	12.2	13.3
13	Other	0.2	40.3	8.9
98	Unspecified	0.0	54.6	3.9

SAS codes

SAS Code 1: Frame Construction

```
Libname Msc "C:\";

/*Sorting all persons attributes*/
Proc Sort Data=Msc.population_group; By SAL_Code;Run;
/*Sorting all households attributes*/
Proc Sort Data=Msc.Households; By SAL_Code;Run;

/*Combine all the attributes that are required to create a
   sampling frame*/
Data Msc.MSC_Frame001;
merge   Msc.age_group Msc.gender Msc.population_group
   Msc.Employment
        ................................................;
By sal_code;
        Total_Pers = G01 + G02;
        Label Total_HH = "Total Households";
        Label Total_Pers = "Total Population";
        Label G01 = "Gender: Male";
        Label G02 = "Gender: Female";
        Label PG01 ="Race: Black African";
    ..................................;
Run;

/*Add Geography base on 2011 boundaries*/
Data geo11;
Set Msc.geohierarchy_ea2011_2016 (Keep =sal_code pr_code_2011
   DC_MN_C_2011
        MN_CODE_2011 EA_TYPE_C EA_GTYPE_C);
                Label sal_code ="Small Area Layer";
                Label pr_code_2011 ="Province";
                Label DC_MN_C_2011 ="District Municipality";
                Label MN_CODE_2011 ="Local Municipality";
                Label EA_TYPE_C ="EA Type";
                Label EA_GTYPE_C ="Geography Type";
Run;

Proc Sort Data= geo11 nodupkey;
        By sal_code;
Run;
Proc Sort Data= Msc.msc_frame001 Out=Msc.msc_frame_sorted;
        By sal_code;
Run;

Data Msc.msc_frame002;
merge geo11(in=a) Msc.msc_frame_sorted(in=b);
        By sal_code;
        if b;
Run;
Proc Sort Data= Msc.msc_frame002(Drop=Total_Pers Total_HH)
   Out=SAL_MDG;
        By sal_code;
Run;

/**Define SAL Exclutions from the frame and the remaining in
   100s**/
Data Msc.msc_frame003;
Length metro_id 3. metro $1.;
Set Msc.msc_frame002;
/*Determine and flag exclutions*/
if ea_type_c in (’5’,’7’, ’9’,’10’) then excl=’1’;
else excl=’0’;
Label excl = "Excluded Small Areas";
/*Group SAL my measure of size*/
        if total_hh <50 then SALgrp=0;
        else if total_hh < 100 then SALgrp=1;
        else if total_hh < 150 then SALgrp=2;
        else if total_hh < 200 then SALgrp=3;
        else if total_hh < 250 then SALgrp=4;
        else if total_hh < 300 then SALgrp=5;
        else if total_hh < 350 then SALgrp=6;
        else if total_hh < 400 then SALgrp=7;
        else if total_hh < 500 then SALgrp=8;
        else if total_hh < 1000 then SALgrp=9;
        else SALgrp=10;
Label salgrp = "Small Area Layer Groups";
/*Define Metro and Non-Metro*/
if DC_MN_C_2011 in (’199’, ’260’, ’299’, ’499’, ’599’, ’797’,
   ’798’, ’799’) then Metro_ID = DC_MN_C_2011;
```

```
if DC_MN_C_2011 in (’199’, ’260’, ’299’, ’499’, ’599’, ’797’,
   ’798’, ’799’) then Metro= 1;
else Metro_ID = 0 ;
Run;

Data metro non_metro;
Length str $4. metro_id 3.;
Set Msc.msc_frame003;
/*Separte metros and non-metros*/
Str= compress(pr_code_2011||DC_MN_C_2011);
        if metro = 1 then output metro;
        else output non_metro;
Run;

Proc Sort Data=non_metro nodupkey Out=non_metro_dc;
        By str;
Proc Sort Data=non_metro_dc;
        By pr_code_2011;
Run;

/*Sequence non-metro DCs*/
Data non_metro2;
Length metro_id 3. metro $1.;
Set non_metro_dc(Keep=pr_code_2011 DC_MN_C_2011 str);
        By pr_code_2011;
                retain seq 0;
                if first.pr_code_2011 then seq=0;
                seq=seq+1;
                DCNO= put(seq,z2.);
                Metro_ID= compress(pr_code_2011||DCNO);
                Metro = 2;
Drop seq dcno str;
Run;

Proc Sort Data=non_metro Out=non_metro1(Drop=metro_id metro);
        By DC_MN_C_2011;
Proc Sort Data=non_metro2;
        By DC_MN_C_2011;
Run;
Data non_metro3;
merge non_metro2(in=a) non_metro1(in=b) ;
        By DC_MN_C_2011;
        if b;
Run;
Proc Sort Data=non_metro3;
        By sal_code;
Run;
```

```
Proc Sort Data=metro;
        By sal_code;
Run;

Data Msc.msc_frame004;
Set metro non_metro3;
By sal_code;
        Label metro_id = "Metro and Non-Metro individual
           identifier";
        Label metro = "Metro and Non-Metro grouping";
Drop str;
Run;

Proc Sort Data=Msc.msc_frame004;
By sal_code;
Run;

Data Msc.msc_frame005;
Length ex_strata $3.;
Set Msc.msc_frame004;
        ex_strata = compress(pr_code_2011||Metro || EA_GTYPE_C);
        Label ex_strata = "Explicit Strata";
Run;
/*Define Explicit strata without taking into account other
   characteritics*/

Proc Means Data=Msc.msc_frame005 Sum Missing;
Var Total_Pers Total_HH
        G01 G02
        PG01 PG02 PG03 PG04 PG05
        AG01 AG02 AG03 AG04
        ES01 ES02 ES03 ES04 ES05
        ........................;
Run;

Proc Means Data= Msc.msc_frame005 Sum n;
        Class  ex_strata excl;
        Var total_hh;
Run;

Proc freq Data= Msc.msc_frame005;
        Tables excl salgrp salgrp*excl pr_code_2011*ex_strata
           ex_strata ex_strata*excl/nocol norow nopercent;
        Weight total_hh;
Run;

/* Creation of file for sampling*/
```

```
Data Msc.msc_frame006;
Set Msc.msc_frame005;
        If excl= "0";
Drop excl;
Run;

Data Msc.msc_frame_Collapse Msc.msc_frame007 Msc.msc_frame_Split;
Set Msc.msc_frame006;
        If SALgrp in ("0", "1") then output
           Msc.msc_frame_Collapse;
        If SALgrp in ("2", "3", "4", "5", "6", "7") then output
           Msc.msc_frame007;
        If SALgrp in ("8", "9", "10") then output
           Msc.msc_frame_Split;
Run;

/*Creation of split files*/
Data Msc.msc_frame_Split01;
Format SAL_Code_S $8.;
Set Msc.msc_frame_split;
If SALgrp = "8";
        Total_HH        =       Total_HH*0.5;
        Total_Pers      =       Total_Pers*0.5;
        G01     =       G01*0.5;
        G02     =       G02*0.5;
        PG01    =       PG01*0.5;
        PG02    =       PG02*0.5;
    .....................;

Split = 1;
SAL_Code_S = compress(SAL_Code||Split);
Run;

Data Msc.msc_frame_Split02;
Format SAL_Code_S $8.;
Set Msc.msc_frame_Split;
        if SALgrp = "8";
        Total_HH = Total_HH*0.5;
        Total_Pers = Total_Pers*0.5;
        G01     =       G01*0.5;
        G02     =       G02*0.5;
        PG01    =       PG01*0.5;
        PG02    =       PG02*0.5;
        PG03    =       PG03*0.5;
    .....................;
Split = 2;
SAL_Code_S = compress(SAL_Code||Split);
```

```
Run;

Data Msc.msc_frame_Split03;
        Set Msc.msc_frame_Split01 Msc.msc_frame_Split02;
Run;

Proc Sort Data = Msc.msc_frame_Split03;
        By SAL_Code_S;
Run;

Data Msc.msc_frame_3_Split01;
Format SAL_Code_S $8.;
Set Msc.msc_frame_Split;
        If SALgrp in ("9", "10");
        Total_HH        =       Total_HH/3;
        Total_Pers      =       Total_Pers/3;
        G01     =       G01/3;
        G02     =       G02/3;
        PG01    =       PG01/3;
        PG02    =       PG02/3;
        PG03    =       PG03/3;
    .....................;
Split = 1;
SAL_Code_S = compress(SAL_Code||Split);
Run;
Data Msc.msc_frame_3_Split02;
Format SAL_Code_S $8.;
Set Msc.msc_frame_Split;
        if SALgrp in ("9", "10");
        Total_HH        =       Total_HH/3;
        Total_Pers      =       Total_Pers/3;
        G01     =       G01/3;
        G02     =       G02/3;
        PG01    =       PG01/3;
        PG02    =       PG02/3;
        PG03    =       PG03/3;
    .....................;
Split = 2;
SAL_Code_S = compress(SAL_Code||Split);
Run;

Data Msc.msc_frame_3_Split03;
Format SAL_Code_S $8.;
Set Msc.msc_frame_Split;
        if SALgrp in ("9", "10");
        Total_HH        =       Total_HH/3;
        Total_Pers      =       Total_Pers/3;
        G01     =       G01/3;
        G02     =       G02/3;
