PSPPD II – a partnership between the Presidency,
the Republic of South Africa and the European Union

Service Contract No. DCI-AFS/2013-333-292

PROVISION OF PROJECT MANAGEMENT CONSULTANCY AND TECHNICAL ASSISTANCE SERVICES

The Learning Facility for the Programme to Support Pro-Poor Policy Development

PSPPD II

South Africa Microdata Scoping Study 2016

January 2017

Dr David McLennan
Professor Michael Noble
Ms Michell Mpike
Dr Gemma Wright
Ms Christine Byaruhanga

saspri
Southern African Social Policy Research Insights

This project is implemented by a Consortium led by Hulla & Co. Human Dynamics KG

Consortium partners

This project is funded by the European Union.

Acknowledgements

SASPRI would like to thank all those who kindly contributed information for this report. We are particularly grateful to the data repositories from where the majority of the census and survey metadata information was drawn for this report, and to the representatives from government who shared their knowledge of administrative datasets. SASPRI would like to thank DPME for commissioning this study and PSPPD2 for providing the funding. We would like to express particular thanks to Harsha Dayal, Carin Van Zyl and Nollen Mdhlovu at DPME for their support throughout the project, including the requesting of administrative datasets for case study purposes.

Disclaimer

Table of Contents

Acronyms

ADRN	Administrative Data Research Network (UK)
ANA	Annual National Assessment
CDR	Child Death Review
CHET	Centre for Higher Education Trust
DBE	Department of Basic Education
DG	Disability Grant
DHET	Department of Higher Education and Training
DHS	Department of Human Settlements
DPME	Department for Planning, Monitoring and Evaluation
DQAT	Data Quality Assessment Team
DSD	Department of Social Development
EMIS	Education Management Information System
GCRO	Gauteng City-Region Observatory
HEMIS	Higher Education Management Information System
HERANA	Higher Education Research and Advocacy Network in Africa
HSDG	Human Settlement Development Grant
HSRC	Human Sciences Research Council
HSS	Housing Subsidy System
IEC	Independent Electoral Commission
IES	Income and Expenditure Survey
LCS	Living Condition Survey
LURITS	Learner Unit Record Tracking System
LMIP	Labour Market Intelligence Partnership
MTSF	Medium Term Strategic Framework
NIDS	National Income Dynamics Study
NISIS	National Integrated Social Information System
NISPIS	National Integrated Social Protection Information System
NPC	National Planning Commission
NSS	National Statistical System
OAG	Old Age Grant
PALMS	Post-Apartheid Labour Market Series
PIT	Personal Income Tax
PSPPD-2	Programme to Support Pro-Poor Development, phase 2
SAIMD	South African Index of Multiple Deprivation
SALDRU	Southern Africa Labour Development Research Unit
SA-SAMS	South African School Administration and Management System
SASPRI	Southern African Social Policy Research Insights
SASSA	South African Social Security Agency
SARS	South African Revenue Service
SAPS	South African Police Service
SASQAF	South African Statistical Quality Assessment Framework
StatsSA	Statistics South Africa
UCT	University of Cape Town
UJ	University of Johannesburg
UKSA	United Kingdom Statistics Authority
WVG	War Veteran's Grant

1 Introduction

This report, by Southern African Social Policy Insights (SASPRI), provides information on the availability of social and economic microdata resources in South Africa up until the end of 2016. The objective was to collate information about the main data holders and data sets that are available in South Africa, both those that are easy to access and those where access may need to be negotiated. In both cases, procedures for access are given where possible. The report has been produced for the South African Department for Planning, Monitoring and Evaluation (DPME) under the Programme to Support Pro-poor Policy Development 2 (PSPPD 2) Learning Facility.

This report is an update of a study carried out in 2007 by researchers at the Centre for the Analysis of South African Social Policy (CASASP) at the University of Oxford for the UK Economic and Social Research Council (Barnes et al., 2007). **Where still relevant, the original text from the 2007 report is retained, with permission of the original lead authors**.

What is microdata?

Microdata are data about individual objects (such as persons, companies, events, transactions). Objects have properties which are often expressed as values of variables of the objects. For example, a 'person' object may have values of variables such as 'name', 'address', 'age', 'income'. Microdata represent observed or derived values of certain variables for certain objects. National microdata is usually available from censuses, surveys and administrative and register data. These data are most commonly collected by the national government or statistical office and access provided by the statistical office or the national archive. The data are collected at an individual, household, or institution level as appropriate (Desai and Cowell, 2006).

Prior to release to researchers for analytical purposes, microdata are typically anonymised to prevent the identification of individual objects (e.g. the identification of individual survey respondents) based upon their reported properties (e.g. survey respondents' age, sex, geographical location). Due to the often sensitive nature of information contained within microdata resources, confidentiality is a high priority.

In contrast, *macrodata* are data aggregated to a country or regional level. Macrodata are estimated values of statistical characteristics concerning sets of objects (or 'populations'). A statistical characteristic is a measure that summarises the values of a certain variable of the objects in a population (Desai and Cowell, 2006). Macrodata are typically more readily accessible by researchers than microdata as aggregated statistics typically do not allow for the identification of individual objects and therefore there are fewer issues concerning confidentiality. It is important when reviewing the potential utility of a macrodata resource for an analytical purpose to consider any issues concerning the collection of underlying microdata that are used as the basis for constructing the aggregate macrodata statistics. As such, this Microdata Review is of relevance to researchers who actively use (or wish to use) macrodata statistics as well as those who use (or wish to use) the individual level microdata.

In Chapter 2, census and survey data are discussed in broad terms, key data producers and data repositories are highlighted, and the census/survey datasets are categorised into a number of policy themes (with document links to the relevant technical details in Appendix 1). Chapter 3 focuses on administrative data and includes a discussion of how administrative data differs from census and survey data (including strengths and weaknesses), an overview of the administrative data landscape in South Africa, and three case studies discussing selected administrative datasets in more detail. Chapter 4 concludes the main body of the report by highlighting key developments and making

recommendations for future priorities. Finally, Appendix 1 contain technical details of each census and survey microdata resource that is referenced in this report; Appendix 2 presents an international example of a data quality assessment framework; and Appendix 3 presents an example of ongoing collaborations between Statistics South Africa and other government departments to enhance the quality of government administrative datasets.

The authors recognise that a substantial body of metadata already exists in the public domain concerning national surveys and censuses in South Africa. The purpose of the survey and census component of this Microdata Review 2016 study is to bring together key summary information about the survey and census microdata resources into a single easily accessible document. Rather than attempt to re-write existing official metadata, the authors of this Microdata Review have reproduced the existing metadata and referenced it accordingly within Appendix 1. The intention is that researchers will use this Microdata Review as an initial entry point to appreciating the wide range of survey/census microdata in existence in South Africa, and would then seek further detailed information from the referenced metadata sources for those datasets of particular interest.

In particular, the authors wish to express particular thanks to the **DataFirst** and **NESSTAR** data repositories from where much of the information concerning surveys and census microdata has been drawn. Further details of South African data repositories are provided in Section 2 of this document.

The authors also acknowledge that several other reviews of microdata resources have been undertaken over recent years. These reviews differ in terms of their purpose, sphere of coverage, type of content, date of compilation and target audience. For example, key data producers and data repositories contributed documentation to the government's 20 Year Review. In addition, reviews of microdata have been undertaken on particular data themes (for example, on labour market-related microdata (Woolfrey, 2013), and education-related microdata (van Wyk, 2015; Gustafsson, 2016a)). The Microdata Review 2016 report should be regarded as complementary to the other documents of this type that are already available or which are currently being produced.

Details concerning government administrative datasets are typically not made publicly available. It was therefore not possible to draw upon existing data repositories in the same way as for the survey/census datasets. The approach taken to document details of administrative microdata therefore required engagement with government departments on a dataset-by-dataset basis. Further details of this are provided in Section 3 of this document.

Lastly, while every effort has been made to provide a comprehensive account of the availability of social and economic microdata in South Africa, it is inevitable that some datasets and information will have been unintentionally missed out. It is recommended that this document is regularly updated and therefore any additions or amendments could be welcomed and incorporated.

2 Census and survey data

2.1 Key microdata producers

Statistics South Africa (www.statssa.gov.za)

Statistics South Africa (Stats SA) is the national statistics agency in South Africa and is mandated to collect and process data and produce official statistics. Stats SA's mission is "to lead and partner in statistical systems for evidence-based decisions"[1]. Stats SA seeks the broadest possible dissemination of the statistical data it collects, and the services it offers.

Stats SA produces a variety of statistics (macrodata) and microdata. In terms of macrodata, areas of focus include demography, health and vital statistics, national accounts, labour market, employment, industry and trade, prices, public sector spending, private sector finances and transport. The main microdata produced are the Population Census and a range of household surveys.

Many statistical publications can be downloaded from the Stats SA website free of charge[2], including the annual *Statistics in Brief* publication which presents an overview of data garnered from a variety of Stats SA publications released during the year in question. Many commonly cited statistical indicators relating to employment, unemployment, the labour force, labour force participation rates, household access to services and experience/fear of crime originate from household surveys conducted by Stats SA.[3]

Stats SA also established the NESSTAR data repository (see section on data repositories below). A wide range of Stats SA's publicly available datasets are available from NESSTAR or via a link on the Stats SA home page. Some of Stats SA's datasets can also be accessed from the South African Data Archive and DataFirst (see section on data repositories below).

Under section 7 (3) (d) of the South African Statistics Act 1999, most datasets are provided free of charge from Stats SA. Stats SA states that 'As a general principle, Stats SA does not seek to recover any of the costs of data collected, products developed or standard services provided, as those costs are met from an allocation voted by Parliament' (Statistics South Africa, 2002, p.1).[4]

Once a copy of a dataset has been secured, it is possible to disseminate the data further, providing no charge is made and Stats SA is acknowledged as the supplier and owner of the data and copyright.

For Stats SA data that is not publicly available, access would need to be negotiated with the Statistician General.

[1] See http://www.statssa.gov.za/?page_id=5360

[2] See http://www.statssa.gov.za or http://www.statssa.gov.za/publications/findpublication.asp to search for publications.

[3] See http://www.statssa.gov.za/publications/StatsInBrief/StatsInBrief2015.pdf for the latest Statistics in Brief publication.

[4] Typically, the only charges are those required to cover the costs of the medium of dissemination if provided on CD or DVD rather than downloaded from the internet.

Human Sciences Research Council (www.hsrc.ac.za)

According to its website, the HSRC "was established in 1968 as South Africa's statutory research agency and has grown to become the largest dedicated research institute in the social sciences and humanities on the African continent, doing cutting-edge public research in areas that are crucial to development"[5]. Their mandate is "to inform the effective formulation and monitoring of government policy; to evaluate policy implementation; to stimulate public debate through the effective dissemination of research-based data and fact-based research results; to foster research collaboration; and to help build research capacity and infrastructure for the human sciences. The Council conducts large-scale, policy-relevant, social-scientific research for public sector users, non-governmental organisations and international development agencies. Research activities and structures are closely aligned with South Africa's national development priorities"[6].

The HSRC aims "to serve as a knowledge hub for research-based solutions to inform human and social development in South Africa, the African continent and the rest of the world"[7].
Many of HSRC's microdata resources are catalogued and made available through the HSRC Research Data Service (see section on data repositories below).

Other relevant microdata producing institutions

Many other organisations across South Africa produce survey microdata on a smaller scale than StatsSA and HSRC. These organisations are acknowledged accordingly within the relevant dataset-specific sections of Appendix 1. Some of these data producers work at a national level, whilst others operate on a more localised geographical level. An example of an important producer of microdata with a sub-national focus is the Gauteng City-Region Observatory (www.gcro.ac.za) though the focus of this report is on national microdata.

2.2 Data repositories

DataFirst

Web address: https://www.datafirst.uct.ac.za/

According to their website DataFirst is "a research data service dedicated to making South African and other African survey and administrative microdata available to researchers and policy analysts. We promote high quality research by providing the essential research infrastructure for discovering and accessing data and by developing skills among prospective users, particularly in South Africa. We undertake research on the quality and usability of national data and encourage data usage and data sharing"[8]. Their mission is described by them as follows: "DataFirst is a research data service dedicated to making South African and other African survey and administrative microdata available to researchers and policy analysts. We promote high quality research by providing the essential research infrastructure for discovering and accessing data and by developing skills among prospective users, particularly in South Africa. We undertake research on the quality and usability of

[5] http://www.hsrc.ac.za/en/about/what-we-do
[6] ibid
[7] http://www.hsrc.ac.za/en/about/what-we-do/mission-vision-values
[8] https://www.datafirst.uct.ac.za/

national data and encourage data usage and data sharing"[9].

Via the website, one can also access of a list of where each particular dataset has been cited. Datasets and accompanying documentation can be downloaded from the website free of charge. DataFirst is working on creating an online data depositor function for their data platform; currently researchers can deposit data by contacting DataFirst directly.

NESSTAR

Web address: http://interactive.statssa.gov.za:8282/webview/

Nesstar is Statistics South Africa's primary data repository. The data repository "allows users to browse, analyse, tabulate and download datasets from a wide variety of census and household survey data and metadata in various formats[10]" It contains amongst others, 10% samples of Censuses, Quarterly Labour Force Surveys, General Household Surveys, and Community Surveys.

South African Data Archive

Web address: http://sada.nrf.ac.za/

According to its website: "The South African Data Archive serves as a broker between a range of data providers (for example, statistical agencies, government departments, opinion and market research companies and academic institutions) and the research community. The archive does not only preserve data for future use, but all adds value to the collections. It safeguards datasets and related documentation and attempts to make it as easily accessible as possible for research and educational purposes.

Existing research data can be an invaluable source for further studies. Such data are, however, currently scattered throughout the country. By preserving this research information in a single resource centre like SADA, unnecessary and costly duplication of research are decreased while the quality of the research findings are enhanced by using data from experienced researchers both locally and internationally.

Objectives of SADA are to

- acquire and catalogue survey data and related information.
- preserve such data against technological obsolescence and physical damage.
- re-disseminate such information for use by other researchers, for re-analysis of data, longitudinal and comparative studies, research training, teaching and policy-making decision purposes.
- formulate policies for the scope and content of data and data preservation.
- promote the optimal use of data.

SADA adds value to its collections in the following ways:

[9] https://www.datafirst.uct.ac.za/about-us/mission-statement
[10] http://www.statssa.gov.za/?page_id=1417

- Comprehensive machine-readable codebooks are developed, which include an abstract, sampling methodology and questionnaire. This documentation is published in open access on the data portal.
- Metadata is added to the datasets and made accessible through electronic search and retrieval systems, for example Internet"[11].

HSRC Research Data Service

Web address: http://www.hsrc.ac.za/en/research-data/

According to its website: "The HSRC Research Data Service provides a digital repository facility for the HSRC's research data in support of evidence based human and social development in South Africa and the broader region. Access to data is dependent on ethical requirements for protecting research participants, as well as on legal agreements with the owners, funders or in the case of data owned by the HSRC, the requirements of the depositors of the data. We facilitate data use by preparing comprehensive metadata and disseminating data and related documents to appropriate target audiences. Data sharing is subject to an End User License agreement"[12].

2.3 Key national survey and census datasets by theme

The main Census and survey datasets that have been identified are listed below by theme. The themes comprise: demography, housing, social welfare, economy, labour market, education, transport, crime, and health. Attitude surveys are also included in the Appendix and are listed as a final theme. Certain surveys straddle a number of themes (whereas others only relate to a single theme) but have been listed under each relevant theme for ease of reference.

The organisation listed as the principal investigator is not necessarily the data holder. Full details of the datasets are provided in Appendix 1, and the tables below have been hyperlinked to the relevant section in the Appendix.

Demography

Dataset	Year(s)	Principal investigator
2011 Census 10% Sample	2011	StatsSA
2001 Census 10% Sample	2001	StatsSA
1996 Census 10% Sample	1996	StatsSA
Community Survey 2016	2016	StatsSA
Community Survey 2007	2007	StatsSA
General Household Survey	2002-2015 (annually)	StatsSA

[11] http://sada.nrf.ac.za/
[12] http://www.hsrc.ac.za/en/research-data/

Integrated Planning, Development and Modelling Project	2008 and 2010	HSRC
National Youth Lifestyle Survey	2005 and 2008	Centre for Justice & Crime Prevention
South African Demographic and Health Survey	2016	Department of Health
Survey of Activities of Young People	1999 and 2010	StatsSA

Housing

Data source	**Year(s)**	**Principal investigator**
2011 Census 10% Sample	2011	StatsSA
2001 Census 10% Sample	2001	StatsSA
1996 Census 10% Sample	1996	StatsSA
Community Survey 2016	2016	StatsSA
Community Survey 2007	2007	StatsSA
All Media Products Survey	1995, 2002, 2010, 2011, 2012, 2013, 2014, 2015	South African Audience Research Foundation
General Household Survey	2002-2015 (annually)	StatsSA
Integrated Planning, Development and Modelling Project	2008 and 2010	HSRC
Living Conditions Survey	2008/09 and 2014/15	StatsSA
National Income Dynamics Study	2008, 2010, 2012 and 2014	SALDRU

Social welfare

Data source	Year(s)	Principal investigator
Department of Social Development Survey	2006 and 2008	DSD
General Household Survey	2002-2015 (annually)	StatsSA
Living Conditions Survey	2008/09 and 2014/15	StatsSA
National Income Dynamics Study	2008, 2010, 2012 and 2014	SALDRU

Economy

Data source	Year(s)	Principal investigator
2011 Census 10% Sample	2011	StatsSA
2001 Census 10% Sample	2001	StatsSA
1996 Census 10% Sample	1996	StatsSA
Community Survey 2016	2016	StatsSA
Community Survey 2007	2007	StatsSA
Department of Social Development Survey	2006 and 2008	DSD
Income and Expenditure Survey	1995, 2000, 2005/06 and 2010/11	StatsSA
Integrated Planning, Development and Modelling Project	2008 and 2010	HSRC
Living Conditions Survey	2008/09 and 2014/15	StatsSA
National Income Dynamics Study	2008, 2010, 2012 and 2014	SALDRU
South African National Innovation Survey	2008	HSRC
Survey of Employers and Self-employed	2001, 2005, 2009 and 2013	HSRC

Labour market

Data source	Year(s)	Principal investigator
2011 Census 10% Sample	2011	StatsSA
2001 Census 10% Sample	2001	StatsSA
1996 Census 10% Sample	1996	StatsSA
Community Survey 2016	2016	StatsSA
Community Survey 2007	2007	StatsSA
Department of Social Development Survey	2006 and 2008	DSD
Employment and Learning pathways of Learnership participants in the NSDS phase II (ELL)	2007	HSRC
General Household Survey	2002-2015 (annually)	StatsSA
Living Conditions Survey	2008/09 and 2014/15	StatsSA
National Income Dynamics Study	2008, 2010, 2012 and 2014	SALDRU
Post Apartheid Labour Market Series	1994 to 2015 (annually)	DataFirst
Quarterly Employment Survey	2006-2016	StatsSA
Quarterly Labour Force Survey	2000 onwards (quarterly from 2008)	StatsSA
South African National Innovation Survey	2008	HSRC
Survey of Activities of Young People	1999 and 2010	StatsSA
Survey of Employers and Self-employed	2001, 2005, 2009 and 2013	StatsSA
Time Use Survey	2000 and 2010	StatsSA

Education

Data source	Year(s)	Principal investigator
2011 Census 10% Sample	2011	StatsSA
2001 Census 10% Sample	2001	StatsSA
1996 Census 10% Sample	1996	StatsSA
Community Survey 2016	2016	StatsSA
Community Survey 2007	2007	StatsSA
All Media Products Survey	1995, 2002, 2010, 2011, 2012, 2013, 2014, 2015	South African Audience Research Foundation
Demographic and Health Survey	2016	Department of Health
Employment and Learning pathways of Learnership participants in the NSDS phase II (ELL)	2007	HSRC
General Household Survey	2002-2015 (annually)	StatsSA
HIV Prevalence and Related Factors – Higher Education Sector Study, South Africa	2008/09	Higher Education South Africa
Living Conditions Survey	2008/09 and 2014/15	StatsSA
National Income Dynamics Study	2008, 2010, 2012 and 2014	SALDRU
National Youth Lifestyle Survey	2005 and 2008	Centre for Justice & Crime Prevention
Survey of Activities of Young People	1999 and 2010	StatsSA
Time Use Survey	2000 and 2010	StatsSA
TIMSS: Trends in International Mathematics and Science Study	1995, 1999, 2002, 2011 and 2015	HSRC

Transport

Data source	Year(s)	Principal investigator
2011 Census 10% Sample	2011	StatsSA
2001 Census 10% Sample	2001	StatsSA
All Media Products Survey	1995, 2002, 2010, 2011, 2012, 2013, 2014, 2015	South African Audience Research Foundation
Integrated Planning, Development and Modelling Project	2008 and 2010	HSRC
Living Conditions Survey	2008/09 and 2014/15	StatsSA
National Household Travel Survey (NHTS)	2003 and 2013	Department of Transport
Time Use Survey	2000 and 2010	StatsSA

Crime

Data source	Year(s)	Principal investigator
All Media Products Survey	1995, 2002, 2010, 2011, 2012, 2013, 2014, 2015	South African Audience Research Foundation
National Victims of Crime Survey	1998, 2003, 2007, 2011, 2012, 2013/14 and 2014/15	StatsSA (n.b. ISS in 2003 and 2007)
National Youth Lifestyle Survey	2005 and 2008	Centre for Justice & Crime Prevention

Health

Data source	Year(s)	Principal investigator
2011 Census 10% Sample	2011	StatsSA
2001 Census 10% Sample	2001	StatsSA
1996 Census 10% Sample	1996	StatsSA
Community Survey 2016	2016	StatsSA
Community Survey 2007	2007	StatsSA
General Household Survey	2002-2015 (annually)	StatsSA
HIV Prevalence and Related Factors – Higher Education Sector Study, South Africa	2008/09	Higher Education South Africa
Living Conditions Survey	2008/09 and 2014/15	StatsSA
National Antenatal Sentinel HIV prevalence Survey	1990-2013 (annually)	DoH
National Income Dynamics Study	2008, 2010, 2012 and 2014	SALDRU
National Youth Lifestyle Survey (NYLS)	2005 and 2008	Centre for Justice & Crime Prevention
SAGE: Study on global AGEing and adult health	2007/08	WHO & HSRC
South African Demographic and Health Survey	2016	Department of Health
South African National Health & Nutrition Examination Survey (SANHANES)	2011/12	HSRC
South African National HIV, Behaviour and Health Survey	2002, 2005, 2008 and 2012	HSRC
Time Use Survey	2000 and 2010	StatsSA

Attitude Surveys

Afrobarometer South Africa	2000, 2002, 2004, 2006, 2008, 2011 and 2015	IJR, South Africa
South African Reconciliation Barometer	2003-2011	IJR, South Africa
South African Social Attitudes Survey (SASAS)	Annually since 2003	HSRC

3 Administrative microdata

3.1 Introduction

The main objective of this chapter of the Microdata Review is to focus on administrative data in order to raise its profile as a valuable resource for researchers with an interest in the social and economic challenges faced by South Africa.

Given the current dearth of publicly available information concerning administrative datasets in South Africa, the approach taken in this chapter differs that of the previous chapter on survey and census data. Rather than compiling an exhaustive list of administrative datasets in existence in South Africa at the present time, the approach here has instead focused on promoting administrative data as an important resource, with references given to pertinent examples including three more detailed case studies. The information presented here was mainly collected through personal correspondence with data experts across South African government, unlike the material presented in Chapter 2 (and Appendix 1) which was largely drawn from (and referenced to) publicly available data repositories.

What is administrative microdata?

Administrative data is information collected during the process of and for the primary purpose of delivering a public service. The vast majority of administrative datasets are therefore collected by public bodies, such as national and local government and certain parastatal organisations. Administrative data is therefore *not* collected primarily for research purposes and, as such, it differs in this regard from survey data (which typically *is* collected primarily for research purposes). However, despite not being the main purpose for which it is collected, administrative data holds great potential for social and economic research.

The emphasis within this chapter is on those types of administrative data that have the greatest potential value for supporting evidence-informed decision-making across the key social and economic policy spheres. Many of the administrative datasets in South Africa do indeed have relevance in this regard (e.g. in relation to poverty, the labour market, education and crime), but certain large administrative datasets are less relevant for this purpose (for instance, the PERSA database on payroll records for national and provincial government employees). Whilst all administrative datasets have relevance for the particular operational purpose for which they are collected, the focus here is explicitly on those dataset that have relevance for tackling the country's main socio-economic challenges.

During the early phase of this Microdata Review it became apparent that there is very little information in the public domain concerning major government administrative datasets. It is certainly the case that researchers are able to access metadata about survey and census datasets far more easily than for administrative datasets. Whereas in recent years there has been considerable progress in collating, documenting and facilitating researcher access to survey and census datasets in South Africa (e.g. the services provided by DataFirst), the situation with regards to administrative data lags far behind. Some of the possible reasons for this are discussed below.

The remainder of this chapter is structured as follows. In Section 3.2, an initial broad overview of administrative data is given to illustrate the main way in which these data differ from survey and census data. This includes a discussion of the key strengths and weaknesses of administrative data as compared to survey and census data, and two important issues are discussed that are pertinent to

researchers considering using administrative data, namely data quality and data security. In Section 3.3, a summary of recent and ongoing initiatives within South African government and academia to promote administrative data is presented and a selection of key datasets and their research uses are highlighted.

In Section 3.4, three administrative datasets are examined in somewhat more detail and examples given of how these datasets have been used, are being used and could potentially be used in research to understand social and economic challenges.

Although the focus here is on South African administrative datasets, international experiences are also referenced where relevant to demonstrate common challenges and/or innovative ways in administrative datasets are being used in different international settings.

3.2 The value of administrative data for research

> *"What steam was to the 19th century, and oil has been to the 20th, data is to the 21st. It's the driver of prosperity, the revolutionary resource that is transforming the nature of economic activity, the capability that differentiates successful from unsuccessful societies." (Royal Statistical Society, 2016, p.1).*[13]

Administrative data may be collected at the levels of the individual, household, institution (e.g. school), or event (e.g. a crime). For example, the South African Social Security Agency (SASSA) collects information about individual beneficiaries of social grants; the South African Revenue Service (SARS) collects information about the tax status and tax obligations of both individuals and firms that operate in the formal sector; the Department of Basic Education (DBE) collects information about school-aged pupils and the schools they attend; and the South African Police Service (SAPS) collects information about crimes recorded across the country.

An important feature of administrative datasets is that they should, in theory, capture all individuals/households/institutions/events that are serviced by the relevant public body for that particular purpose. For example, SASSA's administrative database on social grant beneficiaries should, in theory, contain details of every single grant beneficiary in the country and SAPS's crime database should, in theory, contain details of every single crime that was recorded across the country[14]. Unlike survey datasets, which are based on samples of the overall population of interest, administrative datasets are in effect closer to a census of the population. However, whereas the national Censuses of Population (1996, 2001, 2011) aim to enumerate every single person in the country, each separate administrative dataset is required to enumerate only those individuals/households/institutions/events that are pertinent to the purpose of that particular administrative dataset. For example, SASSA's social grant database will explicitly not contain any details of individuals who have never claimed a social grant. Administrative datasets are therefore akin to a census of the *population of interest* for the relevant public service purpose.

Administrative data have a number of important features that make them well-suited for research about social and economic challenges.

[13] The longer (2014) version of the Data Manifesto is available at http://www.rss.org.uk/Images/PDF/influencing-change/rss-data-manifesto-2014.pdf and was 'aimed at helping government understand what it can do to make the most of the data opportunity'.
[14] Although see below where issues of under-reporting and under-recording of crimes are discussed.

***Strength: Provides a census of the population of interest*.** Arguably the greatest strengths of administrative data are due to these datasets being censuses of the populations of interest, rather than based upon population samples as is the case with survey data. Administrative data are therefore not subject to survey sampling error, which means that statistics derived from administrative data can be regarded as truer measures of the indicator in question[15]. For instance, interrogation of SASSA's social grants database should produce accurate counts of the number of grant beneficiaries at the particular point in time. In contrast, because surveys are only based on samples of the population of interest (the sampling frame), statistics derived from survey sources should always be accompanied by statistical confidence intervals to provide the reader/user with a sense of the reliability of the survey estimates. Typically, survey-based estimates become less reliable as the size of the sample reduces, meaning that survey estimates for population sub-groups (e.g. separately by gender/age/population group/economic status etc) may be accompanied by very wide confidence intervals, indicating the survey estimates are not particularly reliable for that level of disaggregation. Administrative data do not suffer from these weaknesses relating to survey sampling, meaning that statistics can be derived for a range of population sub-groups that are typically regarded as truer measures of the indicator in question for each sub-group. For example, SASSA's social grant database contains details of the age and sex of each beneficiary and so it is possible to calculate counts of grant beneficiaries by age/sex sub-group that reflect the actual numbers of such beneficiaries (whereas survey estimates, for example derived from the National Income Dynamics Study, would necessarily be accompanied by confidence intervals reflecting the uncertainty with which surveys measure the true population values of such indicators).

Strength: Enables sub-national level analysis. One particular form of sub-group analysis to which administrative data are particularly well suited (while survey data are not well-suited) is the calculation of indicator values at sub-national level. Most national surveys in South Africa are deigned to produce national level estimates of the research objective in question. Some surveys can also be used to calculate relatively reliable estimates at provincial level (e.g. the Living Condition Surveys and the Income and Expenditure Surveys)[16]. However, for analyses at a smaller spatial scale than provinces, survey datasets are rarely suitable. To date, most of the sub-provincial analyses of social and economic indicators in South Africa have been based on Census data (which, like administrative data, is not subject to survey sampling errors and does not require the presentation of associated confidence intervals[17]). For example, McIntyre et al. (2000) produced four alternative deprivation indices at magisterial district level using data from the 1996 Census in order to explore the relationship between deprivation and health inequalities in South Africa. Similar work was undertaken by researchers in the provincial government of the Western Cape to generate deprivation indices at municipality level across the Western Cape province using data from the 1996 Census (Department of Health and Social Services, 1999) and then subsequently using data from the 2001 Census (Department of the Premier of the Western Cape, 2005). Other examples of socio-economic indicators being constructed at small area level using Census data include the South African Index of Multiple Deprivation 2001 (SAIMD 2001) at Datazone level (Noble et al., 2009) and the South African Index of Multiple Deprivation 2011 (SAIMD 2011) at Ward level (Noble et al., 2013). In addition to supporting the construction of the South African indices of deprivation, StatsSA

[15] Although some would argue that a census/administrative dataset is, in fact, a sample from a super-population.

[16] Uniquely the Community Surveys 2007 and 2016 with their very large samples are designed to allow some analysis at Municipality level.

[17] But see footnote above concerning super populations.

has also recently used the 2011 Census as the basis for the South African Multidimensional Poverty Index (Statistics South Africa, 2014).[18]

Strength: A resource for small area level indicators. Small area level socio-economic indicators can also be produced from administrative datasets if these datasets contain sufficiently accurate details of the geographical locations of the individuals/households/institutions/events contained within the administrative dataset. In the United Kingdom, for example, the small area level indices of deprivation (similar in purpose and methodological approach to the census-based SAIMD referenced above) are based almost entirely on indicators derived from administrative datasets (see, for example, Noble et al. (2000); Noble et al. (2004); Noble et al. (2007), McLennan et al. (2010) and Smith et al. (2015) for examples from England).

Strength: Up-to-date and often continuously updated. A further important strength of administrative data over both survey and census data is that administrative datasets are typically updated on a continuous basis. This continuous updating is often necessitated by the demands of providing the particular service that the administrative dataset supports. For instance, SAPS continuously records criminal events as and when they are reported by the public or detected by police officers in order to enable SAPS to tailor its local policing activity to prevailing priorities. Similarly, SASSA captures new grant beneficiaries continuously in order to ensure that the beneficiaries are paid the grants for which they are eligible. The continuous reporting for operational purposes serves to generate administrative datasets that can be analysed with a far greater degree of temporal detail than is typically possible with surveys (which are often only undertaken yearly or even less frequently) and certainly than is possible with Censuses (which have only been undertaken three times since the advent of democracy in 1994).

Strength: Opportunities for data linkage and longitudinal analysis. The continuous collection of administrative microdata on the entirety of a relevant dataset population also presents the possibility of linking data over time to undertake longitudinal tracking of individuals, or institutions. For instance, by linking SASSA's social grant data over time using a unique personal identifier (such as ID number) it would be possible to track individuals as they make the transition into receipt of social grants (i.e. enter the SOCPEN database), move between social grants (e.g. stop claiming the Disability Grant and start claiming the Old Age Grant), experience a change in household composition (e.g. the addition of a new child eligible for Child Support Grant), or leave the social grants database altogether (for example because they no longer qualify for any grants, or have left the country, or are deceased or have fallen off the system for some administrative reason). Similarly, using a unique individual identifier to link administrative data on individual pupils' educational records over time offers the possibility of tracking pupils' educational achievements and analyse these in terms of the progress they achieved between different stages of their individual educational trajectories. Indeed, the Department of Basic Education (DBE) is currently striving to develop a linked pupil dataset of this very type, LURITS, which holds great potential for research in future years. There are many examples internationally of how longitudinal linkage of administrative datasets has facilitated valuable insights into socio-economic challenges. Examples from the UK include: studies that have tracked individuals as they moved into and out of employment in the UK (and between employment and social security benefits) (e.g. Evans and Noble (2001); Evans et al. (2002); Barnes et al. (2011)); studies that have tracked school-aged pupils as they progressed through the compulsory education system, including tracking them if they move schools, to assess factors that are associated with their educational achievements (e.g. Wilkinson and McLennan (2010); Wilkinson et al. (2010)); and studies that have tracked criminal offenders to assess their employment and social security

[18] Census data has also been used *in combination* with survey data to generate estimates of poverty and deprivation at small area level using small area estimation techniques. See, for example, Alderman et al. (2002); Demombynes and Ozler (2006).

dynamics prior to conviction and post release from prison (Ministry of Justice and Department for Work and Pensions, 2011). The example of data linkage concerning criminal offenders is particularly innovative as it entails the linkage of personal information about offenders from three different government departments and so is a prime example of the added value that data linkage can offer when tackling research questions. The UK's Administrative Data Research Network (ADRN) has identified data linkage as a priority area and is working with data owners and external researchers to establish methodologies for linking datasets over time and from different sources.[19]

Limitation: Narrow content with limited options for additional questions. Although administrative data have many strengths, there are also a number of acknowledged limitations. Arguably the greatest limitation, especially when compared to social survey data, is that administrative datasets are often relatively narrow in their content. The primary reason for this is that the information collected in an administrative dataset will be determined by the operational purpose underpinning the dataset. As such, the content of an administrative dataset may not be exactly what one would hope for when undertaking research about a socio-economic issue. Social surveys, on the other hand, can be designed to be far broader in scope and to ask questions that are likely to be pertinent to tackling the research question in hand. The very large size of many administrative databases (e.g. all social grant recipients in the entire country or all school-aged pupils in the entire country) means that it is often difficult to modify the data collection requirements to capture new variables that might be valuable for research purposes. However, many administrative datasets do contain sufficient information that can be applied to answer socio-economic research questions, even if the indicators derived from the administrative sources are only proxies for the indicators one might ideally wish to construct.

