Kerr, Andrew; Wittenberg, Martin

Working Paper

Earnings and employment microdata in South Africa

WIDER Working Paper, No. 2019/47

Provided in Cooperation with:
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: Kerr, Andrew; Wittenberg, Martin (2019) : Earnings and employment microdata in South Africa, WIDER Working Paper, No. 2019/47, ISBN 978-92-9256-681-4, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki,
https://doi.org/10.35188/UNU-WIDER/2019/681-4

This Version is available at:
https://hdl.handle.net/10419/211277

UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2019/47

Earnings and employment microdata in South Africa

Andrew Kerr and Martin Wittenberg*

May 2019

United Nations University World Institute for Development Economics Research
wider.unu.edu

Abstract: Traditionally, analysts of the South African labour market have used household survey data to describe earnings and employment in the post-Apartheid period. More recently, administrative data from the South African Revenue Service has been made available, which allows for comparisons and an assessment of each source and its strengths and weaknesses. There are a number of sources of data, including household surveys, firm surveys, and administrative data, and it can be hard to keep up with all of them. In this paper we thus provide a summary of the main sources of data on earnings and employment and their strengths and weaknesses, to aid researchers and policymakers who wish to make use of these data in their own analysis.

Keywords: administrative data, data sources, earnings, employment, surveys, South Africa,
JEL classification: J21, J31

Acknowledgements: This paper has been supported as part of the UNU-WIDER 'Southern Africa—Towards Inclusive Economic Development' (SA-TIED) research programme. The programme also supported a separate paper on the South African Revenue Service IRP5 tax admin data, which Section 4.1 below relies on. We would also like to acknowledge the support of the University of Cape Town Vice Chancellor's strategic fund (2011–12), the International Labour Organization (2013), and the Research Project on Employment, Income Distribution and Inclusive Growth, a programme supported by the National Treasury (2013–17). We thank Bruce McDougall for sharing his work on National Income Dynamics Study and Quarterly Labour Force Survey earnings comparisons, which was the topic of his UCT master's thesis, supervised by Martin Wittenberg.

* Both authors: DataFirst, University of Cape Town (UCT), Cape Town, South Africa; corresponding author: andrew.kerr@uct.ac.za.

This study has been prepared as part of the project 'Southern Africa—Towards Inclusive Economic Development (SA-TIED)'.

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9256-681-4

Typescript prepared by Luke Finley.

The United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

Traditionally, analysts of the South African labour market have used household survey data to describe earnings and employment in the post-Apartheid period. More recently, administrative data from the South African Revenue Service (SARS) has been made available, which allows for comparisons and an assessment of each source and its strengths and weaknesses. There are a number of sources of data, including household surveys, firm surveys, and administrative data, and it can be hard to keep up with all of them. In this paper we thus provide a summary of the main sources of data on earnings and employment and their strengths and weaknesses, to aid researchers and policymakers who wish to make use of these data in their own analysis.

The household survey data sources to be described and analysed include the Quarterly Labour Force Survey (QLFS) and the older Labour Force Surveys (LFS) and October Household Surveys (OHS), all conducted by Statistics South Africa (Stats SA), starting in 1994 and ending with the most recent QLFS. They also include the Post-Apartheid Labour Market Series (PALMS; Kerr et al. 2019), available through DataFirst, which is a harmonized version of all the Stats SA QLFSs, LFSs, and OHSs, as well as the 1993 Project for Statistics on Living Standards and Developments (PSLSD), conducted by the Southern Africa Labour and Development Research Unit (SALDRU).

The other Stats SA household survey data source which we describe is the General Household Survey. Although it lacks the detailed questions about each individual's employment that are found in the QLFS, it does have a roughly consistent question on earnings and employment going back to 2002. Given some of the issues with the QLFS, discussed below, this alternative source of employment and earnings data also needs to be assessed. Finally, the National Income Dynamics Study (NIDS), undertaken by SALDRU, is a source of earnings and employment data not produced by Stats SA, which is a useful check on the other surveys, and we discuss it below.

The data we have mentioned thus far are all publicly available. We also describe two sources of data either that are not in the public domain or to which there is some limited access. The first of these is the SARS IRP5 data set—which contains the tax records of all employees of tax-registered companies who earned more than R2,000 in each tax year between 2011 and 2016 (newer data may be made available). This is currently available to researchers approved by the National Treasury. The second data source is the Stats SA Quarterly Employment Statistics (QES) survey. This has not been made available to researchers, although Kerr et al. (2014) used the data to explore job creation and destruction, and in the process described the employment data (but did not analyse earnings). We also briefly mention a third source of microdata which has not been used in any research that we are aware of—the firm-level data from the Unemployment Insurance Fund (UIF) submissions to the Department of Labour. Such data has often been used in other countries in labour market analysis, and so we note that it may be a useful source.

2 Household Survey Data

2.1 OHS, LFS, and QLFS

The household surveys from Stats SA are the starting point for any analysis of earnings and employment in South Africa. The October Household Surveys began with OHS 1993 and continued until OHS 1999. The Labour Force Surveys replaced the OHS labour market data collection and were run biannually in March and September until September 2007. The Quarterly

Labour Force Surveys were then introduced in February 2008, and they continue to be undertaken every quarter. Any analysis of these three sets of household survey data we undertake below is carried out on PALMS version 3.3 (Kerr et al. 2019). PALMS is a compilation of all the OHSs, LFSs, and QLFSs, as well as SALDRU's PSLSD conducted in 1993, and several versions have been released by DataFirst since 2013. The most recent version (PALMS v3.3) contains QLFS Q2 2018, but earnings data only up until 2017.[1]

We begin by briefly reviewing known issues relating to employment and earnings in these three sets of surveys, then provide some descriptive analysis showing these issues before discussing other household surveys that can be used to investigate earnings and employment in South Africa.

2.1.1 October Household Surveys

The OHSs were the first attempt by Stats SA (which changed its name in 1997 from the Central Statistical Service) to collect nationally representative household survey data and to release them publicly for analysis.

The first OHS was conducted in 1993 but was not nationally representative, since it did not cover some homeland areas (Wittenberg 2008). This survey has thus been mostly overlooked by researchers, and not much is known about the strengths and weaknesses of the data other than that they exclude homelands, which were and are areas with low earnings and low employment. This renders it not very useful for any analysis of earnings and employment.

OHS 1994 was the first nationally representative household survey of the post-Apartheid period. OHSs 1994, 1995, and 1996 have the weakness that the sample frame of enumeration areas (EAs) for the first stage of the two-stage cluster sample came from the 1991 census. The 1991 census did not cover homelands, but these areas were included in the sample frame (Central Statistical Service 1998). Thus it is possible that despite covering all of South Africa, the sample frame was not as reliable as later ones, which used censuses that covered the entire country.

In OHS 1994, analysts have noted other issues that also relate to how the sample frame was constructed and whether that meant it was not truly nationally representative. These include too many whites relative to their share in the population, too few domestic workers, and too much employment (Branson and Wittenberg 2007).

OHS 1995 has been the basis of much work that has sought to describe changes over time in a number of employment-related issues (Branson and Wittenberg 2007). Wittenberg (2014b) has noted several issues in OHS 1995—employment was too high, unemployment too low, and the earnings gap between men and women too low. Branson and Wittenberg (2007) argue for the use of all possible sources of data, rather than just one at the start and one at the end of any period under investigation.

Stats SA faced budget constraints in conducting OHS 1996, so the sample size was smaller. In addition, in both 1996 and 1997 political violence meant that residents in hostels in KwaZulu-Natal and Gauteng were not enumerated (Kerr and Wittenberg 2015). Since hostels have a large share of individuals in mining employment, mining employment was estimated to be much lower than it actually was in those two years.

[1] This is being released in June 2019.

In OHS 1999, a master sample of EAs was used based on the 1996 population census (Statistics South Africa 2000). This was thus the first time that a sample of EAs was drawn using one nationally representative source of EA data—the 1996 population census. The master sample was so-named because the same sample of EAs was to be used to sample households in a number of surveys for several years. Thus OHS 1999, all the LFSs until 2004, the 2001 Income and Expenditure Survey, and the General Household Survey (GHS) 2002–04 were conducted by sampling households from the same set of EAs.