        PG01    =       PG01/3;
        PG02    =       PG02/3;
        PG03    =       PG03/3;
        .....................;
Split = 3;
SAL_Code_S = compress(SAL_Code||Split);
Run;

Data Msc.msc_frame_Split04;
        Set Msc.msc_frame_3_Split01 Msc.msc_frame_3_Split02
           Msc.msc_frame_3_Split03;
Run;

Proc Sort Data = Msc.msc_frame_Split04;
        By SAL_Code_S;
Run;

Data Msc.msc_frame_Split_Final;
        Set Msc.msc_frame_Split03 Msc.msc_frame_Split04;
Run;

Proc Sort Data = Msc.msc_frame_Split_Final;
        By SAL_Code_S;
Run;

/* Combining the final split and the base sampling frame*/

/*Prepare the sampling frame*/

Data msc_frame007;
Format SAL_Code_S $8. SAL_C $7. ;
Set Msc.msc_frame007;
        SAL_C=SAL_Code;
        SAL_Code_S = compress(SAL_C || "0");
Drop SAL_C;
Run;
Proc Sort Data = msc_frame007;
        By SAL_Code_S;
Run;
Data Msc.msc_frame008;
        Set msc_frame007 Msc.msc_frame_Split_Final;
By SAL_Code_S;
Run;
Proc Sort Data = Msc.msc_frame008;
        By SAL_Code_S;
```

```
Run;

/**********Collapsing of Small Area Layers*******************/
Proc Sort Data = Msc.msc_SAL_pooling Out=msc_SAL_pooling;
        By SP_Code EA_GTYPE_C SAL_Code Total_HH;
Run;

Data msc_SAL_pool01;
Length key1 $10.;
Set msc_SAL_pooling;
        key1= compress(SP_Code||EA_GTYPE_C);
Run;

Proc Sort Data = msc_SAL_pool01;
        By key1;
Run;

Proc summary Data=msc_SAL_pool01 Sum Nway;
Class SP_Code EA_GTYPE_C;
        Var Total_HH;
        output Out=msc_SAL_poolx Sum=Tot;
Run;

/*Get the links for collapsed small areas*/
Proc Transpose Data=msc_SAL_pool01 Out=Msc.SAL_Pool(Keep= Key1
   COL1-COL6);
By key1;
        Var SAL_Code;
Run;

Proc Transpose Data=msc_SAL_pool01 Out=Msc.SAL_Pool_Tot(Keep=
   Key1 COL1-COL6);
By key1;
        Var Total_HH;
Run;
Data Msc.SAL_Pool_Tot1;
Set Msc.SAL_Pool_Tot;
        Total_HH = Sum(COL1,COL2, COL3, COL4, COL5, COL6);
Drop COL1-COL6;
Run;

Proc Sort Data = Msc.SAL_Pool_Tot1;
        By key1;
Run;
Proc Sort Data = Msc.SAL_Pool;
        By key1;
Run;
```

```
Data msc_final_pool_tot;
merge Msc.SAL_Pool_Tot1(in=a) Msc.SAL_Pool(in=b);
        By key1;
        if a and b;
Run;

Data SAL_base;
Format SAL_Code_S $8.;
Format Key1 SAL_Code_S SAL_Code SAL2 SAL3 SAL4 SAL5 SAL6;
Set msc_final_pool_tot(Keep= Key1 Total_HH COL1-COL6);
        SAL_Code = COL1;
        SAL2= Col2;
        SAL3= Col3;
        SAL4= Col4;
        SAL5= Col5;
        SAL6= Col6;
        SAL_Code_S = compress(SAL_Code||"0");
Drop COL1 COL2 COL3 COL4 COL5 COL6;
Run;

Proc Sort Data = SAL_base;By key1;Run;

Proc Transpose Data=SAL_base Out= SAL_Pool_Trans;
        By key1;
        Var SAL_Code SAL2 SAL3 SAL4 SAL5 SAL6;
Run;

Data SAL_Pool_Trans1;
Set SAL_Pool_Trans;
        if col1 ne .;
        SAL_Code = col1 ;
Drop _NAME_ col1;
Run;

/* Bring all the Data into pooled EAs*/
Proc Sort Data = SAL_Pool_Trans1;
        By SAL_Code;
Proc Sort Data = Msc.msc_frame006;
        By SAL_Code;
Run;

Data Msc.msc_frame006x;
merge SAL_Pool_Trans1(in=a) Msc.msc_frame006(in=b);
        By SAL_Code;
        if a and b;
Run;
```

```
Proc Means Data=Msc.msc_frame006x Nway noprint Sum;
Class key1 ex_strata metro_id metro MN_CODE_2011
DC_MN_C_2011 PR_CODE_2011 EA_GTYPE_C EA_TYPE_C;
Var Total_Pers Total_HH
        G01 G02
        PG01 PG02 PG03 PG04 PG05
        AG01 AG02 AG03 AG04
    ........................;
output Out = msc_frame006x(Drop=_freq_ _type_) Sum= ;
Run;

Proc Sort Data = SAL_base(Drop=Total_HH SAL_Code) Out=sal_base1;
        By Key1;
Proc Sort Data = msc_frame006x;
        By Key1;
Run;

Data Msc.SAL_frame_base;
merge SAL_base1(in=a) msc_frame006x(in=b);
        By key1;
        if a;
        if Total_HH>100 then output Msc.SAL_frame_base;
Drop key1;
Run;

Proc Sort Data = Msc.SAL_frame_base nodupkey;
        By SAL_Code_S;
Run;

/* Adding Pooled SALs to the final SAL Frame*/
Data Msc.msc_frame009;
Set Msc.msc_frame008 Msc.SAL_frame_base;
By SAL_Code_S;
        Label SAL_Code_S ="Small Area Layer";
        Label pr_code_2011 ="Province";
        Label DC_MN_C_2011 ="District Municipality";
        Label MN_CODE_2011 ="Local Municipality";
        Label EA_TYPE_C ="EA Type";
        Label EA_GTYPE_C ="Geography Type";
    ..................................;
Drop SAL_Code split;
Run;

Proc Sort Data = Msc.msc_frame009;
        By SAL_Code_S;
Run;
```

```
/*Final Msc Sampling Frame*/
Proc Means Data=Msc.msc_frame009 Sum Missing;
*Class pr_code_2011;
Var Total_Pers Total_HH
        G01 G02
        PG01 PG02 PG03 PG04 PG05
        AG01 AG02 AG03 AG04
        ES01 ES02 ES03 ES04 ES05
    ............................;
Run;

Proc Means Data= Msc.msc_frame009 Sum;
Class ex_strata;
        Var total_hh;
Run;
```

SAS Code 2: Sampling Preparation

```
Libname MSC "C:\";
Libname MSSAMP "C:\";

/*Sample allocation based on people with ages between 14 and 65*/
Proc Summary  Data=msc.msc_frame009  Missing NWAY;
Class pr_code_2011;
        Var AG02 AG03 Total_HH;
        Output Out=PR_Tot(Drop=_Type_) Sum= N=;
Format pr_code_2011 prov_.;
Run;

/*Allocation to provinces using square root allocation*/
Data Prov_Total;
Set PR_Tot;
Pers_15_64 = AG02+AG03;
SAL_Count = _Freq_;
        RT_PR_PERS= SQRT (PERS_15_64);
        SUM_RT_PR_PERS = 16003.81395;
        PR_PORTION=RT_PR_PERS/SUM_RT_PR_PERS;
        TOTAL_SAMP=15000;/*Nationals household sample size*/
        PR_HH_SAMP=INT(TOTAL_SAMP * PR_PORTION + 0.5);
        PR_HH_Tot= Total_HH;
Keep pr_code_2011 Pers_15_64 PR_PORTION PR_HH_SAMP PR_HH_Tot ;
Run;

/*Calculate provincial ISR assuming 100% response rate in 2011
   without taking growth into account*/
Data msc.Prov_ISR;
Set Prov_Total;
Prov_ISR = round(PR_HH_Tot/PR_HH_SAMP,.1);
Keep pr_code_2011 Prov_ISR PR_HH_SAMP ;
Run;

/*Sample allocation to strata based on people with ages between
   14 and 64*/
Proc Summary  Data=msc.msc_frame009  Missing NWAY;
Class pr_code_2011 ex_strata;
        Var AG02 AG03 Total_HH Total_Pers;
        Output Out=Strat_Tot(Drop=_Type_) Sum= N=;
Run;

Proc Sort Data= Strat_Tot; By pr_code_2011;
Proc Sort Data= msc.Prov_ISR; By pr_code_2011; Run;
```

```
Data Stratification;
Merge Strat_Tot msc.Prov_ISR;
By pr_code_2011;
        Pers_15_64 = AG02+AG03;
        RT_Strat_PERS= SQRT (PERS_15_64);
Run;

/*Allocation to stratum using square root allocation*/
Proc Means Data=Stratification noprint;
        Var RT_Strat_PERS;
        Output Out=RT Sum=;
Run;

Data Strat_Alloc;
Set Stratification;
        SAL_Count = _Freq_;
        SUM_RT_PERS = 28296.12;
        Strat_Prop=RT_Strat_PERS/SUM_RT_PERS;
        TOTAL_Strat_SAMP=PR_HH_SAMP;/*Provincial household
           sample size*/
        Strat_HH_SAMP=int(TOTAL_HH * Strat_Prop + 0.5);
        SAL_SAMP=INT(TOTAL_Strat_SAMP * Strat_Prop + 0.5);
        Strat_HH_Tot= Total_HH;
Keep ex_strata SAL_SAMP Strat_HH_Tot TOTAL_Strat_SAMP
   Strat_HH_SAMP SAL_Count;
Run;

Proc Sort Data= Strat_Alloc; By ex_strata;
Proc Sort Data= msc.msc_frame009; By ex_strata; Run;

Data Stratification2;
Merge Strat_Alloc(in=a) msc.msc_frame009(in=b);
        By ex_strata;
        If a & b;
        Pers_15_64 = AG02+AG03;
Run;

/*Allocation using square root allocation*/
Data Strat_Total;
Set Stratification2;
/*Calculate proportions of Census characteristics*/
        p_G01 =G01/Total_Pers;
        p_G02 =G02/Total_Pers;
        p_PG01 =PG01/Total_Pers;
        p_PG02 =PG02/Total_Pers; .........;
Keep
SAL_SAMP  TOTAL_Strat_SAMP Strat_HH_SAMP .....;
```

```
Run;

/*Calculate stratum ISR assuming 100% response rate in 2011
   without taking growth into account*/
Data mssamp.Prop_ISR;
Set Strat_Total;
ISR = round(Strat_HH_Tot/Strat_HH_SAMP,1);