Limitation: Linkage challenges. A further limitation of administrative data is that the linkage of two (or more) different datasets is rarely straightforward. There are a number of factors that may hinder data linkage, such as: differences in the time point/period covered by the respective datasets; differences in the unique identifier variables needed to match cases between datasets; differences in the structure and format of the datasets; and differences in the type of object for which the data are collected (e.g. difficulties in matching individual level records to household level records, or difficulties linking the recorded criminal events to the individuals suffering the victimisation). However, as will be discussed below, South Africa's use of national ID numbers means that this should offer a means of linkage that is more straightforward than in countries lacking national ID numbers. Indeed, one example is presented below where the national ID numbers of individuals have been used to link two separate administrative databases from SASSA and IEC to generate a combined dataset. The combined dataset has enabled researchers to undertake types of analysis that were not possible using either of the two input datasets alone.

Data quality considerations

While administrative data clearly hold great potential for social and economic research and statistics, the very nature of these data – that they are explicitly *not* collected for research purposes – means that users should exercise a degree of caution when using such data. Awareness of data quality issues should inform the choice of administrative dataset utilised, the types of analyses performed, and interpretation of the results. The varied nature of administrative data sources and the varied analytical applications of these data mean that the options for dealing with the data quality issues may be both *dataset-specific* and *user-specific*. However, a number of common themes can be identified across international literature concerning the quality of administrative data. In the UK, for example, the United Kingdom Statistics Authority (UKSA) produced a list of factors that data producers (primarily government departments) should be wary of when using

[19] See https://adrn.ac.uk/getting-data/de-identification/data-linkage/

administrative data as the basis of statistics or research[20]. This list of factors is also relevant to researchers outside of the data production process who are considering using administrative data in their work. Box A is reproduced from the UKSA report:

Box A: Data production issues of possible concern in relation to administrative data

Box A

Lack of standardised application of data collection:

- inconsistencies in how different suppliers interpret local guidance
- differences in the use of local systems for the intended administrative function
- the distortive effects of targets and performance management regimes
- differing local priorities, data suppliers might require higher levels of accuracy for certain variables (for example payments) but less so for other aspects that are important to the statistical producer (for example demographics)

Variability in data suppliers' procedures:

- statistical producers typically do not have direct control over the development of guidance for data entry
- local checking of the data can be variable and might not identify incorrect coding or missing values
- local changes in policy could impact on how the data are recorded or on the coverage of the statistics

Quantity of data suppliers:

- there can be a large number of data suppliers, often spread geographically
- there can be many data collectors providing their data to an intermediary organisation for supply to a statistical producer

Complexity and suitability of administrative systems:

- administrative datasets can be complex containing large numbers of variables; it takes time, and therefore resource, to extract the necessary data required by the statistical producer
- data collation can be hampered by IT changes at the data supplier level
- data might need to be manipulated by the data supplier to meet the structural requirements of the statistical producer, leading to potential for errors

Public perceptions:

- lack of knowledge about use of personal data for statistical purposes
- concern that personal data should be sufficiently anonymised and secured

Source: (UKSA, 2014)

In recognition of the relatively common set of data quality themes that are relevant to administrative data, a number of guidance documents have been released internationally in recent years that can aid users in reviewing administrative datasets in terms of data quality. For example, Daas et al. (2012) developed a comprehensive data quality assessment framework for administrative data and this framework has been utilised by a number of national statistical institutions (e.g. Netherlands, Sweden, Australia). This framework distinguishes between three different views on administrative data quality, referred to a 'hyperdimensions': Source; Metadata; and Data. The Source hyperdimension relates to issues concerning the data sharing process, while the Metadata

[20] UKSA is a statutory body in the UK and is independent of government departments. It reports directly to parliament and has responsibility for, amongst other things: "regulating quality and publicly challenging the misuse of statistics" See https://www.statisticsauthority.gov.uk/about-the-authority/

hyperdimension relates to issues concerning the quality and comprehensiveness of any accompanying metadata, and the Data hyperdimension focuses on possible quality concerns with the actual data content. Each hyperdimension of data quality is measured by a series of component dimensions, which are themselves each measured using a series of component indicators of quality. For example, the Data hyperdimension is conceptualised as consisting of five component dimensions of data quality: Technical checks; Accuracy; Completeness; Time-related; and Integrability. Each of these five dimensions of data quality is measured by a series of indicators against which a dataset is assessed. Many counties internationally are now using some form of data quality framework (not necessarily the one proposed by Daas et al) to make assessments of their administrative data sources and to help support data producers in maximising data quality.

In South Africa, the Statistics Act (no.6 of 1999) requires the Statistician General at Statistics South Africa to coordinate statistical production within the country, not only within Statistics South Africa itself, but also across other public bodies. In working to fulfil this role, StatsSA has developed and continues to develop an evolving set of guidance documents on issues concerning statistical standards. An important component of this is StatsSA's development and implementation of the South African Statistical Quality Assessment Framework (SASQAF). In the most recent version of the SASQAF document, StatsSA state that:

> "The main purpose of SASQAF is to provide a flexible structure for the assessment of statistical products. SASQAF can be used for:
>
> - self-assessment by producers of statistics;
> - reviews performed by a Data Quality Assessment Team (DQAT) in the context of the National Statistical System (NSS) work;
> - assessment by data users[...]based on the producing agency's quality declaration;
> - assessment by international agencies (e.g. the International Monetary Fund) based on the quality declaration." (Statistics South Africa, 2010a, p.2)[21]

Whilst the SASQAF is of relevance to all sources of statistics (i.e. administrative, survey and census sources) produced by StatsSA and other public bodies, it is referenced here in light of the important role this framework can play in supporting other government departments to maximise the quality of the *administrative* data that they routinely collect.

Eight dimensions of data quality are listed within the SASQAF:

- Relevance;
- Accuracy;
- Timeliness;
- Accessibility;
- Interpretability;
- Comparability and Coherence;
- Methodological soundness; and
- Integrity.

The SASQAF also highlights important 'prerequisites of quality', which refer to "the institutional and organisational conditions that have an impact on data quality. These include the institutional and legal environment, and availability of human, financial and technological resources" (Statistics South Africa, 2010a, p.4).

The SASQAF provides guidance on how each of the eight dimensions of data quality listed above should be measured. Each dimension is composed of multiple indicators of data quality, and each

[21] http://www.statssa.gov.za/standardisation/SASQAF_Edition_2.pdf

indicator is associated with one or more data quality standards. The guidance sets out a four-tier assessment of quality for each data standard within each indicator of data quality, from Level 1 ('poor statistics'), Level 2 ('questionable statistics'), Level 3 ('acceptable statistics') and Level 4 ('quality statistics'), with each level associated with a set of data quality benchmarks. To support the implementation of the SASQAF, StatsSA has also published the SASQAF Operational Standards and Guidelines document (Statistics South Africa, 2010)[22] which goes into more detail on each of the constituent data quality indicators and standards and provides guidance on the application of the assessment framework.

Although it is outside the scope of this Microdata Review to go into detail on how the SASQAF is being implemented across all statistical producers in South Africa, it is important for researchers and other users of statistics to be aware of this framework and its role in guiding data quality assessments within the country.[23] When considering the strengths of weaknesses of particular South African administrative datasets for research and statistical purposes, the SASQAF provides a useful structure, and has relevance for administrative, survey and census data.[24]

Data security considerations and implications for data sharing

Administrative microdata often contain personal identifiable information, such as names and addresses, and may also contain highly sensitive information, such as information concerning a person's health or income status or educational achievements. Protecting the confidentiality of objects contained within administrative datasets is of critical importance to the organisations that collect these data. However, this understandable emphasis on protecting personal information inevitably acts as a barrier to making full use of the data for research purposes. Crucially, unlike survey data – which consists of questionnaire responses of people who have *willingly consented* to their (anonymised) data being used for research – the information collected on individuals in administrative databases are collected for the operational purpose, not for research. The issue of consent may present legal barriers to sharing administrative microdata with other parties and for making use of administrative data for purposes other than the operational purpose for which it was collected.

Best practice is still being developed for enabling data sharing and data linkage without compromising data security. An important recent study has developed an 'anonymisation decision-making framework' (Elliot et al., 2016). The freely downloadable and lengthy document raises the key issues for consideration to ensure that data is adequately anonymised.

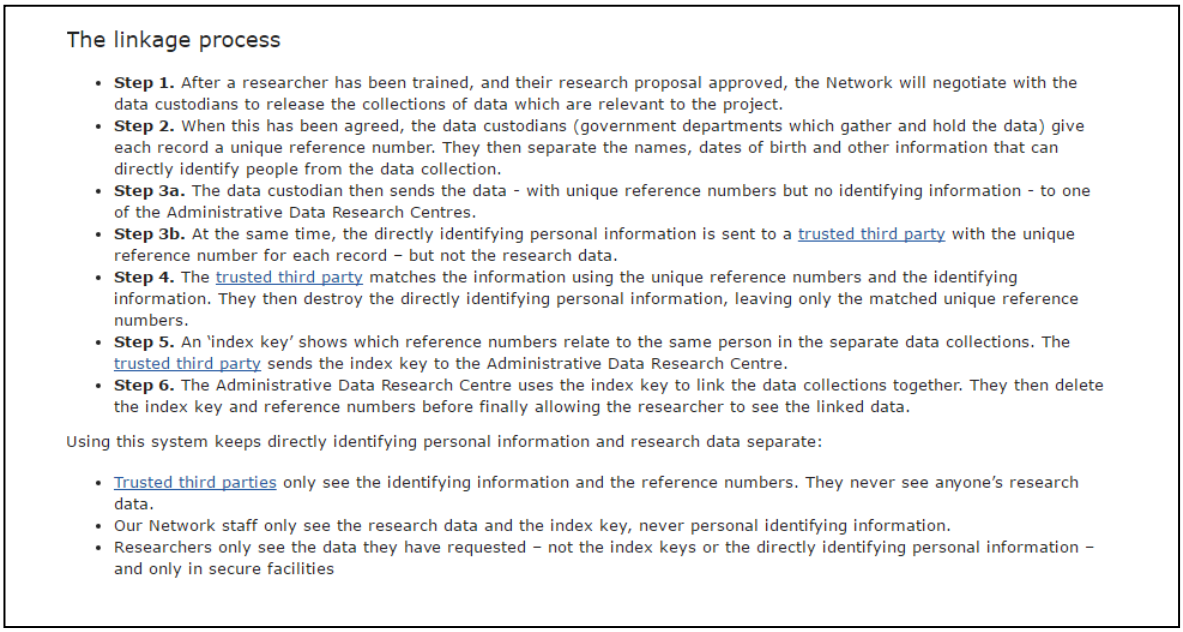
As a second example, the box below summarises the ADRN's procedure for undertaking data linkage of two datasets securely, which highlights the highly technical data processes that can be required, and which also need to be underpinned by legal and ethical compliance.

[22] http://www.statssa.gov.za/standardisation/SASQAF_OpsGuidelines_Edition_1.pdf

[23] In the case study on SAPS recorded crime data in Chapter 3, reference is made to a recent assessment of data quality by the Statistician General using the SASDQF.

[24] See also Western Cape Government (2013) for a provincial example of efforts to increase administrative data quality and usage.

Box B: The Administrative Data Research Network's summary of a secure data linkage process



The linkage process

- **Step 1.** After a researcher has been trained, and their research proposal approved, the Network will negotiate with the data custodians to release the collections of data which are relevant to the project.
- **Step 2.** When this has been agreed, the data custodians (government departments which gather and hold the data) give each record a unique reference number. They then separate the names, dates of birth and other information that can directly identify people from the data collection.
- **Step 3a.** The data custodian then sends the data - with unique reference numbers but no identifying information - to one of the Administrative Data Research Centres.
- **Step 3b.** At the same time, the directly identifying personal information is sent to a trusted third party with the unique reference number for each record – but not the research data.
- **Step 4.** The trusted third party matches the information using the unique reference numbers and the identifying information. They then destroy the directly identifying personal information, leaving only the matched unique reference numbers.
- **Step 5.** An 'index key' shows which reference numbers relate to the same person in the separate data collections. The trusted third party sends the index key to the Administrative Data Research Centre.
- **Step 6.** The Administrative Data Research Centre uses the index key to link the data collections together. They then delete the index key and reference numbers before finally allowing the researcher to see the linked data.

Using this system keeps directly identifying personal information and research data separate:

- Trusted third parties only see the identifying information and the reference numbers. They never see anyone's research data.
- Our Network staff only see the research data and the index key, never personal identifying information.
- Researchers only see the data they have requested – not the index keys or the directly identifying personal information – and only in secure facilities

Source: https://adrn.ac.uk/getting-data/de-identification/data-linkage/

In South Africa, the University of Cape Town has established a secure research data centre in order to facilitate researchers' access to sensitive data in a controlled, security conscious, environment. A number of security measures have been put in place to minimise the risks of accidental (or deliberate) disclosure of sensitive datasets:

> "The data owners do not relinquish ownership of the data or control over access. They will still give individual permission for particular academic researchers to be granted access and for particular projects to go ahead. Researchers will have to comply with the terms of their end-user agreements. DataFirst will prescreen projects and applicants and then pass the necessary documentation and signed confidentiality agreements to the data owner for sign off.
>
> The protection of the data itself is based around the following principles:
>
> - Secure physical access
> - Secure data access
> - Security conscious researchers and
> - Security conscious staff." [25]

For example, with regards to the NIDS dataset, it is possible to obtain extra variables via the secure data centre[26]. As well as accessing the sensitive survey variables, it is also possible to access certain information derived from administrative data, such as details about the child's school as this been linked to the individuals in the survey.[27]

There are many initiatives underway to promote the principle of access to data, termed 'open data'.[28] For example, the Global Open Data Index[29] is an annual effort to measure the status of open

[25] https://www.datafirst.uct.ac.za/documentation/13-sds-brochure/file
[26] http://www.nids.uct.ac.za/documentation/faqs/secure-data
[27] For a list of the administrative data variables that have been linked, please see p.5-8 of: http://www.nids.uct.ac.za/images/documents/wave4/W4_SecureDataVariables.pdf
[28] See the *International Open Data Charter* at hhtp://opendatacharter.net/
[29] See http://index.okfn.org/

government data around the world, using crowd-sourced surveys. Countries are given a score based on the amount of government data that exists and how accessible it is. In 2015, Taiwan, the UK[30] and Denmark were ranked as having the most open data out of 122 countries, while Libya, Syria and Myanmar were ranked as having the least open data out of the 122 countries. South Africa was ranked 54th.[31]

Another initiative, the Open Data Inventory (ODIN) assesses countries on the basis of data coverage (indicators, frequency, disaggregation) and data openness (download format, metadata available, and licensing terms). The inventory is collated by an international not-for-profit called Open Data Watch, and on this ranking system, South Africa ranks 43rd out of 173 countries.[32]

In support of open data, the African Development Bank (AFDB) has launched the African Information Highway which it describes as 'a mega network of live open data platforms [..] linking all African countries and 16 regional organizations. The overall objective is to significantly increase public access to official and other statistics across Africa, while at the same time supporting African countries to improve data quality, management, and dissemination'.[33] The data portal provides access to an extensive amount of open data about South Africa.

3.3 The administrative data landscape in South Africa

Administrative data is increasingly recognised as a potentially valuable resource for research and statistics to inform policy decisions in South Africa. This Microdata Review is just one of a growing number of studies that seeks to raise awareness of administrative data, both through documenting selected administrative datasets and through directing potential data users to sources of further, more detailed information.

The Twenty Year Review: South Africa 1994-2004. During the process of undertaking of this extensive review of South Africa's progress and challenges in the period 1994-2004, , DPME and the NPC liaised with key data producing organisations to compile a list of data sources that might provide valuable evidence to inform the review process. Although the focus was mainly on survey and census datasets, administrative data were also acknowledged as contributing to the data landscape.

In addition to the broad data scoping exercises such as informed the *Twenty Year Review*, a number of focused data audits have been undertaken for particular policy themes, such as education, skills and the labour market. This chapter does not duplicate the detailed content of these other data reviews/audits but rather presents brief summaries and refers readers to the relevant sources of detailed documentation.

Review of Basic Education administrative data. Basic education is one policy sphere within which the data audit, data management and data utilisation functions are being closely scrutinised. As part of this process, DPME recently commissioned an audit of educational data sources, covering

[30] One of the developments in the UK that will have contributed to its ranking is extensive consultation with the public about the use of Government data. See for example this report which summarises responses received by the Cabinet Office to consultation about how to improve the use of data in government: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/535063/better_use_of_data_in_government_response_final.pdf

[31] See http://bit.ly/2k3eBNL.

[32] See odin.opendatawatch.com

[33] See http://dataportal.opendataforafrica.org/apps/atlas/South-Africa

administrative sources and survey sources (van Wyk, 2015)[34]. In addition a programme of work is being undertaken for the National Treasury to ascertain the extent to which data is being effectively utilised in the basic education sector (Gustafsson, 2016a). Basic education data is considered in more detail below as one of the three case studies in Chapter 4 of this Microdata Review.

Review of Skills Planning administrative data. Skills planning is another policy area to have benefited from a recent data audit. Members of the Labour Market Intelligence Partnership (LMIP), led by the Human Sciences Research Council (HSRC) conducted a High-level Audit of Administrative Datasets (Paterson et al., 2015; Paterson and Visser 2016), in which they investigated the relevance of a range of government (and non-government) databases for the particular purpose of skills planning. The LMIP research team liaised with a range of government departments and other entities and, for each of these, documented the organisational objectives and mandates, organisational structures and capacities, and key features of selected datasets held by the organisation that were of relevance to skills planning. A further component of this research was an assessment of options for linking key datasets with other relevant databases as part of the skills planning mechanism. As noted, this LMIP study had an explicit focus on skills planning rather than the broader focus adopted in this Microdata Review, and so the datasets that were reviewed were mainly related to personnel issues. For example, with regards to the South African Police Service datasets, the LMIP study focuses on issues concerning staff recruitment and training, and does not refer to data about police recorded crime. As such, the LMIP study had a distinctly different objective to this Microdata Review and readers are referred to the LMIP outputs for detailed consideration of data relating to skills planning.

Review of Labour Market administrative data. As part of the review of labour market microdata resources referred to above (Woolfrey, 2013), six administrative datasets of central relevance to labour market research were listed and documented. Of these six datasets, one is administered by the Department of Social Development (National Integrated Social Information System (NISIS) database), one is administered by the South African Social Security Agency on behalf of the Department of Social Development (SOCPEN – see case study section below), and four databases are administered by the Department of Labour (Compensation Fund Database; Employment Equity Submissions Database; Employment Services of South Africa Database; and Unemployment Insurance Fund database). Woolfey provides a brief overview of each of the six administrative datasets in terms of data content and structure, and selected data quality issues. Readers are referred to the Woolfrey report for detailed consideration of these labour market administrative datasets.

The data reviews referenced above each make important contributions to the knowledge base concerning data availability and data usage within and outside of South Africa government. An objective of this Microdata Review study is to draw readers' attention to these existing data reviews but also to contribute to this existing body of knowledge by highlighting a selection of specific administrative datasets collected by a range of data producers. The datasets referred to below represent selected examples of administrative data, and are not intended to be comprehensive or exhaustive. The aim here is to use these examples to demonstrate how various types of administrative data can be used in different and innovative ways to inform thinking around key social and economic challenges.

Example 1: Developing a unified Social Protection dataset. The National Integrated Social Information System (NISIS) database referenced by Woolfrey (2013) represents one strand of DSD's efforts to coordinate the development of a National Integrated Social Protection Information System (NISPIS). NISPIS was initiated to help the government to meet its targets under Outcome 13 "*An*

[34] The van Wyk (2015) study was funded by the European Union as part of the PSPPD-2 programme.

inclusive and responsive social protection system" of the Medium Term Strategic Framework (MTSF). The objectives underpinning the development of NISPIS centre on the need to move from the current situation of multiple 'silo' information management systems for social protection, to a unified electronic reporting database / data warehouse. As such, although DSD is the lead department in the development of NISPIS, various other government departments and agencies are also involved due to having mandates that either directly or indirectly relate to social protection. This is an ongoing programme of work and readers are referred to DSD for future progress updates.

Example 2: Pioneering the use of taxation data. The South African Revenue Service (SARS) is collaborating with the National Treasury (Economic Policy Division) to enable the use of administrative SARS microdata to support economic policy analysis. The data sets utilised include personal income tax, corporate income tax, and value-added tax. The data cover all firms and employees operating within the formal sector. These tax-related microdata permit analysis at the level of the individual firm, rather than aggregate sector, thereby providing new insights into the extent of heterogeneity across firms within any given sector. In a recently published paper by the National Treasury and UNU-WIDER, the authors state that "exploitation of tax administrative record data has clearly become global best practice, and South Africa is, to our knowledge, the first country on the African continent to mount a serious effort to employ these data for the purposes of policy analysis" (National Treasury and UNU-WIDER, 2016, p.1).

Researchers from SALDRU have already commenced analysis of the SARS data for the purpose of the REDI3x3 project (funded by the National Treasury). Orthofer (2016) examines wealth inequality in South Africa using both survey data (NIDS) and administrative tax records from SARS. The SARS data consisted of a previously unpublished dataset of approximately 1.2 million Personal Income Tax (PIT) records for the 2010-2011 tax year. Orthofer highlights some of the advantages of using administrative data to capture high income individuals, but also acknowledges some of the weakness of administrative tax data to capture low income individuals:

> "Although the PIT should provide better information on the top of the [wealth] distribution than the NIDS, the data have other limitations. First, the PIT provides no information on forms of wealth that do not generate taxable investment incomes to the tax filer, such as owner-occupied housing, pension assets or assets held in trusts. Second, the PIT excludes all individuals whose incomes are below the filing thresholds. While non-filers are not of much concern to researchers in advanced economies, they constitute the majority of the population in developing countries. Less than 20 percent of the South African adult population are liable to file income taxes, and less than a tenth of these filers—about one percent of the total adult population—declared any investment incomes at all." (Orthofer, 2016, p.3)

Example 3: Linking housing subsidy and deeds registry datasets. Housing Subsidy System (HSS) data is collected by the Department of Human Settlements (DHS). A housing subsidy is a grant by government to qualifying beneficiaries for housing purposes. One of the DHS's areas of responsibility in the delivery of human settlements relates to the lower end of the market, where it provides housing subsidies to poor people. The funding for the subsidies is provided through the Human Settlement Development Grant (HSDG), which is allocated to provinces. The provincial nature of the housing subsidy allocation process is reflected in the provincial structuring of the HSS data collection and management instrument. The provincial DHS are the data owners for their respective provinces. The provincial HSS databases record a large array of details concerning housing projects supported under the national human settlement programmes. Information is collected about the housing subsidy applicant, the type of subsidy application, the project for which the subsidy is sought, and outputs of the project, including further details of the beneficiaries assisted by the subsidy. The

national DHS does not own the HSS data, but rather has the ability to view copies of the nine provincial datasets and use these copies for analysis at national level. Gordon et al. (2011) were provided with an extract of HSS data from a time point in September 2010 which they matched with housing deed records to explore whether houses funded through subsidies were subsequently registered in the Deeds Registry and, as such, could be used as assets by the beneficiaries of the subsidies. The authors concluded that between 1994 and 2009 almost 3 million subsidy houses were reported as being completed or under construction but, of these, only half were registered with the Deeds Registry (Gordon et al., 2011). The implications of not registering with the Deeds Registry is that half of the HSS beneficiaries were unable to use their property as an asset for any future transactions.

Example 4: Exploring relationships between criminal offender data and the local neighbourhood. Recorded crime data concerning individual criminal events are discussed in a case study in Chapter 4. However, other forms of criminal justice administrative data also exist in South Africa. One particularly pertinent dataset in the context of researching socio-economic challenges is individual level offender data. Breetzke and Horn (2006) secured access to geocoded offender data for the province of Tshwane and used this to calculate a Crime Offender Index at suburb level. The offender data was obtained during March 2006 from the management information systems of all five correctional centres operational in Tshwane at that time, and contained details of the residential address of each individual offender (1,870 offenders in total). The authors examined the spatial associations between their Crime Offender Index and other social, economic and demographic indicators. They conclude that:

> "The location of offenders within Tshwane appears to be associated with the spatial incidence of four broad factors - low social status and income, a large and young family, unskilled earners and high residential mobility. To refer to these factors as criminogenic risk factors would be presumptuous and it is perhaps wiser to suggest that these factors create a more favourable environment for offending, or increase probabilities associated with risk factors." (Breetzke and Horn, 2006, p.187)

Example 5: Investigating intimate partner and child homicides. A certain amount of administrative data on criminal offenders is also collected by SAPS and this has been used in conjunction with administrative data from mortuaries to examine intimate partner homicides. In a recent study on intimate partner homicide, Abrahams et al. (2013) collected information on cause of death from autopsy reports undertaken within mortuaries during the 2009 calendar year, and linked this to information provided by SAPS about the offender (where known), the case outcome, any previous history of inter-partner violence, and the relationship of the victim with the offender. The authors looked particularly at female victims and assessed the overall female homicide rate, the proportion of female homicides that involved intimate partner violence, the proportion that involved rape, and the homicide method (e.g. gunshot). Results were compared against an equivalent earlier study which used mortuary data for 1999 (Abrahams et al., 2009). The authors conclude that the rate of overall female homicide was significantly lower in 2009 than in 1999, but the rate of intimate partner femicide and suspected rape homicide were not significantly different between the two studies.

A similar methodology was adopted by Mathews et al. (2013) in their study of child homicides in which they estimated that South Africa's child homicide rate is more than twice the global estimate. Mathews et al. (2016) further utilised this form of data linkage in their establishment of the Child Death Review (CDR) pilots during 2014. The authors highlight the importance of sharing and linking data from different administrative sources for the purposes of better understanding the processes that lead to child homicide and informing policy responses to tackling this crime:

> "At the core of the CDR process is the multidisciplinary team, comprising representatives from law enforcement, social services, health, forensic pathology and prosecution services who meet retrospectively to share case-specific information and review the circumstances of child deaths. The aim is to gather information about all factors contributing to a child's death, to systematically identify modifiable or remediable factors. CDR teams have been shown to be effective in improving the identification of child deaths due to maltreatment, identifying modifiable causes of death, and using these findings to strengthen policy and service provision." (Mathews et al., 2016, p.895)

This research emphasises the potential value that can be gained from sharing administrative data on individuals (i.e. microdata) and it further emphasises the importance of securing the support of key stakeholders to facilitate such data linkage.

Example 6: Championing higher education data – and use of data in higher education.
The Higher Education Management Information System (HEMIS) is managed by the Department of Higher Education and Training (DHET) and consists of socio-demographic information on enrolled students and on academic staff; student enrolments by qualification type and field of study; staff numbers by qualification and rank; and information on research outputs such as published articles. A key function of this dataset is to inform the subsidy allocations for universities and for planning and monitoring of the higher education sector. As such, DHET uses HEMIS data to validate its Annual Performance Plan and to track progress against the relevant Medium Term Strategic Framework (MTSF) targets. DHET publishes aggregate macrodata through its departmental website[35], but the underlying microdata are not routinely made available for research purposes. The Centre for Higher Education Trust (CHET) has made use of the macrodata and also provides links on the CHET website to enable researchers to download the DHET macrodata in easily accessible forms[36]. In November 2016, CHET convened the third meeting of the Higher Education Research and Advocacy Network in Africa (HERANA). This meeting, which took place in South Africa, was attended by representatives of seven of the eight participating universities: Botswana, Cape Town, Dar es Salaam, Eduardo Mondlane in Mozambique, Ghana, Nairobi, Mauritius, and Makerere in Uganda. The primary objective of HERANA is improve data collection across a group of research-intensive universities in order to strengthen the processes underpinning the generation of new academic knowledge.

Example 7: Harnessing data for infrastructure development in Gauteng. The Gauteng Province has recently established the Lutsinga Infrastructure House in 2016 which they describe as a 'state-of-the-art business intelligence hub that – in line with UN Sustainability Goal 9 – harnesses the power of big data to monitor and ensure the effective delivery of public infrastructure in the Gauteng Province.' It has already 'enabled real-time monitoring of over 430 infrastructure delivery projects' and contains detailed information about construction, refurbishment and maintenance work, as well as the number of jobs created by infrastructure projects, which enables the impact and progress of the work to be monitored.

The selected examples of specific administrative datasets provided above offer a sense of the wide range of research applications that could benefit from enhanced sharing of administrative data. These are examples only, and do not represent an exhaustive list. In the following chapter of this report, three case study datasets are discussed in more detail in order to provide further insights into the content, structure and potential uses of administrative data.

[35] http://www.dhet.gov.za/SitePages/UniversityEducation.aspx
[36] https://chet.org.za/data/sahe-open-data

3.3 Three administrative data case studies

Three case study administrative datasets are selected for further discussion within this Microdata Review:

i. individual beneficiary-level administrative data on recipients of social grants;
ii. individual pupil-level administrative data on children in basic education; and
iii. individual event-level administrative data on recorded crimes.

These three datasets were selected to cover a number of related research objectives. One objective was to ascertain the legal and technical processes that underpin the sharing of sensitive administrative microdata across South African government. To address this first objective, DPME submitted formal data requests for two of the selected administrative datasets: social grant beneficiaries and pupils in basic education. Another objective of the case study section was to document the data structure and data content of one or more large administrative datasets in a way that had previously not been possible. Achieving this second objective was therefore dependent upon DPME being granted permission to access one or more administrative microdata resources for the purpose of this Microdata Review. The third objective was to highlight they ways in which publicly available administrative *macrodata* aggregate statistics can be used in research where microdata are not currently available, and to comment on the key issues relating to the underpinning microdata that researchers should be mindful of. Recorded crime data was selected as a case study to enable this third objective to be addressed as macrodata statistics are regularly published by SAPS but the underlying microdata is not currently shared with researchers.

A formal request was submitted by DPME to the South African Social Security Agency (SASSA) requesting access to a copy of the social grants administrative database at individual beneficiary level. This dataset was chosen for this purpose because: (i) it is a large government administrative database with substantial potential as a research resource but very little information exists in the public domain about data content, structure or quality; (ii) the authors of this Microdata Review have previously been granted access to the data for a specified research project, so an historical precedent exists for data sharing; and (iii) other researchers have also more recently been granted access to the database, so additional precedents for data sharing exist. As discussed below, SASSA provided a copy of the social grants database to DPME for the purpose of this Microdata Review, and the case study section therefore presents details of the variables contained within the provided database.

A formal data request was also submitted by DPME to the Department of Basic Education (DBE) requesting access to a pupil level administrative data on learners schools. This dataset was chosen for this purpose because: (i) valuable literature already exists which details the content and structure of the educational datasets, including the study commissioned by DPME which comprises an audit of educational datasets (van Wyk, 2015); and (ii) the authors of this Microdata Review have prior experience of analysing pupil level administrative data in other international contexts and so are aware of some of the common strengths and weaknesses of educational data. DPME submitted a formal request to DBE requesting access to the educational microdata but this data request was still in process at the time of completion of this Microdata Review report. It was therefore not possible to give a detailed description of the content of the variables in the basic education data in this current Microdata Review.

A formal data request was not submitted to SAPS for the individual level crime data, but the authors have prior experience of submitting an equivalent data request and the processes through which such requests are required to follow. The authors have considerable experience of working with

individual level crime microdata in other international contexts and have worked extensively with the SAPS aggregated macrodata on recorded crime which is freely available via the SAPS website.

Before turning to the three case studies in detail, it is first instructive to consider the official processes through which DPME submitted the data requests to SASSA and DBE. The following text has been provided by DPME:

Box C: Statement from DPME on the process of requesting administrative microdata from other government departments

Considering the processes involved in accessing data from other departments, there currently exists no formal process or protocol at the Department of Planning, Monitoring and Evaluation (DPME) that regulates data requests etc.

DPME adheres to the different processes in place at the individual departments it collaborates with – these processes also vary according to the urgency and nature of the requests. Some departments are more formal and strict when allowing access to their data. For example, while the Department of Basic Education (DBE) has a set of guidelines available on their website for researchers in conducting research in their department, they also consider requests that are more urgent.