Casale et al. (2004) note that in OHSs 1997 and 1999 the questionnaires included as examples of employment those involved in subsistence agriculture, but that earlier surveys and OHS 1998 did not. They also document that despite the prompt, almost no individuals were recorded as employed in subsistence agriculture in OHS 1997. We discuss employment in subsistence agriculture in further detail below.

In OHS 1999, Stats SA also changed the method of sampling to what are called multiple household dwelling points (Kerr and Wittenberg 2015). These are places where the listing of the enumerator area/cluster suggested that there was only one household, but there was actually more than one. The common example is a backyard shack which was not noticed by the enumerator when the listing was undertaken. In OHS 1998 and earlier, only one household was enumerated, reducing the probability of selection of particularly small households. Kerr and Wittenberg (2015) show that this meant that small households were under-represented in the weighted data and suggest that one outcome of this was an undercount of employment in 1998 and earlier, since those in the small households that were missed had better employment outcomes than those in larger households.

2.1.2 Labour Force Surveys

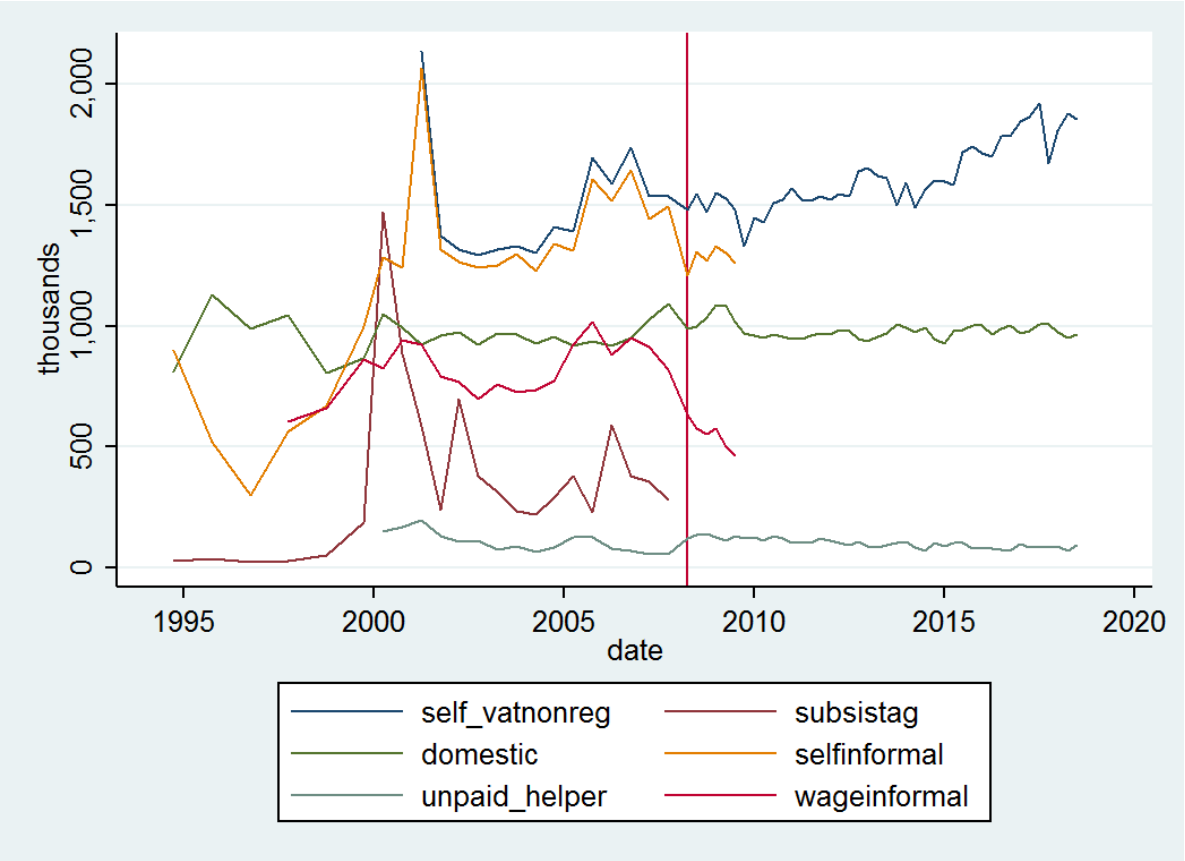
The LFS collected detailed information on employment and earnings for all employed individuals. It continued to use the master sample that was first used in OHS 1999 to sample households. But the LFS also incorporated a rotating panel design, meaning that a panel of dwellings was created (Statistics South Africa 2001). Every time a new LFS was undertaken, 80 per cent of the dwelling units in the sample were reinterviewed, while 20 per cent were rotated out and replaced with a new sample of dwelling units from the same Primary Sampling Unit (PSU). This rotation was supposed to begin in the third round of the LFS (2001: March). However, the LFS 2000: September had the same sample as the Income and Expenditure Survey of 2000, which led to larger non-response rates, and thus the first wave of the panel was 2001: September. It ran until LFS 2004: March. LFS 2004: September was then the first survey to use a new master sample that was based on the 2001 census list of EAs. This master sample was used until 2007.

The first three LFSs (February 2000, September 2000, and March 2001) all had substantial problems that meant total employment was overestimated in these three waves. This point is important to take note of for any longer-run analysis of employment trends. The large increases actually began in the last OHS in 1999. This was partly because OHS 1999 improved small-household coverage and thus found more employment, since small households have higher proportions of employed individuals (Kerr and Wittenberg 2015). But in the first three LFSs, measured employment is simply too high. We make this conclusion based on Figure 1, which shows that the first two LFSs measured substantial amounts of subsistence agriculture that was not replicated in any subsequent survey.

The third LFS (March 2001) was linked to the first Survey of Employers and Self-Employed (SESE). As described by Kerr and Wittenberg (2015) and in more detail by Kerr (2015a), the enumerators for the March 2001 LFS were required to reinterview the owners of any non-VAT-

registered businesses identified in this LFS, and enumerators were paid for each of these surveys. This led to very strong financial incentives for enumerators to 'find' these types of owners and interview them. The result is the massive spike in informal (non-VAT-registered) self-employment seen in Figure 1. Any research that analyses changes in employment and begins or ends with these surveys will thus find either massive increases or massive decreases in employment. This is another reason for researchers to include as much data as possible in their analyses, so that results are not an artefact of the particular surveys used (Branson and Wittenberg 2007).

Figure 1: Informal employment since 1994



Source: Authors' construction based on calculations from PALMS.

2.1.3 Quarterly Labour Force Surveys

The QLFS began in February 2008 and has run every quarter since then. Employment estimates are available for each quarter. Earnings data were not collected for the first six quarters but were then reinstated in quarter 3 of 2009, although these earnings data have only been made available since 2010. Earnings data are not released every quarter: rather, once a year a new publication and data release occurs. This is called 'Labour Market Dynamics', but is really a way of releasing the earnings data that are collected in the QLFS. The main reason for this convoluted data release seems to be that Stats SA does not have the capacity to prepare earnings data for release every quarter, partly because the data undergo some substantial imputation, which we discuss next.

Earnings in the QLFS

The QLFS earnings data that are released in ' Labour Market Dynamics' have earnings imputed for some individuals. As described in Kerr and Wittenberg (2017), up to and including Q2 2012, Stats SA imputed earnings both for those that gave bracket responses and those that refused to give any answer. There is no way to distinguish the actual rand value responses from the brackets

or complete refusals, and no way to distinguish between brackets and complete refusals. The imputation process was changed by Stats SA in 2012 Q3. Complete refusals are no longer imputed, but bracket responses are, and there is again no way of telling bracket responses from the actual rand amount responses.

These imputations are problematic for two important reasons. The first is that descriptions of earnings over time are going to be using three very different types of earnings data—unimputed earnings in the OHS and LFS, completely imputed earnings in the QLFSs from 2010 to 2012 Q2, and then partially imputed earnings from Q3 2012 onwards. This makes it possible that measured changes over time are the result of imputation effects and changes rather than actual real changes. It is hard to know how reliable the imputations are without the ability to distinguish between actual and imputed earnings, which is why researchers have been asking for imputation flags in the QLFS since at least 2012.

Kerr and Wittenberg (2017) used unimputed data from the 2011 QLFS, obtained from Stats SA, to check on the imputations, and found that there were substantial differences in some regression results when using the imputed public data and unimputed data, particularly for variables which are correlated with earnings but that are likely not included in the list of variables used in the imputation by Stats SA.