Keep Ex_Strata Strat_HH_Tot Strat_HH_SAMP ISR Pers_15_64
   Total_pers     .........;
Run;

/*Clustering is done for each Primary Stratum at a time*/
Data data1;
Set mssamp.Prop_ISR;
If ex_Strata= "923";
Run;

Proc standard Data=data1 Mean=0 Std=1 Out=stan2;
        Var Pers_15_64 p_G01 p_G02 p_PG01 p_PG02 p_PG03 p_PG04
           p_ES01 p_ES02
        p_IN01 p_IN02   p_IN03 p_IN04 p_IN05 p_IN06     p_IN07
           p_TF01 p_TF02 p_TF03
        p_TF04  p_TF05  p_TF06  p_TF07  p_PW01  p_PW02  p_PW03
           p_PW04  p_PW05
        p_PW06  p_PW07;
Run;

/*Proc Fastclust require a licence to be executed in base SAS*/
Proc fastclus Data=data1  Delete=20 Maxclusters=20 Maxiter=100
Converge= 0.0001 Out=stratum_cluster Least=2 Short;
Var Pers_15_64 p_G01 p_G02 p_PG01 p_PG02 p_PG03 p_PG04  p_ES01
   p_ES02
        p_IN01 p_IN02   p_IN03 p_IN04 p_IN05 p_IN06     p_IN07
           p_TF01 p_TF02
        p_TF03  p_TF04  p_TF05  p_TF06  p_TF07  p_PW01  p_PW02
           p_PW03
        p_PW04  p_PW05  p_PW06  p_PW07;
                ID Strat_HH_SAMP;
                ID Total_Pers;
                ID Strat_HH_Tot;
                ID Strat_HH_Tot;
                ID ex_strata;
                ID SAL_Code_S;
                ID ISR;
Run;

Data stratum_cluster;
```

```
Set stratum_cluster;
        new_cluster=cluster ;
Keep sal_code_s ex_strata cluster new_cluster;
Run;

Proc append Base=mssamp.stratum_cluster Data=stratum_cluster;
Run;

Proc Sort Data=strata; By ex_strata cluster SAL_Code_S; Run;

Proc freq Data=strata;
        tables ex_strata*cluster/List Missing Nopercent;
Run;

Proc Summary Data=strata Missing;
By ex_strata ;
        Class cluster;
        ID ISR;
        Var Total_Pers Pers_15_64 Total_HH;
Output Out=out1
Sum=
n=n1;
Run;

/* Sample calculated for the new clusters*/
Data out1;
Set out1;
        sample=int(Total_HH/ISR + 0.5);
Drop n1 _type_;
Run;

Data mssamp.p_stratum_SAL;
Set strata;
        new_cluster=cluster;
        sample=int(Total_HH/Strat_ISR  + 0.5);
Keep ex_strata cluster new_cluster SAL_Code_S;
Run;

/*Collapsing of small cluster*/
Data mssamp.p_stratum_cluster;
Set out1;
        *new_cluster=cluster;
        sample=int(Total_HH/ISR + 0.5);
        If sample < 40 Then new_cluster= cluster+1;
    else  new_cluster= cluster;
Keep ex_strata cluster new_cluster Total_HH Strat_ISR sample;
Run;
```

```
Proc Means Data=mssamp.p_stratum_cluster Noprint Nway;
        Class ex_strata new_cluster;
        Var Total_HH sample;
        Output Out=mssamp.p_stratum_cluster2(Drop=_type_ _freq_)
           Sum =;
Run;

Proc append base=mssamp.Final_stratum_cluster
   Data=mssamp.p_stratum_cluster2;
Run;

Data mssamp.Final_stratum_cluster;
        Set mssamp.Final_stratum_cluster;
Run;

Data mssamp.Final_stratum_cluster002;
Format cluster $2.;
Set mssamp.Final_stratum_cluster001;
        cluster = put(new_cluster,z2.);
        Stratum= compress(ex_strata||cluster);
Drop cluster;
Run;

Proc Sort Data= mssamp.Final_stratum_cluster002;
        By ex_strata new_cluster;
Proc Sort Data=mssamp.stratum_cluster ;
        By ex_strata new_cluster;
Run;

Data mssamp.All_stratum_cluster;
Merge mssamp.Final_stratum_cluster002(in=a)
   mssamp.stratum_cluster(in=b);
        By ex_strata new_cluster;
        If b;
Run;

Data mssamp.All_stratum_cluster2;
Set mssamp.All_stratum_cluster;
        If stratum = " " Then new_Cluster = cluster + 1;
        original_cluster = cluster;
Drop cluster;
Run;

Data mssamp.All_stratum_cluster2;
Format cluster $2.;
Set mssamp.All_stratum_cluster2;
```

```
        cluster = put(new_cluster,z2.);
        Stratum= compress(ex_strata||cluster);
Keep ex_strata SAL_Code_S Stratum;
Run;

/* Creation of the final sampling frame*/
Proc Sort Data=mssamp.All_stratum_cluster2; By SAL_code_s;
Proc Sort Data=msc.msc_frame009; By SAL_code_s;
Run;

Data mssamp.MSC_Frame010;
Format Pr_Code Stratum SAL_Code SALgrp Prov_Sample MOS;
Merge mssamp.All_stratum_cluster2(in=a) msc.msc_frame009(in=b);
        By SAL_code_s;
        If a;
        MOS=round(Total_HH,1);
        SAL_Code = Substr(SAL_Code_S,1,7);
        pr_code = compress(substr(stratum,1,1));
                If Pr_code = ’1’ Then Prov_Sample= 1834 ;
                If Pr_code = ’2’ Then Prov_Sample= 1691;
                If Pr_code = ’3’ Then Prov_Sample= 776;
                If Pr_code = ’4’ Then Prov_Sample= 1201;
                If Pr_code = ’5’ Then Prov_Sample= 2305;
                If Pr_code = ’6’ Then Prov_Sample= 1377;
                If Pr_code = ’7’ Then Prov_Sample= 2712;
                If Pr_code = ’8’ Then Prov_Sample= 1471;
                If Pr_code = ’9’ Then Prov_Sample= 1633;
Keep Pr_Code Stratum SAL_Code SAL_Code_S SALgrp Prov_Sample MOS;
Run;

Proc Means Data=mssamp.msc_frame010 nway noprint;
class stratum;
        Var MOS;
        Output out = STR_Sum(drop =_type_)
        Sum=STR_MOS ;
Run;

Data STR_Sum1;
Set STR_Sum;
TOT_SAL= _freq_;
pr_code = compress(substr(stratum,1,1));
        If Pr_code = ’1’ Then Prov_Sample= 1834 ;
        If Pr_code = ’2’ Then Prov_Sample= 1691;
        If Pr_code = ’3’ Then Prov_Sample= 776;
        If Pr_code = ’4’ Then Prov_Sample= 1201;
        If Pr_code = ’5’ Then Prov_Sample= 2305;
        If Pr_code = ’6’ Then Prov_Sample= 1377;
        If Pr_code = ’7’ Then Prov_Sample= 2712;
        If Pr_code = ’8’ Then Prov_Sample= 1471;
        If Pr_code = ’9’ Then Prov_Sample= 1633;
drop _freq_;
Run;

/*calculation only*/
Proc Means Data=STR_Sum1 noprint sum;
Var RT_STR_MOS;
Output out=STR sum=Str_Sum_MOS;
Run;

Proc Sort Data=mssamp.msc_frame010; By stratum;
Proc Sort Data=STR_Sum1; By stratum; Run;

Data mssamp.MSC_Frame011;
Format Pr_Code Stratum SAL_Code SALgrp Prov_Sample TOT_SAL MOS
   STR_MOS;
Merge STR_Sum1(in=a) mssamp.msc_frame010(in=b);
        By Stratum;
        If  a & b;
Keep Pr_Code Stratum SAL_Code SAL_Code_S SALgrp Prov_Sample
   TOT_SAL MOS STR_MOS ; /*SAL_Samp*/
Run;

/*Allocation to stratum using square root allocation*/
Proc Means Data=MSC_Frame000 sum;
Var RT_Str_MOS;
Run;

Data MSC_Frame000;
Set mssamp.MSC_Frame011;
        RT_Str_MOS =SQRT(MOS);
        RT_SUM_MOS = 1017857.32;
        Prop=RT_Str_MOS/RT_SUM_MOS;
        n_sal=int(Prov_Sample * Prop + 0.5);
Keep Pr_Code Stratum SAL_Code SALgrp Prov_Sample TOT_SAL MOS
   Prop n_sal RT_Str_MOS;
Run;

/*Extract the complete small area group*/
Data AreaGrp;
Set  msc.MSC_FRAME005;
        SAL_Code1 = put(SAL_Code,$8.);
Keep SAL_Code1 SALgrp;
Run;
Data AreaGrp1;
```

```
Format SAL_Code $8.;
Set  AreaGrp;
        SAL_Code = compress(SAL_Code1);
        SAL_Group = SALgrp;
Keep SAL_Code SAL_Group;
Run;

Proc Sort Data=mssamp.msc_frame011 out=msc_frame011; By SAL_Code;
Proc Sort Data=AreaGrp1; By SAL_Code; Run;

Data mssamp.MSC_Frame012;
Format Pr_Code Stratum SAL_Code SAL_Code_S SAL_Group Prov_Sample
   ISR TOT_SAL MOS STR_MOS;
Merge AreaGrp1(in=a) msc_frame011(in=b);
By SAL_Code;
If a and b;
        If Pr_code = ’1’ Then ISR= 895 ;
        If Pr_code = ’2’ Then ISR= 860;
        If Pr_code = ’3’ Then ISR= 376;
        If Pr_code = ’4’ Then ISR= 651;
        If Pr_code = ’5’ Then ISR= 1072;
        If Pr_code = ’6’ Then ISR= 764;
        If Pr_code = ’7’ Then ISR= 1465;
        If Pr_code = ’8’ Then ISR= 716;
        If Pr_code = ’9’ Then ISR= 838;
Keep Pr_Code Stratum SAL_Code SAL_Code_S SAL_Group Prov_Sample
   ISR TOT_SAL MOS STR_MOS ;
Run;
```

SAS Code 3: Sample Allocation and Selection of Small Area Layers

```
Libname MSC "C:\";
Libname MSSAMP "C:\";
Libname MSSL "C:\";

Proc Means Data=mssamp.msc_frame012 Sum;
        Class stratum;
        Var mos;