In the case of this study, a request was directed to the Director-General of DBE by means of a formal letter signed by the DPME Director-General. Thereafter, relevant persons were put in contact with the research team at DPME, and a meeting was arranged to discuss the project and the need for the data. Following the meeting the Director-General of the DBE wrote, signed and sent a letter addressed to the DPME Research Director in which he granted permission to access the data, provided that the team submitted to general research ethical guidelines, kept the information confidential and shared the research outcomes with them.

From this study it is thus clear that, even though the Promotion of Access to Information Act No 2 of 2000 (PAIA) provides general guidance on access to information, specific consideration should perhaps also be given towards developing a formal protocol and policy document within the DPME that can regulate processes such as these while also protecting the data involved.

DPME, 2016

There are many examples from other international contexts of approaches taken to establish clear guidelines and processes to govern the sharing of sensitive administrative microdata about individual people, households, institutions and events etc[37]. Such processes typically involve a form of contractual agreement between the parties engaging in data sharing, and the need for researchers working with the sensitive data to sign undertakings not to use the data for any purpose other than that for which access has been approved. These contractual undertakings typically also specify in detail the data security measures that must be adhered to in order to satisfy the terms of the data sharing agreements. These examples from international contexts are designed to safeguard the data and protect the confidentiality of the individual people/households etc that form the content of the data. The importance of contractual arrangements governing data sharing is addressed further in the final section of this Microdata Review.

[37] See, for instance, guidelines from the National Institute for Health and Care Excellence (NICE) https://www.nice.org.uk/Media/Default/About/Who-we-are/Policies-and-procedures/Data-sharing-protocols.pdf

Case study 1: Social grants

Information on the beneficiaries of social grants in South Africa is held in the SOCPEN database. SOCPEN is currently owned and managed by the South African Social Security Agency (SASSA), having previously been owned and managed by the Department of Social Development (DSD).

SOCPEN is a very large database containing details of every single beneficiary of a social grant in South Africa. It contained approximately 17 million cases as of 30th September 2016. SOCPEN is a dynamic system in time in that it is continuously updated through the addition of new cases who have commenced receiving social grants and the removal of cases who have ceased receiving social grants.

SASSA provided an anonymised extract of SOCPEN to DPME for the purpose of informing this Microdata Review study. Details of the application process and contractual arrangements are discussed below. However, before turning to consider these processes, it is first helpful to review the content and structure of the SOCPEN database.

Table 3.1 shows the number of social grants in payment as of 30th September 2016.

Table 3.1: Numbers of Social Grants in payment in South Africa, September 2016

Grant type		Number	Percentage of total	Cumulative percentage
Child	CSG	12,026,000	70.8%	70.8%
	FCG	497,000	2.9%	73.7%
	CDG	138,000	0.8%	74.5%
Adult	OAG	3,247,000	19.1%	93.6%
	DG	1,082,000	6.4%	100.0%
	WVG	<1,000	0.0%	100.0%
TOTAL		**16,990,000**		

Source: Authors calculations using SOCPEN database.
Acronyms: DG=Disability Grant; OAG=Old Age Grant; CSG=Child Support Grant; FCG= Foster Child Grant; CDG=Care Dependency Grant; WVG=War Veterans Grant.
Notes: All numbers rounded to nearest thousand; there were less than 1000 cases of War Veterans Grant.

The total of CSG, FCG and CDG amounted to 12.7 million cases of child grants in payment at the end of September 2016, while the total of OAG, DG and WVG equated to 4.3 million cases of adult grants in payment at that same time point. It is therefore evident that approximately three-quarters of the total social grants in payment at the end of September 2016 were for child grants, with the CSG accounting for the vast majority of these child grants. Adult grants accounted for the remaining quarter of total grants, with OAG being considerably more numerous than Disability Grant. There were, in addition, a very small number of beneficiaries of War Veterans Grant. An adult can receive no more than one of these three adult grants at any given time (though could additionally receive Grant in Aid if in need of full-time care due to illness or disability). Children in receipt of CSG cannot receive FCG or CDG; but children in receipt of FCG can additionally receive CDG.

The extract of SOCPEN provided to DPME by SASSA for the purpose of this Microdata Review contained the variables as listed in Table 3.2. It should be noted that only a selection of variables from the entire SOCPEN database was requested by DPME for the purpose of this Microdata Review. For example, certain variables, such as address fields were not requested and were not supplied. In the discussion below on research uses of SOCPEN data, reference is made to additional variables in the main database that were not contained in the extract provided to DPME.

Table 3.2: Variables contained within the SOCPEN extract provided to DPME

Variable name	Variable description
PENSION_NO	Pension No. of the **adult** beneficiary or primary caregiver
BIRTHDATE	Date of birth of the **adult** beneficiary or primary care giver
BEN_AGE	Age of the **adult** beneficiary or primary care giver
BEN_GENDER	Gender of the **adult** beneficiary or primary care giver
GRANTTYPE1	Type of **adult** grant received (OAG, DG, WVG)
CAREGIVER_APPLIC_DATE	Application date of **adult** grant
IdNo	**Child** ID number
Granttype	Type of **child** grant received (CSG, FCG, CDG)
CHILD_GENDER	Gender of **child**
CHILD_APPLIC_DATE	Application date of **child** grant
CHILD_DOB	Date of birth of **child**
REGION	Province of registration
PP_NAME	Primary paypoint
SECONDARY_PAYPOINT	Secondary paypoint

Each row in the SOCPEN database relates to an individual person for whom a social grant is in payment. In the case of an adult receiving OAG, DG or WVG, variables concerning the adult beneficiary are populated, as are the geographical variables (i.e. province and paypoints), but the variables concerning children are not populated. In the case of a child receiving CSG, FCG or CDG, the payment of the grant is made to the adult primary caregiver. As such, the rows in the database concerning child grants contain populated variables for the primary caregiver (PENSION_NO, BIRTHDATE, BEN_AGE, BEN_GENDER) *as well as* populated variables concerning the child in question, plus the geographical variables. As such, the SOCPEN database concerns key demographic variables on the primary caregiver of children receiving child grants, irrespective of whether the primary caregiver is receiving an adult grant. In contrast, the rows in the database concerning adult grants do not contain any information about any children that the adult may be responsible for.

The following variable descriptions relate to the extract of SOCPEN data provided to DPME by SASSA, with an extraction date of 30th September 2016.

PENSION_NO
All rows in the database contained a populated pension number. The pension number uniquely identifies each adult in the SOCPEN dataset and it can be used to link child recipients of child grants to their primary caregiver.

BIRTHDATE and *BEN_AGE*
All rows in the database contained populated variables on date of birth and age for the adult beneficiary/primary caregiver. Across the 10.5 million unique adults/primary caregivers in the database, the mean age was calculated at 46 years (median of 42 years). The minimum age recorded for adult beneficiaries is 16 years. Ninety-nine per cent of the adult beneficiaries were aged 88 years or younger, but approximately 6000 cases reported ages of 100 years or more. The majority of these cases were recorded as receiving OAG, so the ages of most of these cases were plausible. However, at the highest extreme the ages range up to a maximum of 244 years, which is clearly implausible.

BEN_GENDER
All rows in the database contained a populated variable on gender of adult beneficiary / primary caregiver. There were no missing values. As can be seen from Table 3.3, 65% of adult beneficiaries of the OAG were female and 35% male, while the gender split for DG was much closer to parity. The figures for WVG are presented here and show that almost three-quarters of beneficiaries are male

but, as noted above, there were very few cases of WVG in total so these figures should be treated with caution.

Table 3.3: Gender of adult grant beneficiaries, by grant type

Adult Grant	Beneficiary gender (% of total per grant type)		
	Female	Male	Total
OAG	65%	35%	100%
DG	52%	48%	100%
WVG	27%	73%	100%

In addition to revealing the gender differentials in terms of adult grant beneficiaries, SOCPEN also contains details of the gender of primary caregivers receiving child grants on behalf of eligible children. Table 3.4 shows the breakdown for each of the three child grants, and clearly shows that the overwhelming majority of primary caregivers are female.

Table 3.4: Gender of primary care giver recipients of child grants, by grant type

Child Grant	Primary caregiver gender (% of total per grant type)		
	Female	Male	Total
CSG	98%	2%	100%
FCG	94%	6%	100%
CDG	97%	3%	100%

GRANTTYPE1
This variable identifies which of the three adult grants (OA, DG, WVG) is being received by the adult beneficiary. As noted above, the rules of social grant eligibility state that an adult can receive no more than one of these three grants at any given time.

CAREGIVER_APPLIC_DATE
The caregiver application date variable records the date at which the adult beneficiary registered an application for an adult grant (OA, DG, WVG). Of the 4.3 million cases related to adult grant receipt, all but 17 cases have a populated value for caregiver application date. A very small number of cases have application dates recorded as being earlier than or equal to the adult beneficiary's date of birth. It is difficult to validate the caregiver application date values, but a summary of the duration of receipt is provided in Table 3.5, separately for OAG and DG. WVG is excluded due to very low numbers of cases.

The data contained in Table 3.5 show that a quarter of DG beneficiaries have been in receipt of the grant for less than one year, and over half (53%) of DG beneficiaries have been in receipt of the grant for up to and including five years duration. The concentration of beneficiaries with relatively new grant records is likely a function of the way this grant is structured, consisting of a 'temporary disability grant' and a 'permanent disability grant', with people receiving the temporary grant limited to a duration of 12 months grant receipt[38]. A sizeable proportion of the beneficiaries of DG report receipt of the grant for durations of ten years and over. Specifically, 22% of the DG caseload report receipt of the grant for between 10-14 years duration, while a further 7% of the caseload report receipt of the grant for between 15-19 years duration, and 4% report receipt for 20 or more years duration.

[38] http://www.gov.za/services/social-benefits/disability-grant

Table 3.5: Duration of grant receipt by adult grant type

Duration (completed years)	DG			OAG		
	Number	Percentage	Cumulative percentage	Number	Percentage	Cumulative percentage
0	275,000	25%	25%	243,000	7%	7%
1	64,000	6%	31%	231,000	7%	15%
2	72,000	7%	38%	227,000	7%	22%
3	59,000	5%	43%	219,000	7%	28%
4	50,000	5%	48%	198,000	6%	34%
5	49,000	5%	53%	180,000	6%	40%
6	42,000	4%	57%	204,000	6%	46%
7	40,000	4%	60%	188,000	6%	52%
8	39,000	4%	64%	140,000	4%	56%
9	36,000	3%	67%	110,000	3%	60%
10 to 14	238,000	22%	89%	601,000	19%	78%
15 to 19	77,000	7%	96%	412,000	13%	91%
20+	40,000	4%	100%	295,000	9%	100%
TOTAL	1,082,000	---	---	3,247,000	---	---

Note: numbers rounded to nearest thousand

With regard to OAG, the distribution of durations is more evenly spread across the lower durations, with 7% of the total OAG caseload reporting receipt of less than one year, and each addition year of duration accounting for either 6% or 7% of the total caseload up to seven years of duration. Again, there are considerable numbers of beneficiaries that report much longer durations of OAG receipt, with 40% of the total OAG caseload reporting durations of ten years or more.

IdNo

The IdNo variable contains an identification number for each child for whom a child grant is received. All children recorded as receiving a child grant have a populated value in the IdNo field.

Granttype

This variable relates to the type of child grant in payment. All child cases in the database have a populated value in this field. The numbers of child grants recorded in the dataset are shown above in Table X. whilst it is possible for a child to legitimately receive FCG and CDG at the same time, only 0.02% of the children in the SOCPEN database are recorded as receiving these two grants coterminous. The overwhelming majority of children on the SOCPEN database (99.98%) receive only one child grant at the time the data extract was undertaken (30th Sept 2016).

CHILD_GENDER

As can be seen from Table 3.6, the numbers of male and female children receiving child grants are almost identical. Just less than 1% of children in dataset are recorded as gender 'unknown'.

Table 3.6: Gender distribution of child grant recipients

Gender	Number	Percent
Female	6,261,000	49%
Male	6,292,000	50%
Unknown	109,000	1%
Total	12,662,000	100%

Note: numbers rounded to nearest thousand

CHILD_APPLIC_DATE

Of the 12.7 million cases relating to payment of child grants, all but thirteen cases contained a populated date in the child grant application date field. Table 3.7 shows the durations in completed years of the child grants listed in the database. As noted above, CSG has by far the largest caseload and so the columns for total child grants are primarily driven by the CSG figures. However, it is notable from looking across the three grant types that CSG and CDG tend to have relatively similar numbers of grant recipients for the first five or so years of grant duration, whereas FCG exhibits a notably higher number of grant recipients with durations of less than one year when compared to other durations of FCG receipt.

Table 3.7: Duration of grant payment for child grants

duration (completed years)	CSG	FCG	CDG	Total		
	Number	Number	Number	Number	Percentage	Cumulative percentage
0	1,022,000	108,000	19,000	1,150,000	9%	9%
1	1,124,000	53,000	17,000	1,194,000	9%	19%
2	1,270,000	51,000	16,000	1,337,000	11%	29%
3	938,000	45,000	14,000	997,000	8%	37%
4	962,000	49,000	12,000	1,023,000	8%	45%
5	899,000	45,000	11,000	955,000	8%	53%
6	800,000	38,000	10,000	848,000	7%	59%
7	763,000	36,000	9,000	808,000	6%	66%
8	677,000	26,000	7,000	710,000	6%	71%
9	638,000	18,000	6,000	662,000	5%	76%
10 to 14	2,613,000	27,000	16,000	2,656,000	21%	97%
15 to 19	320,000	1,000	1,000	322,000	3%	100%
20+	0	0	0	0	0%	100%
Total	12,026,000	497,000	138,000	12,662,000	100%	---

Note: all figures rounded to nearest thousand

By comparing the child's date of birth with the child grant application date it is possible to examine the length of time between birth and the commencement of the grant application process. Table 3.8 shows time from birth to grant application for all child grants linked to children of under five years of age as at the time of the SOCPEN extract (30th September 2016). All three child grants are considered here together, but it would equally be possible to look at the three grants separately. The numbers show that, perhaps unsurprisingly, relatively few grant applications are made within the first week of a child's life, but that the numbers then increase week-on-week to a peak of 219,00 applications in the week that the child reaches one month old. Almost half (45%) of child grant applications are submitted before the child is ten weeks old, and the vast majority (86%) of applications are submitted before the child reaches one year of age. The remaining 14% of cases applications are almost entirely made within three years of the child's birth. It should be noted, of, course, that the circumstances specified in the eligibility tests for the three child grants (e.g. means-test; foster-child status; disability status) might not be satisfied at the time of birth and therefore there may necessarily be a period of time between birth and grant application. However, statistics such as these – with more detailed interrogation – can potentially help to identify groups of individuals who may not make their application as soon as is legitimately possible. Evidence of this can be used to help support people to make their applications so as to ensure families with eligible children do not miss out on an important income stream.

Table 3.8: Time from birth of child to date of application for child grant, for grants applied for within the last five years

Time from birth to grant application (weeks/years)	Number	Percentage	Cumulative percentage
0w	17,000	0%	0%
1w	99,000	3%	3%
2w	174,000	5%	8%
3w	206,000	6%	14%
4w	219,000	6%	20%
5w	204,000	6%	26%
6w	180,000	5%	31%
7w	175,000	5%	36%
8w	155,000	4%	41%
9w	137,000	4%	45%
10w to 19w	781,000	22%	67%
20w to 29w	324,000	9%	76%
30w to 1yr	332,000	9%	86%
1yr-2yr	310,000	9%	95%
2yr-3yr	121,000	3%	98%
3yr-4yr	50,000	1%	100%
4yr-5yr	15,000	0%	100%
TOTAL	3,500,000	100%	

Note: all figures rounded to the nearest thousand

CHILD_DOB

The child's date of birth field is completed for all cases of child grant receipt. Using this variable it is possible to derive the age of child at the dataset extract date of 30th September 2016. The ages of the children are shown in Table 3.9, where single-year of ages have been grouped together into five-year age bands for simplicity. There are similar numbers of children in each of the age bands from 0-4 through to 10-14, followed by a sharp drop-off in the 15-19 age band. Analysis of the underlying single-year of age statistics shows that there are in fact similar numbers of children in each single-year age band up to and including age 17 (with each single-year of age accounting for 0.5 million children or more), followed by a very sharp drop at ages 18 and above (with less than 50,000 children per age group up to and including age 21). Children in receipt of FCG are eligible to continue to receive the grant between the ages of 18-21 if in education.

Table 3.9: Numbers of child grants by age group of child

Age of child	Number
0 to 4	3,499,834
5 to 9	3,873,630
10 to 14	3,420,982
15 to 19	1,843,390
20+	23,772

REGION

This variable records the province in which the application for the grant was submitted. All cases in the database contain a populated value in this variable. It is evident from Table 3.10 that the largest number of social grant recipients were registered in KwaZulu-Natal whilst the lowest number of social grant recipients were registered in the Northern Cape. These figures simply reflect the

absolute numbers of individuals receiving the grant, and therefore do not take account of the differences in total population between the nine provinces.

Table 3.10: Number of social grants by province

Province	Number	Percent	Cumulative percentage
KwaZulu-Natal	3,869,000	23%	23%
Eastern Cape	2,734,000	16%	39%
Gauteng	2,468,000	15%	53%
Limpopo	2,377,000	14%	67%
Western Cape	1,490,000	9%	76%
Mpumalanga	1,420,000	8%	85%
North West	1,195,000	7%	92%
Free State	983,000	6%	97%
Northern Cape	455,000	3%	100%
Total	16,991,000	100%	---

PP_NAME & SECONDARY_PAYPOINT

There are two further geographical variables contained within the SOCPEN extract provided to DPME by SASSA for this Microdata Review study. These two variables relate to 'paypoints'. Paypoints were traditionally the physical locations where social grant beneficiaries would receive their monetary payments. In theory, paypoint information should give users an indication of the geographical location of grant beneficiaries. However, in reality, the paypoint information contained within SOCPEN is not appropriate for making geographical assessments because of recent advances in the electronic management and payment of social grants in South Africa. Grant beneficiaries now receive their payments direct into a bank account (from which they can withdraw the money from ATMs or purchase goods using a debit card at particular stores), thereby removing the need for beneficiaries to travel to a physical paypoint location. As such, the paypoint information cannot be used to physically locate grant beneficiaries.

The geographical mapping of the distribution of social grant beneficiaries has long been identified as an important research objective in South Africa. Although the paypoint information is not suitable for this purpose, examples are provided below on successful mapping approaches using SOCPEN data that use other geographical information (both internal to SOCPEN and external via data linkage).

Examples of SOCPEN utilisation in government and/or academic research: Mapping the 'take-up' of social grants

SOCPEN is not routinely available to researchers in other government departments or outside of government. The highly sensitive nature of the data contained within the SOCPEN dataset means that SASSA place great importance on preserving beneficiary confidentiality and ensuring the data are used only for legitimate and justifiable purposes.

There are several examples of SOCPEN data being used for research and policy purposes, and there is great scope for additional research uses in the future, subject to SASSA authorisation.

A key example of how SOCPEN has been used in research to support evidence-informed decision making is in the calculation of 'take-up' rates of social grants. The motivation for this topic of research is to enable policy makers to understand the extent to which people who are eligible for social grants are actually receiving them. Given the important role that social grants are known to

play in helping very poor individuals to purchase the most basic necessities, it is imperative that those people who are eligible for the grants do actually receive the grants. Research into levels of 'take-up' of social grants has aimed to assess the extent of non-take-up nationally and sub-nationally for different social grants. One objective has been to ascertain whether there are certain geographical areas in South Africa where 'take-up' rates are particularly low so that these areas can be targeted by schemes to raise awareness within the community of the social grant application process and to assist eligible people to claim the grants for which they are eligible.

An early example of the research into 'take-up' of social grants using SOCPEN data is a project undertaken between 2004 and 2006 by researchers at the University of Oxford as part of the Strengthening Analytical Capacity for Evidence-based Decision-making (SACED) programme of work. The SACED programme was undertaken for the national Department of Social Development (DSD) with funding from the UK DfID Southern Africa. The researchers developed a methodology for estimating rates of 'take-up' of CSG and OAG at national, provincial and municipal levels of aggregation. The 'take-up' rates were estimated by expressing the numbers of beneficiaries receiving each type of social grant as a percentage of the number of people estimated to be eligible for each type of grant. The numbers of beneficiaries receiving social grants were calculated using SOCPEN whilst the estimates of numbers of people eligible were calculated using the 10% sample of the 2001 Census, with appropriate adjustment to reflect population change between 2001 and 2004/2005. In order to calculate the numbers of beneficiaries for each geographical area using the SOCPEN database it was necessary to interrogate the home address fields collected within the database and use this address information to assign municipality and province identifiers to each grant beneficiary. The research reports produced by the team and published by DSD revealed new insights into geographical patterns of 'take-up' of the CSG in 2004 (Noble et al., 2005b), OAG in 2004 (Noble et al., 2006) and CSG in 2005 (Noble et al., 2005a). Geographical patterns of take-up were assessed by province and also by municipality, and it was identified at the time that the areas with the highest eligibility rates were the areas with the lowest take-up rates. Work was also undertaken to longitudinally track social grant beneficiaries as they moved onto, off and between social grants over a period of time (Anttila et al., 2006).

Although the analyses of take-up of social grants using SOCPEN referenced above revealed geographical patterns at municipality level, attempts were also made during that project to map 'take-up' rates down to *ward* level in order to enable even greater geographical detail. The researchers from Oxford University worked closely with GIS experts in DSD and a number of other government departments and non-governmental organisation to explore different options for mapping SOCPEN beneficiaries down to ward level. As the geographical information contained within SOCPEN was insufficiently detailed to permit confident mapping of beneficiaries at ward level, the researchers considered other options, including using data linkage techniques to take the SOCPEN database and attach geographical data from a separate source. Following an intensive period of discussions during the year 2004, a proposal was formulated that entailed linking individual level data from the Independent Electoral Commission (IEC) with individual level SOCPEN data. The IEC data was identified as the best dataset to match with SOCPEN data because it contained information on all registered voters (and therefore captured a large majority of adults in South Africa at that time) and it contained geographical information on the ward of residence of each registered voter. The proposed method of data linkage would be operationalised using the individual ID number, which was a variable common to both the SOCPEN and IEC datasets. A high-level request was submitted by DSD to IEC proposing this programme of data linkage, however, the proposal did not reach fruition within the time constraints of the SACED research programme.

The year 2016 saw a considerable advance in the measurement of take-up of social grants. In late 2016, DSD, SASSA and UNICEF jointly published a report titled '*Removing barriers to accessing child*

grants: Progress in reducing exclusion from South Africa's Child Support Grant' (DSD et al., 2016). The report presented comprehensive findings from a project commissioned by DSD, SASSA and UNICEF and undertaken by the Economic Policy Research Institute (EPRI). The researchers at EPRI succeeded in coordinating the process of linking SOCPEN data to IEC data that had been initiated but not achieved ten years previously. This new study, in which SOCPEN data from 2014 on CSG recipients (caregivers and their children) was linked to IEC data to reveal the recipients' ward of residence, provides extremely valuable insights into the spatial distribution of grant beneficiaries at a detailed geographical level. Furthermore, when the ward level counts of beneficiaries were combined with estimates of eligible children derived from the 2011 Census, the EPRI team were able to generate ward level measures of 'take-up'. This research shows, for the first time, how 'take-up' rates vary *within* municipalities as well as between municipalities. In addition to the geographical analyses highlighted here, the EPRI team also utilised survey microdata from the General Household Survey 2008 to 2014 and the National Income Dynamics Study 2008 to 2012 to undertake a range of additional non-spatial analyses, such as exploring differences in take-up rates between age, population group and income decile. While the report authors acknowledge that their results at ward level should be regarded as estimates rather than definitive true values (due to the methodological processes involved), this research nevertheless represents an important advance in knowledge on the issue of social grant take-up in South Africa.

Looking forwards, SOCPEN data holds great potential as a means of mapping and tracking individuals as they move through the social grants system. The linkage of SOCPEN to IEC data referenced above demonstrates how new insights can be gained by combining two separate datasets. Future data linkage possibilities might include linking SOCPEN data to education data or health data to examine associations between receipt of income from social grants and positive outcomes in other socio-economic dimensions.

Case study 2: Basic education

Education reform has been a key priority in South Africa since the beginning of democracy in 1994, and has played an important role in redressing the injustices of colonial, segregationist and apartheid rule (OECD, 2008). In terms of present-day policy emphasis, Outcome 1 of the government's Medium Term Strategic Framework (MTSF) relates to improving the quality of basic education in the country. The need to monitor progress in raising educational standards and tackling educational inequalities has led to a number of important developments in data capture and data management.

As noted above, DPME recently commissioned an audit of educational data sources, covering administrative sources and survey sources (van Wyk, 2015)[39]. Van Wyk lists a wide range of relevant datasets and discusses the strengths and weaknesses of each, as well as providing a selection of summary statistics to illustrate the data content. Van Wyk's study focuses solely on education data and considers each dataset in detail.

A further study of importance, which is currently ongoing, is the 'Assessment of education department data use in provinces and the formulation of recommendations aimed at improving systems and service delivery outcomes', being undertaken (Gustafsson, 2016a)[40]. Gustafsson's study not only identifies and discusses key educational datasets, but also aims to provide new insights (using a questionnaire approach) into how education data is being used (or why it is not being used)

[39] The study was funded by the European Union as part of the PSPPD-2 programme.
[40] This study forms part of National Treasury's Financial Management Improvement Programme (FMIP) III, funded by the European Union.

to strengthen service delivery within schools (Gustafsson, 2016a). This study builds on earlier work by the Michael & Susan Dell Foundation (in collaboration with national DBE) which examined the extent to which data was being effectively used within South African's school system to drive improved educational performance (Michael and Susan Dell Foundation, 2013). The Michael and Susan Dell Foundation has also developed the 'Data Driven Districts' web portal: https://www.eddashboard.co.za/ which enables users to access various educational statistics and present these graphically to aid interpretation using a 'dashboard' approach.

The national Education Management Information System (EMIS) was introduced in 1995. EMIS is both a management information database and also a physical unit within the national DBE and within each provincial DBE. The keys responsibilities of the national and provincial EMIS units include the production, management and dissemination of basic education data. Until relatively recently, most education data was collected as aggregate data at the school level, for instance through the annual SNAP survey (which captures information from each school on numbers enrolled, numbers of educators and numbers of other adult support staff across various categories) and the Annual School Survey (which also captures information at school level concerning enrolment but additionally includes other information such as numbers of learners who are pregnant). More recently, the emphasis has shifted away from aggregate-level data capture to the new priority of *individual-level* data capture. The need to collect data at the level of the individual has been recognised both in terms of individual *learners* and individual *educators*.

Individual learner data is routinely collected by schools. The national DBE developed an electronic data capture software system, called the South African School Administration and Management System (SA-SAMS), and this has been made freely available to all schools for the purpose of collecting information on learners and on a number of functions of school administration. Currently, this software product is an off-line (i.e. desktop) application. DBE requires schools to submit learner (and other) data to provincial DBE data repositories on a regular basis, but the off-line nature of SA-SAMS necessitates schools sending data on CD/DVD/USB stick. DBE has acknowledged the need to move away from an off-line software product to a web-enabled version, although this is still in the planning stage. Whilst national DBE is making SA-SAMS freely available to all schools, the system has not been adopted across the whole country. A comprehensive set of documentation on SA-SAMS is publicly available via the SA-SAMS website: http://www.sasams.co.za/

In terms of educational achievements of individual learners, SA-SAMS is used in many schools to capture the Annual National Assessment (ANA) scores (which are relatively narrow in scope and are typically used to monitor school outcomes). Many schools also use SA-SAMS as a means of recording the continuous assessment process across the breadth of the curriculum which educators then use to determine the appropriate grade for each learner. It is important to note that the Grade 12 examinations are not currently captured in SA-SAMS, but rather in a completely separate database. Individual learner microdata on Grade 12 examination scores is now available for researchers to access through DataFirst[41]. Gustafsson (2016) used individual learner level Grade 12 examination score data for each year between 2008 and 2015 to examine trends in attainment according to a number of different achievement thresholds (such as the 50%, 60% and 70% mark level). As well as assessing learners' raw test scores, Gustafsson also implemented a mark adjustment technique in an attempt to correct for possible changes in examination difficulty over the analytical time period.

The provincial SA-SAMS databases can be accessed and analysed by national DBE and this has facilitated the establishment of the Learner Unit Record Tracking System (LURITS). The LURITS database consists of a relatively small sub-set of variables about individual learners and their parent(s)/carer(s). LURITS is the mechanism through which national DBE assigns each school-aged

[41] https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/510/related_materials

child a unique individual learner number. The main purpose of LURITS is to enable DBE to track the movement of learners as they transition between schools and to provide accurate up-to-date enrolment numbers and learner data to support evidence-informed strategic planning decisions (van Wyk, 2015)[42]. LURITS also offers the possibility of tracking individual learners' progress over time in curriculum assessments, the ANA and the Grade 12 assessments, although these forms of data linkage are not yet fully established. A number of initiatives are currently underway to improve the quality of the LURITS data, guided by internal DBE data quality reviews (e.g. Gustafsson (2014))

The ability to link learner assessment scores over time (including learners who move school) will greatly expand the analytical opportunities and will therefore contribute to the monitoring of Outcome 1 from the MTSF. For instance, the work of Gustafsson (2016) referenced above may have further benefited from the ability to assess learners' ANA test scores over the period preceding the Grade 12 examinations to enable longitudinal analysis of dynamic learner trends as well as repeated cross-sectional analyses.

Case study 3: Recorded crime

According to the United Nations Office on Drugs and Crime (UNODC), South Africa continues to exhibit one of the highest homicide rates in the world (UNODC, 2011; UNODC, 2013). Unsurprising, as evidenced by Statistics South Africa's recent series of Victim of Crime Surveys (VOCS), crime is persistently highlighted as a major concern amongst the country's population (Statistics South Africa, 2015). The responsibility for tackling South Africa's crime problems does not lie solely with the police service, but rather necessitates contributions from a wide range of government departments and non-governmental organisations to address the causes of crime and the effects of crime.

A great deal of research has been undertaken internationally over the last century or so into the determinants of crime which include a range of other societal challenges such as poverty and inequality. Whilst it is outside the scope of this Microdata Review to critique this extensive criminological literature, there is a strong theme within the literature on the importance of *place* as a determinant of crime. Understanding *where* crimes occur is a first step in seeking to understand why they occur in that location and what can be done to prevent those crimes. As such, *geographically referenced* crime data represents a vital source of evidence in attempts to tackle the problem of crime.

As noted in the earlier sections, crime surveys are not suited to providing evidence on the geographical locations where crimes occur due to difficulty in producing reliable indicators at sub-national level, and the Census contains no information on crime. Geographically referenced crime data is however collected routinely as an administrative microdata source by the South African Police Service (SAPS).

SAPS collects information on a broad spectrum of different crime types, ranging from extremely serious crimes such as murder, to relatively minor crimes such as shoplifting. Each recorded crime is logged on a police database as an individual event and, as such, police recorded crime data represent an *event-based* source of administrative data. The variables in the police database therefore contain details of each criminal event in question, such as the type of crime recorded, the date and time of occurrence, and geographical information on the location of the offence. These

[42] A similar system is already operational in the Western Cape, called the Centralised Educational Management Information System (CEMIS).

data are used by SAPS to identify emerging crime patterns and trends and to guide policing responses[43].

At the time of writing, police recorded crime *microdata* are not made available on a routine basis to researchers or other interested parties. However, SAPS does publish regular *aggregate* crime statistics at police station precinct level for a range of key crime types, and these aggregate statistics are available for download from the SAPS website[44]. Police station precincts represent an operational policing geography rather than a statistical geography and, as such, users should be aware of the implications of this for their analysis. Some of the specific limitations of police precincts as a unit of analysis include: (i) wide variations between precincts in terms of the numbers of people living within the precinct boundaries (ranging from less than 5,000 to over 150,000 people in a precinct) which means it is difficult to compare precincts on a like-for-like basis; (ii) difficulties posed by those large population precincts where important local patterns and trends may be obscured by being aggregated into such a populous unit of analysis; and (iii) changes to police precinct boundaries and names over time in response to policing needs, which poses problems for assessing change over time in crime statistics in these areas of re-alignment of boundaries.

Despite the acknowledged limitations of aggregate crime statistics presented at police station precinct level, these statistics do nevertheless represent a valuable resource for social researchers. For example, Lancaster and Kamman (2016) tested for statistical associations between the murder rate at police precinct level and a range of socio-economic risk factors identified from their review of the criminological literature. Their indicators of socio-economic risk factors were derived from the 2011 Census and were also constructed at police precinct level by apportioning small area level Census statistics to police precincts using a GIS approach. They tested for associations with the murder rate over the 2014/15 year and also over a ten-year period (which required them to deal with the problem of changing police precinct boundaries discussed above). The authors found that, in relation to the single-year murder rate, "police stations in more urban areas, with more informal housing, more people renting property, a higher percentage of orphans, and that are relatively poor compared to the rest of the municipality, tend to have a higher murder rate" (Lancaster and Kamman, 2016, p.32) and, in relation to the ten-year murder rate, "police stations with a higher population density, higher unemployment rates, and lower relative poverty compared to the rest of the municipality, tend to have a higher average murder rate over 10 years" (Lancaster and Kamman, 2016, p.33). McLennan and Noble (forthcoming) also analysed a selection of the crime data and socioeconomic data that had been produced by Lancaster, and supplemented this with a number of additional socio-economic variables, including indicators of spatial inequality[45]. They separately regressed three different measures of violent crime at police precinct level on a range of independent variables. The authors found significant positive associations between all three measures of violent crime and a socioeconomic indicator of 'intensity of exposure to inequality', all other things being equal. In other words, geographical areas where the population suffered high levels of deprivation and experienced high exposure to socioeconomic inequality tended to have higher levels of violence, after controlling for other selected variables.