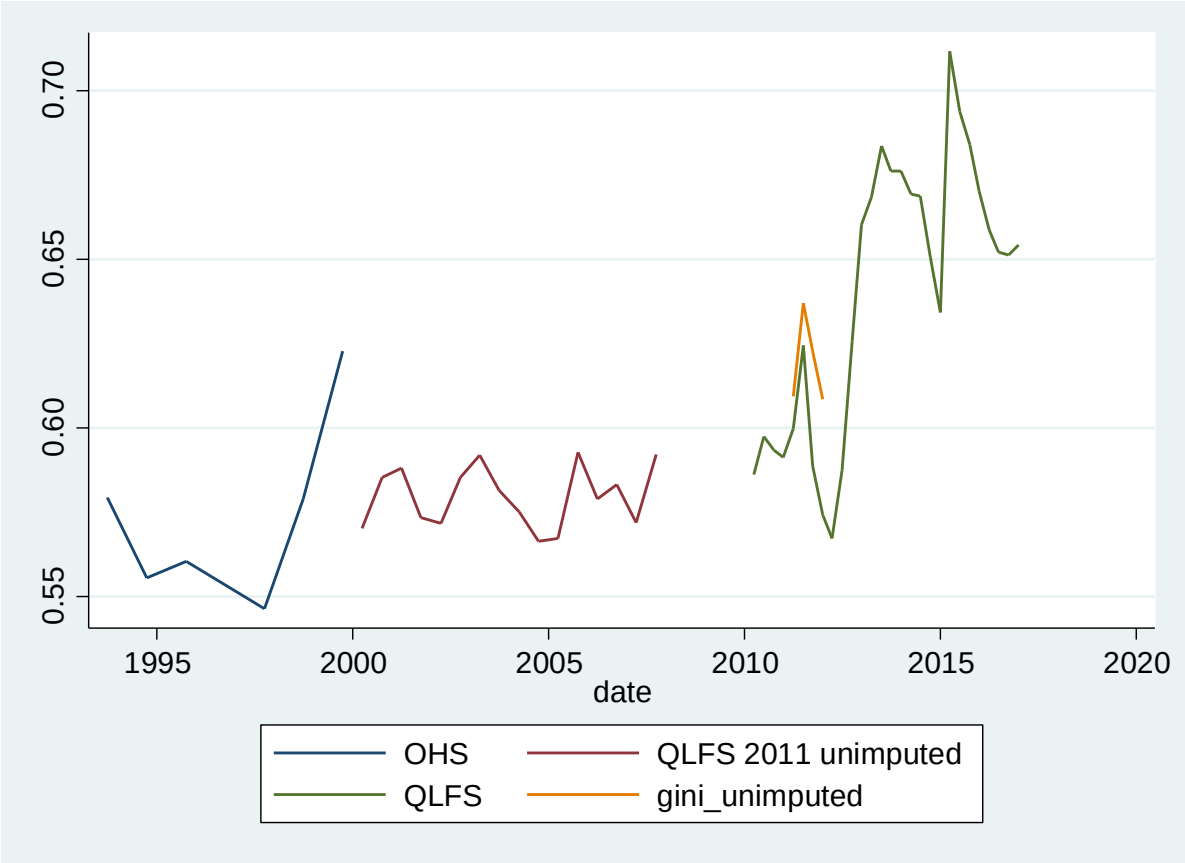
Wittenberg (2016) and Finn and Ranchhod (2017) showed trends in inequality in earnings from the QLFS, as measured by the Gini coefficient, that seem implausible, although neither attributed this to imputation. Figure 2 shows the Gini coefficient of earnings using the imputed QLFS earnings data, as well as for the unimputed QLFS 2011 earnings data to which DataFirst was given access. The Gini estimated from the imputed data fluctuates wildly over the last parts of the QLFS, rising 14 points between 2013 and 2015. This is highly unlikely, and the prime suspect is the imputation undertaken by Stats SA. This is partially confirmed by the unimputed earnings data from the 2011 QLFSs, where there is no large decrease in the last part of 2011 when using the unimputed data. It should be noted that the changes over time in various earnings percentiles shown by Wittenberg (2016) do not show such dramatic fluctuations as the Gini.

The second important issue with the QLFS earnings imputations is that the statistical uncertainty of the earnings numbers obtained from them will be biased downwards, because analysts are treating the data as if they were from actual responses, whereas in reality there is uncertainty about the true values, particularly for those with imputed earnings who refused to answer at all, which occurred from 2010 to 2012 Q2. This is another reason why Stats SA should release imputation flags—so analysts can compute the true estimates of statistical uncertainty for estimates of earnings.

Employment in the QLFS

There are also several issues with employment and some of its subcomponents that are measured in the QLFS. These include the exclusion of subsistence agriculture from the definition of employment from the first QLF onwards, the changing of the definition of the informal sector in QLFS 2009 Q3, and estimation statistical uncertainty in employment totals in the QLFS release documents that are likely to be incorrect, underestimating this uncertainty and also the uncertainty of employment changes over time in the QLFS.

Figure 2: Gini coefficients in earnings in the PSLSD, OHS, LFS, and QLFS, 1994–2017



Source: Authors' construction based on calculations from PALMS and unimputed QLFS 2011 data.

The first important change from the LFS was that subsistence agriculture was no longer counted as employment. The extremely large number of individuals working in subsistence agriculture in the two 2000 LFSs was not replicated in later LFSs, but this still resulted in estimates of between 250,000 and 750,000 individuals employed in subsistence agriculture. Any analysis of changes in agricultural employment over periods that include both the LFSs and the QLFSs will thus overestimate any decline in agricultural employment. A recent World Bank report states that agricultural employment halved between 2005 and 2010 (World Bank 2018: 79), but most of this decline is due to the disappearance of the estimated 300,000 individuals working in subsistence agriculture in 2005. In PALMS (Kerr et al. 2019), a data set that includes all the OHS, LFS, and QLFS microdata, we have included a new employment variable that excludes the self-employed in agriculture, which we argue is a better variable with which to measure changes in employment, given the issue with subsistence agriculture discussed here.

Beginning in OHS 1997, Stats SA asked both employees and the self-employed whether they thought the firm they worked for/owned was in the informal sector, sometimes with prompts from enumerators to give examples of what constitutes the informal sector. Figure 1 shows the different components of informal sector employment over the period 1994–2017. 'Selfinformal' is the estimate from a direct question asking whether the business was formal or informal, while 'self-vatnonreg' is from a question about whether an individual's business was registered for VAT. These track each other very closely from 2001 to 2007 and are somewhat different beginning in the QLFS. Direct questions about whether the individual worked in the formal or informal sector were asked until 2009 Q2. Besides the massive spikes in agriculture and informal self-employment,

the one strange trend is the halving in informal wage employment between 2007 and 2009—a trend that pre-dates the Great Recession.

The direct questions about informality disappeared in the QLFS 2009 Q3. In the absence of a direct question, Stats SA created its own indicator of informal sector employment that was included in the public releases of the microdata. The definition of the informal sector used to create the informal sector indicator is given in the Stats SA QLFS releases but is somewhat ambiguous. Employees are defined as being in the informal sector if they 'are not registered for income tax and … work in establishments that employ less than five persons' (Statistics South Africa 2008: 16). Not being registered for income tax is defined on the following page as 'Income tax [being] deducted by employer' (Statistics South Africa 2008: 17). This definition suggests that anyone under the tax threshold (R75,750 in the 2018 tax year, which is way above median annual earnings of around R42,000) who did not have income tax deducted should be considered to be in the informal sector—or at least it may lead to some ambiguity about this. Other employed individuals (employers, own-account workers, and persons helping unpaid in their household business) are considered to be in the informal sector if they 'are not registered for either income tax or value-added tax'.