Run;

/*Stratum ISR is equal to provincial ISR*/
/*Randomisation of SALs*/
Data one;
Set mssamp.msc_frame012;
        rand_num=uniform(1);
Run;

Proc Sort Data=one;
        By pr_code stratum rand_num;
Run;
Data two;
Set one; By pr_code stratum;
        If first.stratum Then seq_num=1;
        Else seq_num=seq_num + 1;
        retain seq_num;
Keep pr_code stratum SAL_Code_S seq_num mos;
Run;

/*Obtain dwelling count and Number of SALS in each Stratum*/
Proc Summary Nway Data=one Missing;
Class pr_code stratum Prov_Sample;
Var mos;
Output Out=out1
       Sum=sum_mos
       n=n_pop_sal;
Run;

Proc Means Data=MSC_Frame000 Sum;
        Var RT_Str_Sal;
Run;

Data MSC_Frame000;
Set out1;
```

```
        RT_Str_sal =SQRT(n_pop_sal);
        RT_SUM_sal = 2992.21;
        Prop=RT_Str_sal/RT_SUM_sal;
        n_sal=int(prov_sample * Prop + 0.5);
Run;

Proc Sort Data=MSC_Frame000;
        By Pr_Code stratum;
Proc Sort Data= one;
        By Pr_Code stratum;
Run;

Data one_;
merge MSC_Frame000(in=a) one(in=b);
        By pr_code stratum;
Run;

Data sampl2;
Set one_;
        str_isr=isr;
        n_sample_sal= n_sal;
        sum_isr=n_sample_sal*str_isr;
        r1=uniform(1);
        str_rand_start=int(str_isr*r1) + 1;
Keep pr_code stratum str_isr n_sample_sal sum_isr str_rand_start;
Run;

Proc Sort Data=two;
        By pr_code stratum seq_num;
Run;

Data frame1 err1 err2;
merge merg1(in=in1) two(in=in2);
By pr_code stratum;
        If in1 & in2 Then Output frame1;
        Else If in1 Then Output err1;
        Else If in2 Then Output err2;
Run;

Data frame2 err3(Keep=pr_code stratum sal_code_s mos pi);
Set frame1;
        pi=n_sample_sal*(mos/sum_mos);
        If pi ge 1.0 Then Output err3;
        Else do;
                r_isr=pi*str_isr;
                integer=int(r_isr);
                fraction=r_isr-integer;
        If integer=0 Then fraction=0.9999;
Output frame2;
End;
Run;

Proc Summary Nway Data=frame2 Missing;
Class pr_code stratum;
Var integer;
Output Out=out2
       Sum=sum_integer
       n=n_integer;
Run;

Data out2;
Set out2;
Keep pr_code stratum sum_integer n_integer;
Run;

Proc Sort Data=frame2;
        By pr_code stratum fraction;
Proc Sort Data=out2;
        By pr_code stratum;
Run;

Data frame3 err1 err2;
merge frame2(in=in1) out2(in=in2);
By pr_code stratum;
        If in1 & in2 Then Output frame3;
        Else If in1 Then Output err1;
        Else If in2 Then Output err2;
Drop _Type_ _Freq_;
Run;

Data frame4 error;
Set frame3; By pr_code stratum;
If first.stratum Then do;
n_zero=n_pop_sal - (sum_isr - sum_integer);
End;
        If n_zero gt 0 Then do;
        round=0;
        n_zero=n_zero - 1;
        End;
Else round=1;
sal_isr = integer + round;
Output frame4;
        If n_integer ne n_pop_sal Then Output error;
        retain n_zero;
```

```
Drop n_zero round fraction;
Run;

Proc Sort Data=frame4;
        By pr_code stratum seq_num;
Run;

Data frame5 error;
Set frame4; By pr_code stratum;
If first.stratum Then cum_isr=sal_isr;
        Else cum_isr = cum_isr + sal_isr;
        Output frame5;
                If last.stratum Then do;
                If cum_isr ne sum_isr Then Output error;
        End;
retain cum_isr;
Run;

Data MSSL.new_psu_frame_sorted;
Set frame5;
Keep pr_code stratum str_isr str_rand_start
SAL_Code_S seq_num mos sum_mos sal_isr cum_isr
n_pop_sal n_sample_sal ;
Run;

/**********SAMPLE SELECTION***************/
/* Read in SAL Frame File. */
Data data1;
        Set mssl.new_psu_frame_sorted;
Run;

Proc Sort Data=data1;
        By pr_code stratum seq_num;
Run;

Data data2;
Set data1;
By pr_code stratum;
length selection $1. sal_start $3. ;
selection= " " ;
If first.stratum Then do;
sel_isr=str_rand_start;
        If sel_isr le cum_isr Then do;
        selection="*";
        start1=sel_isr + sal_isr - cum_isr ;
        sel_isr=sel_isr + str_isr;
        End;
```

```
End;
Else do;
        If sel_isr le cum_isr Then do;
        selection="*";
        start1=sel_isr + sal_isr - cum_isr ;
        sel_isr=sel_isr + str_isr;
        End;
End;
If start1=. Then sal_start=" ";
Else sal_start=put(start1, Z3.);
retain sel_isr;
Run;

Data mssl.new_sal_frame_sampled;
Set data2;
        Keep pr_code stratum seq_num SAL_Code_S sal_isr cum_isr
           sal_start mos selection;
Run;

Data mssl.MSC_Sal_Sample;
Set mssl.new_sal_frame_sampled;
        If selection="*";
Drop selection;
Run;
```

SAS Code 4: Households sample selection

```
Libname MSC "C:\";
Libname MSSAMP "C:\";
Libname MSSL "C:\";

Data Frame;
Set mssl.DWELLING_FRAME;/*SUPERCROSS census 2011
households count were used with provincial growth applied*/
        du_count = du_count_0001;
Drop du_count_0001;
Run;

Data Sample;
Set mssl.MSC_SAL_SAMPLE;
        SAL_Code = compress(substr(SAL_Code_S,1,7));
Run;

Proc Sort Data=Sample;
        By SAL_Code;
Proc Sort Data=Frame;
        By SAL_Code;
Run;

Data Fr_samp;
Merge Sample(in=a)Frame(in=b);
        By SAL_Code;
        If a;
Run;

Data Fr_samp2;
Set Fr_samp;
        If du_count < mos Then du_count = round((MOS*1.17),1);
Run;

Proc Sort Data= Fr_samp2;
        By pr_code stratum SAL_Code_S;
Run;

Data sal_dwelling_frame(Keep= pr_code stratum Sal_Code_S Sal_isr
   Sal_start
     du_num du_count);
Set Fr_samp2;
        By pr_code stratum SAL_Code_S;
        If first.SAL_Code_S Then do du_num=1 to du_count ;
        Retain du_num;
        Output;
        End;
Run;

Data dwelling_frame;
Format pr_code stratum Sal_Code_S Sal_isr Sal_start1 Sal_flag
     n_dus du_num count SAL_FLAG;
Set sal_dwelling_frame;
        count=1;
        SAL_FLAG = substr(SAL_Code_S,8,1)+1;
        SAL_start1=SAL_start+0;
        n_dus=du_count;
Keep pr_code stratum Sal_Code_S Sal_isr Sal_start1 Sal_flag
     n_dus du_num count;
Run;

Proc Sort Data=dwelling_frame;
        By pr_code stratum Sal_Code_S du_num;
Run;

/* obtain household totals By SAL */
Proc Summary Nway Data=dwelling_frame Missing;
class pr_code stratum SAL_Code_S;
        Var count;
        ID SAL_isr SAL_start1 SAL_flag n_dus;/**/
        Output out=out1
        sum=sum_count;
Run;

Data sample_flag; *error;
Set dwelling_frame; /*out1; */
If SAL_flag GE 2 Then do;
        SAL_start1= (SAL_flag - 1)*SAL_isr + SAL_start1;
        SAL_isr= SAL_flag*SAL_isr;
end;
/*dwelling unit (DU) contains one or more households. In the
   program
  DU and household where used interchangably*/
        r_start=mod(du_num,SAL_isr);
        If r_start = 0 Then r_start = SAL_isr;
        If r_start = SAL_start1 Then sample_flag =1;
        else sample_flag = 0;
Keep pr_code stratum SAL_Code_S SAL_flag SAL_isr SAL_start1
   sample_flag
     du_num n_dus;
Run;
```

```
Data mssl.SAL_Dwelling_Frame2;
        Set sample_flag;
Run;

Data dwelling_sample;
Set sample_flag;
        hhno=du_num;
        If sample_flag = 1;
Run;

Data mssl.MSC_dwelling_sample;
Set dwelling_sample;
Run;

Proc freq Data= mssl.MSC_dwelling_sample;
        table SAL_Code_S;
Run;
```

SAS Code 5: Survey data simulation

```
Libname MSSAMP "C\:";
Libname MSSL "C\:";
Libname MSSV "C\:";
/*Read person Data*/
Data mssv.SAL_Household;
        Set mssv.head_of_household_001 mssv.head_of_household_002
        mssv.head_of_household_003 mssv.head_of_household_004
        mssv.head_of_household_005;
Run;

Proc freq Data = mssv.SAL_Household;
        Tables ’’’Age␣of␣household␣head’’n’n /List Missing;
Run;

Data SAL_HH_Test;
Set mssv.SAL_Household;
        rand1=uniform(1);
Run;
Proc Sort Data=SAL_HH_Test;
        By geography rand1;
Run;

Data Test1;
Set SAL_HH_Test;
By geography;
        If first.geography Then hh_no = 1;
        Else hh_no=hh_no + 1;