Although the microdata on recorded crime are not made *routinely* available for research, there have been a small number of instances when a subset of police microdata has been shared with researchers. For example, Horn and Breetzke (2009) were provided with an extract of approximately 1 million individual instances of serious crime across the province of Tshwane covering the period

[43] In addition to collating administrative microdata on recorded crimes, SAPS also collates similar administrative microdata on public order policing interventions and records these events in the Incident Registration Information System (IRIS) database.

[44] See http://www.saps.gov.za/services/crimestats.php

[45] See McLennan et al. (2015), for further details of the inequality measures.

2001-2007 for the purpose of examining the incidence of crime around the Loftus Versfeld stadium in the lead up to the 2010 FIFA World Cup. The authors mapped the crimes over this period of time and examined the patterns and trends in key crime types in both the immediate area of the sports stadium and in proximate surrounding areas. Breetzke (2010) also used individual level recorded crime microdata over the period 2001-2003 for the province of Tshwane to test the social disorganisation theory from the criminological literature. Breetzke provides the following description of the data he received from SAPS: "The information provided [by SAPS] included the geographic location, date and time of day, and type of violent crime committed in Tshwane for the years 2001-2003" (Breetzke, 2010, p.448). The types of analyses undertaken by Breetzke and colleagues highlights some of the many ways that SAPS recorded crime microdata can facilitate innovative research to support evidence-informed decision making, both within the police service and across government and non-governmental bodies more broadly.

More recently, pioneering work has been undertaken at the University of Cape Town in collaboration with the Western Cape Provincial Government, the City of Cape Town and Statistics South Africa to collate SAPS data (mostly for 2014-15) in relation to young people who are victims of or accused of contact crime, as well as area-level incidence of contract crime and property crime per 10,000 population. This information can be accessed via a dedicated portal and can be obtained down to ward level.[46]

Although there are a number of recognised limitations with police recorded crime data (common to all countries, including South Africa), with the two main issues being under-reporting and under-recording of crime, these administrative data nevertheless represent a powerful resource for researchers who wish to examine sub-national crime patterns and trends. In light of the importance of crime data for supporting evidence-informed decision making in South Africa, StatsSA has been working with SAPS since 2011 to improve the quality of these data. This relationship was further cemented in April 2015 with the signing of a Memorandum of Understanding between StatsSA and SAPS for this purpose. As part of this relationship, the Statistician General at StatsSA recently published a statement (reproduced here as Appendix 2 of the Microdata Review) on the 2015/16 SAPS crime statistics following a review of these statistics against selected indicators from the SASQAF under the dimensions of: Methodological Soundness; Accuracy; Comparability and Coherence; Integrity; and Timeliness. This review was undertaken by a Clearance Committee on behalf of the Statistician General. The Statistician General concludes the statement as follows:

> "*My assessment of the [SAPS 2015/16 Crime Statistics] publication taking into account the recommendations of the Clearance Committee is that whilst the publication has not reached the level of official statistics, it is compliant with national statistics and I thus endorse the 2015/16 crime statistics publication and encourage its use by stakeholders. I also thank the leadership of SAPS for the ambition of producing crime statistics quarterly and for consistently aspiring for high quality crime statistics in the country. To this end, as the Statistician General, I stand ready to work with them on assessing capacity and resource requirements for achieving this ambition.*" (Statistics South Africa, 2016)

The collaboration between StatsSA and SAPS on issues of data quality is of relevance to researchers in this field as it may lead not only to improved crime statistics, but also potentially to renewed discussions on how the underlying microdata may be better utilised to support evidence-informed decision making.

Looking forwards, there are some important developments internationally that have relevance for SAPS and the sharing of recorded crime microdata with the research community. For example, in the

[46] See https://youthexplorer.org.za

UK, recorded crime microdata are now available for access by the public via the *police.uk* website: https://www.police.uk/. All police forces in England, Wales and Northern Ireland are required to submit regular extracts of microdata to feed into a database underpinning the *police.uk* website. These individual crime records are then subjected to a process of geographical anonymisation to deliberately introduce a degree of error to the geographical coordinates associated with each crime. The process of anonymisation acts to preserve the confidentiality of victims whilst still enabling users to map the approximate location to a reasonable level of accuracy. Working towards a solution such as this in South Africa would result in a powerful crime dataset which could be made available to the research community.

4 Summary and recommendations

The aim of this Microdata Review 2016 study was to bring together up-to-date information on South African censuses, surveys and administrative datasets. This review builds upon the earlier study undertaken by researchers from the Centre for the Analysis of South African Social Policy (CASASP) at the University of Oxford, which was published in 2007.

The compilation of information concerning *census* and *survey* datasets was undertaken almost entirely as a desk-based exercise, and consisted primarily of searching through existing data repositories and internet resources to obtain the details about the datasets. In this regard, it is evident that considerable advances have been made since the publication of the 2007 report in terms of the comprehensiveness of metadata that is readily available to researchers through existing repositories. Particular recognition should be afforded to DataFirst at the University of Cape Town, from which a sizeable quantity of metadata was extracted and reproduced for this Microdata Review. The South African Data Archive, Statistics South Africa's NESSTAR, and HSRC's Research Data Service also contributed much valuable information concerning a wide range of datasets.

It is important to note that the purpose of this Microdata Review is not to compete with existing sources of metadata, such as DataFirst, but rather to assist researchers and policy makers by bringing together a selection of key features on an array of microdata resources in one easily accessible report. This Microdata Review should therefore be seen as one way for researchers and policy makers to engage with the vast array of census and survey microdata that is currently available in South Africa. Prior to submitting a data request and then working with any particular census or survey microdata resource, it is recommended that the user first consults the relevant data repositories for a more in depth assessment of the dataset (where the user may find more detailed technical notes than was appropriate for inclusion in this Microdata Review 2016 report).

The compilation of information concerning *administrative* datasets did not following the same desk-based approach as for censuses and surveys because there is far less publicly available metadata relating to administrative data. The approach taken with regard to administrative data was therefore to liaise with selected data experts from across government in an attempt to scope out a selection of datasets currently being used to inform decision making and to learn about the strengths and weaknesses of those datasets.

The process of identifying suitable administrative data experts across government and establishing effective lines of communication posed a number of challenges for the research team working on this Microdata Review. These challenges have also been raised in other administrative data audits that have been referenced here, such as the review of datasets relevant to skills planning undertaken by Paterson et al. (2015). One recommendation emerging from this Microdata Review is that further efforts should be made to clarify the primary contact point for each administrative dataset. This would greatly ease the challenges of establishing the effective lines of communication that are needed as a first step to facilitating increased data sharing of between government departments (and indeed with external researchers).

The SOCPEN case study provided in Chapter 3 offers researchers some valuable insights into the content of the database and the possible analytical uses for which it could be utilised. A simple variable list and short variable description can often provide researchers with sufficient background information to enable them to make an initial judgement as to whether the dataset is likely to be suitable for their research objective. This sort of basic data content information is now available in easily accessible forms for a whole multitude of *survey* and *census* datasets through DataFirst etc, and a recommendation of this Microdata Review is that similar efforts should be made to compile

and publish equivalent metadata for key government *administrative* datasets. For instance, for the purpose of this Microdata Review, DBE shared with DPME a list of variables in the LURITS database. This variable list, accompanied by short variable descriptions, would represent an important piece of information for researchers considering the potential utility of LURITS for their research. Ideally all government departments would work towards compiling up-to-date variable lists and variable descriptions for their key administrative datasets and publishing these on their departmental websites.

In terms of recommendations relating to particular administrative datasets, one recommendation is that SASSA and the IEC should continue to collaborate to undertake linkage of the SOCPEN data with IEC geographical codes on ward of residence and that these linked datasets should be made more widely available to researchers via suitable secure settings (e.g. UCT's secure data centre). These linked microdata hold great potential for research and should be highlighted as an example of the added value that can be obtained by combining two separate datasets at individual level.

An additional recommendation relating to a particular administrative dataset is that further work should be undertaken to consolidate the learner level educational microdata across the various different sources, including Grade 12 examinations data. Many new research opportunities will become possible if researchers can access a linked dataset at individual learner level containing attainment data and background socio-economic data at each grade throughout a learner's transition through the stages of basic education. For example, assessing learners' Grade 12 examination scores in the context of their prior attainment in earlier years offers the chance to produce measures of contextual value added as have been developed in other international research (e.g. Wilkinson and McLennan (2010)). It may also be possible to link learners' basic education data with information on their tertiary education trajectories.

The third recommendation in relation to a particular administrative dataset is that SAPS should strive to highlight the potential value of its recorded crime microdata for research purposes and further strive to make these data available for appropriate research projects. The example of the *police.uk* website in the UK referenced above shows how it is possible to make sensitive recorded crime microdata available to researchers in a way that preserves victim confidentiality. The process of geographical anonymisation applied to the crime data can be designed in such a way as to ensure SAPS does not disclose any sensitive victim information while ensuring the data are sufficiently accurate to permit rigorous analytical applications.

The final recommendation of this Microdata Review is that this document should ideally represent the start of a process rather than an end-point of a discrete piece of work. The challenges of establishing lines of communication with data experts can best be overcome through ongoing liaison and partnership working, so efforts should be made to consolidate the existing relationships with data experts and form new relationships with experts from additional government departments. The experience of undertaking this Microdata Review, along with past experiences in South Africa and other international contexts, has highlighted one particular factor that is of central importance for the successful development of administrative data usage and sharing: the importance of *personal interactions* between *data literate* representatives of key stakeholder departments. The process of establishing these personal interactions is ongoing, both between DPME and other government departments, and also between the other government departments themselves. If the Microdata Review can be updated on a regular, ideally annual, basis, this will provide added motivation and justification for engaging with key data experts across government which will, in itself, help to formalise the process of data review and data audit, and contribute towards the goal of facilitating increased researcher access to the multitude of valuable administrative datasets that currently exist (and may be developed in the future) in South Africa.

References

ABRAHAMS, N., JEWKES, R., MARTIN, L. J., MATHEWS, S., VETTEN, L. & LOMBARD, C. 2009. Mortality of women from intimate partner violence in South Africa: a national epidemiological study. *Violence and victims,* 24**,** 546-556.

ABRAHAMS, N., MATHEWS, S., MARTIN, L. J., LOMBARD, C. & JEWKES, R. 2013. Intimate partner femicide in South Africa in 1999 and 2009. *PLoS Med,* 10**,** e1001412.

ALDERMAN, H., BABITA, M., DEMOMBYNES, G., MAKHATHA, N. & OZLER, B. 2002. How low can you go? Combining census and survey data for mapping poverty in South Africa. *Journal of African Economies,* 11**,** 169-200.

ANTTILA, C., BARNES, H., COVIZZI, I., NOBLE, M. & WRIGHT, G. 2006. Dynamics of Social Grant Receipt: The Child Support Grant and Foster Child Grant between 2004 and 2005. Department of Social Development.

BARNES, H., GARRATT, E., MCLENNAN, D. & NOBLE, M. 2011. Understanding the worklessness dynamics and characteristics of deprived areas. Department for Work and Pensions.

BARNES, H., NOBLE, M., DIBBEN, C., METH, C., WRIGHT, G. & CLUVER, L. 2007. South Africa Microdata Scoping Study. Oxford: Centre for the Analysis of South African Social Policy, University of Oxford.

BREETZKE, G. D. 2010. Modeling violent crime rates: A test of social disorganization in the city of Tshwane, South Africa. *Journal of Criminal Justice,* 38**,** 446-452.

BREETZKE, G. D. & HORN, A. C. 2006. Crossing the racial divide: a spatial-ecological perspective of offenders in the City of Tshwane Metropolitan Municipality, South Africa. *GeoJournal,* 67**,** 181-194.

DAAS, P. J., OSSEN, S. J., TENNEKES, M. & BURGER, J. Evaluation and visualisation of the quality of administrative sources used for statistics. Paper for the European Conference on Quality in Official Statistics, 2012.

DEMOMBYNES, G. & OZLER, B. 2006. Crime and local inequality in South Africa. *In:* BHORAT, H. & KANBUR, R. (eds.) *Poverty and Policy in Post-Apartheid South Africa.* Cape Town: HSRC Press.

DEPARTMENT OF HEALTH AND SOCIAL SERVICES 1999. Poverty in the Western Cape: An analysis of poverty in the Western Cape as enumerated in teh 1996 Census. Cape Town: Department of Health and Social Services.

DEPARTMENT OF THE PREMIER OF THE WESTERN CAPE 2005. Measuring the state of development in the Province of the Western Cape. Cape Town: Western Cape: Department of the Premier.

DESAI, T. & COWELL, F. 2006. ESRC Review of Data Resources and Needs
Economic and Social Research Council.

DSD, SASSA & UNICEF 2016. Removing barriers to accessing Child Grants: Progress in reducing exclusion from South Africa's Child Support Grant. . Pretoria: UNICEF South Africa.

ELLIOT, M., MACKEY, E., O'HARA, K. & TUDOR, C. 2016. The Anonymisation Decision-Making Framework. UKAN, University of Manchester.

EVANS, M., NOBLE, M., WRIGHT, G., SMITH, G. A. N., LLOYD, M. & DIBBEN, C. 2002. Growing Together or Growing Apart? Geographic Patterns of Change in IS and JSA-IB Claimants in England 1995-2000. Bristol: The Policy Press.

EVANS, M. C. & NOBLE, M. 2001. *Changing Fortunes: geographic patterns of income deprivation in the late 1990s*, Department for Transport, Local Government and the Regions.

GORDON, R., BERTOLDI, A. & NELL, M. 2011. Exploring the Performance of Government Subsidised Housing in South Africa.: Finmark Trust.

GUSTAFSSON, M. 2014. Report on the state of the Limpopo LURITS data 2011 to 2013. Department of Basic Education.

GUSTAFSSON, M. 2016a. Towards better generation and use of data within the basic education sector: Literature review and interview tool.

GUSTAFSSON, M. 2016b. Understanding trends in high-level achievement in Grade 12 mathematics and physical science.: Department of Basic Education.
HORN, A. & BREETZKE, G. Informing a crime strategy for the FIFA 2010 World Cup: a case study for the Loftus Versfeld Stadium in Tshwane, South Africa. Urban forum, 2009. Springer, 19-32.
LANCASTER, L. & KAMMAN, E. 2016. Risky localities: Measuring socioeconomic characteristics of high murder areas. *SA Crime Quarterly***,** 27-35.
MATHEWS, S., ABRAHAMS, N., JEWKES, R., MARTIN, L. J. & LOMBARD, C. 2013. The epidemiology of child homicides in South Africa. *Bulletin of the World Health Organization,* 91**,** 562-568.
MATHEWS, S., MARTIN, L. J., COETZEE, D., SCOTT, C., NAIDOO, T., BRIJMOHUN, Y. & QUARRIE, K. 2016. The South African child death review pilot: A multiagency approach to strengthen healthcare and protection for children. *SAMJ: South African Medical Journal,* 106**,** 895-899.
MCINTYRE, D., MUIRHEAD, D., GILSON, L., GOVENDER, V., MBATSHA, S., GOUDGE, J., WADEE, H. & NUTUTELA, P. 2000. Geographic patterns of deprivation and health inequalities in South Africa: Informing public resource allocation strategies. University of Cape Town, University of Witwatersrand, London School of Tropical Hygiene and Medicine, and National Department of Health.
MCLENNAN, D., BARNES, H., NOBLE, M., DAVIES, J., GARRATT, E. & DIBBEN, C. 2010. English Indices of Deprivation 2010. London, UK: Department for Communities and Local Government.
MCLENNAN, D. & NOBLE, M. forthcoming. The relationship between violence, poverty and exposure to socio-economic inequality at a local level across South Africa.
MCLENNAN, D., NOBLE, M. & WRIGHT, G. 2015. Developing a spatial measure of exposure to socio-economic inequality in South Africa. *South African Geographical Journal***,** 1-21.
MICHAEL AND SUSAN DELL FOUNDATION 2013. Success by Numbers: How using data can unlock the potential of South Africa's R-12 public school system. Michael and Susan Dell Foundation.
MINISTRY OF JUSTICE & DEPARTMENT FOR WORK AND PENSIONS 2011. Offending, employment and benefits - emerging findings from the data linkage project.
NATIONAL TREASURY & UNU-WIDER 2016. Growth and development policy: New data, new approaches, and new evidence. Part I: South Africa.: National Treasury and UNU-WIDER.
NOBLE, M., BARNES, H., WRIGHT, G., MCLENNAN, D., AVENELL, D., WHITWORTH, A. & ROBERTS, B. 2009. The South African Index of Multiple Deprivation 2001 at Datazone level. Pretoria: Department of Social Development.
NOBLE, M., BARNES, H., WRIGHT, G. & NOBLE, S. 2006. The Old Age Grant: A Sub-Provincial Analysis of Eligibility and Take Up in January 2004. Department of Social Development.
NOBLE, M., MCLENNAN, D., WILKINSON, K., WHITWORTH, A., BARNES, H. & DIBBEN, C. 2007. The English Indices of Deprivation 2007. London: Communities and Local Government.
NOBLE, M., SMITH, G., PENHALE, B., WRIGHT, G., DIBBEN, C., OWEN, T. & LLOYD, M. 2000. Measuring multiple deprivation at the small area level: The Indices of Deprivation 2000. London: Department of the Environment, Transport and the Regions.
NOBLE, M., WRIGHT, G., BARNES, H., NOBLE, S., NTSHONGWANA, P., GUTIERREZ-ROMERO, R. & AVENELL, D. 2005a. The Child Support Grant: A Sub-Provincial Analysis of Eligibility and Take Up in January 2005. Department of Social Development.
NOBLE, M., WRIGHT, G., BARNES, H., NOBLE, S., NTSHONGWANA, P., GUTIERREZ-ROMERO, R., MCLENNAN, D. & AVENELL, D. 2005b. The Child Support Grant: A Sub-Provincial Analysis of Eligibility and Take Up in January 2004. Department of Social Development.
NOBLE, M., WRIGHT, G., DIBBEN, C., SMITH, G. A. N., MCLENNAN, D., ANTTILA, C., BARNES, H., MOKHTAR, C., NOBLE, S., AVENELL, D., GARDNER, J., COVIZZI, I. & LLOYD, M. 2004. The English Indices of Deprivation 2004. London: Neighbourhood Renewal Unit, Office of the Deputy Prime Minister.
NOBLE, M., ZEMBE, W., WRIGHT, G. & AVENELL, D. 2013. Multiple Deprivation and Income Poverty at Small Area Level in South Africa in 2011. *Cape Town: Southern African Social Policy Research Institute and Southern African Social Policy Research Insights (SASPRI)*.

OECD 2008. Reviews of National Policies for Education: South Africa. OECD.
ORTHOFER, A. 2016. Wealth inequality in South Africa: evidence from survey and tax data. *REDI3x3 Working Papers.* 15 ed.
PATERSON, A. & VISSER , M. 2016. Utilisation of administrative and research databases in government departments: Providing the platform for skills planning. *Development Southern Africa,* 33**,** 328-342.
PATERSON, A., VISSER, M., ARENDS, F., MTHETHWA, M., TWALO, T. & NAMPALA, T. 2015. High-Level Audit of Administrative Datasets.: Labour Market Intelligence Partnership (LMIP).
ROYAL STATISTICAL SOCIETY 2016. The Data Manifesto (Short Version) Royal Statistical Society.
SMITH, T., NOBLE, M., NOBLE, S., WRIGHT, G., MCLENNAN, D. & PLUNKETT, E. 2015. The English Indices of Deprivation 2015. London: Department for Communities and Local Government.
STATISTICS SOUTH AFRICA 2010a. South African Statistical Quality Assessment Framework (SASQAF). 2 ed. Pretoria: Statistics South Africa.
STATISTICS SOUTH AFRICA 2010b. South African Statistical Quality Assessment Framework (SASQAF): Operational Standards and Guidelines. 1 ed. Pretoria: Statistics South Africa.
STATISTICS SOUTH AFRICA 2014. The South African MPI: Creating a multidimensional poverty index using Census data. Pretoria: Statistics South Africa.
STATISTICS SOUTH AFRICA 2015. Victims of Crime Survey 2014/15. Pretoris: Statistics South Africa.
STATISTICS SOUTH AFRICA 2016. Statistician General's statement on the 2015/16 Crime Statistics Processes based on the Clearance Committee Assessment Report. *In:* AFRICA, S. S. (ed.).
UKSA 2014. Exposure draft of a report from the UK Statistics Authority: Quality Assurance and Audit Arrangements for Administrative Data. United Kingdom Statistics Authourity.
UNODC 2011. UNODC Global Study on Homicide 2011. United Nations publication
UNODC 2013. UNODC Global Study on Homicide 2013. United Nations publication
VAN WYK, C. 2015. An overview of Education data in South Africa: an inventory approach. *Stellenbosch Economic Working Papers: 19/15.* University of Stellenbosch.
WESTERN CAPE GOVERNMENT 2013. Promoting the use and Sharing of Administrative data.
WILKINSON, K. & MCLENNAN, D. 2010. Narrowing the gap? Analysing the impact of the New Deal for Communities Programme on educational attainment. Communities and Local Government, UK.
WILKINSON, K., WHITWORTH, A. & MCLENNAN, D. 2010. Evaluation of the National Strategy for Neighbourhood Renewal: Improving educational attainment in deprived areas. London: Communities and Local Government.
WOOLFREY, L. 2013. South African Labour Market Microdata Scoping Study. University of Stellenbosch.

Appendix 1: Census and Survey dataset descriptions

Introduction

This Appendix contains detailed information on the key South African Census and survey datasets. The section on Census data also contains details of the two inter-censal Community Surveys. The section on survey data contains only those surveys that have been updated or have been introduced since the 2007 study. Please see the original 2007 report (Barnes et al, 2007) for full details of the historical datasets. . The descriptions of new and updated survey data are organised in alphabetical order. Where possible the following information is provided for the new and updated surveys:

- Name and principal investigator;
- Year(s);
- Area(s) of interest
- Brief description of the data source
- Availability of associated documentation and data descriptions;
- Nature of any conditions that are stipulated by data owners
- Contact details for information and data.

This information reflects the current state of South African microdata and documentation. All documentation is in English. The main focus of this current report is on survey and census micro-datasets that have become available since the publication of the Microdata Review 2007 study (Barnes et al, 2007); this includes data surveys that have been in existence since before 2007 but which have subsequently been updated post-2007, plus new surveys that have been introduced since 2007. Though every effort has been made to provide a comprehensive list of datasets, there may be some additional surveys that have not been listed here. It is hoped that this document will continue to be regularly updated and that any additions or amendments could therefore be incorporated.

The contents of the pages below are largely drawn from documentation that is in the public domain relating to the listed datasets. Most of the sections in each of the tables are direct quotes from the official metadata or related documentation. For ease of use, quotations and page numbers are not given, though references to the sources of information are supplied.

Census and Community Survey data

In this section the focus is explicitly on the microdata resources available to researchers from the national Censuses and inter-censal Community Surveys since 1994. The Census microdata resources detailed here consist of the 10% sample of the Census dataset. In addition to the 10% sample of the Census microdata resource, StatsSA also publishes a wealth of *macrodata* derived from the full 100% Census microdata. Please see www.statssa.gov.za for further details of the Census macrodata and indeed macrodata derived from the inter-censal Community Surveys.

Name	***Census 2011: 10% Sample***
Principal investigator	Stats SA
Year(s)	2011
Area(s) of interest	Demography; housing; labour market; economy; education; transport; health
Source(s) of data description provided here	StatsSA 2011 Census website: http://www.statssa.gov.za/?page_id=3839
Brief description	*Summary* The aim of a census is to provide government, the private sector, and academic and research institutions with information which is essential for policy development, planning, monitoring, and evaluation of development projects and informed decision-making. Census 2011 was the third democratic census to be conducted in South Africa. Census 2011 specific objectives included: • To provide statistics on population, demographic, social, economic and housing characteristics; • To provide a base for the selection of a new sampling frame; • To provide data at lowest geographical level; and • To provide a primary base for the mid-year projections. *Methodology* The adopted enumeration method for Census 2011 was canvassing; whereby the enumerator conducted face-to-face interviews with the respondent while simultaneously completing the questionnaires. In exceptional circumstances, however, households that preferred to enumerate themselves were given self-enumeration guides that outlined procedures on how to complete the household questionnaire. Self-enumeration guides were provided in various languages to guide all households that chose this enumeration method. The data processing included the storage of boxes, data capturing, editing, tabulation and analysis. Information received from questionnaires collected during fieldwork was converted into data represented by numbers or characters. The two processes used for this conversion were manual (key-entry) and scanning. The majority of census questionnaires were scanned. Manual entry was used only in instances of damaged questionnaires that could not be scanned. Census data are characterised by numerous errors ranging from content to data processing. In order to detect and minimise some of the errors, the automated error detection and correction method was used based on a predefined set of editing rules

	(specifications). The purpose of editing was to make processed data complete and internally consistent, while making a minimum number of changes. *Sampling* No details given in Metadata *Weighting* No details given in Metadata *Confidentiality* The confidentiality clause found in the Statistics Act (Act No. 6 of 1999) was printed on the covers of the three types of Census 2011 questionnaires and in all manuals of census field staff. The clause states that any person disclosing confidential information will be liable to a fine of up to R10,000, imprisonment, or both. Statistics South Africa has the responsibility to ensure that all information collected from the households is not disclosed to any unauthorised persons. To uphold this responsibility, every Statistics South Africa official, including census field staff, is legally bound to sign the Oath of Confidentiality, which states that they are never to disclose any information gathered in the course of their duties to Statistics South Africa to any unauthorised person, even after their employment is terminated. Furthermore, the information collected is aggregated into tables and statistical information that cannot be traced back to any individual. For micro data, the respondent details are removed, and the content of the information is reduced and modified. For data that are to be tabulated, cells are collapsed or suppressed, particularly when they are sensitive. *Data structure* The Census 2011 data will be in four files: • Person • Household • Mortality • Questionnaire information The files are flat, ASCII, fixed-field files, with one line of given length per record. This format was chosen so as to make the data usable with as many statistical programs as possible, and thus accessible to as wide a range of people as possible.
Availability of data descriptions	A wide variety of supporting technical documentation, including questionnaires and metadata, is available from the Census 2011 website: http://www.statssa.gov.za/?page_id=3839
Conditions	The information products and services of Stats SA are protected in terms of the Copyright Act, 1978 (Act 98 of 1978). As the State President is the holder of State copyright, all organs of State enjoy unhindered use of the Department's information products and services, without a need for further permission to copy in terms of that copyright. Where a copy of the information is made available to any third party outside the State, the third party must be made aware of the existence of State copyright and ownership of the information by the State. The State (through Stats SA) retains the full ownership of its information, products and services at all times. Access to information does not give ownership of the information to the client. The use of any data is subject to acknowledgement of Stats SA as the supplier and owner of copyright. Users may apply or process the data, provided Stats SA is acknowledged as the original source of the data; that it is specified that the analysis is the result of the user's independent

	processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. SADA conditions also apply to data obtained from SADA.
Contact for Information and Data Supply	For information: info@statssa.gov.za http://www.statssa.gov.za/ For Data http://interactive.statssa.gov.za:8282/webview/

Name	***Census 2001 10% Sample***
Principal investigator	Stats SA
Year(s)	2001
Area(s) of interest	Demography; housing; labour market; economy; education; transport; health
Source(s) of data description provided here	Drawn from StatsSA web pages (e.g. http://www.statssa.gov.za/?page_id=3905)
Brief description	*Summary* A 10% unit level sample drawn from Census 2001. People present in the country who were living in households or communal living quarters or who were homeless on the night of 9-10 October were counted. *Dataset* The 10% sample comprises six files: households, persons, mortality, geography, household imputation flags and person imputation flags. All variables as per the questionnaire are included in the 10% sample, as well as derived variables and imputation flags (see below). Enumeration Area numbers are excluded to preserve confidentiality. Geographic type is excluded from the final sample. Instead two additional geographical variables are supplied: urban and rural Census '96 classification and size and density of locality. *Sampling* The sample was drawn from the full Census as follows: *Households* (948,592 records) - a 10% sample of households in housing units, and a 10% sample of collective living quarters (both institutional and non-institutional) and the homeless. *Persons* (3,725,655 records) - a sample consisting of all persons in the households and collective living quarters, and the homeless, drawn from the samples. *Mortality* (36,267 records) - a sample consisting of all mortality information for the households in housing units drawn in the 10% sample of households. *Weighting* Both the 10% household and person sample files contain a weight variable. This weight variable is the adjustment factor for undercount (for households or persons as appropriate) multiplied by 10 to inflate the 10% samples to the relevant population. In the person records, aggregated totals of sparsely populated codes, such as very old ages, might differ substantially from real totals due to sampling fluctuations – no scaling of the weights was done. In the household records aggregated totals will be approximately equal to real totals. Mortality was not adjusted for undercount and therefore there is no weight variable. An Excel file is available with four worksheets showing the adjustment factors for persons and households at municipality and provincial level, which can be used to calculate the 100%. If required, standard errors for each variable can be calculated by Stats SA. *Geography* The South African geographical structure for the 10% sample consists of the following geographical entities, which fit into different geographical hierarchical levels: South

	Africa, Province, District Council (Category C) or Metropolitan Area (Category A), Magisterial Districts, Local Municipality (Category B), or District Management Area. While the structure is intended to be hierarchical, South Africa's geography has cross-boundary entities, which complicate the picture. For example, there are eight municipalities which lie across provincial boundary lines. Users are advised to bear this in mind when choosing the appropriate hierarchy. For example, for the City of Tshwane, which lies in two provinces, one would not use the provincial hierarchy. *Imputation* Imputation was used to allocate values for unavailable, unknown, incorrect or inconsistent responses. A combination of both logical imputation and hot deck imputation (dynamic imputation) was used. Undetermined values were used for only a few variables in a few cases, such as industry and occupation. *Confidentiality* In order to preserve confidentiality the lowest geographical level that unit records can be linked to is municipality. As further assurance of the confidentiality of the data, municipalities with 200 or fewer households are logically grouped with adjacent municipalities.
Availability of data descriptions	The following detailed documentation is available from the Census 2001 website http://www.statssa.gov.za/?page_id=3905 Metadata - for households, persons, mortality, geography and imputation files Code lists - country of birth and citizenship, religion, occupation and industry Questionnaires Record layouts Concepts and definitions How the count was done
Conditions	The information products and services of Stats SA are protected in terms of the Copyright Act, 1978 (Act 98 of 1978). As the State President is the holder of State copyright, all organs of State enjoy unhindered use of the Department's information products and services, without a need for further permission to copy in terms of that copyright. Where a copy of the information is made available to any third party outside the State, the third party must be made aware of the existence of State copyright and ownership of the information by the State. The State (through Stats SA) retains the full ownership of its information, products and services at all times. Access to information does not give ownership of the information to the client. The use of any data is subject to acknowledgement of Stats SA as the supplier and owner of copyright. Users may apply or process the data, provided Stats SA is acknowledged as the original source of the data; that it is specified that the analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. SADA conditions also apply to data obtained from SADA.
Contact for Information and Data Supply	For information: info@statssa.gov.za http://www.statssa.gov.za/ For data http://interactive.statssa.gov.za:8282/webview/

Name	***Census 1996***
Principal investigator	Stats SA
Year(s)	1996
Area(s) of interest	Demography; economy; education; labour market; housing; health
Brief description	*Summary* A 10% unit level sample drawn from Census '96, the first count of all citizens of South Africa. All people present in the country on the night of 9-10 October were counted. Questionnaires were made available in all 11 official languages. *Dataset* A 10% unit level sample of all households (excluding special institutions and hostels) and all persons as enumerated in Census '96 in South Africa. This sample of actual Census records reflects some 50 categories at individual level and 25 at household level, including weights. *Sampling* The sample was drawn as a 10% systematic sample of households from the Census household file. The 10% person level sample was obtained by including all persons in these households plus the persons drawn in independent 10% systematic samples of all persons in special institutions and hostels. The Census household records were explicitly stratified according to province and District Council. Within each District Council the records were further implicitly stratified by local authority and EA type. Within each implicit stratum the household records were ordered according to the unique seven-digit census Enumerator Area number, of which the first three digits are the (old) Magisterial District number. Different terms are used for the local authority boundaries in different parts of the country. There are Transitional Local Councils (TLCs); Transitional Rural Councils (TRCs); Local Authority Councils; Metropolitan Sub-Structures; Metropolitan Local Councils; Rural Local Councils; District Councils (DCs); Transitional District Councils and Regional Councils. To ensure confidentiality within the 10% sample, a local authority had to have a minimum of 2,000 households. As many local authorities had fewer than this number, they had to be grouped together to ensure that the minimum number of households was met. For this purpose, hostel dwellers were treated as single person households. Local authorities with less than 2000 households were pooled with other local authorities based on the following principles: • *All provinces except KwaZulu/Natal and North West*: A TLC with less than 2,000 households was grouped with the TRC within which the TLC was located. In cases where the TRC was big enough to stand on its own but the TLC's within its boundaries were too small, the sample was drawn in such a way that the TRC can be analysed either on its own or together with other TLCs within its boundaries. Where a TRC plus all the TLCs within its boundaries were less than the minimum of 2,000 households the TRC (including the TLCs within its boundaries) was pooled with the adjacent TRC. In a few cases the required minimum of 2,000 households could not be achieved when all the local authorities within a DC were pooled together. In such a case no further implicit stratification within the DC was done.