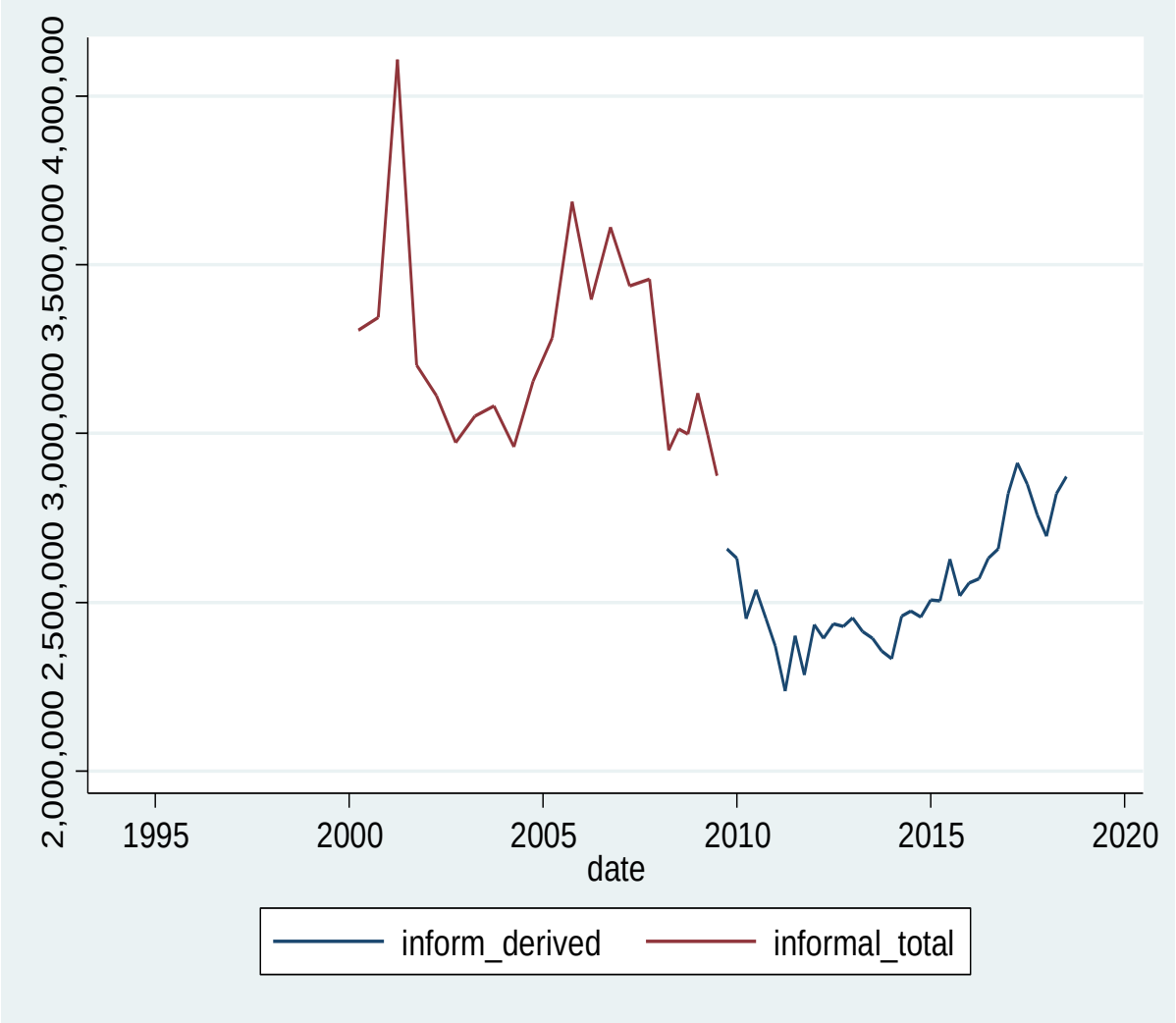
As well as being somewhat ambiguous and possibly incorrect on how the informal sector is defined, the definitions also lead to some strange possibilities. Imagine a taxi driver who works for a boss who owns five taxis and employs five drivers to drive these taxis. We could assume that the taxi owner is not registered for tax and the drivers are not either. Because of the size criterion in informal sector employment that is only used for employees, this would lead to the employees being classified as in the formal sector, since the firm they work for has five employees, while the owner would be classified as informal, since he is not registered for income tax or VAT.

The change in definition of 'informal sector' is important for any estimates of the change in informal sector employment using pre- and post-2009 Q3 data: these will not be comparable. Figure 3 shows estimates of total informal sector employment before and after 2009 Q3. In the initial period we used four components to estimate informal employment—wage informal, self informal, unpaid family workers, and domestic work. We excluded subsistence agriculture since this was not counted as employment at all from 2008 onwards. Figure 3 shows a declining trend in informal sector employment at the time of the change in definition, which continued when the new definition was introduced, although there was a shift down by around 10 per cent at the change in definitions. Any analysts should be aware of this change when measuring changes in the size of the informal sector.

Statistical uncertainty in the QLFS

Statistical uncertainty is inherent in any household or firm survey because these surveys cover only a small part of the population. There are methods to reduce statistical uncertainty, primarily through the use of stratification, which Stats SA implements in all of its household (and firm) surveys. This method takes independent samples from each of a predefined set of groups, or strata. In the OHSs and earlier these were urban and rural strata for each province. In the master sample introduced in LFS 2004 these were changed to the 53 district councils, and in the master sample introduced in QLFS 2008 Q1 a more complex set of strata was used.

Figure 3: Changes in the definition of the informal sector, 1999–2018



Note: 'inform_derived' is Stats SA's own variable derived from a number of questions; 'informal_total' is the sum of informal sector self-employment, informal sector wage employment, domestic work, and unpaid family helpers. Informal sector self-employment and informal sector wage employment were obtained from direct questions to individuals if the firm they owned/worked for was in the informal sector. These questions were not asked from 2009 Q3 onwards, hence the change to the derived variable at that point.

Source: Authors' construction based on calculations from PALMS.

Statistical uncertainty is increased, relative to a simple random sample, by the method of sampling used by Stats SA and pretty much all organizations implementing household surveys. This is two-stage cluster sampling, where only some clusters, small physical areas demarcated in the previous population census, are drawn. Stats SA has always undertaken cluster sampling in the household surveys it runs.

In a recent unpublished paper, Kerr and Wittenberg (2019) document that the statistical uncertainty reported in the QLFS release documentation is incorrect and always understated. The likely reason seems to be an odd method of calculating the statistical uncertainty that uses incorrectly aggregated primary sampling units to calculate the standard errors reported in the QLFS release document. In this paper the authors note that they can almost perfectly replicate the uncertainty estimates from Stats SA in the LFS but cannot replicate them in the QLFS. Estimated uncertainty that is too low is very important for any estimate of changes in employment from year to year or quarter to quarter. These numbers are regularly reported on in the media, but generally

the statistical uncertainty is not discussed, and the implication is that these numbers are the truth. The statistical uncertainty of the numbers is important, because if it is too high it is then not actually possible to say with any certainty whether the measured changes in employment are real, or simply an artefact of the particular sample of households that were chosen.

Other relevant issues in the QLFS

We noted above that Stats SA introduced a master sample of EAs based on the 1996 population census in the 1999 OHS and that a new master sample based on the 2001 census was introduced in 2004. This master sample was used up until LFS 2007: September. Another master sample was introduced in 2008 and ran until the QLFS 2014 Q4, but was also based on the 2001 population census enumeration areas.

In QLFS 2015 Q1, a new master sample based on the 2011 census EAs was introduced. Before this, all Stats SA surveys had been drawn from a list of Eas from the 2001 population census. Substantial population changes between 2001 and 2014 would have rendered the list of areas less and less accurate. Areas with substantial population growth would be under-represented in the samples drawn from a list based on the 2001 population census. If high population growth areas also had higher than average levels of unemployment, as seems likely, then unemployment would be underestimated until 2015, when a more representative sample was drawn.

Thus it is not surprising that QLFS 2015 Q1, the first QLFS based on the new master sample, found a substantially higher unemployment rate. Due to its rapid population growth, Gauteng comprised a larger share of the clusters sampled using this new master sample, as did other metros to a lesser extent. But the effect is not only related to the provincial composition of the sample: Gauteng also had a very large increase in the narrow unemployment rate and the total number of people unemployed using the narrow definition—these increased by 15 per cent and 22 per cent respectively. It is likely that this is due to changes in the within-province composition—for example informal areas that did not exist in 2001 would have been completely excluded from possible selection into the sample until 2015. We think that this is a composition issue and not a generic change in the measurement of unemployment because two provinces (Western Cape and Free State) actually had decreases in the unemployment rate across the change in master samples.