        Retain hh_no;
        Output;
Run;

Proc Sort Data=Test1;
        By geography hh_no;
Run;

Data mssv.SAL_HH;
Set Test1;
By geography hh_no;
        If first.hh_no Then do persno = 1 to ’’’household␣
           size’’n’n;
        Retain persno;
        Output;
End;
Run;
```

```
/*******Person file*********/
Data SAL_Pers;
Set mssv.SAL_Person;
        If Person_weighted < 2 Then Flag = 1;
        Else If Person_weighted < 3  Then Flag = 2;
        Else If Person_weighted < 4  Then Flag = 3;
        Else If Person_weighted < 5  Then Flag = 4;
        Else If Person_weighted < 6  Then Flag = 5;
        Else If Person_weighted < 7  Then Flag = 6;
        Else If Person_weighted < 8  Then Flag = 7;
        Else If Person_weighted < 9  Then Flag = 8;
        Else Flag = 9;
        If Person_weighted;
Run;

Proc Sort Data= SAL_Pers ;
        By Age_in_completed_years Educational_institution
           Enumeration_area_type Gender
        Geo_type Geography Highest_educational_level Language
           Population_group Present_school_attendance
        Person_weighted;
Run;

Data SAL_Pers1;
Set SAL_Pers;
By Age_in_completed_years Educational_institution
   Enumeration_area_type
Gender Geo_type Geography Highest_educational_level Language
Population_group Present_school_attendancePerson_weighted;
        If first.Person_weighted Then do Pers = 1 to Flag;
        Retain pers;
        Output;
        End;
Run;

Data SAL_Pers2;
Set SAL_Pers1;
        new_weight = Person_weighted/Flag;
Run;

Proc means Data= SAL_Pers2 Sum;
        Var new_weight;
Run;

Data SAL_HHL;
Set mssv.SAL_HH;
```

```
        Age_in_completed_years = ’’’Age␣of␣household␣head’’n’n;
        Enumeration_area_type = ’’’Enumeration␣area␣type’’n’n;
        Gender = ’’’Gender␣of␣head␣of␣household’’n’n;
        Geo_type = ’’’Geo␣type’’n’n;
        Population_group = ’’’Population␣group␣of␣hhead’’n’n;
Run;

Proc Sort Data= SAL_HHL;
        By geography Geo_type Enumeration_area_type Gender
           Population_group;
Run;

Proc Sort Data= SAL_Pers2;
        By geography Geo_type Enumeration_area_type Gender
           Population_group;
Run;

Data SAL_Hpers;
merge SAL_HHL(in=a) SAL_Pers2(in=b);
        By geography Geo_type Enumeration_area_type Gender
           Population_group;
Run;

Proc Sort Data= SAL_Hpers nodupkey Out=SAL_Hpers2;
        By geography hh_no persno;
Run;

Data SAL_Hpers3;
Set SAL_Hpers2;
        If geography ne . or hh_no ne . or persno ne .;
Run;

Proc Sort Data= SAL_Hpers3;
        By geography hh_no descending age_in_completed_years;
Run;

Data mssv.SAL_Hpers4;
Set SAL_Hpers3;
By geography hh_no descending age_in_completed_years;
        If first.hh_no Then Persno2 = 1;
        Else Persno2=Persno2+1;
        Retain Persno2;
        Output;
Run;

Proc freq Data= mssv.SAL_Hpers4;
Table Age_in_completed_years;
```

```
Run;

/*Survey Data linked to sample*/
Data sal_hh_person2;
format sal_code $7.;
Set mssv.SAL_Hpers4;
        sal_code = geography;
Run;

Data Msc_dwelling_sample2;
format sal_code $7.;
Set mssl.Msc_dwelling_sample;
        sal_code = compress(substr(sal_code_s,1,7));
        hh_no = du_num;
        dwelling_id= du_num;
Run;

Proc Sort Data = Msc_dwelling_sample2;
        By sal_code hh_no;
Proc Sort Data = sal_hh_person2;
        By sal_code hh_no;
Run;

Data mssv.sample_survey_data_final;
format gender1 $1.;
merge Msc_dwelling_sample2(in=a) sal_hh_person2(in=b);
        By sal_code hh_no;
        If a;
        gender1=compress(gender);
        pr_code1=compress(pr_code);
Run;

Proc Sort Data=mssv.sample_survey_data_final Out=check1;
        By sal_code sal_code_s hh_no descending
           Age_in_completed_years;
Run;

Proc freq Data=Check1;
        Table Sal_Code_S/nocol norow nopercent;
Run;

Data check3;
Set check1;
        gender=compress(gender1);
        pr_code=pr_code1;
        If personno > 2 Then Age = int(Age_in_completed_years*1);
        Else Age = Age_in_completed_years;
```

```
Run;

Proc Sort Data=check3 nodupkey
   Out=mssv.sample_survey_data_person(keep=sal_code
        pr_code
        stratum
        Sal_Code_S
        hh_no
        Enumeration_area_type
        Geo_type
        persno
        Age_in_completed_years
        Age
        Educational_institution
        Gender
        Highest_educational_level
        Language
        Population_group
        Present_school_attendance);
By sal_code hh_no persno;
Run;

Proc Sort Data=mssv.sample_survey_data_final nodupkey
Out=mssv.sample_survey_data_household(drop=
        Age_in_completed_years Educational_institution Gender
        Highest_educational_level Language Population_group
        gender1 rand1 pr_code1 Present_school_attendance persno
        person_weighted flag pers new_weight);
By sal_code hh_no;
Run;

/*Sort*/
/*Allocate responses and non responses*/

Proc Sort Data = mssv.sample_survey_data_household nodupkey
Out=sample_survey_data_household;
        By sal_code_s;
Run;

Data sample_survey_data_household1;
Set mssv.sample_survey_data_household;
        If Pr_code = ’1’ Then Resp= 1891;
        If Pr_code = ’2’ Then Resp= 1679;
        If Pr_code = ’3’ Then Resp= 818;
        If Pr_code = ’4’ Then Resp= 1321;
        If Pr_code = ’5’ Then Resp= 2409;
        If Pr_code = ’6’ Then Resp= 1457;
        If Pr_code = '7' Then Resp= 2516;
        If Pr_code = '8' Then Resp= 1632;
        If Pr_code = '9' Then Resp= 1773;
        rand_resp =uniform(1);
Run;

Proc Sort Data = sample_survey_data_household1;
        By pr_code rand_resp;
Run;

Data sample_survey_data_household2;
Set sample_survey_data_household1;
By pr_code;
        If first.pr_code and resp ne . Then seq=1;
        Else seq=seq+1;
        Retain seq;
        Output;
Run;

Data mssv.sample_survey_data_household3;
Set sample_survey_data_household2;
        If seq le resp and Result = . Then Result = 1;
        Else Result = 2;
Run;

Proc freq Data = mssv.sample_survey_data_household3;
        Tables pr_code*result stratum*result/nocol norow
           nopercent;
Run;

Proc freq Data = mssv.sample_survey_data_household3;
        Tables pr_code/nocol norow nopercent;
Run;
```

SAS Code 6: Sampling parameters

```
Libname MSSAMP "C:\";
Libname MSSL "C:\";
Libname MSSV "C:\";
Libname MSSW "C:\";

Title1 'Sample␣DUs␣from␣sampled␣SALs';
Options Nofmterr;

Data Parms;
Set mssl.Msc_dwelling_sample;
        n_starts_a=SAL_isr - SAL_start1;
        If (SAL_isr - SAL_start1 + 1) ge 4 Then n_starts_o =0;
        Else n_starts_o = SAL_start1 - SAL_isr + 3;
        If n_starts_a < 3 Then n_starts_a = 3;
        n_starts_u= 1;
        Dwelling_id=du_num;
Keep pr_code stratum SAL_Code_S SAL_isr SAL_start1 SAL_flag
      n_starts_a n_starts_u n_starts_o sample_flag Dwelling_id;
Run;

Data dwelling_frame;
Set mssl.Sal_dwelling_frame2;
        TotalPDs=n_dus;
        Dwelling_id=du_num;
        If TotalPDs > 0;
Keep pr_code stratum  SAL_Code_S TotalPDs Dwelling_id;
Run;

Proc Sort Data=dwelling_frame;
        By pr_code stratum SAL_Code_S Dwelling_id;
Proc Sort Data=parms;
        By pr_code stratum SAL_Code_S Dwelling_id;
Run;

Data merg1 merg2 merg3;
Merge dwelling_frame (in=a) parms (in=b);
        By pr_code stratum SAL_Code_S dwelling_id;
        If a & b Then Output merg1;
        Else If a Then Output merg2;
        Else If b Then Output merg3;
Run;

Data growth;
Set merg1;
```

```
        If SAL_flag GE 2 Then yield=totalPDS/(SAL_flag*SAL_isr);
        Else yield= totalPDS/SAL_isr;
        factor= (yield/10);
        If factor LT 1.3 Then factor = 1;
        SAL_isr1=int(factor*SAL_isr);
Keep pr_code stratum SAL_Code_S SAL_isr1 factor dwelling_id;