	• *KwaZulu/Natal*: The equivalent to a DC in KwaZulu/Natal is known as a Regional Council. There are no rural councils in KwaZulu/Natal. Smaller local authorities could therefore not be pooled with the rural council within which boundaries it falls. Where such TLCs were adjacent to another TLC they were pooled to form one stratum. In two cases three TLCs were pooled to form one stratum. In all cases the TLCs that were pooled are adjacent to each other. • *North West*: The TRCs in the North West do not encompass TLCs as is the case in other provinces. The area between TLCs/TRCs in the North West is part of the relevant DC. Small TLCs/TRCs in the North West were either pooled with adjacent local authorities or they were pooled with the relevant DC. *Weighting* Both the 10% household sample file and the 10% person sample file contain a weight variable. This weight variable is the adjustment factor for undercount (for households or persons as appropriate) multiplied by 10 to inflate the 10% sample to the population.
Availability of data descriptions	The 1996 Census pages have now been archived by StatsSA, although they are still currently accessible via this web link: https://apps.statssa.gov.za/census01/Census96/HTML/default.htm
Conditions	The information products and services of Stats SA are protected in terms of the Copyright Act, 1978 (Act 98 of 1978). As the State President is the holder of State copyright, all organs of State enjoy unhindered use of the Department's information products and services, without a need for further permission to copy in terms of that copyright. Where a copy of the information is made available to any third party outside the State, the third party must be made aware of the existence of State copyright and ownership of the information by the State. The State (through Stats SA) retains the full ownership of its information, products and services at all times. Access to information does not give ownership of the information to the client. The use of any data is subject to acknowledgement of Stats SA as the supplier and owner of copyright. Users may apply or process the data, provided Stats SA is acknowledged as the original source of the data; that it is specified that the analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. SADA conditions also apply to data obtained from SADA.
Contact for Information and Data Supply	For information: info@statssa.gov.za http://www.statssa.gov.za/ For Data: http://interactive.statssa.gov.za:8282/webview/

Name	***Community Survey 2016***
Principal investigator	Statistics South Africa
Year(s)	2016
Area(s) of interest	Demography; housing; labour market; economy; education; transport; health
Source(s) of data description provided here	http://www.statssa.gov.za/?page_id=6283 http://cs2016.statssa.gov.za/
Brief description	*Summary* The Community Survey 2016 (CS) is a large-scale survey that happens in between Censuses 2011 and 2021. The main objective is to provide population and household statistics at municipal level to government and the private sector, to support planning and decision-making. The last Community Survey was conducted in 2007. This household based survey is one of the few available data sources providing data at municipal level. The survey remains one of the main data sources that provide indicators at national, provincial and municipal levels for planning and monitoring the performance of specific development programmes such as education, health, sanitation, water supply, housing and transport. In addition, the survey provides demographic information critical in understanding population-development nexus. The objective of the community survey was thus to provide population estimates as well as household characteristics. The information will be used to inform Integrated Development Plans and infrastructure investment budgeting. *Methodology* Stats SA is visited approximately 1.3 million sampled households across the country. The target population for CS 2016 was non-institutional population residing in private dwellings in the country. *Sampling* The sampling methodology is set out in the CS2016 Technical Report (http://cs2016.statssa.gov.za/?portfolio_page=cs-2016-technical-report-web). The sample design for CS 2016 was a stratified single stage sample design. At EA level, all in scope EAs were included in the sample and a sample of dwelling units was taken within each EA (i.e. there was no sub-sampling of EAs). The EA frame was based on the Census 2011 information. The updated dwelling unit (DU) frame was constructed by the Geography team using geo-referenced spatial systems. The final dataset comprises 3,328,867 persons in 984,627 households.
Availability of data descriptions	Metadata is available at http://interactive.statssa.gov.za:8282/metadata/surveys/CS2016/Community%20Survey%202016%20Metadata.pdf For data descriptions Visit http://interactive.statssa.gov.za:8282/webview/
Conditions	Visit http://interactive.statssa.gov.za:8282/webview/
Contact for Information and Data Supply	For Information CS2016@statssa.gov.za http://cs2016.statssa.gov.za/ For Data http://interactive.statssa.gov.za:8282/webview/

Name	***Community Survey 2007***
Principal investigator	Statistics South Africa
Year(s)	2007
Area(s) of interest	Demography; housing; labour market; economy; education; transport; health
Source(s) of data description provided here	http://www.statssa.gov.za/?page_id=3914
Brief description	*Summary* The Community Survey 2007 is a large-scale survey that was conducted between the 2001 and 2011 Censuses. The purpose of Community Survey 2007 was to collect inter-censal information on the trends and level on demographic and socio-economic data; the extent of poor households; access to facilities and services; levels of employment/unemployment; in order to assist government and private sector in planning, evaluation and monitoring of programmes and policies. Approximately 280,000 households nationwide were enumerated. Enumeration for the purpose of the Community Survey is a process of collecting demographic and other information from individuals within an enumeration area. An enumerations area (EA) is the smallest geographical unit (piece of land) into which the country is divided for enumeration purposes. Enumeration areas contain between 100 and 250 households. The main objectives of the survey were to: • provide data at lower geographical levels than existing household-based surveys; • build human, management and logistical capacities for Census 2011; and • establish a primary base for a mid-year population projection. The project strove for maximum coverage and an acceptable degree of quality data. Quality was determined by the extent to which the listing was done correctly, information was correctly recorded on the questionnaires by enumerators, the extent of coverage, and how the data were captured, and how all inconsistencies were eliminated through editing. *Methodology* The design of the CS questionnaire was household-based and intended to collect information on up to 10 people per household. It was developed in line with the household-based survey questionnaires conducted by Stats SA. The questions were based on the data items generated out of a consultation process (see CS2007 Technical Report for more details of this consultation process). Both the design and questionnaire layout were pre-tested in October 2005 and adjustments were made for the pilot in February 2006. Further adjustments were done after the pilot results had been finalised. *Sampling* The sampling approach consisted of two stages, namely the selection of enumeration areas, and the selection of dwelling units. Each municipality was considered a unique stratum. The stratification is done for those municipalities classified as category B

municipalities (local municipalities) and category A municipalities (metropolitan areas) as proclaimed at the time of Census 2001. However, the newly proclaimed boundaries as well as any other higher level of geography such as province or district municipality, were considered as any other domain variable based on their link to the smallest geographic unit – the enumeration area.

The Census 2001 enumeration areas were used as the sampling frame because they gave a full geographic coverage of the country without any overlap. Although changes in settlement type, growth or movement of people have occurred, the enumeration areas assisted in getting a spatial comparison over time. Out of 80,787 enumeration areas countrywide, 79,466 were considered in the frame. A total of 1,321 enumeration areas were excluded (919 covering institutions and 402 recreational areas). On the second level, the listing exercise yielded the dwelling frame which facilitated the selection of dwellings to be visited. The dwelling unit is a structure or part of a structure or group of structures occupied or meant to be occupied by one or more households. Some of these structures may be vacant and/or under construction, but can be lived in at the time of the survey. A dwelling unit may also be within collective 5 living quarters where applicable (examples of each are a house, a group of huts, a flat, hostels, etc.).

The EAs within each municipality were ordered by geographic type and EA type. The selection was done by using systematic random sampling. The criteria used were as follows:

- In municipalities with fewer than 30 EAs, all EAs were automatically selected.
- In municipalities with 30 or more EAs, the sample selection used a fixed proportion of 19% of all sampled EAs. However, if the selected EAs in a municipality were less than 30 EAs, the sample in the municipality was increased to 30 EAs.

The second level of the sampling frame required a full re-listing of dwelling units. The listing exercise was undertaken before the selection of DUs. The adopted listing methodology ensured that the listing route was determined by the lister. This approach facilitated the serpentine selection of dwelling units. The listing exercise provided a complete list of dwelling units in the selected EAs. Only those structures that were classified as dwelling units were considered for selection, whether vacant or occupied. This exercise yielded a total of 2,511,314 dwelling units. The selection of the dwelling units was also based on a fixed proportion of 10% of the total listed dwellings in an EA. A constraint was imposed on small-size EAs where, if the listed dwelling units were less than 10 dwellings, the selection was increased to 10 dwelling units. All households within the selected dwelling units were covered. There was no replacement of refusals, vacant dwellings or non-contacts owing to their impact on the probability of selection. Concerted efforts were made to improve the response rates through multiple visits.

Weighting

The CS 2007 sample has equal probabilities for all elements in the cluster which make it a self-weighting systematic random sample. Since the sample is stratified by municipalities as demarcated at the time of Census 2001, the inclusion probability of selection of an EA at the first level of selection, and the dwelling unit at the second level of selection, is the product of first and second-level probabilities. Also, since all 8 households within the dwelling unit are considered, their probability of being in the dwelling unit is always one.

	It is important to note that non-responses in the CS can occur at EA level, at dwelling-unit level and at household level. For instance, there were two EAs in Western Cape where fieldworkers were unable to gain access because of political unrests. On dwelling-unit level, only 238,067 out of 274,348 sampled dwelling units returned completed questionnaires. This means that out of 16,255 EAs with listed dwelling units that were visited for enumeration, the completed questionnaires came from only 16,173 EAs. There are also non-responses at household level which occur inside the dwelling unit. However, the undercount of households in the dwelling unit as well as the undercount of some persons in the households was not easy to account for without any dual estimation approach in place, such as the post-enumeration survey. In general, non-responses can be dealt with as either non-coverage, undercount, or proper non-responses, depending on the situation. The adjustment of non-responses is based on the classification of dwelling units or households, based on their enumeration status (enumeration completed, partially completed, non-contact, refusal, unusable information, listing error, unoccupied dwelling, demolished dwelling, vacant dwelling, and others).
Availability of data descriptions	StatsSA: http://www.statssa.gov.za/?page_id=3929
Conditions	The information products and services of Stats SA are protected in terms of the Copyright Act, 1978 (Act 98 of 1978). As the State President is the holder of State copyright, all organs of State enjoy unhindered use of the Department's information products and services, without a need for further permission to copy in terms of that copyright. Where a copy of the information is made available to any third party outside the State, the third party must be made aware of the existence of State copyright and ownership of the information by the State. The State (through Stats SA) retains the full ownership of its information, products and services at all times. Access to information does not give ownership of the information to the client. The use of any data is subject to acknowledgement of Stats SA as the supplier and owner of copyright. Users may apply or process the data, provided Stats SA is acknowledged as the original source of the data; that it is specified that the analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA.
Contact for Information and Data Supply	For information and data http://interactive.statssa.gov.za:8282/webview/

Survey data

This section contains details of survey data that has been updated since the 2007 report and new survey data.

Name	***Afrobarometer South Africa***
Principal investigator	Institute for Justice and Reconciliation in South Africa (IJR, South Africa)
Year(s)	2000, 2002, 2004, 2006, 2008, 2011 and 2015
Area(s) of interest	Social attitudes
Source(s) of data description provided here	*Afrobarometer* website: http://www.afrobarometer.org *DataFirst* web portal: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central
Brief description	*Summary* A comparative series of public attitudes surveys measuring the social, political and economic atmosphere in African countries, including South Africa. The objectives of Afrobarometer surveys in general are: • To produce scientifically reliable data on public opinion in sub-Saharan Africa. • To strengthen institutional capacity for survey research in Africa. • To broadly disseminate and apply survey results. *Methodology* There have been 6 full rounds of Afrobarometer for South Africa plus an additional round in 2004: Round 1 - 2000 Round 2 - 2002 Round 2.5 - 2004 Round 3 – 2006 Round 4 – 2008 Round 5 – 2011 Round 6 - 2015 Afrobarometer surveys are face-to-face interviews by trained interviewers in the language of the respondent's choice. National probability samples that represent an accurate cross section of the voting age population are used. Random selection is used at every stage of sampling and the sample is stratified to ensure that all major demographic segments of the population are covered. For South Africa, the sample size is 2,400 people (2,200 in Round 1). *Sample design* There is a standard protocol for drawing a national probability sample for an Afrobarometer survey and a new sample has to be drawn for each round of Afrobarometer surveys. The sample is designed as a representative cross-section of all citizens of voting age in South Africa. The goal is to give every adult citizen an equal and known chance of selection for interview. This is achieved by (a) strictly applying random selection methods at every stage of sampling and by (b) applying sampling with probability proportionate to population size wherever possible. A randomly selected sample of 2,400 cases allows inferences to national adult populations with a margin of

	sampling error of no more than plus or minus 2 percent with a confidence level of 95 percent. Excluded are areas determined to be either inaccessible or not relevant to the study, such as those experiencing armed conflict or natural disasters, as well as national parks and game reserves. People living in institutionalised settings, such as students in dormitories and persons in prisons or nursing homes are excluded. The sample design is a clustered, stratified, multi-stage, area probability sample. In a series of stages, geographically defined sampling units of decreasing size are selected. To ensure that the sample is representative, the probability of selection at various stages is adjusted as follows: The sample is stratified by key social characteristics in the population such as sub-national area (e.g. region/province) and residential locality (urban or rural). The area stratification reduces the likelihood that distinctive ethnic or language groups are left out of the sample. And the urban/rural stratification is a means to make sure that these localities are represented in their correct proportions. Wherever possible, and always in the first stage of sampling, random sampling is conducted with probability proportionate to population size. The purpose is to guarantee that larger (i.e., more populated) geographical units have a proportionally greater probability of being chosen into the sample. The sampling design has four stages: 1. Stratify and randomly select primary sampling units. 2. Randomly select sampling start-points. 3. Randomly choose households. 4. Randomly select individual respondents. Each interviewers alternates in each household between interviewing a man and interviewing a woman to ensure gender balance in the sample. To keep the costs and logistics of fieldwork within manageable limits, eight interviews are clustered within each selected PSU. *Weighting* The data are weighted to correct for either deliberate (e.g. to provide an adequate sample of specific sub-groups for analytical purposes) or inadvertent over- or under-sampling of particular sample strata. In these cases, a weighting variable is included as the last variable in the data set, with details described in the codebook. These weighting factors should be used when calculating all national-level statistics.
Availability of data descriptions	The following documentation relating to South Africa is available from the Afrobarometer website (http://www.afrobarometer.org): Codebooks from all six rounds (from 2000 to 2015) - for each question/field in the dataset, the following information is provided: question number, question, variable label, values, value labels, source, notes Questionnaires from rounds 1, 4, 5 and 6. A selection of codebooks and questionnaires for South Africa are also available on the DataFirst website (https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central).
Conditions of obtaining microdata	Because the Afrobarometer is funded from public resources, its data are a public good. All data are released via the Afrobarometer website and other outlets, along with

	relevant codebooks. However, to allow initial in-house analysis and publication, the datasets are release for public use one year after the completion of fieldwork in the relevant country. Afrobarometer data are protected by copyright. Authors of any published work based on Afrobarometer data or papers are required to acknowledge the source including, where applicable, citations to data sets posted on the Afrobarometer website.
Contact for Information and Data Supply	For information: *Afrobarometer* website**:** http://www.afrobarometer.org For data: *Afrobarometer* website: http://www.afrobarometer.org *DataFirst* web portal: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	**All Media and Products Survey *(AMPS®)***
Principal investigator	South African Audience Research Foundation (SAARF) (formerly the South African Advertising Research Foundation (SAARF))
Year(s)	1995, 2002, 2010, 2011, 2012, 2013, 2014, 2015
Area(s) of interest	Housing; transport; crime; education, media use
Source(s) of data description provided here	www.datafirst.uct.ac.za
Brief description	*Summary* The survey collects demographic data on the surveyed households, including data on race, sex, age, income, education level and home language. Data is also collected on media used by households, including newspapers and magazines, television, and radio, as well as cinema attendance. The survey also collects data on ownership and usage of products and services. *Sampling* The universe from which the AMPS sample is drawn, comprises adults aged 15 years or older in South Africa. In the case of each racial group, certain areas were excluded from consideration, as containing no persons or a negligible number of persons in a given group. A multistage, stratified, quasi-probability design was employed. This study is based on a full annual sample.
Availability of data descriptions	Not currently available
Conditions of obtaining microdata	"Licenced data, available under conditions"
Contact for Information and Data Supply	South African Audience Research Foundation : http://www.saarf.co.za/amps/amps-evolution.asp

Name	***Department of Social Development Survey***
Principal investigator	Department of Social Development
Year(s)	2006 and 2008
Area(s) of interest	Economy; Labour Market; Social Welfare; Service provision and usage
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys
Brief description	*Summary* The Department of Social Development (DSD) commissioned a set of socio-economic and demographic baseline studies in the 22 nodes that make up the Integrated Sustainable Rural Development Programme (ISRDP) and Urban Renewal Programme (URP), coupled to a management support programme that ran from 2006 to 2008. The nodes – 14 of which fall under the ISRDP and 8 of which fall under the URP – were selected because of the deep poverty in which many of their citizens live. Two surveys were commissioned: a larger baseline in 2006 and a smaller measurement survey in 2008. In the interim, the Department implemented a national, provincial and nodal support programme while considering and reacting to the findings of the first phase of background reports and qualitative nodal-level evaluations. The second survey sought to detect changes (good or bad) that occurred in the interim period. The ISRDP and URP aimed to transform their respective nodes into economically vibrant and socially cohesive areas initially through anchor projects to kick-start the programmes, and then through better co-ordination between departments geared to providing an integrated suite of services to all citizens, especially those living in poverty. The point of both programmes is the more efficient and effective use of existing government resources, rather than operating as standard, stand-alone programmes with a dedicated budget. *Methodology* *Sample design* The Baseline Survey The 2006 baseline survey sought to conduct 400 interviews in each of the 14 ISRDP nodes and the 8 URP nodes. The adult population aged 18 and older according to the Census 2001 was used as the sample frame. For the ISRDP nodes, the sample was stratified by local municipalities to ensure sufficient interviews were conducted in each municipality. According to the principles of probability proportional to size sampling (PPS), a list of place names in each of the local municipalities was then generated as a starting point for the fieldwork. At each starting point in the ISRDP nodes five interviews were conducted. For the URP nodes, detailed maps at a ward level were generated from the Municipal Demarcation Board website. Again using the principles of probability proportional to size sampling (PPS), starting points across the different wards were identified on the maps. At each starting point in the URP nodes four interviews were conducted. At the end of the fieldwork phase a total of 8,387 interviews across the 22 nodes had been conducted. Once the information from each interview had been coded and captured on computer, the realised samples in each of the ISRDP nodes were weighted back to the actual population figures across each local municipality. It should be noted

	that on the one hand, 8,400 is a very large sample with a margin of sampling error of only 1.1%. However, when the data are analysed at nodal level, each of the 22 samples of 400 have a larger sampling error of 4.9%. The Measurement Survey The 2008 measurement survey sought to conduct 250 interviews in each of the 14 ISRDP nodes (except in Bushbuckridge and Maruleng, where 250 interviews were divided across the two nodes according to population size) and the 8 URP nodes. In order to allow for comparisons with the 2006 baseline survey, the 250 interviews for Maluti-a-Phofung were spread across the whole district municipality of Thabo Mofutsanyane. For comparative purposes, the sample frame (the adult population aged 18 and older according to the Census 2001) and list of starting points from the 2006 baseline survey was used. For the ISRDP nodes, the following steps were followed: The sample for each node was firstly stratified by local municipalities (to ensure sufficient interviews were conducted in each municipality). Within each municipality, the sample was then stratified by settlement type (rural versus urban). According to the principles of probability proportional to size sampling (PPS), a random list of place names in each municipality was then generated. At each place name, the fieldworkers were instructed to find a school (if multiple starting points at one place, subsequent starting points were at different schools or crèches). From the school, they then walked in the direction of dwellings and started at first dwelling - thereafter, every fifth dwelling was selected. The birthday rule was used to select the respondent at each selected dwelling - this random process seeks to interview the adult in the household whose birthday is next. For the ISRDP nodes, five interviews were conducted per starting point. For the URP nodes, the following steps were followed: The sample for each node was firstly stratified by wards. Within each ward, the sample was then stratified by settlement type (formal versus informal types). Detailed maps at a ward level were generated from the Municipal Demarcation Board website. According to the principles of probability proportional to size sampling (PPS), a random series of starting points in each ward were then generated using a random grid of points. From the identified starting point, the fieldworkers proceeded in the direction of the centre of the node and interviewed at the first dwelling they came to - thereafter, every fifth dwelling was selected. The birthday rule was again used to select the respondent at each selected dwelling. For the URP nodes, four interviews were conducted per starting point. At the end of the fieldwork phase a total of 5,232 interviews across the 22 nodes had been conducted. Note, while 5,250 is a large sample with a margin of sampling error of only 1.4%, a nodal sample of 250 has a far larger sampling error of 6.2%. *Weighting* The samples for each of the URP nodes were self-weighting, therefore no weighting needed to be applied to these samples. The data should be seen as representative of the adult population in each of the 22 nodes.
Availability of data descriptions	Documentation is also available on the DataFirst website (http://www.datafirst.uct.ac.za/surveys)
Conditions of obtaining microdata	Online Application for Access to a Public Use Dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give

	assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: support@data1st.org For data: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central/about

Name	***Domestic Tourism Survey***
Principal investigator	StatsSA
Year(s)	Annually 2008 to 2015 inclusive
Area(s) of interest	Tourism
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys) Statistics South Africa (http://interactive.statssa.gov.za:8282/webview/)
Brief description	*Summary* Data on the travel behaviour and expenditure of South African residents travelling within and outside the borders of South Africa. One of the objectives of the DTS is to obtain the estimates that indicate the Tourism contribution toward the South Africa economy. Data on domestic tourism is also needed to measure its contribution to the national economy. The Domestic Tourism Survey (DTS) is aimed at addressing this need by collecting data on the travel behaviour and expenditure of South African residents travelling within and outside the borders of South Africa *Methodology* The target population of the survey consists of all private households in all nine provinces of South Africa and residents in workers' hostels. The survey does not cover other collective living quarters such as students' hostels, old age homes, hospitals, prisons and military barracks, and is therefore only representative of non-institutionalised and non-military persons or households in South Africa. The survey is conducted using face to face interviews. *Sample design* The sample design for the DTS carried out between 2008 and 2013 was based on the StatsSA master sample (MS). The master sample used a two-stage, a stratified design with probability–proportional-to-size (PPS) sampling of PSUs from within strata, and systematic sampling of dwelling units (DUs) from the sampled primary sampling units (PSUs). A self-weighting design at provincial level was used and MS stratification was divided into two levels, primary and secondary stratification. Primary stratification was defined by metropolitan and non-metropolitan geographic area type. During secondary stratification, the Census 2001 data were summarised at PSU level. The following variables were used for secondary stratification; household size, education, occupancy status, gender, industry and income. The sample design for the DTS 2014 was based on the StatsSA master sample (MS) that was originally designed for the Quarterly Labour Force Survey (QLFS). This master sample is shared by the QLFS, GHS, Living Conditions Survey (LCS), Domestic Tourism Survey (DTS) and the Income and Expenditure Survey (IES). The master sample used a two-staged, stratified design with probability–proportional-to-size (PPS) sampling of PSUs from within strata, and systematic sampling of dwelling units (DUs) from the sampled primary sampling units (PSUs). A self-weighting design at provincial level was used and MS stratification was divided into two levels, primary and secondary stratification. Primary stratification was defined by metropolitan and non-metropolitan geographic area type. During secondary stratification, the Census 2001 data were summarised at PSU level. The following variables were used for secondary stratification: household size, education, occupancy status, gender, industry and income.

	For all the Domestic Tourism Surveys, census enumeration areas (EAs) as delineated for Census 2001 formed the basis of the PSUs. Where possible, PSU sizes were kept between 100 and 500 dwelling units (DUs). A randomised Probability Proportional to Size (RPPS) systematic sample of PSUs was drawn in each stratum, with the measure of size being the number of households in the PSU. Approximately 3,080 PSUs were selected. In each selected PSU a systematic sample of dwelling units was drawn. The number of DUs selected per PSU varies from PSU to PSU and depends on the Inverse Sampling Ratios (ISR) of each PSU. *Sample design of the DTS 2015 Quarter 1* One of the objectives of the DTS is to obtain the estimates that indicate the Tourism contribution toward the South Africa economy. However this can be obtained by calculating estimates within the calendar year. Due to the rolling three month recall period, there was a need to supplement the DTS 2014 data with the DTS 2015 Quarter 1 data. This first quarter of the DTS 2015 contain data from October, November and December 2014. It should be noted that the DTS 2015 PSU's were drawn from the new master sample (2013). Approximately 3,324 PSU's were selected. Similar sampling method that was used for DTS 2014 was applied to the DTS 2015. One of the weaknesses of the 2014 DTS was that the dwellings were not assigned to survey months within the quarters. We selected the DU sample for the 2015 DTS from the redesigned master sample, and considered two scenarios to assign dwellings to the survey months within the quarters: (1) Assign 1/3rd of the sampled DUs from each of the master sample PSUs to each of the 3 survey months within the quarter, and (2) Assign all sampled DUs from 1/3rd of the master sample PSUs to each of the 3 survey months within the quarter. Based on the cost-variance analysis the scenario 1 was found to be more efficient. Therefore, the scenario 1 was implemented to assign sampled DUs to the survey months within the quarters. *Sampling and the interpretation of the data* Caution must be exercised when interpreting the results of the DTS at low levels of disaggregation. Revisions to the DTS data sets based on the new population estimates involved benchmarking at national level in terms of age, sex and population group while at provincial level, benchmarking was by population group only. The sample and reporting are based on the provincial boundaries as defined in December 2005. The DTS 2014 data was collected from January to December 2014 and therefore takes seasonality into consideration. Its unique weighting method also results in the data not being comparable with the data collected during the previous years (see below). *Weighting* For the 2008 to 2013 Domestic Tourism Surveys, sampling weights for the data collected from the sampled households are constructed so that responses can be expanded appropriately to represent the entire population of South Africa. The weights are the result of calculations involving several factors, including design weights, adjustment for non-response, and benchmarking to known population estimates from the Demographic Analysis division of Stats SA. The final survey weights are constructed by calibrating the adjusted base weight to the known population counts at national and provincial levels (which are supplied by the Demographic Analysis division of Statistics SA), cross-classified by 5-year age groups gender and race. The calibrated weights are constructed to ensure that all persons in a household have the same final weight (integrated weighting). .. (Data First)

	DTS 2014 was collected continuously, based on a rolling three month recall period. This meant that to reconstruct travel for the period January to December 2014 the data from the 2014 DTS and the DTS 2015 January to March collection to create the final data set. Thus 14 survey months from February 2014 to March 2015 were used to create the 12 monthly data files for the calendar year 2014. The sampling weights were then constructed for the 12 reference months. These included both full sample and replicate weights for each of the 12 reference months. The full sample and replicate weights were calibrated using the population control totals for the cells defined by the cross classification Age-Group x Population-Group x Gender at the national level, and broad Age Groups at the province level. These are the same population control totals that are used for constructing the calibrated weights for the Quarterly Labour Force Survey and the other household surveys. The weighted monthly data files can be combined to produce annual and bi-annual estimates with the WESVAR software.
Availability of data descriptions	The following documentation is available from the NESSTAR (hhttp://interactive.statssa.gov.za:8282/webview/): Questionnaire Metadata Concepts and Definitions Documentation also available on the DataFirst website (http://www.datafirst.uct.ac.za/surveys).
Conditions of obtaining microdata	Stats SA: Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. Data First: Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: info@statssa.gov.za http://www.statssa.gov.za/ For data: http://interactive.statssa.gov.za:8282/webview/ https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	***Employment and Learning pathways of Learnership participants in the NSDS phase II (ELL)***
Principal investigator	HSRC
Year(s)	2007
Area(s) of interest	Labour market; Education
Source(s) of data description provided here	HSRC (http://curation.hsrc.ac.za/index.php?module=pagesetter&tid=125&tpl=projects)
Brief description	*Summary* In 2006 the Department of Labour (DoL) requested that the Human Sciences Research Council (HSRC) undertake a study that evaluated the effectiveness of learnerships in terms of internal efficiency and the labour market outcomes of learnership participants. It was decided that the focus of this HSRC research would be on investigating the extent and ways in which learnerships are equipping the employed to advance through the formal labour market with enhanced skills and capacities, or equipping the young unemployed to find jobs, or create self-employment, or to advance to further education and training. Such empirical research required a clear focus on the experience of individual participants in learnership programmes, rather than on the programmes themselves or on the Sector Education and Training Authorities (SETAs) that host them. The data set consists of information on persons who registered for a learnership qualification and who enrolled in the first year of the National Skills Development Strategy (NSDS) Phase II. A total number of 6,815 valid surveys were returned. This represents a total return rate of 85.2%. *Methodology* A sample of Learnership participants under National Skills Development Strategy (NSDS) Phase II were selected and a telephonic survey was conducted using a Computer Aided Telephonic Interview (CATI) tool developed by the HSRC using Microsoft Access. *Sample design* The aim was to use this sample frame to obtain 8,000 responses, proportionately spread across the 22 SETAs according to the size of each SETA. PSETA provided no information on their learners registered in the NSDS Phase II and was the only SETA excluded from the sampling frame. Each data record within each SETA database was allocated a random number. Each data set was then sorted in ascending order according to the random number. The call centre operators proceeded by telephoning the learners from the top to the bottom of the list for each SETA separately. *Weighting* The database of returns consisted of a sample of learners. Hence, statistical weights were calculated for each sample cell to adjust the number of responses in a particular cell to the original number of learnership participants in the sample frame or population, that is, those enrolled in the first financial year of NSDS Phase II.
Availability of data descriptions	Documentation is available from the HSRC website http://curation.hsrc.ac.za/Dataset-322.phtml

Conditions of obtaining microdata	By accessing the data, the user gives assurance that The data and documentation will not be duplicated, redistributed or sold without prior approval from the rights holder. The data will be used for scientific research or educational purposes only. The data will only be used for the specified purpose. If it is used for another purpose the additional purpose will be registered. Redundant data files will be destroyed. The confidentiality of individuals/organisations in the data will be preserved at all times. No attempt will be made to obtain or derive information from the data to identify individuals/organisations. The HSRC will be acknowledged in all published and unpublished works based on the data according to the provided citation. The HSRC will be informed of any books, articles, conference papers, theses, dissertations, reports or other publications resulting from work based in whole or in part on the data and documentation. For archiving and bibliographic purposes an electronic copy of all reports and publications based on the requested data will be sent to the HSRC. To offer for deposit into the HSRC Data Collection any new data sets which have been derived from or which have been created by the combination of the data supplied with other data. The data team bears no responsibility for use of the data or for interpretations or inferences based upon such uses.
Contact for Information and Data Supply	For information: Tel: +27 (0)12 3022000 datahelp@hsrc.ac.za For data: http://curation.hsrc.ac.za/index.php?module=pagesetter&tid=125&tpl=projects datahelp@hsrc.ac.za

Name	**General Household Survey *(GHS)***
Principal investigator	Stats SA
Year(s)	2002 to 2015 (annually)
Area(s) of interest	Demography; housing; labour market; education; health; social welfare
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys) Statistics South Africa (hhttp://interactive.statssa.gov.za:8282/webview/)
Brief description	*Summary* A national annual survey designed to measure various aspects of the living circumstances of South African households. The five broad areas covered by the GHS are: education, health, activities related to work and unemployment, housing and household access to services and facilities. *Methodology* Information was collected on various aspects of the living circumstances of members from over 30,000 households across the country. The sampled dwelling units in each of the nine provinces were visited by field staff employed and trained by Stats SA, and a questionnaire was completed through face-to-face interviews for each household visited. The target population of the survey consists of all private households in all nine provinces of South Africa and residents in workers' hostels. The survey does not cover other collective living quarters such as students' hostels, old-age homes, hospitals, prisons and military barracks, and is therefore only representative of non-institutionalised and non-military persons or households in South Africa. *Sample design* For 2002 to 2007, enumeration Areas (EAs) that had a household count of less than 25 were omitted from the Census frame that was used to draw the sample of Primary Sampling Units (PSUs) for the Master Sample. Other omissions from the Master Sample frame included all institution EAs except workers' hostels, convents and monasteries. EAs in the Census database that were found to have less than sixty dwelling units during listing were pooled together to form PSUs. The Master Sample was a multi-stage stratified sample. The overall sample size of PSUs was 3,000. The explicit strata were the 53 district councils. The 3,000 PSUs were allocated to these using the power allocation method. The PSUs were then sampled using probability proportional to size principles. The measure of size used was the number of households in a PSU as calculated in the Census. The sampled PSUs were listed with the dwelling unit as the listing unit. From these listings systematic samples of dwelling units were drawn. These samples of dwelling units formed clusters. The size of the clusters differed depending on the specific survey requirements. The GHS used one of the clusters that contained ten dwelling units. From 2008 onwards, the sample design for the GHS was based on a master sample (MS) that was originally designed for the Quarterly Labour Force Survey (QLFS). This master sample is shared by the QLFS, GHS, Living Conditions Survey (LCS), Domestic Tourism Survey (DTS) and the Income and Expenditure Survey (IES). The master sample used a two-stage, stratified design with probability-proportional-to-size (PPS) sampling of

	primary sampling units (PSUs) from within strata, and systematic sampling of dwelling units (DUs) from the sampled PSUs. A self-weighting design at provincial level was used and MS stratification was divided into two levels. Primary stratification was defined by metropolitan and non-metropolitan geographic area type. During secondary stratification, the Census 2001 data were summarised at PSU level. The following variables were used for secondary stratification: household size, education, occupancy status, gender, industry and income. Census enumeration areas (EAs) as delineated for Census 2001 formed the basis of the PSUs. The following additional rules were used: • Where possible, PSU sizes were kept between 100 and 500 DUs; • EAs with fewer than 25 DUs were excluded; • EAs with between 26 and 99 DUs were pooled to form larger PSUs and the criteria used was same settlement type; • Virtual splits were applied to large PSUs: 500 to 999 split into two; 1,000 to 1,499 split into three; and 1,500 plus split into four PSUs; • Informal PSUs were segmented. A randomised-probability-proportional-to-size (RPPS) systematic sample of PSUs was drawn in each stratum, with the measure of size being the number of households in the PSU. Altogether, approximately 3,080 PSUs were selected. In each selected PSU a systematic sample of dwelling units was drawn. The number of DUs selected per PSU varies from PSU to PSU and depends on the Inverse Sampling Ratios (ISR) of each PSU. *Weighting* From 2008, the Statistics Canada software StatMx has been used for constructing calibration weights. The population controls at national and provincial levels were used for the cells defined by cross-classification of Age by Gender by Race. Records for which the age, population group or sex had item non-response could not be weighted and were therefore excluded from the dataset. No imputation was done to retain these records. *Geography* In early surveys, data are available for analysis at national and provincial level, with a rural/urban split. From 2005 onwards the GHS also presents data at district municipality (42 areas), cross-border district municipality (5) and major metropolitan area (6) levels.
Availability of data descriptions	The following documentation is available from the NESSTAR website (http://interactive.statssa.gov.za:8282/webview/): Questionnaire Metadata Concepts and Definitions
Conditions of obtaining microdata	Stats SA: Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA.