One unfortunate side effect of the (correct) change in the sample design to increase the fraction of the sample coming from Gauteng is that the overall non-response rate from QLFS 2015 Q1 onwards is now substantially higher, since Gauteng residents, and urban residents more generally, are more likely to refuse to be interviewed than rural residents from around the country, who are now a smaller fraction of both the population and the sample than before. The sample design for the 2015 master sample implied a sample size from 2015 Q1 onwards that was 10 per cent higher than that of previous surveys, but because of the decline in the response rate the realized sample actually decreased compared with the pre-2015 surveys. Gauteng households made up 15.5 per cent of the 2014 Q4 sample but 23.7 per cent of the realized 2015 Q1 sample. One result of the decrease in the realized sample is that the measured uncertainty in total employment (and any other statistic) will be higher than when the realized sample was larger, all else being equal.

Several commentators have noticed that the change to using the 2015 master sample has impacted both employment and unemployment. Makgetla (2016) discusses the large estimated decreases in employment between 2015 and 2016 and argues that these may be the result of the new master sample in 2015, supposedly because 2015 estimated too much employment, and the decline in 2016 is actually a correction. Our explanation is that this is simply statistical variation due to relatively small samples and the associated statistical uncertainty. We base this conclusion on the

fact that the standard error of the change between 2015 Q1 and 2016 Q1 is large enough that this yearly change is not statistically significant,[2] informed by our work in Kerr and Wittenberg (2019).

Arrow (2018) discusses the large increase in the unemployment rate in QLFS 2015 Q1 and puts the blame on the master sample. Arrow (2018) incorrectly argues—based on declining realized sample sizes, declining average household sizes, and increases in one-person households—that in fact the unemployment rate increase he documents is *understated* because of too many one-person households in the samples realized from the 2015 master sample. Our explanation is simpler: the master sample used until 2014 was very out of date and was not representative of South Africa at that time. The new master sample based on the 2011 census correctly sampled areas where unemployment is higher, and single-person households are also much more common (for example Gauteng and other urban areas); this explains the jump in the unemployment rate.

2.1.4 General issues in the OHS, LFS, and QLFS surveys

Measurement error in earnings in the Stats SA household surveys

Wittenberg (2017b) argues that respondents in household surveys under-report earnings and that rich individuals are more likely to refuse to respond to surveys, and, if they do respond, to refuse to answer questions about their earnings. Wittenberg (2017a) undertook a comparison between individual-level SARS tax assessment records from tax year 2011 and the QLFS over the same period. He showed that earnings reported in household surveys are on average 40 per cent lower than those in the tax assessment data, with larger differences at the top of the distribution and only an 18 per cent premium at bottom end of the tax data (this was at the three-millionth employee, i.e. the data covered approximately the top 25 per cent of the earnings distribution). The premium was declining monotonically from close to 60 per cent at rank 500,000 to 20 per cent for the last person in the tax assessment data—approximately the three-millionth highest earner. The extent of under-reporting is likely to be lower and declining for the bottom 75 per cent of the distribution, where benefits and tax are less likely to be paid. But this is an open and extremely important question.

Earnings mismeasurement matters for a number of policy-related reasons. If individuals in the bottom half of the earnings distribution under-report their earnings, this would affect any estimates of, for example, the proportion of earners likely to be affected by a minimum wage or an increase in the minimum wage. Further work is being undertaken by one of the authors in comparing the IRP5 tax data—which goes much further down the earnings distribution than the assessment data used by Wittenberg (2017b)—and the QLFS.

A further argument for measurement error in the LFS and QLFS is made by Seekings and Nattrass (2015: 58), who stated, without much evidence, that 'it is likely that household surveys under recorded earnings and did so by a rising margin in the 2000s'. Kerr and Wittenberg (2017) provided some evidence against this assertion. They compared the total value of earnings for employees reported in the OHS, LFS, and QLFS surveys, imputing for missing earnings due to refusals and bracket responses, to the total compensation in the national accounts, as reported in the South African Reserve Bank Quarterly Bulletin. They found that total earnings (summing over all individuals and using weights to estimate the population total) are always lower in the household surveys than in the national accounts data, providing support for the argument that earnings are under-reported in total (although this could include the situation where only the very rich under-

[2] The p-value is 0.136 according to our calculations. The QLFS 2016 Q1 release document (Statistics South Africa 2016) indicates that the change was statistically significant, with a p-value of 0.00. We think this is incorrect, as discussed in the section above on 'Employment in the QLFS' and in more detail in Kerr and Wittenberg (2019).

report their earnings and the poor do not). The ratio is not, however, declining over time; if anything, there is a small increase from 2000 to 2014, contradicting Seekings and Nattrass (2015).

Thus, while there is evidence that earnings in the household surveys are under-reported at the top end of the distribution, the evidence does not suggest that this has increased over time. Whether there is under-reporting of earnings in household surveys for the middle and bottom of the earnings distribution is, for now, an open question.

Weighting and changes over time

Household surveys include survey weights to gross up sample statistics to estimate population statistics. These weights are a function of the probability of selection of each household, the extent of non-response, and the calibration done to make the population totals match demographic models of the population. If the demographic models are updated with new information the weights from new surveys will be adjusted, but the weights of older surveys are often not, even if the new demographic model suggests that the previous population totals were incorrect. This can result in incorrect weights and erratic estimates of totals that can change quite substantially over a relatively short period. Branson and Wittenberg (2014) suggested using cross entropy weighting with a consistent demographic model to adjust weights in all waves of a series of household surveys, and, as an example, reweighted the OHS and LFS surveys using the Actuarial Society of South Africa's 2003 demographic model. Stats SA does periodically revise the weights of some of its older surveys when new demographic information, for example from the 2011 population census, is obtained.

The implication of this discussion is that any estimate of changes in employment or earnings over time should use weights that are themselves calibrated on a consistent demographic model. In previous versions of PALMS we undertook this using first the 2003 Actuarial Society of South Africa (ASSA) model and then the 2008 ASSA model, which was last updated in 2011. As we document in the PALMS v3.3 guide, the weights calibrated to the ASSA 2008 demographic model underestimate the population by several million people compared with the Stats SA mid-year population estimates by 2018, largely as a result of overestimating mortality rates.

In PALMS v3.3 we have thus shifted, from 2018, to using the Stats SA mid-year population estimates, which run back to 2002. Before this time, PALMS used population totals projected back using simple exponential growth rates. In a related project funded by the UNU-WIDER SA-TIED programme, we aim to improve on this simple expansion backwards by updating the Centre for Actuarial Research demographic model, in collaboration with Professor Rob Dorrington, the author of the model.