Run;

Proc Sort Data=parms Nodupkey Out=parms1;
        By pr_code stratum SAL_Code_S Dwelling_id;
Proc Sort Data=growth Nodupkey Out=growth1;
        By pr_code stratum SAL_Code_S Dwelling_id;
Run;

Data merg_parms;
Merge growth1(in=a) parms1(in=b);
        By pr_code stratum SAL_Code_S Dwelling_id;
        If b;
Run;

Data merg_parms2;
Set merg_parms;
        If factor ne . Then SAL_adj=SAL_isr1/SAL_isr;
        Else SAL_adj = 1;
        VS_rand=uniform(1);
Run;

Proc Sort Data=merg_parms2 Nodupkey Out=merg_parmsx;
        By Sal_Code_S;
Proc Freq Data = merg_parmsx;
        Table stratum/Noprint Nopercent Out=SAL1;
Run;

Data SAL1;
Set SAL1;
        Half=round((count/2),1);
Drop percent count;
Run;

Proc Sort Data= merg_parmsx;
        By stratum;
Proc Sort Data= SAL1;
        By stratum;
Run;

Data merg_parms3;
Merge merg_parmsx(in=a) SAL1(in=b);
```

```
        By stratum;
        If a and b;
Run;

Proc Sort Data= merg_parms3;
        By stratum VS_rand;
Run;

Data merg_parms4;
Set merg_parms3;
By stratum VS_rand;
        If first.stratum Then seq1=1;
        Else seq1=seq1+1;
        retain seq1;
        If seq1 le half Then VarUnit = 1;
        Else VarUnit = 2;
Run;

Proc Freq Data = merg_parms4;
        Table VarUnit;
Proc Sort Data= merg_parms2;
        By SAL_Code_S;
Run;

Proc Sort Data= merg_parms4;
        By SAL_Code_S;
Run;

Data merg_parms5;
Merge merg_parms2(in=a) merg_parms4(in=b);
        By SAL_Code_S;
        If a and b;
Run;

Proc Freq Data = merg_parms5;
        Table VarUnit;
Run;

Data mssl.Msc_sampling_parms;
Set merg_parms5;
Keep pr_code stratum SAL_Code_S SAL_isr VarUnit SAL_start1
   SAL_flag
n_starts_a n_starts_u n_starts_o SAL_adj Sample_flag ;
Run;
```

SAS Code 7: Sample weights adjustment

```
Libname MSSAMP "C:\";
Libname MSSL "C:\";
Libname MSSV "C:\";
Libname MSSW "C:\";

/************************************************************/
/*    Create Base Weights and apply non-response adjustment. */
/************************************************************/

Data Du_Sample;
Set mssl.Msc_dwelling_sample;
        hh_no = du_num;
Keep Pr_code stratum SAL_Code_S hh_no;
Run;

Proc Sort Data=Du_Sample;
        By SAL_Code_S HH_no;
Run;

Data hhld_Sample;
format uqno $12.;
Set mssv.sample_survey_data_household3;
        hh_no1=put(hh_no,z4.);
        uqno=compress(SAL_Code_S||hh_no1);
        If hh_no = . Then delete;
        If uqno = " " Then delete;
Run;

Proc Sort Data = hhld_Sample;
        By uqno;
Run;

Data hhld_sample;
Set hhld_sample; By uqno;
        If first.uqno;
Run;

Proc Sort Data = hhld_Sample;
        By SAL_Code_S HH_No;
Run;

Data Merg1  (Keep = pr_code stratum SAL_Code_S hh_no uqno result)
         Error1 (Keep = SAL_Code_S hh_no uqno);
Merge Du_Sample (in=a) Hhld_Sample(in=b);/*Totnohh*/
```

```
        By SAL_Code_S HH_No;
        If a Then Output Merg1;
        Else If b Then Output Error1;
Run;

Data merg1;
Set merg1;
length Response_Code $1.;
        If result = 1 Then Response_Code = "1";
        Else Response_Code = "2";
Keep pr_code stratum SAL_Code_S hh_no uqno result response_code
   uqno;
Run;

Proc freq Data=merg1; /*hhld_Sample*/
        Tables Response_Code*result/list Missing;
Run;

Data merg2;
Set merg1; If response_code = "1" or response_code = "2" ;
        If Pr_code = "1" Then B_wgt = 895;
        If Pr_code = "2" Then B_wgt = 860;
        If Pr_code = "3" Then B_wgt = 376;
        If Pr_code = "4" Then B_wgt = 651;
        If Pr_code = "5" Then B_wgt = 1072;
        If Pr_code = "6" Then B_wgt = 764;
        If Pr_code = "7" Then B_wgt = 1465;
        If Pr_code = "8" Then B_wgt = 716;
        If Pr_code = "9" Then B_wgt = 838;
Run;

/*Get sampling parameters*/
Data sampling_parms;
Set mssl.MSC_sampling_parms;
Keep pr_code stratum SAL_Code_S varunit SAL_Adj;
Run;

Proc Sort Data = Merg2;
        By pr_code stratum SAL_Code_S;
Proc Sort Data = sampling_parms;
        By pr_code stratum SAL_Code_S;
Run;

Data Base_Wgts No_Sample;
Merge Merg2 (in=a) sampling_parms (in=b);
By pr_code stratum SAL_Code_S;
If a Then do;
```

```
                If SAL_Adj = . Then SAL_Adj = 1;
                Base_Wgt=SAL_Adj*B_Wgt;
                Output Base_Wgts;
End;
Else If b Then Output No_Sample;
Run;

Proc freq Data=base_wgts;
        tables response_code/list Missing;
Run;

/* Check that all base weights are > 0. */
Data  check;
Set base_wgts;
        If base_wgt = . or base_wgt = 0;
Run;

/* Apply Non-Response Adjustment. */
Proc summary Nway Data=base_wgts Missing;
Class Pr_Code Stratum SAL_Code_S Response_Code;
ID varunit;
        Var Base_Wgt;
        Output Out=out1
        Sum=sum_Wgts
N=n1;
Run;

Data resp (rename=(n1=n_resp)) non_resp (rename=(n1=n_nresp))
   error1;
Set out1;
        If response_code = "1" Then Output resp;
        Else If response_code = "2" Then Output non_resp;
        Else Output error1;
Keep pr_code stratum SAL_Code_S varunit n1;
Run;

Proc Sort Data=resp;
        By Pr_Code stratum SAL_Code_S;
Proc Sort Data=non_resp;
        By Pr_Code stratum SAL_Code_S;
Run;

Data nresp_adj1;
merge resp(in=a) non_resp(in=b);
By Pr_Code stratum SAL_Code_S;
        If a or b;
        If n_nresp = . Then n_nresp=0;
        If n_resp = . Then n_resp=0;
        If n_resp > 0 Then adj_factor= (n_resp + n_nresp)/n_resp;
        Else adj_factor = 99;
Keep pr_code stratum SAL_Code_S varunit adj_factor;
Run;

Data nresp_adj2;
Set nresp_adj1;
        If adj_factor GE 1.5;
Run;

/*Compute non-response adjustment at the var_unit level for high
   non-response SALs.*/
Proc Sort Data = nresp_adj2;
        By pr_code stratum varunit ;
Run;

Data nresp_adj2;
Set nresp_adj2;
        By pr_code stratum varunit;
        If first.varunit;
Keep pr_code stratum varunit;
Run;

Proc Sort Data = base_wgts;
        By pr_code stratum varunit;
Run;

Data Adj2;
Merge base_wgts (in=a) nresp_adj2 (in=b);
        By pr_code stratum varunit;
If a & b ;
Run;

/* Compute Non-Response Adjustment at Var_Unit Level. */
Proc summary Nway Data=Adj2 Missing;
Class Pr_Code Stratum varunit Response_Code;
        Var Base_Wgt;
        Output Out=out2
        Sum=sum_Wgts
N=n2;
Run;

Data resp (rename=(sum_wgts=n_resp)) non_resp
   (rename=(sum_wgts=n_nresp)) error2;
Set out2;
        If response_code = "1" Then Output resp;
        Else If response_code = "2" Then Output non_resp;
        Else Output error2;
Keep pr_code stratum varunit sum_wgts;
Run;

Data nresp_adj2;
merge resp(in=a) non_resp(in=b);
By pr_code stratum varunit;
        If a or b;
        If n_nresp = . Then n_nresp=0;
        If n_resp = . Then n_resp=0;
        If n_resp > 0 Then adj_factor2= (n_resp +
           n_nresp)/n_resp;
        Else adj_factor2 = 0;
Keep pr_code stratum varunit adj_factor2;
Run;

Proc Sort Data=nresp_adj1;
        By pr_code stratum varunit;
Proc Sort Data=nresp_adj2;
        By pr_code stratum varunit;
Run;

Data nresp_adj;
merge nresp_adj1(in=in1) nresp_adj2(in=in2);
By pr_code stratum varunit;
        If in1 ;
        If adj_factor2 ne . Then adj_factor=adj_factor2;
        If adj_factor2 = 0 Then delete;
Keep pr_code stratum SAL_Code_S adj_factor;
Run;

Data base_wgts;
Set base_wgts;
        If response_code = "1";
Run;

Proc Sort Data=base_wgts;
        By pr_code stratum SAL_Code_S;
Proc Sort Data=nresp_adj;
        By pr_code stratum SAL_Code_S;
Run;

Data non_unique;
Set nresp_adj;
By pr_code stratum SAL_Code_S;
        If not (first.SAL_Code_S & last.SAL_Code_S);
```

```
Run;

Data adjusted_wgts;
merge base_wgts (in=a) nresp_adj(in=b);
By pr_code stratum SAL_Code_S;
        If a;