	Data First: Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: Tel: (012) 310 8600 Fax: (012) 310 8944 info@statssa.gov.za http://www.statssa.gov.za/ For data: http://interactive.statssa.gov.za:8282/webview/ https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	***HIV Prevalence and Related Factors – Higher Education Sector Study, South Africa***
Principal investigator	Higher Education South Africa
Year(s)	2008/09
Area(s) of interest	Health; Education
Source(s) of data description provided here	http://heaids.org.za/
Brief description	*Summary* The purpose of this study was to enable the higher education sector to understand the threat posed by the epidemic to its core mandate. This was done through determining, at the institutional and sector level, the prevalence and distribution of HIV and associated risk factors among the staff and students at public, higher education institutions (HEIs) in South Africa. The results were used to conduct an assessment of the risks posed by the HIV epidemic to the sector and their respective populations and make recommendations to mitigate potential impacts. *Methodology* Questionnaires were designed to be completed in approximately 30 minutes. Questionnaires were designed to be self-completed by participants and comprised 10 pages of questions with multiple choice options that were marked using pencils provided. Questionnaires were available in English and Afrikaans with separate translations available if this was required. Three separate questionnaires were available, depending on the participant category: student, academic staff, or admin/service staff. The questionnaires differed on a small introductory subset of questions related to participant institutional data – for example, course of study versus teaching field versus administrative or service job focal area. The remainder of the questions related to knowledge, attitude, behaviour and practice (KABP) and were common to all questionnaires. Given that the population at HEIs is overall literate, self-completion of questionnaires was possible in nearly all instances. The exceptions included some staff who were less literate, or some of those who were not sufficiently conversant in English or Afrikaans. Assistance was provided to such participants by field workers. Selected participants who wished to opt out were free to leave at any point. Participants were advised of the confidential nature of the questionnaires and the importance of being seated in such a way that their answers could not be seen by other participants. The intention of the self-completion process was to ensure that sensitive questions about sexual behaviour could be answered confidentially. Five blood spots were also gathered from a single finger. In addition to the quantitative questionnaire, the study also included a qualitative component which was designed to understand contextual factors underpinning the risk of HIV infection at HEIs, as well as factors related to the effectiveness of prevention, support, treatment and impact mitigation efforts. It was also intended that this qualitative component of the study would capture perspectives of members of each HEI on existing responses and perspectives on what further responses were needed. The qualitative component of the research is not

	Sample design In order to improve reliability of the data, staff were over-sampled relative to population sizes, with approximately 20% of all staff being sampled and approximately 4% of all contact students being sampled. Universities were grouped into large, medium, and small categories based upon numbers of staff and students so as to allocate sample sizes among the universities. The Higher Education Management Information System HEMIS database 2006 was used to estimate the student and staff populations. Actual numbers sampled from each HEI varied from these target numbers, depending upon requirements identified during the execution of fieldwork A sample of 15,728 students (3.4% of the population) were allocated. The minimum sample size per institution would be 562 in small institutions; 737 in medium institutions and 1,053 in large institutions. With a sample of this size, the national HIV prevalence could be estimated to within 0.6%; for small institutions to within 5.2% and for medium and large institutions to within 2.3% assuming an HIV prevalence of 10%, a design effect of 1.5 (given the cluster design) and a confidence level of 95%. According to HEMIS 2006, there are a total of 39,154 permanent and contracted staff at tertiary institutions. The institutions were stratified by size into three categories and sampled proportionately. Staff were further stratified by job category: academic; administrative and service staff. A sample of 8,786 staff were to be allocated. The sample size per institution was 162 in small institutions (52 academic; 110 admin/service); 344 in medium institutions (127 academic; 217 admin/ service) and 674 in large institutions (236 academic; 438 admin/service). The 21 universities that were part of the study were considered as strata, which were divided into two sub-populations of students and staff. In cases of multiple campuses within an HEI, if student populations were expected to be substantially different at different campuses and if the campuses were sufficiently large to justify presence of a sampling team for a full day, the campuses were separated into strata and departments randomly selected within each campus stratum. For HEIs composed of multiple previously independent campuses, with each campus too small to justify the presence of a sampling team for a full day, approximately three campuses were randomly sampled. Faculties were combined into groups that were treated as strata. These groupings were intended to facilitate data collection. The faculty groups were based on HEMIS categories, with adjustment to facilitate efficient sampling. The groupings used were: (1) natural sciences, engineering, and agricultural sciences; (2) arts, education, and theology; (3) law and economics/ management. A list of departments and numbers of students registered for courses offered by each department was obtained from the registrar at each HEI. One department was randomly selected from each faculty stratum (with probability of selection proportional to department size). For each selected department, courses offered by the department were randomly ordered, with larger classes having greater probability of being near the beginning of the list and smaller classes having greater probability of being near the end. Classes were oversampled in order to ensure that the minimum sample size would be obtained at each institution. Information was communicated to heads of the Faculties and Departments selected urging them to encourage all academic staff whose class had been selected to assist

	with the survey. Samples were drawn from lectures or tutorials and practical sessions. Although lecturers were briefed about selection of a lecture/tutorial/practical session, students were not made aware of this selection, and were only advised at the start of the session. At the selected lecture/tutorial/practical sessions, less than 50 students were typically selected.
Availability of data descriptions	Documentation is available on the SADA website (http://sada-data.nrf.ac.za/)
Conditions of obtaining microdata	Complete an ORDER FORM (download in Microsoft Word or Adobe Acrobat format) and faxed, e-mailed or posted the request to SADA before the data can be made available. The user should abide by the following regulations when using data and documentation from SADA: The data and documentation will be used only for research, research training, teaching and policy decisions. No attempt will be made to use the data to derive information on specifically identified individuals or households in the data. The data and documentation received from SADA will not be duplicated without prior approval of the Director of SADA. Both SADA and the Depositor will be acknowledged in all published works based on the data and documentation. The text for the citation is now part of the Study Description and metadata that is available on the Data Portal. SADA and the Depositor of the data will not be held liable for the accuracy or comprehensiveness of the data.
Contact for Information and Data Supply	For information: sada@nrf.ac.za +2712 481 4120 +2712 481 4016 For data: http://sada.nrf.ac.za/order.html

Name	**Income and Expenditure Survey *(IES)***
Principal investigator	Stats SA
Year(s)	1995, 2000, 2005/06, 2010/11
Area(s) of interest	Economy
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys) Statistics South Africa (hhttp://interactive.statssa.gov.za:8282/webview/)
Brief description	*Summary* A national household survey tracking all the details of receipts of cash, goods and services and those related to the purchase of goods and services for the household's own consumption. The survey forms the basis for the determination of the basket of consumer goods and services used for the calculation of the Consumer Price Index (CPI). Although primarily intended to provide weights for the CPI, the IES gathers so much information on incomes and expenditure, that, inevitably, much use of it has been made in poverty and inequality studies. The survey covered private dwellings, workers' hostels, residential hotels, and nurses' and doctors' quarters, but excluded hospitals and clinics, hotels and guest houses, prisons, schools and student hostels and old-age homes. *Methodology* The 1995 and 2000 IES studies were based on the sample for the rotating panel of the twice yearly Labour Force Survey. The surveys were done by means of an interview with the household head or a responsible adult and the questionnaire was completed by the enumerator during this interview. In cases where the household requested to complete the questionnaire themselves, it was dropped off by the enumerator, and the completed questionnaire was collected at a second visit. Depending on the nature of the transaction, respondents were required to recall over periods ranging from 1 month (non-durable consumption, for example) to 12 months (durables and other major expenditure), The IES 2005/2006 was based on the diary method was the first of its kind to be conducted by Stats SA. A fieldworker administered a main questionnaire to a selected household over five separate visits during which households were required to account for all income and for acquisitions of goods and services over the 11 months prior to the survey. During the four weeks of the survey month, households were also given diaries and were required to record their daily acquisitions in a diary on a daily basis. The diaries were collected on a weekly basis for a period of a month. The purpose of the diary was to try to minimise or eliminate the recall problem over the four weeks of the survey month so that the information collected was as accurate as possible. The IES 2010/ 2011 is conducted in the same way as the IES 2005/2006, although in the IES 2010/2011 the diary period was shortened from one month to two weeks to reduce respondent fatigue. *Sample design* The 2000 IES used a Master Sample based on the 1996 Census of enumeration areas (EA) and the estimated number of dwelling units from the 1996 Census. All 3,000 PSUs included in the Master Sample were used in the IES. A PSU is either one EA or several EAs when the number of dwelling units in the base or originally selected EA was found to have less than 100 dwelling units. Each EA had to have approximately 150 dwelling units but it was discovered that many contained less. Thus, in some cases, it has been found necessary to add EAs to the original EA to ensure that the

	minimum requirement of 100 dwellings, in the first stage of forming the PSUs, was met. The size of the PSUs in the Master Sample varied from 100 to 2,445 dwelling units. Special dwellings such as prisons, hospitals, boarding houses, hotels, guest houses (whether catering or self-catering), schools and churches were excluded from the sample. Explicit stratification of the PSUs was done by province and area type (urban/rural). Within each explicit stratum, the PSUs were implicitly stratified by District Council, Magisterial District and, within the magisterial district, by average household income (for formal urban areas and hostels) or EA. The allocated number of EAs was systematically selected with probability proportional to size in each stratum. Once the PSUs included in the sample were known, their boundaries had to be identified on the ground. After boundary identification, the next stage was to list accurately all the dwelling units in the PSUs. The second stage of the sample selection was to draw from the dwelling units listing whereby a systematic sample of 10 dwelling units was drawn from each PSU. As a result, approximately 30,000 households (units) were interviewed. However, if there was growth of more than 20% in a PSU, then the sample size was increased systematically according to the proportion of growth in the PSU. For the 2005/06 IES, a newly designed Master Sample, consisting of 3000 Primary Sampling Units (PSUs), based on the 2001 Population Census Enumeration Areas, was used as the sampling frame. The Master Sample is used for all household surveys conducted by Statistics South Africa (Stats SA). The 3,000 primary sampling units (PSUs) from the Master Sample were representatively divided into four quarterly allocations of 750 each. Within each quarterly allocation, a random sample of 250 PSUs was selected every month. Eight dwelling units were systematically selected from each of the sampled PSUs for fieldwork. In total, 24,000 dwelling units were covered during the twelve months of data collection for the IES 2005/2006. This process ensured that the sample was evenly spread over the twelve months, while it remained nationally representative in each quarter. The sampling frame for the IES 2010/2011 was obtained from Statistics South Africa's Master Sample (MS) based on the 2001 Population Census enumeration areas (EAs). The scope of the Master Sample (MS) is national coverage of all households in South Africa and the target population consists of all qualifying persons and households in the country. In summary, it has been designed to cover all households living in private dwelling units and workers living in workers' quarters in the country. The IES 2010/2011 sample is based on an extended sample of 3,254 PSUs, which consists of the 3,080 PSUs in the Master Sample and a supplement of 174 urban PSUs selected from the PSU frame. The IES sample file contained 31,419 sampled dwelling units (DUs). The 31,419 sampled DUs consist of 31,007 DUs sampled from the 3,080 design PSUs in the Master Sample and 412 DUs from the supplemented 174 urban PSUs. In the case of multiple households at a sampled DU, all households in the DU were included. From the 31,419 dwelling units sampled across South Africa, 33,420 households were identified. Out of these, there was a sample realisation of 27,665 (82.8%) households, with the remaining 5,755 (17.2%) households being classified as out of scope. For the 2010/2011 IES, the sample for the survey used a two-stage stratified design with probability-proportional-to-size (PPS) sampling of primary sampling units from strata in the first stage, and systematic sampling (SYS) of dwelling units from the sampled PSUs. The MS stratification was divided into two levels: (1) the primary stratification was defined by metropolitan and non-metropolitan geographic area type; (2) during the second stratification, the

	Census 2001 data were summarised at PSU level using the following variables: household size, education, occupancy status, gender, industry and income.
Availability of data descriptions	The following documentation is available from the Nesstar (http://interactive.statssa.gov.za:8282/webview/): Questionnaire Metadata Concepts and Definitions Metadata is also available on the DataFirst website (http://datafirst.uct.ac.za/).
Conditions	Stats SA: Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. Data First: Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: Tel: (012) 310 8600 Fax: (012) 310 8944 info@statssa.gov.za http://www.statssa.gov.za/ For data: http://interactive.statssa.gov.za:8282/webview/ https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	***Integrated Planning, Development and Modelling Project***
Principal investigator	HSRC
Year(s)	2008 and 2010
Area(s) of interest	Demography; Housing; Economy; Transport
Source(s) of data description provided here	HSRC (http://curation.hsrc.ac.za/index.php?module=pagesetter&tid=125&tpl=projects)
Brief description	*Summary* The IPDM questionnaire survey data is part of the IPDM/STEPSA project's spatial planning exercise, and was aimed at supplying detailed migration history data for 8 provinces (Gauteng, Limpopo, North West, Mpumalanga, KZN, Eastern Cape, Western Cape and Free State), together with respondent data on household structure and economy, transport access and needs, and housing access and needs. *Purpose* The main purpose is to set individual migration into the economic and spatial context in which it takes place, so as to provide policy-relevant migration information in reasonable depth, with particular reference to where housing and infrastructure delivery should best be situated for youth, women and couples with children. The key question which the survey work addressed was how best to use spatial planning of housing and infrastructure to promote better access to the labour market for socially excluded groups, with special reference to youth and women. *Methodology* The survey data itself was collected in two tranches, with the first survey of 2865 cases in 2008 and the second survey of 3051 cases in 2010. The two component surveys are aimed at providing in-depth data that can then potentially be used in conjunction with national Census data for wider coverage. These two survey data sets have been combined into one composite migration data set, which caters for the differences in the questionnaires between the two surveys and allows for the lapse of time; however, the individual 2008 and 2010 surveys can also be accessed separately in the data set, and contain some information that the combined data set does not include due to the divergence of the detailed questionnaire content in the two surveys. The 2008 Phase 1a data set contains more detailed migration and housing information, while 2010 Phase 1b data contains proportionately more transport information and devotes less space to migration. Major variables include standard respondent household information linked to migration history, housing particulars, and transport activity and costs. Migration data is provided in some depth with reference to spatial location and is broken down by distance zone relative to metro city centres. The survey was conducted using face-to-face interviews. *Sample design* The two surveys (Phases 1a and 1b) used similar stratified, clustered, random sampling frames with 2001 census EAs as primary sampling units (PSUs), which were randomly selected from the list of all EAs from within the universe. The strata used were (a) weighted distance from the nearest large central business district (CBD), taking into account the distances from other 'competing' major CBDs, (b) age of the

	settlement (ensuring sufficient representation of both older and younger settlements), and (c) city type (metro, secondary city or non-city). Within every PSU a sample of six eligible households -- the ultimate sampling units -- was drawn by means of systematic sampling with a random starting point. Within households an informed adult member (preferably the head or spouse) being present at the time of the survey provided some information for the household in general and on behalf of all its members and also provided some information for himself/herself specifically). *Weighting* The weights calculated were based on the 2001 census populations of the various spatial entities concerned, but these weights have since been re-weighted down to the sample size to avoid unnecessary problems with inferential statistics. It is suggested that the former be used for descriptive work and the latter for statistical analyses.
Availability of data descriptions	Documentation is available from the HSRC website: http://curation.hsrc.ac.za/Dataset-323.phtml
Conditions of obtaining microdata	By accessing the data, the user gives assurance that The data and documentation will not be duplicated, redistributed or sold without prior approval from the HSRC. The data will be used for statistical and scientific research purposes only and the confidentiality of individuals/organisations in the data will be preserved at all times and that no attempt will be made to obtain or derive information relating specifically to identifiable individuals/organisations. The HSRC will be informed of any books, articles, conference papers, theses, dissertations, reports or other publications resulting from work based in whole or in part on the data and documentation. The HSRC will be acknowledged in all published and unpublished works based on the data according to the citation as stated in the study information file or the web page metadata field, citation. For archiving and bibliographic purposes an electronic copy of all reports and publications based on the requested data will be sent to the HSRC. The collector of the data, the HSRC, and the relevant funding agencies bear no responsibility for use of the data or for interpretations or inferences based upon such uses. By retrieval of the data you signify your agreement to comply with the above-stated terms and conditions and give your assurance that the use of statistical data obtained from the HSRC will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: Tel: +27 (0)12 3022000 datahelp@hsrc.ac.za For data: http://curation.hsrc.ac.za/index.php?module=pagesetter&tid=125&tpl=projects datahelp@hsrc.ac.za

Name	***Living Conditions Survey***
Principal investigator	StatsSA
Year(s)	2008/09 and 2014/15
Area(s) of interest	Housing; Social welfare; Economy; Labour Market; Education; Transport; Health
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys) Statistics South Africa (hhttp://interactive.statssa.gov.za:8282/webview/ (http://www.statssa.gov.za/publications/P0310/P03102008.pdf)
Brief description	*Summary* The main aim of this survey is to provide data that will contribute to better understanding of living conditions and poverty in South Africa for monitoring levels of poverty over time. The 2014/2015 data and official metadata will be released on the 27th of January 2017. The text below currently relates only to the 2008/09 round of data collection *Methodology* The Living Conditions Survey 2008/2009 used a household questionnaire, a weekly diary, and a survey assessment questionnaire. The data were collected over a 12 month time period, between September 2008 and August 2009. The household questionnaire was a booklet of questions administered to respondents during the course of the survey month. There were seven modules in the household questionnaire. The first module dealt with establishing the composition and structure of the household, as well as capturing particulars of all household members. The second module collected information on health, disability, education and employment. The third module dealt with welfare, assets and information on dwellings and services. Modules four and five collected information on the different categories of consumption expenditure (including housing, clothing, furniture, appliances, transport, computer and telecommunication equipment, etc.), as well as information on subsistence and living circumstances. The sixth module dealt with savings, investments, debt, remittances and income. The seventh and last module collected anthropometric measurements (height, weight and waist) for all household members. The Weekly diary was a booklet that was left with the responding household to track all acquisitions made by the household during the survey month. The household (after being trained by the Interviewer) was responsible for recording all their daily acquisitions as well as information about where they purchased the item (source) and the purpose of the item. A household completed a different diary for each of the four weeks of the survey month. Interviewers then assigned codes for the classification of individual consumption expenditure according to purpose (COICOP) to reported items recorded in the weekly diary, using a code list provided to them. Finally the survey included a survey assessment questionnaire that was administered to households after the survey month was complete by either the district survey coordinator or provincial quality monitor. In addition to serving as a control questionnaire to verify information collected by the interviewers, the instrument was

	designed to evaluate data collection processes and the respondent's perceptions of Stats SA and the survey. *Sample design* The sampling frame for the LCS was obtained from Statistics South Africa's Master Sample (MS) based on the 2001 Population Census Enumeration Areas. The scope of the Master Sample (MS) is national coverage of all households in South Africa. It was designed to cover all households living in private dwelling units and workers living in workers' quarters in the country. The MS consists of 3080 primary sampling units (PSUs) made up of enumeration areas. The PSU coverage comprises all settlement types, including urban formal, urban informal, rural formal and tribal areas. For the LCS, 3065 PSUs were sampled from the MS and roughly ten dwelling units (DUs) were sampled on average per PSU. In the case of multiple households, all households in the DU were included. The sample was evenly split into four rotations (quarters) with national representativeness in each rotation. Each rotation (consisting of a sample for three months) was then evenly split into monthly samples. Ultimately, the sample was evenly spread over the 12 survey periods (one month each). *Weighting* Sample weights for the collected data are constructed in such a way that the responses could be properly expanded to represent the entire civilian population of South Africa. The weights are the results of calculations involving several factors such as design weights adjustments, non-response adjustments and the calibrations process.
Availability of data descriptions	Documentation is also available on the DataFirst website (http://www.datafirst.uct.ac.za/surveys).
Conditions of obtaining microdata	Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: support@data1st.org +2721 650 5708 http://www.datafirst.uct.ac.za/surveys For data: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central/about

Name	***National Antenatal Sentinel HIV prevalence Survey***
Principal investigator	National Department of Health
Year(s)	Annually 1990 to 2013 (No information as to surveys beyond 2013)
Area(s) of interest	Health
Source(s) of data description provided here	https://www.health-e.org.za/wp-content/uploads/2016/03/Dept-Health-HIV-High-Res-7102015.pdf
Brief description	*Summary* The purpose of the survey is to assess the HIV sero-prevalence amongst first time antenatal clinic attendees (seen as a particularly suitable "sentinel" group to represent most closely the HIV prevalence of the generally sexually active part of the population) and to assess trends in HIV prevalence over time. The general objective of the HIV surveillance is to determine the distribution of HIV infection among pregnant women attending public health antenatal clinics at national, province and district level, disaggregated by demographic factors and age of the participants. Other objectives include monitoring HIV over time among women attending public antenatal clinics; to use this data for estimation and projection of HIV sero-prevalence trends and the burden of AIDS in the general population; to provide scientific evidence to measure progress towards meeting the Millennium Development Goal 6 and to estimate the national prevalence of HIV infection among the adult. *Methodology* The survey targets 36,000 pregnant women recruited from 1,497 Primary Sampling Units compared with 16,000 women recruited from 461 clinics in 2005. This has expanded the geographic coverage considerably to include a representative sample from all 52 health districts in all the nine provinces as well as urban, peri-urban and rural comparisons. A total of 36,000 first time pregnant antenatal care bookers served in public health clinics are targeted in a single month of October annually since 1990. In 2013, a total of 33,077 pregnant women participated compared to 34,260 in 2012 and 33,446 in 2011. The survey is used as a proxy to estimate the trend in the prevalence of HIV among pregnant first bookers aged 15–49 years served in public health facilities. The 2013 National Antenatal Sentinel HIV Prevalence Survey, South Africa are used as the target population as they are sexually active, constitute an easily accessible and stable population and are more likely than other groups to be representative of the general population. In addition, they obtain antenatal care at facilities that draw blood as part of routine medical services offered to this group. It is also unlikely that they can participate in the same survey twice in the same year. All other pregnant women who had previously visited antenatal clinics during their current pregnancy during the survey period were excluded (to avoid duplicate sampling during the same month). No pregnant women were excluded from participation on the basis of their HIV status. The basic goal was to select sentinel surveillance sites representative of the population size estimate of the area to be surveyed. Sentinel sites were selected using the 'Probability Proportional to Size' (PPS) method as this combines a random approach with a bias towards the larger clinics. The clinics had to provide pregnancy testing and antenatal care. The District Health Information System

	(DHIS) antenatal data was used to ensure that the participating sentinel sites had a minimum of 20 first time bookers per month.
Availability of data descriptions	https://www.health-e.org.za/wp-content/uploads/2016/03/Dept-Health-HIV-High-Res-7102015.pdf
Conditions of obtaining microdata	Not available
Contact for Information and Data Supply	www.nmc.gov.za www.doh.gov.za

Name	**National Household Travel Survey *(NHTS)***
Principal investigator	Department of Transport (DoT)
Year(s)	2003, 2013
Area(s) of interest	Transport
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys) Statistics South Africa (hhttp://interactive.statssa.gov.za:8282/webview/)
Brief description	*Summary* The national survey of the travel habits of individuals and households. Stats SA executed the survey on behalf of the DoT. The purpose of the survey is to: • Assist with the effective targeting of subsidies for public transport. • Assist in identifying disadvantaged regions for investment in transport infrastructure. • Measure the Key Performance Indicators for land passenger transport as required by the National Land Transport Transition Act (Act No. 22 of 2000) and the National Land Transport Strategic Framework. • Understand the transport needs and habits and/or behaviour of all household members at all times of day and for all purposes. • Ascertain the cost of transport for individuals and households and to assess the extent to which they can afford to pay for the mobility which is essential for their survival. • Assess customer attitudes towards transport services, service providers and the quality of transport facilities which they are required to use. • Measure existing car ownership and uses. • Understand the travel choices of different market segments. • Determine the extent of accessibility to opportunities such as work, health facilities, education and markets for social interaction and all other social needs. *Methodology* The NHTS 2003 sampled approximately 0.5% of all households in South Africa in May/June 2003. Metropolitan and district municipal boundaries, as determined by the Demarcation Board in 2000, were used as the basis for the determination of the analysis zones. The sample was made proportional to the population of each of the municipalities. In order to measure movements between different parts of a metropolitan or district municipality, each municipality was divided into a number of travel analysis zones. The minimum number of households per analysis zone was 100 households. The NHTS relied on selected household members (worker and learner) recalling all trips taken on the day prior to the survey. The NHTS 2013 data collection exercise took place between January and March 2013, and a total of 51,341 households and/or dwelling units were sampled, using a random stratified sample design. The findings are representative of the population of South Africa and can be analysed and reported on at provincial, municipal and Transport Analysis Zone (TAZ) levels. The NHTS 2013 questionnaire was largely based on the 2003 questionnaire. However, it was revised based on emerging information needs, the need

	to standardise certain questions from a Stats SA perspective and the technological requirements for scanning and processing. *Sample design* In 2003 the explicit strata were the 342 Travel Analysis Zones (TAZ). A sample of 5,000 Enumeration Areas (EAs) was allocated using the power allocation method. The first step was to take out vacant, industrial, institution, and recreational EAs. EAs were selected with probability proportional to size, using the total number of households as enumerated during Census 2001 as a measure of size. EAs which had less than 80 dwelling units were pooled together with another EA with similar characteristics to form primary sampling units (PSU). An EA with 80 or more dwellings automatically qualified to become a PSU. Census listings of the selected PSUs were updated where necessary and then a systematic sample of 10 dwelling units was selected in each PSU. Because there was sometimes more than one household at each dwelling unit, the sample of 50,000 dwelling units produced a sample of 52,376 households. Section 7 of the questionnaire required the selection of one person aged 15 years and above to answer the attitude questions. This person was randomly selected using a grid. The sample design for the National Household Travel Survey (NHTS) 2013 was based on the Census 2011 enumeration areas (EAs) frame and was based on two-staged random stratified sampling. Firstly, a sample of 5,034 primary sampling units (PSUs) was selected from the Census dwelling frame, with stratification at TAZ and provincial levels. Twenty-two of these PSUs were vacant and 51,341 dwelling units (DUs) were sampled from the remaining 5,012 PSUs. Of the sampled DUs, there were 849 DUs for which no questionnaires were received or completed. There were 4,957 PSUs that had at least one responding household. Furthermore, 5 PSUs had all sampled DUs with 'out-of-scope' households, while the remaining 50 PSUs had sampled DUs without responding households. More details about this can be found in the technical report *Weighting* For the 2003 survey, a two stage weighting procedure was used: adjusting for non-response and benchmarking where population totals were adjusted at municipality level and gender, five-year age group and race were taken into consideration at national level. The adjusted weights for the National Household Travel Survey (NHTS) 2013 full sample were obtained by applying three adjustments to the base-weights (also known as design weights). The first adjustment was applied to account for PSU natural growth; the adjustment factors were truncated at the 99th percentile (which was 2.32432) in an attempt to minimise the sample variation. The second adjustment was applied to account for the EAs with fewer than 25 households excluded during the survey design (i.e. adjustment for the Take-none portion), and the third was the non-response adjustment. There were two types of non-response adjustments: PSU nonresponse adjustment and household nonresponse adjustment. The PSU non-response adjustment was applied at the stratum level, whereas the household nonresponse adjustment was applied at the PSU level. The final calibrated weights were constructed by calibrating the adjusted design weights to the known population estimates as control totals using the 'Integrated Household Weighting' method. The lower bound for

	the calibrated weights was set equal to 50 when computing the calibrated weights with the StatMx software.
Availability of data descriptions	The following documentation is available from the Nesstar (http://interactive.statssa.gov.za:8282/webview/): Questionnaire Metadata Concepts and Definitions Metadata is also available on the DataFirst website (http://datafirst.uct.ac.za/).
Conditions of obtaining microdata	Stats SA: Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. Data First: Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: info@statssa.gov.za http://www.statssa.gov.za/ For data: http://interactive.statssa.gov.za:8282/webview/ https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	***National Income Dynamics Study***
Principal investigator	Southern Africa Labour and Development Research Unit (SALDRU), University of Cape Town
Year(s)	2008, 2010, 2012, 2014
Area(s) of interest	Labour market; health; education; housing; economy; social welfare
Source(s) of data description provided here	NIDS website: http://www.nids.uct.ac.za/index.php
Brief description	*Summary* The National Income Dynamics Study (NIDS) is the first national household panel study in South Africa. It is part of an intensive multi-million rand effort on the part of the government to track and understand the shifting face of poverty. The National Income Dynamics Study is implemented by the Southern Africa Labour and Development Research Unit (SALDRU) based at the University of Cape Town's School of Economics. The study began in 2008 with a nationally representative sample of over 28,000 individuals in 7,300 households across the country. The survey continues to be repeated with these same household members every two years. • 2008: Wave 1 • 2010: Wave 2 • 2012: Wave 3 • 2014: Wave 4 NIDS examines the livelihoods of individuals and households over time. It also provides information about how households cope with positive or negative shocks, such as a death in the family or an unemployed relative obtaining a job. Other themes include changes in poverty and well-being; household composition and structure; fertility and mortality; migration; labour market participation and economic activity; human capital formation, health and education; vulnerability and social capital. The study captures information on: • 'Continuing Sample Members' (CSMs) who are all resident members of the original selected Wave 1 households (including children) and any children born to female CSMs in subsequent waves, and • 'Temporary Sample Members (TSMs), who are people that are not a CSM but are co-resident with a CSM at the time of the interview. *Methodology* NIDS uses a combination of household and individual level questionnaires. The questionnaires are administered through face-to-face interviewing. The data from the different questionnaires are recorded in separate data files with one row per record (individual or household). A set of files is released for each wave, but they can be combined across waves using the unique identifier for the individual. In each wave, four types of questionnaires are administered: • Household questionnaire: One Household questionnaire is completed per household by the oldest woman in the household or another person knowledgeable about household affairs and particularly household spending.

Household questionnaires take approximately 39 minutes in non-agricultural households and 50 minutes in agricultural households to complete.

- Adult questionnaire: The Adult questionnaire is applied to all present CSMs and other household members resident in their households that are aged 15 years or over. This questionnaire takes an average of 38 minutes per adult to complete.
- Proxy questionnaire: Should an individual qualifying for an Adult questionnaire not be present, then a Proxy questionnaire (a much reduced Adult questionnaire using third party referencing in the questioning) is taken on their behalf with a present resident adult. On average, a Proxy questionnaire takes 12 minutes to complete. Proxy questionnaires are also asked for CSMs who have moved out of scope (out of South Africa or to a non- accessible institution such as prison), except if the whole household has moved out of scope, and can therefore not be tracked or interviewed directly.
- Child questionnaire: This questionnaire collects information about all CSMs and residents in their household younger than 15. Information about the child is gathered from the care- giver of the child. The questionnaire focuses on the child's educational history, education, anthropometrics and access to grants. This questionnaire takes an average of 16 minutes per child to complete.

The NIDS team state that every effort has been made to be consistent in the data collection methodology applied across waves, while also paying attention to being more efficient in field operations. From Wave 2 onwards, all data have been collected using Computer Assisted Personal Interviewing (CAPI) software, which has been extended and improved upon over time. Use of paradata to monitor interviewer performance has also been developed in order to improve the quality of data collected and so reduce interviewer effects.

Sampling

A stratified, two-stage cluster sample design was employed in sampling the households to be included in the base wave.

In the first stage, 400 Primary Sampling Units (PSUs) 2 were selected from Stats SA's 2003 Master Sample of 3000 PSUs. This Master Sample was the sample used by Stats SA for its Labour Force Surveys and General Household Surveys between 2004 and 2007 and for the 2005/06 Income and Expenditure Survey. Each of these surveys was conducted on non-overlapping samples drawn within each PSU. The target population for NIDS was private households in all nine provinces of South Africa and residents in workers' hostels, convents and monasteries. The frame excludes other collective living quarters such as students' hostels, old age homes, hospitals, prisons and military barracks.

The sample of PSUs for NIDS is a subset of the Master Sample. The explicit strata in the Master Sample are the 53 district councils (DCs). The sample was proportionally allocated to the strata based on the Master Sample DC PSU allocation and 400 PSUs were randomly selected within strata. It should be noted that the sample was not designed to be representative at provincial level, implying that analysis of the results at province level is not recommended.

At the time that the Master Sample was compiled, 8 non-overlapping samples of dwelling units were systematically drawn within each PSU. Each of these samples is

called a “cluster” by Stats SA. These clusters were then allocated to the various household surveys that were conducted by Stats SA between 2004 and 2007. However, two clusters in each PSU were never used by Stats SA and these were allocated to NIDS.

All resident household members in the sampled dwelling units became NIDS sample members. In addition, non-resident members that were “out of scope” at the time of the survey also became NIDS sample members. Out-of-scope household members were those living in institutions (such as boarding school hostels, halls of residence, prisons or hospitals) which were not part of the sampling frame. These individuals had a zero probability of selection at their usual place of residence and were thus included in the NIDS sample as part of the household that had listed them as non-resident members. These two groups constitute the permanent sample members (PSMs) and should have had an individual questionnaire (adult, child or proxy) completed for them. These individuals are PSMs even if they refused to be interviewed in the base wave.