2.2 Project for Statistics on Living Standards and Developments, 1993

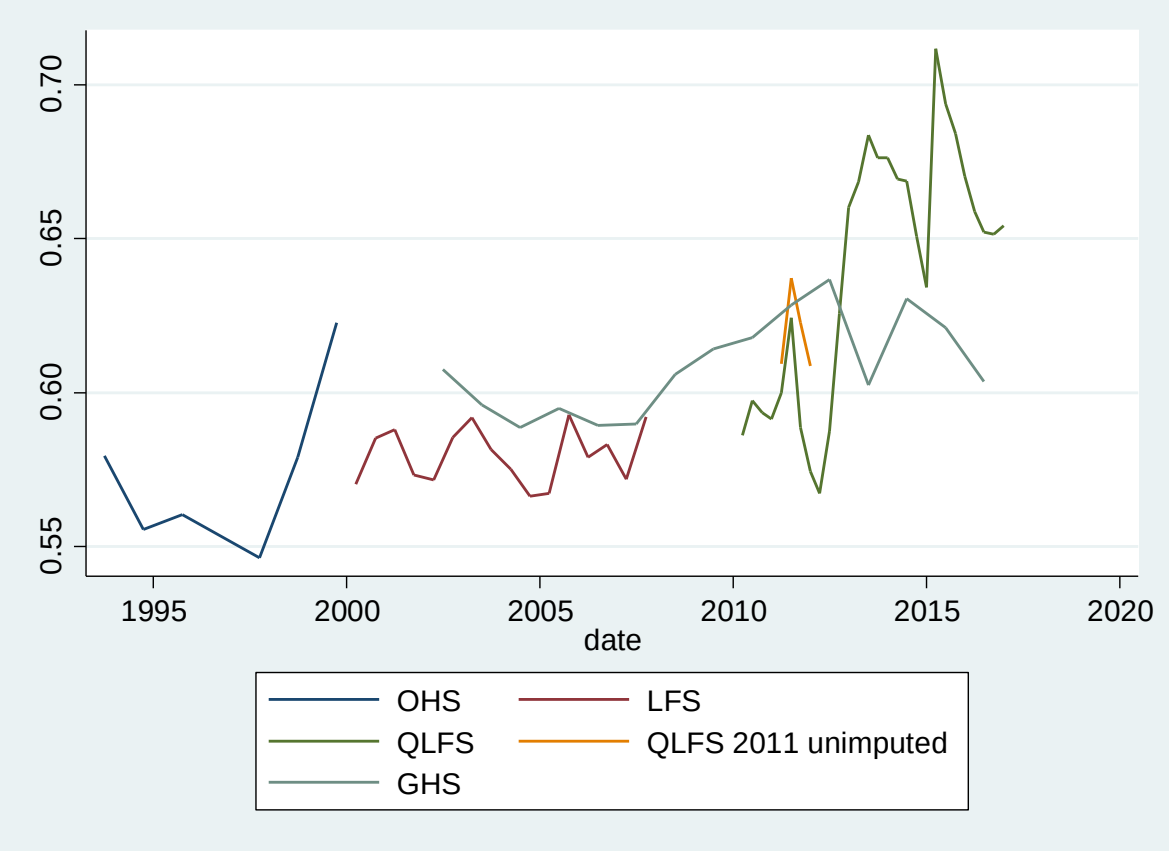
Given the concerns raised above about measurement of some aspects of employment in the early OHSs, it is pertinent to mention an alternative and much-used survey from the end of the Apartheid period. The 1993 PSLSD was run by SALDRU at UCT and based on the World Bank's Living Standards Measurement Surveys. The questions on earnings and employment were not identical to those of the subsequent Stats SA surveys, but they are similar enough to be comparable. Analysts seeking to describe changes over the last 25 years are encouraged to use all available data between the start and end points of any analysis, at least initially, so that any trends are not the result of the particular start and end points chosen. This should include the 1993 PSLSD for analysis of changes in the post-Apartheid period. This is one of the reasons the PSLSD has been included in recent versions of PALMS.

2.3 General Household Survey

After OHS 1999, Stats SA split the content of the OHSs into two. The LFS began in February 2000 and focused on the labour market. The GHS was then aimed at measuring the non-labour market outcomes that had been surveyed in the OHS—for example access to services, fertility, and mortality. Some basic labour market questions were, however, still asked in every GHS, including the employment status of each individual and their earnings. Other questions such as industry or occupation were only asked in some waves.

The GHS can thus be used as a check on the earnings and employment data that are obtained from the LFS and QLFS. This is particularly useful for earnings, because the GHS does not impute earnings for those responding in brackets or complete refusals. This means that it is possible to obtain a series of roughly consistent earnings data between 1993 and the latest GHS. The QLFS and LFS are run by the Labour Statistics section in Stats SA, while the GHSs are run by the Social Statistics section, so this means there are two somewhat independent sources of earnings and employment data. However it should be noted that both the GHS and the LFS/QLFS use the same master sample. Any issues with the master sample, some of which were noted above, will thus affect both the GHS and the LFS/QLFS.

Figure 4: Gini coefficients in earnings in the PSLSD, OHS, LFS, QLFS, and GHS surveys, 1994–2017



Source: Authors' construction based on calculations from PALMS, GHS, and unimputed QLFS 2011 data.

Figure 4 shows the Gini coefficient of earnings from the GHS, as well as the PSLSD, OHS, LFS, QLFS, and unimputed QLFS from 2011 that were shown in Figure 2. The GHS and LFS track each other fairly closely, but the QLFS diverges sharply from the GHS when earnings data began to be collected in the QLFS in 2010. This is further evidence that the imputation procedures in the QLFS are the likely cause of the erratic Gini coefficient in the QLFS. It is also a further reason

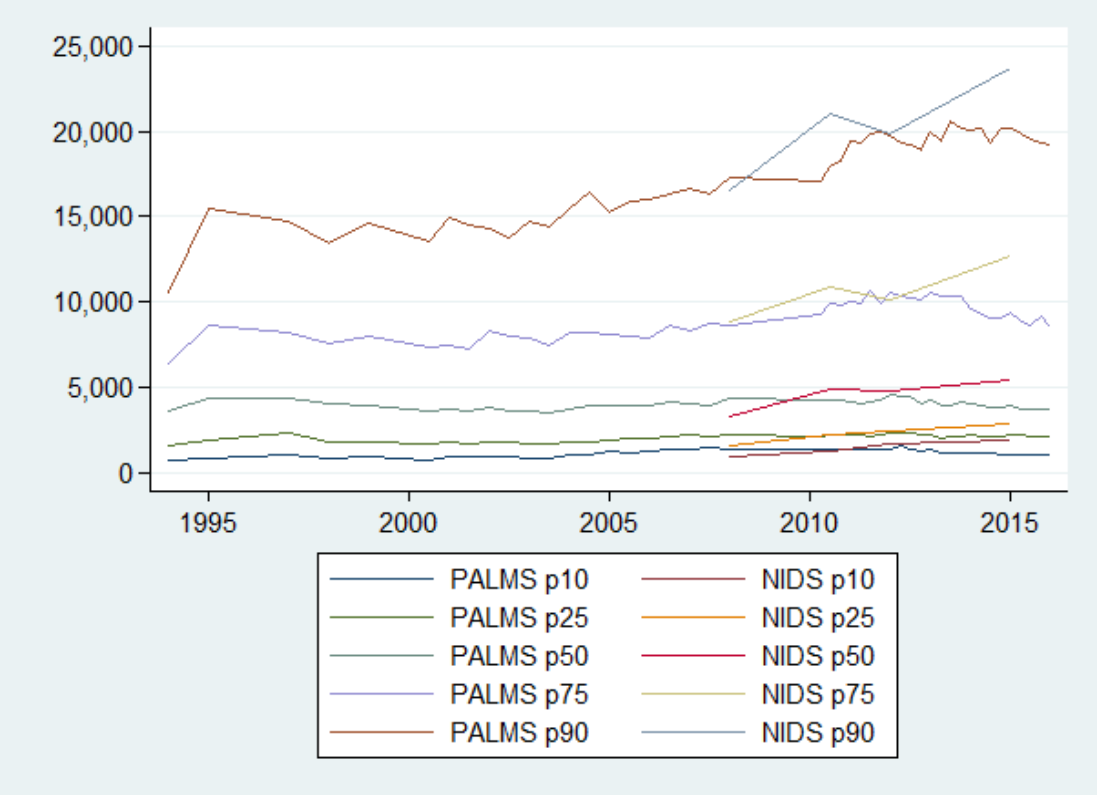
to motivate Stats SA to release unimputed earnings data, or at least indicators of which individuals' earnings are imputed.

2.4 National Income Dynamics Study

The National Income Dynamics Study (NIDS) is a panel survey that began in 2008 and surveyed the same participants every roughly two years, with the most recent wave being 2017. The first wave of NIDS was a representative cross-section, directly comparable with the 2008 QLFSs (though unfortunately no earnings data were collected in these QLFSs). The subsequent waves involved tracking the same individuals, even those that moved. The loss of members of the 2008 wave in subsequent waves means that NIDS is likely to have become less representative of South Africa over time. But the data can be used to estimate total employment and earnings statistics, which we undertake below and compare with the other sources of household survey data.

Figure 5 shows comparisons of various parts of the earnings distribution in both NIDS waves 1–4 and the QLFS. It shows that the NIDS and LFS earnings values at various percentiles are fairly similar at the start of NIDS in 2007/8 but that the earnings values at these various percentiles have increased faster in NIDS than in the QLFS. Figure 6 shows the bottom half of the distribution, which suggests that the NIDS percentiles become quite a lot higher than the QLFS equivalent percentiles by the end of NIDS wave 4.