        If adj_factor = . Then adj_factor=1;
base_wgt_adj=adj_factor*base_wgt;
Run;

Data check2;
Set adjusted_wgts;
        If base_wgt_adj = . or base_wgt_adj = 0 ;
Run;

Proc Sort Data=adjusted_wgts;
        By pr_code stratum varunit;
Run;

Data varunits;
Set adjusted_wgts;
By pr_code stratum varunit;
        If first.varunit;
Keep pr_code stratum varunit;
Run;

Data single_varunit;
Set varunits;
By pr_code stratum;
        If first.stratum & last.stratum;
Keep Pr_code stratum varunit;
Run;

Proc Sort Data = adjusted_wgts;
        By pr_code stratum varunit;
Run;

Data adjusted_wgts2;
merge adjusted_wgts (in=a) single_varunit (in=b);
        By pr_code stratum varunit;
        If a & not b;
Keep pr_code stratum varunit SAL_Code_S hh_no uqno base_wgt_adj;
Run;

Proc Sort Data = adjusted_wgts2;
        By pr_code stratum varunit;
Run;
```

```
Data check_varunit2;
Set adjusted_wgts2;
By pr_code stratum varunit;
        If first.varunit;
Keep pr_code stratum varunit;
Run;

Data check_stratum2;
Set check_varunit2;
By pr_code stratum;
        If first.stratum & last.stratum;
Keep pr_code stratum;
Run;

Data mssw.MSC_base_wgts (rename=(base_wgt_adj = full_wgt));
        Set adjusted_wgts2;
Run;

Data SALs (rename=(base_wgt_adj = full_wgt));
Set adjusted_wgts2;
Keep pr_code stratum varunit SAL_Code_S base_wgt_adj;
Run;

Proc Sort Data=SALs;
By pr_code stratum varunit SAL_Code_S;
Run;

Data SAL_1;
Set SALs; By pr_code stratum varunit SAL_Code_S;
        If first.SAL_Code_S;
Run;

Data mssw.MSC_base_wgts_SAL_1;
        Set SAL_1;
Run;

Proc Sort Data=mssw.MSC_base_wgts Nodupkey Out=test0;
        By pr_code stratum varunit SAL_Code_S;
Run;
Proc Sort Data=mssw.MSC_base_wgts_SAL_1 Nodupkey Out=test1;
        By pr_code stratum varunit SAL_Code_S;
Run;
```

SAS Code 8: Calibration

```
Libname TOOLKIT "C:\" access=readonly ;
options sasmstore=toolkit mstored ;
Libname MSSAMP "C:\";
Libname MSSL "C:\";
Libname MSSV "C:\";
Libname MSSW "C:\";
Libname MSSMX "C:\";
Libname MSWSV "C:\";

/*Create control Totals and get the latest version of statmx*/
/*Read person level weights*/
Data mssv.sample_survey_data_person2;
Format uqno $12. Person_ID $14. Pr_Code1 $1. Age_grp1 $1.
   Age_Grp2 $1. Race $1. Gender $1.;
Set mssv.sample_survey_data_person;
        hh_no1 = put(hh_no, z4.);
        persno1 = put(persno, z2.);
        Person_id = compress(SAL_Code_S||HH_No1||persno1);
        uqno = compress(SAL_Code_S||HH_No1);
        Race=Population_group;
        pr_code1= compress(pr_code);
                If Age le 4 Then Age_Grp1 = '1';
                Else If Age le 9 Then Age_Grp1 = '1';
                Else If Age le 14 Then Age_Grp1 = '2';
                Else If Age le 19 Then Age_Grp1 = '2';
                Else If Age le 24 Then Age_Grp1 = '3';
                Else If Age le 29 Then Age_Grp1 = '3';
                Else If Age le 34 Then Age_Grp1 = '4';
                Else If Age le 39 Then Age_Grp1 = '4';
                Else If Age le 44 Then Age_Grp1 = '5';
                Else If Age le 49 Then Age_Grp1 = '5';
                Else If Age le 54 Then Age_Grp1 = '6';
                Else If Age le 59 Then Age_Grp1 = '6';
                Else If Age le 64 Then Age_Grp1 = '6';
                Else Age_Grp1 = '6';
                If Age le 19 Then Age_Grp2 = '1';
                        Else If Age le 34 Then Age_Grp2 = '2';
                        Else If Age le 49 Then Age_Grp2 = '2';
                        Else Age_Grp2 = '3';
Run;

Data mssv.sample_survey_data_person3 missing_demo;
length pr_code $1.;
Set mssv.sample_survey_data_person2;
```

```
        pr_code = Pr_Code1;
        If gender = ’.’ or race = ’.’ or race = ’5’ or Age_Grp1
           = ’.’ Then Output missing_demo;
Run;

Proc Sort Data = mssv.sample_survey_data_person3;
        By uqno;
Run;

Proc Sort Data = mssw.Msc_base_wgts;
        By uqno;
Run;

Data mssw.Msc_Person_base_wgts;
Merge mssv.sample_survey_data_person3(in=a)
   mssw.Msc_base_wgts(in=b);
        By uqno;
        If a and b;
Keep Uqno Person_ID SAL_Code_S Pr_Code Age_grp1 Age_Grp2 Race
   Gender full_wgt;
Run;

Proc Means Data=mssw.Msc_Person_base_wgts Nway Missing Sum;
Class Pr_Code Age_grp2;
        Var full_wgt;
Run;

Proc Means Data=mssw.Msc_Person_base_wgts Nway Missing Sum;
Class Age_grp1 Race Gender;
        Var full_wgt;
Run;

/* STATMX start here*/
options spool ibufsize=30000 ;
OPTIONs SPOOL;
** Transpose the National Control Totals from the Element
Groups to person Auxiliary Totals: _m1 - _m&m3 ;
Proc Sort Data=mssmx.MSC_CS_CONTROL_TOTALS3_AGEGRP1
   Out=NATIONAL_TOTALS
  (Keep= age_grp1 Race gender pop_count) ;
 By age_grp1 Race gender;
Run;
** Transpose the Provincial Control Totals from the Element
Groups to person Auxiliary Totals: _m1 - _m&m3 ;
Proc Sort Data=mssmx.MSC_CS_PROV3_AGEGRP2 Out=PROVINCIAL_TOTALS
  (Keep= pr_Code age_grp2 pop_count) ;
 By pr_Code age_grp2;
```

```
Run;

Data national_totals( index=( national=( age_grp1 race gender )
   /unique /nomiss ) );
  Set national_totals End=eof;
    By age_grp1 race gender;
      If ( first.gender ) Then _j_ + 1;
      Output;
      If ( eof ) Then call symput( 'm1', put( _j_, 5. ) );
Run;
%let m1 = &m1; Proc print; Run;
Proc transpose Data=national_totals Out=national( drop=_name_ )
   prefix=_n ;
  Var pop_count;
  id _j_;

Data national;
  Retain nation 1;
  length _n1-_n&m1 8;
  Array _total_ _n1-_n&m1;
  Set national;
  do over _total_;
    If ( _total_ < 0 ) Then _total_ = 0;
  End;

Data provincial_totals( index=( provincial=( pr_code age_grp2 )
   /unique /nomiss ) );
Set provincial_totals End=eof;
    By pr_code age_grp2;
      If ( first.age_grp2 ) Then _k_ + 1;
      Output;
      If ( eof ) Then call symput( 'm2', put( _k_, 5. ) );
Run;

%let m2 = &m2;
Proc transpose Data=provincial_totals Out=provincial(
   drop=_name_ ) prefix=_p ;
    Var pop_count;
    id _k_;

Data provincial;
  Retain province 1;
  length _p1-_p&m2 8;
  Array _total_ _p1-_p&m2;
  Set provincial;
  do over _total_;
    If ( _total_ < 0 ) Then _total_ = 0;
  End;

%*-------------------------------------------------------------------
  Obtain a list of the auxiliary variables for the national and
     provincial levels
;
Proc datasets library=work nolist;
  contents data=national   Out=temp1( Keep=name ) noprint;
  contents data=provincial Out=temp2( Keep=name ) noprint;
Run;

Proc Sql noprint;
  select name into :national_auxvars separated By ’␣’
    from temp1
      where substr( name, 1, 2 ) = ’_n’
  ;
  select name into :provincial_auxvars separated By ’␣’
    from temp2
      where substr( name, 1, 2 ) = ’_p’
  ;
  drop table temp1, temp2;
Quit;
%*------------------------------------------------------------------
  Aggregate the element auxiliary data to the cluster
     (household) for both
the national and provincial calibration groups

Required: cluster design weights, element data and element
   design weights
;

Data persons;
        Set mswsv.PERSON_REPLICATES;
Run;

Data households;
Set persons;
Keep uqno rpl_wgt001-rpl_wgt160 full_wgt ;
Run;

Proc Sort Data=households
          Out=hhld_wgts nodupkey;
  By uqno;
Run;

Proc Sort Data=persons
```

```
                  ( Keep=uqno person_id pr_code age_grp1
                     age_grp2 race gender rpl_wgt001-rpl_wgt160
                     full_wgt )
           Out=person_wgts;
  By person_id;
Run;

Data hhld_wgts( Keep=uqno full_wgt rpl_wgt001-rpl_wgt160
   _n0-_n&m1 _p0-_p&m2 nation province);
  Retain nation 1;
  Retain province 1;

  sampled = 0;
  Merge person_wgts( in=sampled ) hhld_wgts;
    By uqno;
      If sampled;
      %*