Weighting
The NIDS team acknowledge (and caution users that) it can be rather difficult to keep track of all the different types of weights that there are in the National Income Dynamics Study. They state that fundamentally there are three types of weights:
a) Design weights (correcting for nonresponse)
b) Calibrated weights
c) Panel weights
The design weights released with Wave 1 are fundamental to every other weight released with NIDS14. They are used to calculate the corresponding design weights for waves 2 and 3.

Together with Wave 4 of the National Income Dynamics Study, updates to Waves 3, 2 and 1 have been released. Since the information on the sample for these waves has changed a little (e.g. age information has been improved, some households have been removed) it has been necessary to recalculate all the weights previously released as well. Indeed, since a few households have been removed from Wave 1, even the “design weights correcting for nonresponse” will be slightly different in the affected clusters. Furthermore, the way deceased respondents are handled has been adjusted. While the initial calculation of weights included a correction for people who died, this was conceptually incorrect, therefore the new set of weights only correct for non-response. Nevertheless, the methods used, i.e. the algorithms underpinning the calculations, have not been changed. This means that the revised weights will be very similar in most cases to the ones released previously. Indeed, because the algorithms have not been changed, the documentation released with previous weights should also be consulted for further information.

The calibrated weights, however, have changed in that all calibration has happened in line with the revised mid-year population estimates as released by Statistics South Africa (StatsSA) in 2015. This was necessary to ensure that the population totals (and totals within particular provinces and age groups) did not jump discontinuously as a result of the upward revision of South Africa’s overall population size. In practice, this means that the calibrated weights for 2008, 2010 and 2012 will now gross up to slightly larger totals than before.

Each of the waves, treated as a cross-section of the South African population, has been separately calibrated to the corresponding population totals as given in the mid-year

	population estimates released in 2015. All waves were calibrated to provincial totals and to gender-race-age group cell totals (with the oldest three age categories for Indian males and Indian females collapsed, as noted in the release notes accompanying the previous release). All individuals within the same household were constrained to get the same weight.
Availability of data descriptions	The NIDS website contains a wealth of detailed technical documentation for all four waves of the study: http://www.nids.uct.ac.za/index.php
Conditions of obtaining microdata	The NIDS data can be downloaded from the DataFirst website: http://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central/about See the "how to register' video can be viewed by clicking here or follow steps below. The steps to follow to gain access to the data are: • Step 1: Register as a user on the DataFirst website. Once you have registered on the DataFirst website the registration details can be used to access datasets from the website. • Step 2: Complete a short online Application for Access to a Public Use Dataset for the NIDS datasets. On the form you will need to provide a short description of your intended use of the data. The information provided here helps us to understand how NIDS data is being used by the research community. The form also asks you to agree to Terms and Conditions related to the use of the NIDS data, namely: a. The data provided by DataFirst will not be redistributed or sold to other individuals, institutions, or organisations without the written agreement of DataFirst. b. The data will be used for statistical and scientific research purposes only. They will be used solely for reporting of aggregated information, and not for investigation of specific individuals or organisations. c. No attempt will be made to re-identify respondents, and no use will be made of the identity of any person or establishment discovered inadvertently. Any such discovery should immediately be reported to NIDS at the following address: nids-survey@uct.ac.za. d. No attempt will be made to produce links among datasets provided by DataFirst, or among data from DataFirst and other datasets that could identify individuals or organisations. e. Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from DataFirst will cite the source of data in accordance with the Citation Requirement provided with each dataset. f. A digital copy of all reports and publications based on the requested data will be sent to DataFirst. g. The original collector of the data, DataFirst, and the relevant funding agencies bear no responsibility for use of the data or for interpretations or inferences based upon such uses. • Step 3: Download the data. Selected coding and syntax files can also be downloaded at this stage.
Contact for Information and Data Supply	Tel: +27 (0)21 650 5968 Fax: +27 (0)21 650 5403 Email: nids-survey@uct.ac.za

Name	***National Victims of Crime Survey***
Principal investigator	Stats SA (n.b. 2003 and 2007 studies undertaken by ISS)
Year(s)	1998, 2003, 2007, 2011, 2012, 2013,/14, 2014/15
Area(s) of interest	Crime
Source(s) of data description provided here	StatsSA VOCS website: http://www.statssa.gov.za/publications/P0341/P03412014.pdf
Brief description	*Summary* The Victims of Crime Survey (VOCS) series is a countrywide household-based survey and examines three aspects of crime: • The nature, extent and patterns of crime in South Africa, from the victim's perspective; • Victim risk and victim proneness, so as to inform the development of crime prevention and public education programmes; • People's perceptions of services provided by the police and the courts as components of the criminal justice system. Data from victimisation surveys can be used to supplement official police recorded crime statistics. StatsSA undertook the first VOCS in 1998. The Institute for Security Studies (ISS) was responsible for conducting the 2003 and 2007 versions of the VOCS. Starting with the VOCS 2011, Stats SA has begun to conduct the VOCS annually. It should be noted that while the question compositions in the pre-2011 surveys were largely similar to the newer surveys from 2011 onwards, the sample sizes were much smaller in the earlier surveys. The focus here is on the VOCS from 2011 onwards. Please see other referenced resources for details of earlier VOCS studies. *Methodology* Stats SA conducted the 2014/15 Victims of Crime Survey in close collaboration with other role players in the Safety and Security cluster between April 2014 and March 2015. Since 2013 the Victims of Crime Survey, the Domestic Tourism Survey(DTS) and the General Household Survey(GHS) have adopted the Continuous Data Collection(CDC) methodology. The Victims of Crime Survey 2014/15 conducted data collection from April to March. In the long run, this methodology will enable data collection to coincide with the financial year and the reporting cycle of administrative data related to crime. Data collection took place from April 2014 to March 2015 with a moving reference period of 12 months. This is different from the 2011 and 2012 collections which were done from January to March and had a fixed reference period from January to December of the previous year. The sample has been distributed evenly over the whole collection period in the form of quarterly allocations. This will provide a guarantee against possible seasonal effects in the survey estimates. It will, in future, provide an opportunity for the production of rolling estimates relating to any desired time period. It has been noted that the change of data collection methodology may cause concerns over the survey estimates, particularly upon comparisons of years before and after the change. Victimisation questions referred to the twelve calendar months ending with the month before the interview.

	Sample design For the 2011 and subsequent VOCS the sample design was based on a master sample (MS) originally designed as the sampling frame for the Quarterly Labour Force Survey (QLFS). The MS is based on information collected during the 2001 Population Census conducted by Stats SA. The MS has been developed as a general-purpose household survey frame that can be used by all household-based surveys, irrespective of the sample size requirement of the survey. The VOCS, like all other household-based surveys, uses a MS of primary sampling units (PSUs) which comprises census enumeration areas (EAs) that are drawn from across the country. The sample for the VOCS used a stratified two-stage design with probability proportional to size (PPS) sampling of PSUs in the first stage, and sampling of dwelling units (DUs) with systematic sampling in the second stage. The sample was designed to be representative at provincial level. A self-weighting design at provincial level was used and MS stratification was divided into two levels. Primary stratification was defined by metropolitan and non-metropolitan geographic area type. During secondary stratification, the Census 2001 data were summarised at PSU level. The following variables were used for secondary stratification: household size, education, occupancy status, gender, industry and income. A randomised probability proportional to size (RPPS) systematic sample of PSUs was drawn in each stratum, with the measure of size being the number of households in the PSU. The sample size of 3 080 PSUs was selected. In each selected PSU a systematic sample of dwelling units was drawn. The number of DUs selected per PSU varies from PSU to PSU and depends on the inverse sampling ratios (ISR) of each PSU. *Weighting* Sampling weights for the data collected from the households sampled in the 2011 and subsequent surveys are constructed in such a manner that the responses could be properly expanded to represent the entire civilian population of South Africa. The design weights, which are the inverse sampling rate (ISR) for the province, are assigned to each of the households in a province. The design weights for the sample were obtained by applying three adjustments to the base-weights. The first adjustment was applied to account for informal and/or growth PSUs. The second adjustment was applied to account for the EAs with less than 25 households and the third was the non-response adjustment. In addition, there were two types of non-response adjustments: PSU non-response adjustment and household non-response adjustment. The PSU non-response adjustment was applied at the stratum level, whereas the household non-response adjustment was applied at the PSU level. The final survey weights were constructed by calibrating the adjusted non-response design weights to the known population estimates as control totals using the 'Integrated Household Weighting' method. The lower bound for the calibrated weights was set equal to 50 when computing the calibrated weights with the StatMx software (Statistics Canada software). The VOCS 2011 sample was weighted using the population estimate of mid-November 2010; population estimates for mid-November 2011 were used for the VOCS 2012; the VOCS 2013/14 sample was calibrated using the Population Estimate of Mid May 2013 and the VOCS 2014/15 sample was calibrated using the Population Estimate of Mid May 2014. The final weights were benchmarked to the known population estimates of 5-year age groups by population groups by gender at national level, and broad age groups at province level. The calibrated weights were constructed in such a way that all persons in a household would have the same final weight. Records for which the age, population

	group or gender had item non-response could not be weighted and were therefore excluded from the dataset. No additional imputation was done to retain these records.
Availability of data descriptions	The following documentation is available from the Nesstar (http://interactive.statssa.gov.za:8282/webview/): Questionnaire Metadata Concepts and Definitions Metadata is also available on the DataFirst website (http://datafirst.uct.ac.za/).
Conditions of obtaining microdata	Stats SA: Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. Data First: Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: Tel: (012) 310 8600 Fax: (012) 310 8944 nfo@statssa.gov.za http://www.statssa.gov.za/ For data: http://interactive.statssa.gov.za:8282/webview/ https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	***National Youth Lifestyle Survey (NYLS)***
Principal investigator	Centre for Justice & Crime Prevention (CJCP)
Year(s)	2005, 2008
Area(s) of interest	Demographics, Crime, Education, Health
Source(s) of data description provided here	www.cjcp.org.za
Brief description	*Summary* The CJCP embarked on a National Youth Victimisation Study (NYVS) in 2005, which involved interviewing 4,409 young people between the ages of 12 and 22 years recruited from all nine provinces of South Africa. The study provided compelling evidence to suggest that young people in South Africa are disproportionately at risk of falling prey to crime compared to their adult counterparts. Since the initial study was intended to provide baseline data on the nature and extent of youth victimisation, the CJCP conducted a second sweep of the study in 2008. Participants in the 2008 NYLS responded to a survey questionnaire exploring various issues such as the extent of crimes they may have experienced in the past twelve months, as well as their exposure to violence in the different social contexts in which they live. The survey sampled 4391 young people in the specified age range. The study also measured self-reported offending and youth engagement in risky behaviour. *Sampling* As with the 2005 study, the sample used here was designed to be proportionately representative in order to make it reflective of the South African population. The sample frame was provided by Statistics South Africa 2001 Census data, and the sample was stratified by province and race. The total population between the ages of 12 and 22 years was identified. Based on this, a sample of 550 enumerator areas (EAs) was randomly selected, with eight households identified to be interviewed in each. Each EA was mapped, each household within the EA assigned a number, and a list of all houses within the EA compiled. Households were then randomly selected from this numbered list and visited by enumerators. Where a youth between the ages of 12 and 22 lived in the household and was available and willing to participate in the study, an interview was conducted. The next house on the list was visited if no respondent falling within the required age cohort lived in the house. *Weighting* The final data was weighted by province, race and gender using the marginal totals drawn from the 2001 Census. This was done to ensure the most accurate representation of the experiences of young people throughout South Africa.
Availability of data descriptions	http://www.cjcp.org.za/uploads/2/7/8/4/27845461/monograph_6_-_running_nowhere_fast_-_2008_youth_lifestyle.pdf http://www.cjcp.org.za/uploads/2/7/8/4/27845461/research_bulletin_3_-_snapshot_results_of_the_2008_youth_lifestyle_study.pdf
Conditions	Not available

Contact for Information and Data Supply	Tel: +27 (0)21 687 9177 Fax: +27 (0)21 685 3284 Email: wendy@cjcp.org.za www.cjcp.org.za

Name	***Post Apartheid Labour Market Series***
Principal investigator	DataFirst
Year(s)	1994-2015
Area(s) of interest	Labour market
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys)
Brief description	*Purpose* The Post Apartheid Labour Market Series (PALMS) 1994-2015 was constructed in order to provide a single repository of harmonised labour market indicators over the post-apartheid period. *Summary* The PALMS version 3.1 stacked cross sectional dataset consists of microdata from 54 household surveys conducted by Statistics South Africa between 1994 and 2015, as well as the 1993 Project for Statistics on Living Standards and Development conducted by the Southern Africa Labour and Development Research Unit (SALDRU) at the University of Cape Town. The Statistics South Africa surveys include the October Household Surveys from 1994 to 1999, the bi-annual Labour Force Surveys from 2000-2007, including the smaller LFS pilot survey from February 2000, and the Quarterly Labour Force Surveys from 2008-2015. The data is at individual level, but household level variables may be created using the unique household identification variable. No attempt has been made to link individuals or households across waves, although there was a panel element to the earlier rounds of the LFS. *Methodology* The OHS, LFS and QLFS data have been prepared separately and then appended together. *Sample design* As PALMS is a harmonised compilation of existing survey microdata, please see the relevant sections relating to the respective input survey data sources. *Weighting* PALMS v3+ includes several weight variables: (i) the person weights released by Statistics South Africa/SALDRU; (ii) cross entropy weights created by Nicola Branson from SALDRU at the University of Cape Town (which uses a model for 2003 from the South African population from the Actuarial Society of South Africa (ASSA)); and (iii) cross entropy weights created by Takwanisa Machemedze at DataFirst, UCT (who uses Branson's method but the 2008 ASS model). Machemedze's cross entropy weights are included and are the weighting variables preferred by DataFirst (please see referenced technical documentation for further details of DataFirst's recommendation in this regard).
Availability of data descriptions	Documentation is also available on the DataFirst website (http://www.datafirst.uct.ac.za/surveys).
Conditions	Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give

	assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: support@data1st.org +2721 650 5708 For data: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central/about

Name	***Quarterly Employment Survey (QES)***
Principal investigator	Stats SA
Year(s)	2006-2016
Area(s) of interest	Economy; Labour Market
Source(s) of data description provided here	Stats SA: http://www.statssa.gov.za/publications/P0277/P0277March2016.pdf
Brief description	*Summary* The Quarterly Employment Survey is a quarterly survey covering a sample of approximately 20,208 private and public enterprises in the formal non-agricultural sector of the South African economy. The survey provides data essential for estimating key economic statistics of employment and gross earnings. These economic data are used by a wide range of private and governmental organisations to monitor South Africa's Economy. Survey results are published in the statistical release P0277 – Quarterly Employment Survey. *Scope of survey:* This survey covers employment statistics of the following industries according to the Standard Industrial Classification of all Economic Activities (SIC), Fifth edition, January 1993: • mining and quarrying; • manufacturing; • electricity, gas and water supply; • construction; • wholesale and retail trade; repair of motor vehicles, motor cycles and personal and household goods; and hotels and restaurants; • transport, storage and communication; • financial intermediation, insurance, real estate and business services; • community, social and personal services The latest QES survey results are based on a sample drawn from the 2015 Business Sampling Frame (BSF) of Stats SA which contains enterprises registered for Value Added Tax (VAT) at the South African Revenue Service (SARS). The BSF is continuously updated by adding new enterprises and ceasing dormant enterprises. The Quarterly Employment Survey complements the Quarterly Labour Force Survey (QLFS). Although the results may differ due to different coverage and conceptual issues, there are some adjustments which can be conducted on the QLFS data to improve the comparability of these surveys. However, in some cases even after the adjustments are made, some differences are difficult to explain due to the business frame instability over time. Over the years the business frame has improved and will continue to improve as administrative systems improve their integrity. Statistics South Africa has embarked on a continuous improvement plan of the frame and this is likely to cause breaks in the series in future. The numerous conceptual and methodological differences between the QLFS & QES based surveys result in important distinctions in the employment estimates derived from the surveys. Among these are:

	• The household survey includes agricultural workers, self-employed workers whose businesses are unincorporated, unpaid family workers, and private household workers among the employed. These groups are excluded from the enterprise based survey. • The household survey is limited to workers 15 years of age and older. The enterprise based survey is not limited by age. • The household survey has no duplication of individuals, because individuals are counted only once, even if they hold more than one job. In the enterprise based survey, employees working at more than one job and thus appearing on more than one payroll are counted separately for each appearance. • QLFS includes income tax, VAT and number of employees in determining the formal sector while QES use only VAT. Statistics based on the household and enterprise based surveys are subject to both sampling and non-sampling error *Statistical unit:* The statistical unit for the collection of information is an enterprise. An enterprise is a legal unit or a combination of legal units that includes and directly controls all functions necessary to carry out its production activities. *Classification:* The Standard Industrial Classification of all Economic Activities (SIC), Fifth edition, January 1993, was used to classify the statistical units in the survey. The SIC is based on the 1990 International Standard Industrial Classification of all Economic Activities (ISIC), with suitable adaptations for local conditions. Statistics in this publication are only presented at the SIC major division (one digit) level. Each enterprise is classified to the industry which reflects the predominant activity of the enterprise.
Availability of data descriptions	Stats SA: http://www.statssa.gov.za/?page_id=1866&PPN=P0277&SCH=6683&page=1
Conditions	Stats SA has copyright on this publication. Users may apply the information as they wish, provided that they acknowledge Stats SA as the source of the basic data wherever they process, apply, utilise, publish or distribute the data; and also that they specify that the relevant application and analysis (where applicable), result from their own processing of the data.
Costs	None
Contact for Information and Data Supply	labourquestions@statssa.gov.za (technical enquiries) info@statssa.gov.za (user information services)

Name	***Quarterly Labour Force Survey***
Principal investigator	StatsSA
Year(s)	2000 – to present
Area(s) of interest	Labour market
Source(s) of data description provided here	StatsSA: http://www.statssa.gov.za/publications/P0211/P02112ndQuarter2016.pdf
Brief description	*Summary* Between 2000 and 2007 the Labour Force Survey consisted of a twice-yearly rotating panel household survey, measuring the dynamics of employment and unemployment in the country. From 2008 onwards the Labour Force Survey become quarterly, and is now widely referred to as the Quarterly Labour Force Survey (QLFS). The focus here is on the methodology from 2008 onwards. The Quarterly Labour Force Survey (QLFS) is a household-based sample survey conducted by Statistics South Africa (StatsSA). It collects data on the labour market activities of individuals aged 15 years or older who live in South Africa. Since 2008, StatsSA have also produced an annual dataset based on the QLFS data, "Labour Market Dynamics in South Africa". The dataset is constructed using data from all four QLFS datasets in the year. The QLFS sample covers the non-institutional population except for workers' hostels. However, persons living in private dwelling units within institutions are also eligible for enumeration. For example, within a school compound, one could enumerate the schoolmaster's house and teachers' accommodation because these are private dwellings. Students living in a dormitory on the school compound would, however, be excluded. *Methodology* The QLFS is conducted as a face to face interview with household respondents. *Sample design* The QLFS frame has been developed as a general purpose household survey frame that can be used by all other household surveys irrespective of the sample size requirement of the survey. The sample size for the QLFS is roughly 30,000 dwellings per quarter. The Quarterly Labour Force Survey (QLFS) uses the StatsSA Master Sample frame. The 2013 Master Sample is based on information collected during the 2011 Census conducted by Stats SA. In preparation for Census 2011, the country was divided into 103,576 enumeration areas (EAs). The census EAs, together with the auxiliary information for the EAs, were used as the frame units or building blocks for the formation of primary sampling units (PSUs) for the Master Sample, since they covered the entire country and had other information that is crucial for stratification and creation of PSUs. There are 3,324 primary sampling units (PSUs) in the Master Sample with an expected sample of approximately 33,000 dwelling units (DUs). The number of PSUs in the current Master Sample (3,324) reflects an 8,0% increase in the size of the Master Sample compared to the previous (2008) Master Sample (which had 3,080 PSUs). The larger Master Sample of PSUs was selected to improve the precision (smaller coefficients of variation, known as CVs) of the QLFS estimates.

	The Master Sample is designed to be representative at provincial level and within provinces at metro/non-metro levels. Within the metros, the sample is further distributed by geographical type. The three geography types are Urban, Tribal and Farms[47]. This implies, for example, that within a metropolitan area, the sample is representative of the different geography types that may exist within that metro. They survey is divided equally into four sub-groups or panels called rotation groups. The rotation groups are designed in such a way that each of these groups has the same distribution pattern as that which is observed in the whole sample. They are numbered from 1 to 4 and these numbers also correspond to the quarters of the year in which the sample will be rotated for the particular group. The sample for the QLFS is based on a stratified two-stage design with probability proportional to size (PPS) sampling of PSUs in the first stage, and sampling of dwelling units (DUs) with systematic sampling in the second stage. Sample rotation: For each quarter of the QLFS, a ¼ of the sampled dwellings are rotated out of the sample. These dwellings are replaced by new dwellings from the same PSU or the next PSU on the list. Thus, sampled dwellings are expected to remain in the sample for four consecutive quarters. It should be noted that the sampling unit is the dwelling, and the unit of observation is the household. Therefore, if a household moves out of a dwelling after being in the sample for, say two quarters (as an example) and a new household moves in, the new household will be enumerated for the next two quarters. If no household moves into the sampled dwelling, the dwelling will be classified as vacant (or unoccupied). *Weighting* The sample weights were constructed in order to account for the following: the original selection probabilities (design weights), adjustments for PSUs that were sub-sampled or segmented, excluded population from the sampling frame, non-response, weight trimming, and benchmarking to known population estimates from the Demographic Analysis Division within Stats SA. Non-response adjustment: In general, imputation is used for item non-response (i.e. blanks within the questionnaire) and edit failures (i.e. invalid or inconsistent responses). The eligible households in the sampled dwellings can be divided into two response categories: respondents and non-respondents. Weight adjustment is applied to account for the non-respondent households (e.g. refusal, no contact, etc.). The adjustment for total non-response was computed at two levels of nonresponse: PSU non-response and household non-response. Final survey weights: In the final step of constructing the sample weights, all individuals within a household are assigned the same adjusted base weight. The adjusted base weights are calibrated such that the aggregate totals will match with independently derived population estimates (from the Demographic Analysis Division) for various age, race and gender groups at national level and individual metropolitan and non-metropolitan area levels within the provinces. The calibrated weights are constructed using the constraint that each person within the same household should have the same calibrated weight, with a lower bound on the calibrated weights set at 50.

[47] These are the terms used by StatsSA at that time to describe area types.

	Estimation The final survey weights are used to obtain the estimates for various domains of interest, e.g. number of persons employed in Agriculture in Western Cape, number of females employed in Manufacturing, etc.
Availability of data descriptions	The following documentation is available from the Nesstar (http://interactive.statssa.gov.za:8282/webview/): Questionnaire Metadata Concepts and Definitions Metadata is also available on the DataFirst website (http://datafirst.uct.ac.za/).
Conditions of obtaining microdata	Stats SA Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA.
Contact for Information and Data Supply	For information: info@statssa.gov.za http://www.statssa.gov.za/ For data: http://interactive.statssa.gov.za:8282/webview/ https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	***SAGE: Study on global AGEing and adult health***
Principal investigator	WHO & HSRC
Year(s)	2007/08 (wave 1)
Area(s) of interest	Health
Source(s) of data description provided here	WHO: http://apps.who.int/healthinfo/systems/surveydata/index.php/catalog/5/related_materials HSRC : http://www.hsrc.ac.za/en/research-outputs/view/6320
Brief description	*Summary* The multi-country Study on Global Ageing and Adult Health (SAGE) is run by the World Health Organization's Multi-Country Studies unit in the Innovation, Information, Evidence and Research Cluster. SAGE is part of the unit's Longitudinal Study Programme which is compiling longitudinal data on the health and well-being of adult populations, and the ageing process, through primary data collection and secondary data analysis. SAGE baseline data (Wave 0, 2002/3) was collected as part of WHO's World Health Survey http://www.who.int/healthinfo/survey/en/index.html (WHS). SAGE Wave 1 (2007/10) provides a comprehensive data set on the health and well-being of adults in six low and middle-income countries: China, Ghana, India, Mexico, Russian Federation and South Africa. The main objectives are: • To obtain reliable, valid and comparable health, health-related and well-being data over a range of key domains for adult and older adult populations in nationally representative samples. • To examine patterns and dynamics of age-related changes in health and well-being using longitudinal follow-up of a cohort as they age, and to investigate socio-economic consequences of these health changes. • To supplement and cross-validate self-reported measures of health and the anchoring vignette approach to improving comparability of self-reported measures, through measured performance tests for selected health domains. • To collect health examination and biomarker data that improves reliability of morbidity and risk factor data and to objectively monitor the effect of interventions. *Methodology* SAGE's first full round of data collection included both follow-up and new respondents in most participating countries. The goal of the sampling design was to obtain a nationally representative cohort of persons aged 50 years and older, with a smaller cohort of persons aged 18 to 49 for comparison purposes. In the older households, all persons aged 50+ years (for example, spouses and siblings) were invited to participate. Proxy respondents were identified for respondents who were unable to respond for themselves. Standardized SAGE survey instruments were used in all countries consisting of five main parts: 1) household questionnaire; 2) individual questionnaire; 3) proxy questionnaire; 4) verbal autopsy questionnaire; and, 5) appendices including showcards. A VAQ was completed for deaths in the household over the last 24 months. The procedures for including country-specific adaptations to the standardized questionnaire and translations into local languages from English follow those developed by and used for the World Health Survey.

	Sample design The SAGE sample design entails a two-stage probabilistic sample that yields national and sub-national estimates to an acceptable precision at provincial level, by residence (urban and rural), and by population group (including Black, Coloured, Indian or Asian and White). The first stage of sampling was the selection of primary sampling units (PSUs), using the 2002 HSRC master sample as the sampling frame (HSRC 2005). The master sample is a probabilistic sample of 1000 enumeration areas (EA) drawn from the South African National Census, conducted by Statistics South Africa in 2001 (Statistics South Africa 2003). For the SAGE study, a total sample of 600 EAs was drawn from the master sample and used as the PSUs. This stage of selection was done centrally at the HSRC. The master sample was stratified by province, residence and race, and the EAs were then selected with a probability proportional to size, with the estimated number of people aged 50 years or older in each EA as a measure of size. Thus, EAs with a larger number of people aged 50 years or older had a higher chance of being selected. The second stage of the sample design was the selection of visiting points (VPs) – in this case, households – which formed the secondary sampling units. To ensure that an adequate number of households with at least one person aged 50 years or older was selected, 30 households were randomly selected from each EA, and screened to identify the presence of a person 50 years or older. If the household had at least one person 50 years or older, then that household was included in the 50 years or older sample. The remaining households (that is, with no member 50 years or older) were used to randomly select two households and, in each of these, one respondent aged 18–49 years was randomly selected using Kish tables. A cohort of younger adults (aged 18–49 years) was included for comparison purposes. The sample contained EAs with different numbers of households containing people aged 50 years or older, and only two households with people aged 18–49 years. Altogether, about 18,000 households were targeted (that is, 600 EAs with 30 households in each). The individual eligible for interview in selected households formed the ultimate sampling unit. The total sample size of individuals was targeted to be 1000 people in the age group 18–49 years, and 5000 people aged 50 years or older. In the sample of households with people aged 50 years or older, anyone aged 50 years or older was eligible for interview. If the household had an eligible member who was unavailable for interview, then up to three revisits were made. In the case that a usual member was at an old-age home or visiting a hospital within 100 km, then an attempt was made to visit the person at that institution for an interview. Although the targeted number of people 50 years or older was 5000 from among the 600 EAs, for the process described above, it was not possible to predict the exact number of older people before fieldwork. In the sample of households with people in the age group 18–49 years, two households per EA were selected: a sample size sufficiently large to allow for a margin of refusals. Thus, for the 18–49 age group, 1200 households were selected, from which about 1000 individuals were eligible for interview The SAGE South Africa study team did not follow up the Wave 0 sample in Wave 1, but will attempt to do so in Wave 2.
Availability of data descriptions	Not available

Conditions of obtaining microdata	Not available
Contact for Information and Data Supply	For information: datahelp@hsrc.ac.za For data: datahelp@hsrc.ac.za

Name	***South Africa Demographic and Health Survey (SADHS)***
Principal investigator	Department of Health (DoH)
Year(s)	2016
Area(s) of interest	Demography; Health
Source(s) of data description provided here	StatsSA: http://www.statssa.gov.za/?p=6039
Brief description	*Summary* The survey aims to provide a better understanding of the health status of the population in South Africa. The information collected will assist the Department of Health to plan and prioritize health programmes and service delivery. It also provides an opportunity for household members to understand their individual health status. The survey was previously conducted in 1998 and 2003 The main purpose of this survey is: ▪ To assist the DoH to plan and prioritize health programmes and service delivery ▪ To gain a better understanding of the health status of the population in South Africa. ▪ To collect data on a variety of demographic, health and nutrition aspects at national and provincial level monitor the health status, coverage and quality of selected health programmes A critical part of the survey will be to update the Dwelling Frame. This process involves the pre-loading of Dwelling Frame data on digital devices and updating information on both residential and non-residential structures in the sampled areas. The Dwelling Frame update will take place from 1 February to 30 March 2016 in 750 sampled areas across the country. The following will be updated: • all dwelling units; • all rooms or units within collective living quarters; • all non-residential buildings; • all vacant stands; • sports fields; • parks; • parking lots; • cemeteries; • demolished structures; and • semi-demolished structures that fall within the sampled area's boundaries *Methodology* A data collection team consisting of seven people will be working in the selected areas to conduct interviews. A trained nurse will be accompanying the field workers and may conduct medical testing. Men, women and caregivers of children will be asked questions about their well-being.