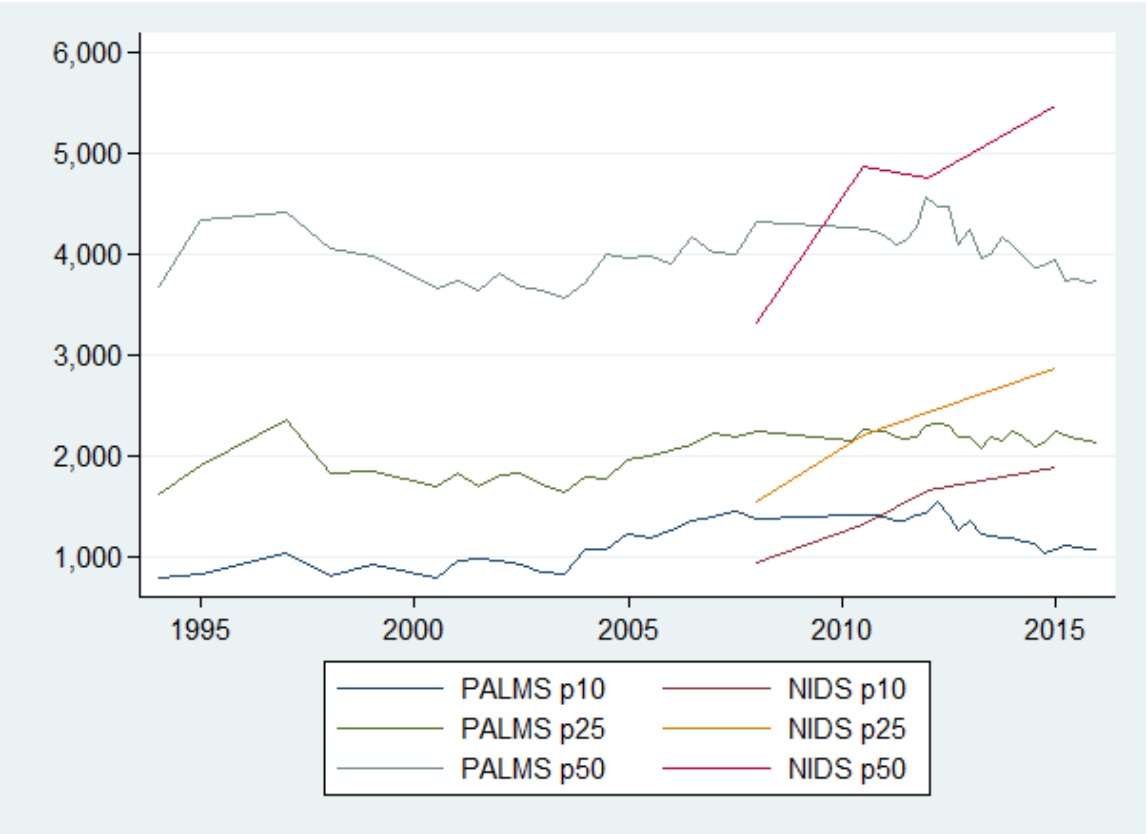
Figure 5: Real monthly earnings at low percentiles of the earnings distribution, NIDS and PALMS



Note: Deflated to June 2016 prices.

Source: Authors' construction, adapted from McDougall (2018).

Figure 6: Lower earnings percentiles in NIDS and QLFS



Note: Deflated to June 2016 prices.

Source: Authors' construction, based on McDougall (2018).

It should be pointed out that NIDS and the QLFS enumerate and process earnings quite differently, and thus the comparison is not directly one of like with like. In later waves of NIDS, if individuals refuse to give an actual earnings amount they are asked a series of unfolding bracket questions and asked to classify in which one they fall. NIDs does not not impute missing earnings in the public data releases. In the QLFS, Stats SA uses a simple set of earnings brackets, and individuals are asked into which one they fall. Like in NIDS, the bracket questions are asked after individuals refuse to give an exact earnings amount. Stats SA does impute earnings for the QLFS public release data, as discussed above. But since these two quite different sets of data are what is available, any comparison between the sources must make use of them. Two potential explanations for the divergence between the NIDS and QLFS relate to measurement issues. Since NIDS is a panel, it is likely to suffer from attrition bias. If that is a concern not resolved using the panel weights then the suggestion is that NIDS attriters were those with lower earnings. The other possibility is that the imputations undertaken by Stats SA imply that the earnings percentiles in the QLFS are incorrect.

3 Firm survey data

Stats SA has undertaken a number of firm surveys and censuses in the post-Apartheid period. These include the Survey of Employment and Earnings (SEE) between 1998 and 2003 (Bhorat and Oosthuizen 2006), which was used by Stats SA to estimate total formal sector earnings and employment and as an input into GDP estimates. The SEE replaced the Manpower Surveys that had run from the 1960s (Kerr 2015b). Manufacturing censuses were conducted regularly in the

Apartheid period, the last one conducted in 1996 (Fedderke and Simbanegavi 2008). Manufacturing censuses were replaced with the Large Sample Surveys, which cover several sectors and are undertaken every three to four years, while the Annual Financial Statistics and Quarterly Financial Statistics Surveys (UNU-WIDER, no date) are mainly used to estimate GDP.

3.1 Quarterly Employment Statistics surveys

The most relevant firm survey for this paper is the QES survey, which began in 2004, replacing the SEE. It surveys firms every quarter and asks questions on the number of employees (part time and full time), the number of new employees, and employees that recently left the firm. The earnings questions are about total gross earnings in the firm, bonuses, overtime pay, and severance packages.

Both the QLFS household survey and QES firm survey are used by Stats SA to estimate different measures of total employment. As Stats SA emphasizes, some important differences between them should be noted. The QLFS covers all employment, whether or not the firm the worker works in is registered for VAT, whereas the sample frame for the QES is the SARS business register, which includes all VAT-registered enterprises. The QES excludes formal sector agriculture. Mining employment data are collected from the Department of Minerals and Energy rather than from the mining companies themselves (Kerr et al. 2014).

The sample designs in the QES and QLFS are also different in important ways. Because a few relatively large firms account for a substantial fraction of total formal sector employment in South Africa, Stats SA designed the QES to be a census of the largest firms. The fraction of firms in the population that are sampled in the QES declines with firm size. This means that the QES samples contain firms that together account for 45–55 per cent of employment (Kerr et al. 2014). The recent QLFSs contain in the sample around 18,000 employed individuals, which is 0.1 per cent of the total employment estimated from this sample.

Given these differences in sector coverage and sample design, as well as the statistical uncertainty inherent in any survey, it is strange that differences in the quarterly results from the QES and QLFS (for example, one shows a small increase in employment and the other a small decrease) have attracted media attention. Our view is that differences are more likely to be the result of the differences discussed above, rather than due to weaknesses in one or both of the surveys.

Earnings data from the QES is not generally discussed in the media. Wittenberg (2014a) undertook a detailed investigation of the QES and QLFS, finding that the average earnings figure in the QES is double that in the QLFS and showing that under-reporting in the QLFS is the likely explanation for the large differences. His conclusion is that it would not be wise to mix and match earnings and employment data from the QES and QLFS.

Stats SA has not released the QES data, or other firm-level data, publicly, as it argues that public release would make the firms less likely to respond. That being said, the data have been used by several authors. We have already mentioned the Kerr et al. (2014) work on the QES. A PhD thesis made use of the Large Sample Surveys mentioned above (Naughtin 2016), while Flowerday et al. (2017) made use of the 1996 manufacturing census and the 2001 Large Sample Survey of manufacturing. Thus, data have made their way to some researchers, albeit in an ad hoc and potentially uncontrolled manner.

Many other countries do make firm data available to researchers in a variety of ways. Some of these are documented in Vilhuber (2013). One possibility is to provide aggregated information in electronic tables/spreadsheets, but to do so at a relatively disaggregated level. Vilhuber (2013) gives

the example of the Quarterly Workforce Indicators data from the US, which provide employment stocks and flows for each of 3,000 counties in the US.

A solution that allows access to microdata is the use of Statistical Data Enclaves—secure facilities at which researchers can access confidential data. Vilhuber (2013) notes that this is undertaken for some data sources by the US Bureau of Labor Statistics and Statistics Canada. DataFirst has provided such a facility at the University of Cape Town since 2013, and hosts confidential firm-level data from the City of Cape Town, household survey data from the Cape Area Panel Study and the NIDS, and UCT admissions applicant data, among others. The South African National Treasury is building such a centre in Pretoria to allow access to confidential SARS data as part of the UNU-WIDER SA-TIED project.