        Determine the conditional weight
      ;
      _cwgt_ = 1;

      Retain _n0-_n&m1 _p0-_p&m2 0;
      Array _count_  _n0-_n&m1 _p0-_p&m2;
      Array _n[ &m1 ];
      Array _p[ &m2 ];
     %*
Lookup the national cell index and aggregate the element
   auxiliary data (counts)
to the cluster level (household)
      ;
Set national_totals( Keep=age_grp1 race gender _j_ )
   key=national /unique ;
      If ( _iorc_ = 0 ) Then _n[ _j_ ] = _n[ _j_ ] + _cwgt_;
      Else
        do;
          _error_ = 0;
          _n0 = _n0 + _cwgt_;
        End;
      %*
Lookup the provincial cell index and aggregate the element
   auxiliary Data (counts)
to the cluster level (household)
      ;
      Set provincial_totals( Keep=pr_code age_grp2 _k_ )
         key=provincial /unique ;
      If ( _iorc_ = 0 ) Then _p[ _k_ ] = _p[ _k_ ] + _cwgt_;
      Else
```

```
        do;
          _error_ = 0;
          _p0 = _p0 + _cwgt_;
        End;

      If last.uqno Then
        do;
          Output;
          do over _count_;
            _count_ = 0;
          End;
        End;
Run;

Data hhld_wgts;
Set hhld_wgts;
        By uqno;
        If first.uqno;
Run;

%CALIBRATION_WEIGHTS
DATALINES;
survey= MSC Univeristy of Limpopo Dissertation,
number_of_input_periods= 1,
input_period1 = Year2017,
design= MSC Complex Survey Design,
  number_of_stages= 2,
  stage1= one-phase element [ppswr],
  stage2= one-phase element [srswor],

  stage2_element_sample_file= work.hhld_wgts,
    element= uqno,

  stage2_element_design_weights_file= work.hhld_wgts,
    element= uqno,
    design_weight= full_wgt,

calibration= simple regression,
  design_block= MSC Complex Survey Design,
    stage2_phase1_element_calibration_weight_bounds_file= ,
      element= ,
      lower_bound= [50],
      upper_bound= ,
  partition1= age_grp1 race gender calibration,
    calibration_groups_file= work.hhld_wgts,
      calibration_group= nation,
      auxiliary_variables= _n1 _n2 _n3 _n4 _n5 _n6 _n7 _n8 _n9
         _n10 _n11 _n12 _n13 _n14 _n15 _n16 _n17 _n18 _n19 _n20
         _n21 _n22 _n23 _n24 _n25 _n26 _n27 _n28 _n29      _n30
         _n31 _n32 _n33     _n34 _n35 _n36 _n37    _n38 _n39
         _n40 _n41     _n42 _n43 _n44 _n45     _n46 _n47 _n48,
    auxiliary_totals_file= work.national,
      calibration_group= nation,
      auxiliary_totals= _n1 _n2 _n3 _n4 _n5 _n6 _n7 _n8 _n9 _n10
         _n11 _n12 _n13 _n14 _n15 _n16 _n17 _n18 _n19 _n20 _n21
         _n22 _n23 _n24 _n25 _n26 _n27 _n28 _n29 _n30 _n31 _n32
         _n33     _n34 _n35 _n36 _n37     _n38 _n39 _n40
         _n41     _n42 _n43 _n44 _n45     _n46 _n47 _n48,

  partition2= pr_code age_grp2 calibration,
    calibration_groups_file= work.hhld_wgts,
      calibration_group= province,
      auxiliary_variables= _p1 _p2 _p3 _p4 _p5 _p6 _p7 _p8 _p9
         _p10     _p11 _p12 _p13 _p14     _p15 _p16 _p17
         _p18     _p19 _p20 _p21 _p22 _p23 _p24 _p25 _p26 _p27,
    auxiliary_totals_file= work.provincial,
      calibration_group= province,
      auxiliary_totals= _p1 _p2 _p3 _p4 _p5 _p6 _p7 _p8 _p9 _p10
         _p11 _p12 _p13 _p14 _p15 _p16 _p17 _p18 _p19 _p20 _p21
         _p22 _p23 _p24 _p25 _p26 _p27,
output_period= Year2017,
  calibration= simple regression,
    stage2_element_calibration_weights_file= work.hhld_calwgts,
      element= uqno,

      calibration_weight= full_calwgt,
      calibration_factor= full_calfactor,
options= run
;
** Merge the household calibration factor with the element Data
   and
calculate the calibration weight for the elements
;
Data mssmx.msc_full_calwgts;
Merge person_wgts hhld_calwgts( in=on_hhld_calwgts );
    By uqno;
      If not( on_hhld_calwgts ) Then full_calfactor = 1;
      full_calwgt = full_calfactor * full_wgt;
Keep person_id full_calwgt full_calfactor full_wgt age_grp1 race
   gender pr_code age_grp2;
Run;

Proc Means Data=mssmx.msc_full_calwgts Missing Sum;
```

```
Class pr_code;
        Var full_calwgt;
Run;

/**************End OF FULL WEIGHT CALIBRATION**************/
/*********BEGINNING OF REPLICATE WEIGHT CALIBRATION*******/

Proc Sort data=mssw.Msc_base_wgts nodupkey Out=mswsv.SAL_RPL;
        By SAL_Code_S;
Proc Sort data=mssw.Msc_base_wgts nodupkey Out=mswsv.HH_RPL;
        By uqno;
Run;

Proc Sort data=mssw.Msc_base_wgts nodupkey Out=Stratum_RPL;
        By stratum;
Run;

Proc freq Data=mswsv.SAL_RPL;
        table varunit/ nocol norow nopercent;
Run;

/******** Combine person data with replicates********/
Proc Sort data=mssw.Msc_Person_base_wgts;
        By SAL_Code_S;
Proc Sort data=mswsv.REPLICATES_SAL_RPL;
        By SAL_Code_S;
Run;

Data mswsv.Person_Replicates;
Merge mssw.Msc_Person_base_wgts(in=a)
   mswsv.REPLICATES_SAL_RPL(in=b);
        By SAL_Code_S;
        If a;
Run;

/*******Calibration of replicates in STATMX********/

filename parmhh "C:\msc_parms.txt";

%macro replicate( n );
  %do i= 1 %to &n;
    %let a = %sysfunc( putn( &i, z3. ) );
    Proc Sql noprint;
      create view work.hhld_wgt1 as
        select *, rpl_wgt&a as rpl_wgt
          from hhld_wgts
          where rpl_wgt&a > 0
      ;
    Quit;
        options nomprint nomlogic;
    %calibration_weights( infile=parmhh )

data work.hhld_calwgts;
      Merge work.hhld_calwgts( in=on_file )
            work.hhld_calwgt1( Keep=uqno rpl_calfactor);
        By uqno;
          If on_file;
                  rpl_calfactor&a = rpl_calfactor;
          If ( rpl_calfactor&a = . ) Then rpl_calfactor&a = 1;
    run;
  %End;
%mend replicate;

%replicate( 160 );

data hhld_calwgts_out (Keep=uqno full_calwgt full_calfactor
   rpl_calfactor001-rpl_calfactor160);
        Set work.hhld_calwgts;
run;

data rpl_Calwgts;
Merge person_wgts (in=a) hhld_calwgts_out (in=b);
By uqno;
        If a;

        Array rpl_calfactors rpl_calfactor001-rpl_calfactor160;
        Array rpl_calwgts rpl_calwgt001-rpl_calwgt160;
        Array rpl_wgts rpl_wgt001-rpl_wgt160;

        do over rpl_calwgts;
        rpl_calwgts=rpl_calfactors*rpl_wgts;
        End;

Keep uqno person_id full_calwgt full_calfactor
   rpl_calwgt001-rpl_calwgt160;
run;

data mssmx.MSC_RPL_Calwgts;
Set rpl_Calwgts;
run;

/****Add Survey Data****/

Proc Sort data= mssmx.MSC_RPL_Calwgts;
```

```
        By person_id;
Proc Sort data= mssv.sample_survey_data_person3;
        By person_id;
run;

Data mssw.MSC_RPL_Calwgts_Person;
Merge mssv.sample_survey_data_person3(in=a)
   mssmx.MSC_RPL_Calwgts(in=b);
        By person_id;
        If a and b;
Run;

Proc Means Data=mssw.MSC_RPL_Calwgts_Person Missing Sum;
        Class Highest_educational_level;
        Var full_calwgt;
Run;

Proc export data=mssw.MSC_RPL_Calwgts_Person
    outfile=’C:\MSC_RPL_Calwgts_Person.csv’
    dbms=csv
    replace;
run;
```

SAS Code 9: Tabulation in WesVar and SAS

```
/*Import Wesvar Files*/
Libname westbl "C:\";

/*Estimate*/
Data estimate(Keep= pr_code gender estimate est_type ’cv(%)’n
   deff);
set westbl.pr_gender;
Run;

Proc Sort Data=estimate; By pr_code; Run;

Proc Transpose Data=estimate Out=pr_gender_est;
By pr_code;
        Var estimate ’cv(%)’n deff;
        ID gender est_type;
Run;

/*Percent*/
Data pr_gender_pct(Keep= Pr_code _Name_ Male Female Total);
set pr_gender_est;
        Male=’1percent’n + 0;
        Female=’2percent’n + 0;
        Total =marginalpercent + 0;
Run;

Proc Sort Data=pr_gender_pct; By pr_code; Run;

Proc Transpose Data=pr_gender_pct Out=pr_gender_pct2;
By pr_code;
        ID _Name_;
        Var male female total;
Run;

Data pr_gender_pct3;
Set pr_gender_pct2;
        Latex= PR_code ||" "|| " & "||Estimate||" "|| " &
           "||’cv(%)’n||" "|| " & "||Deff|| "\\";
Run;
```