	Sample design The 2016 survey will be conducted across all provinces and a sample of 15 000 households will be targeted
Availability of data descriptions	Not yet available
Conditions of obtaining microdata	Not yet available
Contact for Information and Data Supply	www.statssa.co.za www.health.gov.za

Name	***South African National Health & Nutrition Examination Survey (SANHANES)***
Principal investigator	HSRC
Year(s)	2011/12
Area(s) of interest	Non-communicable diseases, adult & child nutrition, tobacco & alcohol use, health, health care services, behavioural and social determinants of health & nutrition, risk factors
Source(s) of data description provided here	www.hsrc.ac.za
Brief description	*Summary* SANHANES aims to assess selected aspects of the health and nutritional status of the South African population. The information gathered from the survey will be used to address the National Department of Health's (NDOH) priority health indicators SANHANES-1 was undertaken during 2011/12. The primary objectives of the SANHANES-1 were to assess defined aspects of the health and nutritional status of South Africans with respect to the prevalence of non-communicable diseases (NCDs) (specifically cardiovascular disease, diabetes and hypertension) and their risk factors (diet, physical activity and tobacco use). Other objectives were to assess the knowledge, attitudes and behaviour of South Africans with respect to non-communicable and communicable infectious diseases. *Methodology* SANHANES-1 included individuals of all ages living in South Africa. All persons living in occupied households (HHs) were eligible to participate, but individuals staying in educational institutions, old-age homes, hospitals, homeless people, and uniformed-service barracks were not eligible to participate in the survey. SANHANES-1 obtained questionnaire-based data through interviews in combination with health measurements obtained through a clinical examination, a selection of clinical tests as well as the collection of a blood sample for selected biomarker analysis. This first round of the SANHANES (SANHANES-1) was a cross-sectional survey providing baseline data for future longitudinal analysis. The SANHANES project also combined longitudinal as well as cross-sectional design elements. A prospective cohort approach addressed the relationships between medical, nutritional and behavioural/societal risk factors assessed in the first survey phase (SANHANES-1) and subsequent morbidity, mortality and changes in risk factors at the national level. *Sampling* A multi-stage disproportionate, stratified cluster sampling approach was applied in the survey. A total of 1,000 census enumeration areas (EAs) from the 2001 population census were selected from a database of 86,000 EAs and mapped in 2007 using aerial photography to create the 2007 HSRC master sample to use as a basis for sampling of households. The selection of EAs was stratified by province and locality type. In the formal urban areas, race was also used as a third stratification variable (based on the predominant race group in the selected EA at the time of the 2001 census). The allocation of EAs to different stratification categories was disproportionate, in other words, over-sampling or over-allocation of EAs

	occurred in areas that were dominated by Indian, Coloured or white race groups to ensure that the minimum required sample size in those smaller race groups was obtained. Based on the HSRC 2007 Master Sample, 500 EAs representative of the socio-demographic profile of South Africa were identified and a random sample of 20 visiting points (VPs) were randomly selected from each EA, yielding an overall sample of 10,000 VPs. EAs were sampled with probability proportional to the size of the EA using the 2001 census estimate of the number of VPs in the EA database as a measure of size (MOS). One of the tasks of the SANHANES-1 was to recruit and establish a cohort of 5,000 households to be followed up over the coming years. Data for this survey were collected in two separate but integrated components. These components included administering questionnaires to participants (conducting interviews) and performing a clinical examination (free-of-charge medical check-up by a doctor, selected measurements by a nurse/clinic assistant and collection of a blood sample for biomarker analysis) on each participant.
Availability of data descriptions	http://www.hsrc.ac.za/en/research-areas/Research_Areas_PHHSI/sanhanes-health-and-nutrition
Conditions of obtaining microdata	Not available
Contact for Information and Data Supply	www.hsrc.ac.za

Name	***South African National HIV, Behaviour and Health Survey***
Principal investigator	HSRC
Year(s)	2002, 2005, 2008 and 2012
Area(s) of interest	Health
Source(s) of data description provided here	http://www.hsrc.ac.za/en/research-areas/Research_Areas_HAST/HAST_National_HIV_Survey
Brief description	*Summary* The 2012 population-based survey of HIV prevalence is the fourth in the series of national HIV-prevalence surveys that have investigated HIV prevalence and behaviour. In 2002, a consortium consisting of the Human Sciences Research Council (HSRC), Medical Research Council (MRC), Centre for AIDS Development, Research and Evaluation (CADRE) and Agence Nationale de Recherche sur le Sida (ANRS) constituted the first research team to conduct a national population-based survey of HIV prevalence in South Africa. Since 2002, the HSRC and its partners, supported by different international and local donors, have conducted several national surveys that have contributed to the country's understanding of the HIV epidemic over time The main objectives of the survey were as follows: • To determine the prevalence and incidence of HIV infection in South Africa in relation to social and behavioural determinants. • To determine the proportion of males in South Africa who are circumcised. The secondary objectives were as follows: • To determine the proportion of people living with HIV and AIDS who are receiving antiretroviral treatment in South Africa. • To determine the extent to which mother-child pairs include HIV-negative mothers and HIV-positive infants. • To describe trends in HIV prevalence, HIV incidence and risk behaviour in South Africa over the period 2002 to 2012. *Methodology* The 2012 survey design was similar to that implemented in the previous surveys. A multi-stage, stratified cluster sampling design was implemented with everyone in the sampled household invited to participate. This approach enabled analyses linking HIV results obtained from co-habiting or married sexual partners and also mother-child pairs. Over 38 000 people were interviewed and almost 29 000 agreed to be tested for HIV. Persons of all ages living in South African households and hostels were eligible to participate. A 'household member' was defined as any person who slept in the household on the night preceding the survey (including visitors who spent the night before the survey in this household). Persons resident in educational institutions, old-age homes, hospitals, correctional facilities and uniformed-service barracks, as well as homeless persons, were excluded from the survey. *Sampling* A total of 1,000 census enumeration areas (EAs) from the 2001 population census were randomly selected using probability proportional to size and stratified by province, locality type and race in urban areas from a database of 86,000 EAs that were mapped

	in 2007 using aerial photography to develop the 2007–2011 HSRC master sample for selecting households. The sampled EAs formed primary sampling units (PSUs). Locality types were defined as urban formal, urban informal, rural formal (including commercial farms) and rural informal (tribal authority) areas. Oversampling of the coloured and Indian/Asian race groups was done to meet the required minimum sample size. Aerial photographs drawn from Google Earth were also employed to ensure that the most up-to-date information was available for the master sample In each sampled EA, a total of 15 visiting points (VPs) or households were used as secondary sampling units (SSUs). Within each household selected for the survey, all household members (including consenting and non-consenting household members) constituted the ultimate sampling unit (USU). A VP was defined as a stand with an address that might have one or more residential household in which a group of people live and eat together 'from the same pot'. If multiple households existed in a visiting point, a Kish grid was used to randomly select a responding household where all members of the selected household were eligible to participate. *Weighting* Sample weights were introduced at the EA, household and individual levels to correct potential bias due to unequal sampling probabilities, and also to adjust for non-response. The final sampling weight was thus equal to the final EA weight multiplied by the final VP sampling weight and adjusted for individual nonresponse. The final individual weights were benchmarked to the 2012 mid-year population estimates by age, race, sex and province. This process produced a final sample representative of the population in South Africa for sex, age, race, locality type and province. *Questionnaires* Four types of questionnaires were administered in the survey, namely: household questionnaires; questionnaires for parent/guardian of children aged 0–11 years; questionnaires for children aged 12–14 years; and, questionnaires for persons aged 15 years and older.
Availability of data descriptions	Not available
Conditions	Not available
Contact for Information and Data Supply	For information: datahelp@hsrc.ac.za For data: datahelp@hsrc.ac.za

Name	***South African National Innovation Survey***
Principal investigator	HSRC
Year(s)	2008
Area(s) of interest	Labour market; Economy
Source(s) of data description provided here	HSRC (http://curation.hsrc.ac.za/index.php?module=pagesetter&tid=125&tpl=projects)
Brief description	*Summary* The South African Innovation Survey follows the international OECD/Eurostat guidelines and methodology and is based on the core EU Community Innovation Survey (CIS) with modifications and a few particular questions for the South African environment. Following international methodology allows the results of the South African Innovation Survey to be usefully compared with the results from other countries. The National innovation Survey 2008 collected primary data from the business sector. Although some general organisational information was collected, the survey focussed on product and process innovation. Innovation Surveys are currently designed to measure the extent of innovative activity in the industry and service sectors of the economy and are based on the guidelines of the OECD/Eurostat Oslo Manual and the core EU Community Innovation Survey (CIS) but are usually adapted by countries to meet particular country needs or conditions. *Methodology* The survey was conducted using a postal questionnaire (with non-responders followed-up by telephone and/or email). *Sample design* The South African Innovation Survey 2008 was based on the guidelines of the Organisation for Economic Cooperation and Development's (OECD) Oslo Manual (OECD 2005) and more specifically the methodological recommendations for round five of the Community Innovation Survey (CIS 2006) for European Union (EU) countries as provided by Eurostat, the Statistical Office of the European Commission. Using these guidelines enabled the production of indicators that were both relevant for South Africa and internationally comparable. The survey design was also informed by the structure of the Business Register of Statistics South Africa (Stats SA), which was used to draw a suitable stratified random sample for the survey. The sample frame from which the original sample was drawn had 30 Standard Industrial Classification (SIC) codes and four size classes, which gave a total of 120 strata. Innovation Surveys are based on a random stratified sample of business enterprises from the national business register (or equivalent) and results are extrapolated to the original population. The South African Innovation Survey 2008 was based on a random stratified sample of 4 000 enterprises obtained from the Statistics South Africa business register. After cleaning the remaining entries in the database totalled 2 836 valid enterprises and after two postal rounds and telephonic and e-mail follow ups and reminders a final response of 757 completed questionnaires was obtained giving a response rate of 26.7%.

	Weighting A non-response survey was conducted, the results of which were subsequently used to adjust the weights of the strata for bias in the estimation of innovation rate that might arise from a low response rate. The results of the survey were extrapolated to the target business population of 22,849 enterprises by applying the weights of 108 realised sample strata based on SIC codes and four size classes (determined on the basis of annual turnover) used at Stats SA in 2007.
Availability of data descriptions	Documentation is available from the HSRC website http://curation.hsrc.ac.za/Dataset-344.phtml
Conditions of obtaining microdata	By accessing the data, you give assurance that The data and documentation will not be duplicated, redistributed or sold without prior approval from the rights holder. The data will be used for scientific research or educational purposes only. The data will only be used for the specified purpose. If it is used for another purpose the additional purpose will be registered. Redundant data files will be destroyed. The confidentiality of individuals/organisations in the data will be preserved at all times. No attempt will be made to obtain or derive information from the data to identify individuals/organisations. The HSRC will be acknowledged in all published and unpublished works based on the data according to the provided citation. The HSRC will be informed of any books, articles, conference papers, theses, dissertations, reports or other publications resulting from work based in whole or in part on the data and documentation. For archiving and bibliographic purposes an electronic copy of all reports and publications based on the requested data will be sent to the HSRC. To offer for deposit into the HSRC Data Collection any new data sets which have been derived from or which have been created by the combination of the data supplied with other data. The data team bears no responsibility for use of the data or for interpretations or inferences based upon such uses.
Contact for Information and Data Supply	For information: Tel: +27 (0)12 3022000 datahelp@hsrc.ac.za For data: http://curation.hsrc.ac.za/index.php?module=pagesetter&tid=125&tpl=projects datahelp@hsrc.ac.za

Name	***South African Reconciliation Barometer 2003-2011***
Principal investigator	Institute for Justice and Reconciliation
Year(s)	2003-2015 (annually)
Area(s) of interest	Attitudes (reconciliation specific)
Source(s) of data description provided here	http://www.ijr.org.za/ http://www.ijr.org.za/publications/pdfs/IJR%20SARB%203%202015%20WEB%20final.pdf
Brief description	*Summary* The South African Reconciliation Barometer (SARB) is an annual public-opinion survey conducted by the IJR. Since its launch in 2003, the SARB has provided a nationally representative measure of citizens' attitudes to national reconciliation, social cohesion, transformation, and democratic governance. The SARB is the only survey dedicated to critical measurement of reconciliation and the broader processes of social cohesion and is the largest longitudinal-data source of its kind globally. *Methodology* The SARB survey was conducted annually between 2003 and 2013 through face-to-face interviews and by using a structured questionnaire. In 2013 and 2014, the SARB survey instrument underwent extensive review in order to improve the survey questionnaire in both its conceptualisation and measurement. This process was concluded in 2015 and the new survey was fielded during August and September 2015. The survey employed a multistage cluster design whereby enumerator areas (EAs) were randomly selected, and, within each of these, households were randomly selected with a view to visiting such households. At each household, a systematic grid system was employed in order to select the specific respondent for an interview. The final sample of 2,219 respondents was then weighted so as to adequately represent the adult population of South Africa.
Availability of data descriptions	Unknown
Conditions of obtaining microdata	Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information, contact: *IJR:* Tel: 021-202 4071 Email: info@ijr.org.za

Name	**South African Social Attitudes Survey *(SASAS)***
Principal investigator	HSRC
Year(s)	Annually since 2003, ongoing
Area(s) of interest	Attitudes (with varying focus according to the survey round)
Source(s) of data description provided here	HSRC's SASAS website: http://www.hsrc.ac.za/en/projects/view/TAAMAA
Brief description	*Summary* The South African Social Attitudes Survey (SASAS) is a nationally representative, repeated cross-sectional survey that has been conducted annually by the Human Sciences Research Council (HSRC) since 2003. The survey series charts and explains the interaction between the country's changing institutions, its political and economic structures, and the attitudes, beliefs and behaviour patterns of its diverse populations. Designed as a time series, SASAS provides a unique, long-term account of the speed and direction of change in underlying public values and the social fabric of modern South Africa. SASAS thus represents a notable tool for monitoring evolving social, economic and political values among South Africans, but it also demonstrates promising utility as an anticipatory, or predictive, mechanism that can inform decision- and policy-making processes. *Methodology* The SASAS questionnaire contains a standard 'core' set of demographic, behavioural and attitudinal variables, which is repeated each round, with the aim of monitoring change and continuity in a variety of social, economic and political values over time. In addition to the core module, each round of interviewing accommodates rotating modules on specific themes, the aim being to provide detailed attitudinal evidence to inform policy and academic debate. In determining the thematic content of the survey, attempts are made to identify key perennial topics that would provide reliable and robust measures to shape our understanding of present-day South Africa and the processes of change within it. SASAS focuses on variations in culture and social structure within the country and aspires to be an instrument for identifying and interpreting long-term shifts in social circumstances and values, rather than simply monitoring short-term changes. *Sampling* Each round of SASAS has been designed to yield a representative sample of between 3500-7000 individuals aged 16 and older, regardless of nationality or citizenship, in households which are geographically spread across the country's nine provinces. The sample has been drawn from the HSRC's Master Sample - a sampling frame that consists of 1 000 Population Census enumeration areas (EAs). Each SASAS round of interviewing consists of a sub-sample of 500 EAs drawn from the master sample, stratified by province, geographical sub-type and majority population group. The sampling frame used for the more recent surveys is based on the 2011 census. Small area layers (SALs) were used as primary sampling units and the estimated number of dwelling units (taken as visiting points) in the SALs as secondary sampling units. In the first sampling stage the primary sampling units (SALs) were drawn with probability

	proportional to size, using the estimated number of dwelling units in an SAL as measure of size. The dwelling units as secondary sampling units were defined as "separate (non-vacant) residential stands, addresses, structures, flats, homesteads, etc." In the second sampling stage a predetermined number of individual dwelling units (or visiting points) were drawn with equal probability in each of the drawn dwelling units. Finally, in the third sampling stage a person was drawn with equal probability from all 16 year and older persons in the drawn dwelling units. Three explicit stratification variables were used, namely province, geographic type and majority population group. Within each stratum, the allocated number of primary sampling units (which could differ between different strata) was drawn using proportional to size probability sampling with the estimated number of dwelling units in the primary sampling units as measure of size. Selection of individuals: For each of the SASAS samples interviewers visited each visiting point drawn in the SALs (PSU) and listed all eligible persons for inclusion in the sample, that is all persons currently aged 16 years or older and resident at the selected visiting point. The interviewer then selected one respondent using a random selection procedure based on a Kish grid. *Weighting* The data were weighted to take account of the fact that not all units covered in the survey had the same probability of selection. The weighting reflected the relative selection probabilities of the individual at the three main stages of selection: visiting point (address), household and individual. In order to ensure representivity of smaller groups, i.e. Northern Cape residents or Indian/Asian people, weights needed to be applied. Person and household weights were benchmarked using the SAS CALMAR macro for province, population group, gender and 5 age groups (i.e. 16-24, 25-34, 35-49, 50-59 and 60 and older).
Availability of data descriptions	Details (including questionnaires) are available from HSRC's SASAS website: http://www.hsrc.ac.za/en/projects/view/TAAMAA
Conditions	Please note: Some data sets are only available to SASAS project team members. Please contact the SASAS team for further information, including conditions of access.
Contact for Information and Data Supply	For information: http://www.hsrc.ac.za/en/projects/view/TAAMAA For data: http://www.hsrc.ac.za/en/projects/view/TAAMAA

Name	**Survey of Activities of Young People *(SAYP)***
Principal investigator	Stats SA
Year(s)	1999 and 2010
Area(s) of interest	Education; Demographics, Labour market;
Source(s) of data description provided here	StatsSA: http://www.statssa.gov.za/publications/P0212/P02122010.pdf
Brief description	*Summary* Statistics South Africa was commissioned by the Department of Labour (DoL), to conduct the first Survey of Activities of Young People in 1999. Stats SA was responsible for data collection and processing, while the analysis and report writing was the responsibility of DoL. In the third quarter of 2010 Stats SA conducted the second Survey of Activities of Young People (SAYP) as a supplement to the Quarterly Labour Force Survey (QLFS). However, there should be no comparisons made between the 1999 SAYP and 2010 SAYP because of differences in methodologies followed in the two surveys. The main aim of the survey was to collect data on educational activities, economic activities, noneconomic activities, health and safety issues, and household tasks of individuals aged 7–17 years who live in South Africa. The specific objectives of SAYP are: • To understand the extent of children's involvement in economic activities; • To provide users with a statistical base regarding the number of working children; • To supply information for the formulation of an informed policy to combat child labour within the country; and • To monitor the CLAP (Child Labour Action Plan). *Methodology* SAYP 2010 is a household-based sample survey that collects data on the activities of children aged 7 to 17 years living in South Africa. This information is gathered from respondents who are members of households living in dwellings that have been selected to take part in the QLFS and have children aged 7–17 years. The survey covers market production activities, production for own final consumption, household chores as well as activities that children engaged in at school. The reference period for some activities is the week preceding the survey interview and for others it is the past twelve months. The report does not attempt to classify children according to whether they are in child labour or not, but rather identifies children who are involved in economic activities. *Sampling* The Survey of Activities of Young People (SAYP) comprised two stages. The first stage involved identifying households with children aged 7–17 years during the Quarterly Labour Force Survey (QLFS) data collection that took place in the third quarter of 2010. The second stage involved a follow-up interview with children in those households to establish what kind of activities they were involved in and several other aspects related to the activities they engaged in. *Weighting*

	During the Quarterly Labour Force Survey (QLFS) of quarter three 2010, children aged 7 to 17 years were screened and later interviewed for Survey of Activities of Young People (SAYP). The SAYP interviews were not conducted at the same time with QLFS. This resulted in the reduction of SAYP persons as compared to the ones identified during QLFS screening. This was due to persons refusing to participate in SAYP, persons not at home during SAYP interviews, demolished structures, vacant dwellings, etc. The SAYP weight adjustment accounts for those persons who qualified for SAYP, but refused to take part or were not available for interviews and those that were considered to be other non-response. The non-response adjustment is done through the creation of adjustment classes. The adjustment classes are created using Response Homogeneity Groups (RHGs), where respondents have the same characteristics with non-respondents in the group. The response rate (which is the ratio of responses to all eligible units in the sample) is calculated within each class. The inverse of the response rate (adjustment factor) is calculated within each class, and the result is multiplied by the QLFS 2010 person's weights of the responding units to get the adjusted SAYP person weights for responding units. Children identified as ineligible for SAYP were not considered when calculating weights adjustment. In short, the weights of responding children are inflated to account for eligible children that did not respond during SAYP data collection. The final SAYP weight assigned to each responding unit is computed as the product of the QLFS person weight and the non-response adjustment factor. The sum of QLFS person weight qualifying for SAYP (for both respondents and non-respondents, excluding the out-of-scope persons) must be equal to the sum of final SAYP person weight.
Availability of data descriptions	http://www.statssa.gov.za/publications/P0212/P02122010.pdf
Conditions of obtaining microdata	"The data and metadata set from the Survey of Activities of Young People, 2010 will be available on CDROM. A charge may be made according to the pricing policy, which can be accessed on the website."
Contact for Information and Data Supply	info@statssa.gov.za

Name	***Survey of Employers and Self-Employed***
Principal investigator	StatsSA
Year(s)	2001, 2005, 2009 and 2013
Area(s) of interest	Labour market, Economy
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys) Statistics South Africa (hhttp://interactive.statssa.gov.za:8282/webview/)
Brief description	*Summary* The Surveys of Employers and Self-Employed (SESE) are a set of surveys of non-VAT registered businesses in South Africa. The surveys have been undertaken by Statistics South Africa in September 2001, September 2005, September 2009 and September 2013. The primary aim of these surveys is to obtain estimates of the size of value added in the informal sector for the compilation of the national accounts. The surveys are undertaken to: collect data on micro- and small-businesses in South Africa and their operations; to collect income tax information on non-VAT paying businesses, and; to determine the contribution of these businesses to the economic growth of the country *Methodology* Currently, there is no sampling frame on which to base weights and raising factors for small unregistered businesses in South Africa. As a result, the research design used for SESE was a household based survey, consisting of two stages. The first stage involved using the Labour Force Surveys (LFS) (in 2001 and 2005) and the Quarterly Labour Force Survey (QLFS) (in 2009 and 2013) enumeration to identify individuals who were running unregistered businesses. The second stage involved follow-up interviews with the owners of these businesses by QLFS enumerators. For each SESE survey, the criterion for inclusion depends on whether or not the business is registered for Value Added Tax (VAT). Only persons who had businesses which were not registered for VAT were included. These businesses are generally excluded from the Business Frame which is used by Stats SA during surveys to assess the formal economy. In 2001, SESE was conducted in March and the SESE interview was undertaken immediately after the LFS interview while the enumerator was still at the dwelling unit. In 2005 the data collection for the SESE occurred over a two-week period during the month of September after the LFS interviews had been concluded. In both 2009 and 2013, data collection for the QLFS occurred during the middle two weeks of the month throughout the quarter, while SESE data collection was undertaken in the last week of the month, also throughout the quarter. Because of these changes in the methodology, comparisons should be interpreted with caution.

	Weighting The sampling weights for the data collected from the sampled dwelling units are constructed in such a manner that the responses could be properly expanded to represent the entire civilian population of South Africa. The weights are the result of calculations involving several factors, including original selection probabilities, adjustment for non-response, and benchmarking to known population estimates from the Demography Division of Stats SA (SESE main report) The non-respondent adjustment is done through the creation of adjustment classes. The adjustment classes are created using Response Homogeneity Groups (RHGs), where respondents have the same characteristics with non-respondents in the group. The response rate (which is the ratio of responses to all eligible units in the sample) is calculated within each class. The inverse of the response rate (adjustment factor) is calculated within each class, and the result is multiplied by the person weight of the QLFS for the responding units to get the adjusted SESE person weight for non-responding units. In essence, the weights of responding persons are inflated to account for those that did not respond during SESE. The final SESE weight assigned to each responding unit is computed as the product of the QLFS person weight and the non-response adjustment factor. The sum of QLFS person weight qualifying for SESE (for both respondents and non-respondents, excluding the out-of-scope persons) must be equal to the sum of final SESE person weight. The final SESE business weights were calculated as the ratio of final adjusted SESE person weight to the number of businesses a person is running.
Availability of data descriptions	The following documentation is available from the NESSTAR (hhttp://interactive.statssa.gov.za:8282/webview/): Questionnaire Metadata Concepts and Definitions Documentation is also available on the DataFirst website (http://www.datafirst.uct.ac.za/surveys).
Conditions of obtaining microdata	Stats SA: Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. Data First: Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: Tel: (012) 310 8600 Fax: (012) 310 8944

	info@statssa.gov.za http://www.statssa.gov.za/ For data: http://interactive.statssa.gov.za:8282/webview/ https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	***Time Use Survey***
Principal investigator	Stats SA
Year(s)	2000, 2010
Area(s) of interest	Labour market; health; education; transport
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys) Statistics South Africa (hhttp://interactive.statssa.gov.za:8282/webview/)
Brief description	*Summary* A national sample survey, using a diary methodology, to investigate how people aged 10 years and above spend their time. Such information contributes to greater understanding of policymakers on the economic and social well-being of different societal groups; to provide new information on the division of both paid and unpaid labour between women and men, and greater insight into less well understood productive activities such as subsistence work, casual work and work in the informal sector and to improve concepts, methodology and measurement of all types of work and work-related activity. *Methodology* In 2000, the fieldwork for the study was conducted in three rounds - February, June, and October - so as to catch possible seasonal variations in time use. The study used a 24-hour diary, divided into half-hour slots, as the core instrument to record activities. In each slot, a maximum of three activities could be recorded. The diary was administered face-to-face to the respondent by means of an interview. In addition to the diary, the questionnaire contained many of the standard questions of Stats SA household surveys. Thus one member per household provided basic information about the household as a whole, and, before administration of the diary, the respondent was asked for basic demographic information about themselves, such as age, sex, children and work situation. The questionnaire for the time use survey comprised five sections. Section one covered details of the household. Section two covered demographic details of the first person selected as a respondent in that household. Section three consisted of a diary in which to record the activities performed by the first person selected during the 24 hours between 4 am on the day preceding the interview and 4 am on the day of the interview. Sections four and five were for the second selected person in the household but were otherwise identical to sections two and three respectively. In 2010, data collection for the time use survey was conducted in the fourth quarter (October to December) of 2010 by survey officers employed to do data collection for the Quarterly Labour Force Survey (QLFS). In the middle two weeks of each month these survey officers collected data for the QLFS. They then utilised the last week of the month throughout the quarter to administer the Time Use Survey questionnaire. Face-to-face interviews, including the administration of the diary, were conducted in the language preferred by the respondent. The details of all household members were collected, and the number of persons eligible (those 10 years or older) for selection for the TUS was established and recorded on the questionnaire. The instruction was to select two eligible persons in each household. If more than two persons were eligible for TUS, the survey officer selected two household members for inclusion in the TUS using a selection grid; in cases where there were only two eligible persons in the household, they were both interviewed. If there was only one eligible person, then that person was interviewed. The

	survey officer then recorded the activities undertaken by the respondents in 30-minute time slots on the 24-hour diary retrospectively. *Sample design* The sample for the time use survey was chosen so as to be representative of the country's population. Each round included households from all nine provinces and from four different strata, or types of settlement area. The strata were formal urban settlements, informal urban settlements, commercial farming areas, and other rural areas. The TUS sample covered the non-institutional population except for workers' hostels. However, persons living in private dwelling units within institutions were eligible for enumeration. For example, within a school hostel/dormitory, the principal's house and teachers' accommodation could be enumerated because they are private dwellings. Students living in a dormitory of the school hostel would not be enumerated. *Weighting* The raw data were weighted so as to adjust the responses collected to be representative of the underlying sample frame. Because the sample frame itself was designed so as to reflect the population of South Africa aged 10 years and above, the results reported should reflect the proportions in the total population of this age in terms of sex, population group, age group and settlement type.
Availability of data descriptions	The questionnaire and accompanying metadata is available on the DataFirst website (http://www.datafirst.uct.ac.za/surveys.html).
Conditions for obtaining microdata	Stats SA: Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. Data First: Online application for access to a public use dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from
Contact for Information and Data Supply	For information: info@statssa.gov.za http://www.statssa.gov.za/ For data: http://interactive.statssa.gov.za:8282/webview/ https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	***TIMSS: Trends in International Mathematics and Science Study***
Principal investigator	HSRC
Year(s)	1995, 1999, 2002, 2011 and 2015
Area(s) of interest	Education
Source(s) of data description provided here	South African TIMSS website: http://www.timss-sa.org.za/
Brief description	*Summary* TIMSS was first administered in South Africa in 1995 and subsequently in 1999. In 2002 it was administered both to Grade 8 and 9 learners, and in 2011 to Grade 9 learners. Together, the assessments provide data to analyse trends in South African education for over a decade. For TIMSS 2011 in South Africa, the Human Sciences Research Council (HSRC) conducted the study in 285 schools among 11,969 learners. *Methodology* During TIMSS administration a total of four questionnaires are administered in addition to the achievement assessment instruments: • The Learner Background Questionnaire which is completed by the learner who completed the assessment and asks about aspects of the learners' home and school lives: their home environment, school climate for learning and their perceptions and attitudes towards mathematics and science. • The Teacher Questionnaire is administered to the mathematics and science teachers of the learners who write the assessment tests. The questionnaire was designed to gather information on teacher characteristics, as well as classroom contexts for teaching and learning mathematics and science. • The School Questionnaire is administered to the principal in all sampled schools. It asks about school characteristics like instructional time, resources and technology, as well as parental involvement. • The Curriculum Questionnaire is completed by the National Research Coordinator who is required to complete information pertaining to the curriculum which is followed by South African public schools. The most recent round of TIMSS was administered in August 2015.TIMSS 2015 is the sixth cycle of the IEA Trends in International Mathematics and Science Study (TIMSS). The study was initiated in February 2013 at the first national research coordinators meeting. Framework and instrument development work was carried out in 2013. The field test was conducted in March–April 2014. The data collection for the main survey was carried out in August 2015. The international reports will be released in December 2016, followed by the international database and user guide in February 2017.
Availability of data descriptions	http://www.timss-sa.org.za/
Conditions	Not available
Contact for Information and Data Supply	http://www.timss-sa.org.za/

Name	***Volunteer Activities Survey***
Principal investigator	StatsSA
Year(s)	2010
Area(s) of interest	Volunteering
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys) Statistics South Africa (hhttp://interactive.statssa.gov.za:8282/webview/)
Brief description	*Summary* This is a household-based sample survey that collects data on the volunteer activities of individuals aged 15 years and older who live in South Africa. This information is gathered from respondents who are members of households living in dwellings that have been selected to take part in the Quarterly Labour Force Survey (QLFS). VAS covers activities willingly performed for little or no payment to provide assistance or promote a cause in the four weeks preceding the survey interview. These activities can be performed either through an organisation or directly for someone outside one's own household. They take many forms, from picking up groceries for a disabled neighbour to participating in community policing. *Methodology* The VAS questionnaire was administered in each selected QLFS household to individuals aged 15 years and older. The VAS questionnaire collected data on demographic characteristics, type of volunteer activity, reasons for volunteer activity, time spent on volunteer activity, and monetary value of volunteer activity. Please see the relevant appendix relating to the QLFS for methodological details of that survey, including sampling and weighting.
Availability of data descriptions	The following documentation is available from the NESSTAR website (http://interactive.statssa.gov.za:8282/webview/): Questionnaire Metadata Concepts and Definitions Documentation is also available on the DataFirst website (http://www.datafirst.uct.ac.za/surveys).
Conditions of obtaining microdata	Stats SA: Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or offered for sale in any form whatsoever without prior permission from Stats SA. Data First: Online Application for Access to a Public Use Dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.

Contact for Information and Data Supply	For information: Tel: (012) 310 8600 Fax: (012) 310 8944 info@statssa.gov.za http://www.statssa.gov.za/ For data: http://interactive.statssa.gov.za:8282/webview/ https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central

Name	***Women and Land Survey***
Principal investigator	Community Agency for Social Enquiry
Year(s)	2009-2010
Area(s) of interest	Land rights
Source(s) of data description provided here	DataFirst (http://www.datafirst.uct.ac.za/surveys)
Brief description	*Summary* The overarching aim of the research was to investigate the nature of women's land rights in three rural ex-homeland areas of South Africa. In particular, the survey aimed to explore how women access land (including different types of land such as residential and fields), their actual use of the different types of land, their decision making capacity in relation to the different categories of land, and the extent of their security or vulnerability to eviction. The survey also aimed to explore the impact of marital status on the nature and content of women's land rights. *Methodology* This survey was conducted using face-to-face interviews.
Availability of data descriptions	Documentation is available on the DataFirst website (http://www.datafirst.uct.ac.za/surveys).
Conditions of obtaining microdata	Online Application for Access to a Public Use Dataset. One must provide a short description of the research project (project question, objectives, methods, expected outputs, partners) and agree to comply with the stated terms and conditions and give assurance that the use of statistical data obtained from DataFirst will conform to widely-accepted standards of practice and legal restrictions that are intended to protect the confidentiality of respondents.
Contact for Information and Data Supply	For information: support@data1st.org +2721 650 5708 For data: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central/about

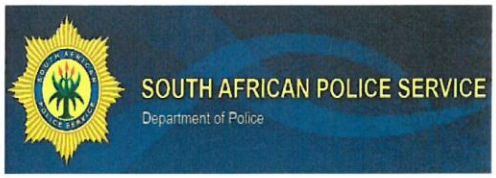
Appendix 2: Data Quality Framework (Daas et al, 2012)

Table 1. Quality dimensions and indicators for administrative input data used for statistics

Dimension Indicators	Description
1. Technical checks	*Technical usability of the file and data in the file*
1.1 Readability	Accessibility of the file and data in the file
1.2 File declaration	Compliance of the data in the file to the metadata
1.3 Convertability	Conversion of the file to the NSI-standard format
2. Accuracy	*The extent to which data are correct, reliable and certified*
Objects	
2.1 Authenticity	Legitimacy of objects
2.2 Inconsistent objects	Extent of erroneous objects in source
2.3 Dubious objects	Presence of untrustworthy objects
Variables	
2.4 Measurement error	Deviation of actual value from ideal error-free value, occurring during reporting, registration, or processing of data
2.5 Inconsistent values	Extent of inconsistent values for combinations of variables
2.6 Dubious values	Presence of implausible values or combinations of values
3. Completeness	*Degree to which a data source includes data describing the corresponding set of real-world objects and variables*
Objects	
3.1 Undercoverage	Absence of target objects (missing objects) in the source
3.2 Overcoverage	Presence of non-target objects in the source
3.3 Selectivity	Statistical coverage and representativity of objects
3.4 Redundancy	Presence of multiple registrations of objects
Variables	
3.5 Missing values	Absence of values for (key) variables
3.6 Imputed values	Presence of values resulting from imputation actions by DSH[a]
4. Time-related dimension	*Indicators that are time and/or stability related*
4.1 Timeliness	Time lag between the end of the reference period in the source and the moment of receipt
4.2 Punctuality	Time lag between the settled date and actual delivery date
4.3 Overall time lag	Time lag between the end of the reference period in the source and the moment NSI concluded the data can be used
4.4 Delay	Time lag between an actual change in the real-world and its registration in the source
Objects	
4.5 Dynamics	Changes in the population of objects (births/deaths) over time
Variables	
4.6 Stability	Changes of variables or values over time
5. Integrability	*Extent to which the data source is capable of undergoing integration or of being integrated in the statistical system*
Objects	
5.1 Comparability of objects	Similarity of objects in source -at the proper level of detail- with objects used by NSI
5.2 Alignment of objects	Linking-ability (align-ability) of objects with those of NSI
Variables	
5.3 Linking variable	Usefulness of linking variables (keys) in source
5.4 Comparability of variables	Proximity (closeness) of variable values in different sources

[a] DSH, Data Source Holder

Appendix 3: StatsSA and SAPS collaboration on crime data quality



Statistician General's statement on the 2015/16 Crime Statistics Processes based on the Clearance Committee Assessment Report

Page | 1

Statistician General's statement

Statistics Act (no.6 of 1999) requires the Statistician General (SG) to coordinate statistical production in the country, beyond the confines of Statistics South Africa (Stats SA). In this respect, Stats SA has been working with the South African Police Service (SAPS) on improving the quality of crime statistics since 2011. The collaboration between the SAPS and Stats SA has culminated in the two organisations entering into a Memorandum of Understanding (MoU) in April 2015.

As a consequence of the MoU, the Statistician General (SG) constituted a Clearance Committee to evaluate and authenticate the quality of crime statistics in line with South African Statistical Quality Assessment Framework (SASQAF), using SASQAF Lite. The main focus of the assessment was on the processes of compiling the crime statistics using selected indicators within the following SASQAF dimensions: Methodological Soundness; Accuracy; Comparability and Coherence; Integrity and Timeliness. During 2015/2016 an assessment of the SASQAF measures not previously done namely Relevance, Accessibility and Interpretability were performed.

The assessment outcome indicated compliance of SAPS processes with most aspects of the selected SASQAF quality dimensions. The challenges identified through the assessment process, were such that the SAPS could effect the necessary improvements without any major difficulty to ensure compliance by the next annual publication. Once all recommendations have been implemented, SAPS may request a full independent assessment to establish its level of readiness to produce official statistics.

Finally, having given due consideration to the intended introduction of a Quarterly Crime Statistics Series, I fully support this initiative. However this should be preceded by an exercise that focuses on risk analysis and human resource requirements. Such an exercise will ensure that the production system is both stable and sustainable before actual quarterly publication can occur. In order to ensure full benefit from the criminal justice value chain, the Statistician General will coordinate a process of aligning statistics production throughout the criminal justice system. This will require amongst others the implementation of standardized classifications, standards and policies that covers the whole statistical value chain.