An extended version of the Statistical Data Enclave is to have a central processing facility with centres located at universities or at regional offices of the organization providing the data. Vilhuber (2013) gives examples from the US Census Bureau's 29 Research Data Centres at universities and institutions around the US and the French Centre d'accès sécurisé distant aux données, which gives access to specific institutions through a piece of hardware that is installed on a computer located at the institution. Access to the US Census Bureau Research Data Centre requires swearing to protect the confidentiality of the data for life, and there are substantial financial and legal penalties for failure to uphold this.

The lack of any use of the methods of access discussed for the QES microdata unfortunately limits both the analysis of employment and earnings that can be conducted on the data, and the ability to check the quality of the data. It should be noted that Kerr et al. (2014) concluded that the QES data that the authors were able to access were of fairly good quality. The QES also has important strengths relative to the tax data, to which we now turn.

4 Administrative data

Administrative data are data collected for administrative processes in government or by private companies or non-governmental organizations (NGOs). The two relevant sources of admin data that we discuss are both collected in the process of government administration of the tax system and the unemployment insurance system.

4.1 South African Revenue Service tax data

SARS collects information from all those employed in pay as you earn (PAYE) tax-registered firms who earn more than R2,000 a year, so this is almost a census of those employed in the formal sector. Earnings information is contained on the IRP5 certificate that firms issue their employees. So, in theory, the IRP5 certificates can be used to obtain estimates of employment and earnings comparable with the QES and the formal sector part of the QLFS (although the discussion of the QLFS above suggests that identifying formal sector employed individuals in the QLFS would require some approximations).

Limited access to the SARS data has been given in two ways. As part of the Research Project on Employment, Income Distribution and Inclusive Growth (REDI) funded by the South African National Treasury, a sample of assessed tax records was made available to a few researchers. Assessment data are for those in the top 25 per cent of the earnings distribution. As discussed above, Wittenberg (2017a) used the assessed data from tax year 2011 to compare the earnings in the QLFS and the assessed tax records from SARS.

The second source of tax data has been made available through a project funded by UNU-WIDER, SARS, and the National Treasury (Pieterse et al. 2018). As part of this project, all (anonymized) IRP5 certificates have been made available in the National Treasury to researchers whose research proposals are accepted.

Given the evidence that there is likely to be measurement error in the earnings data from household surveys, the SARS IRP5 data set is thus a potentially very valuable source of data on earnings, since it is believed that it is subject to much less measurement error. It can be used to extend the Wittenberg (2017a) analysis further down the earnings distribution, which one of the authors is working on.

Unfortunately there are several reasons why the tax data and the household survey cannot easily be compared. Firstly, the IRP5 certificate issued by an employer gives the earnings of an individual in that company for the entire tax year. It should be possible to obtain an estimate of monthly earnings to compare with the household survey data, since the period worked during the tax year is included in the tax certificate. But Kerr (2018) has noted that the 'period employed' variable in the SARS IRP5 data is unreliable and has some strange trends over time.

More recent (as yet unpublished) work on the IRP5 data by one of the authors suggests that this 'period employed' variable is unreliable enough that the Gini coefficient calculated using monthly earnings derived from the total earnings in the year and the period of employment is around 0.9—an unbelievable number.

If one does not want to use period of employment then one could use earnings for the entire tax year (even summing up across different certificates from different companies for the same individual) compared with monthly earnings multiplied by 12, as in Wittenberg (2017b). But this is also not a good method when one wants to compare across the entire earnings distribution. The reason for this is that many individuals are not employed for the entire year, but rather enter and exit employment for relatively short periods. One would then be comparing the group in the household surveys employed at a particular time with the earnings of all those employed at any point in the tax year. This might be reasonable at the top end, where most people would be employed for most of the year, but there are much higher separation rates from employers, which includes movement into non-employment, for those at the bottom end of the earnings distribution in the tax data (Kerr 2018).

The tax data thus need much more careful analysis, coupled with household panel data work describing the movement into and out of employment by individuals at various points in the earnings distribution, before one can usefully use the IRP5 earnings data as a reliable check on the household survey earnings data.

That the IRP5 data are a yearly rather than a point-in-time measure and that period of employment seems mis-measured also affect comparisons with household surveys on employment. One cannot compare formal sector employment in the IRP5 with the household surveys if the period of employment is not correctly measured. Kerr (2018) noted a further difficulty, which was that in 2010–12, SARS did not separate pension income and earnings income. This means that one cannot use the IRP5 data to estimate total employment, since one would be including several million pensioners. Kerr (2018) did attempt to identify what were probable pension funds, but noted that there were around 500,000 pension income certificates that could not be identified. For estimating total employment this is not a serious issue, but any analysis using the data would thus have a substantial number of bogus 'employees' in this period.

The IRP5 data are broad, but they are also shallow—there are only a few individual characteristics, including age and gender. Anonymized passport numbers and the lack of an identity document allows for the identification of foreigners, and it would be possible for SARS to distinguish between citizens and permanent residents using the ID numbers provided on the IRP5 certificates. This must be done by SARS, since researchers (for good reason) do not have access to non-anonymized ID numbers. Characteristics that would allow for more interesting analysis or for comparison with household surveys (particularly race, education, occupation) are missing. As discussed above there are a number of shortcomings of the earnings and employment data obtainable from the IRP5. A much better grip on these is necessary for anyone using the earnings data from SARS and for any meaningful comparisons between the tax and household surveys.

4.2 Unemployment Insurance Fund data

The Department of Labour (DoL) requires each tax-registered firm to submit monthly reports on the number of employees and their earnings, so that UIF contributions can be paid by the firms to the DoL. These data are thus potentially another source of admin data that can be used for analysis and to compare with the household and tax data discussed above. They could also be a measure of employment put out with a much lower lag than the QES data from Stats SA.

Bhorat et al. (2013) used the individual-level claimant data (individuals who became unemployed), but there has not been any research conducted on the firm-level data. There have been suggestions that these data are messy and would thus be hard to analyse. We mention them because they are a possible source of (formal sector) earnings and employment data that could be made available for research.

5 Conclusion

In this paper we have documented a number of sources of microdata on earnings and employment in South Africa. Such sources of data are crucial for any research about the South African labour market, and thus it is important to know both which data can be used and the strengths and weaknesses of each source.

The household surveys undertaken by Stats SA have traditionally been the source of data for analysis of earnings and employment in South Africa. Their main strength is that they are publicly available and well used, and thus the weaknesses are fairly well known in most cases. We have tried to highlight those weaknesses that are important for any analysis of earnings and employment. Probably the key weakness is the imputation of earnings in the QLFS, which substantially hampers any efforts to understand earnings in the period covered by the QLFS (2008–present). We have also documented the 1993 PSLSD and the five waves of NIDS from 2008 to 2017 as alternative sources of household survey data on earnings and employment.

We have not discussed the Income and Expenditure Surveys or the Living Conditions Surveys, which are also possible sources of data on earnings and employment, since these have not been widely used in labour market research in South Africa.

We have documented the Stats SA firm survey, the QES, the results from which are regularly discussed in the media as a measure of formal sector employment. Unfortunately, it does not appear likely that this data source will be made more widely available to researchers and policymakers, after the initial work by Kerr et al. (2014) documenting the data quality and estimating the extent of job and worker flows. This data could be used to better understand labour

demand, worker flows, and labour productivity, amongst other things. It could also be used as a check on the SARS administrative data currently being used by researchers, since Stats SA does substantial work in checking aspects of the data that SARS, which is purely interested in tax receipts, does not. Combined with the other firm surveys undertaken by Stats SA, there would be a wealth of research waiting to be done if researchers were given permission to access these firm data in a secure manner.

Administrative data are an exciting new source for social scientists. The UIF has been briefly discussed speculatively, since this type of data has been an important part of economics research in recent years in other countries. But it has not been used in any analysis that we are aware of.

The recent availability of the SARS tax admin data is thus an important step forward for analysts of the labour market that was made possible through the collaboration of SARS, the National Treasury, and UNU-WIDER. Information on earnings and employment is available from individual IRP5 certificates, although we have noted a number of issues that are encountered when the IRP5 data are used, and when they are compared with the household survey data. This suggests that admin data are not a panacea, but should rather be used in conjunction with household surveys and firm data to better understand the South African labour market.

Current analysts of the South African labour market have available to them a number of possible sources of earnings and employment data. We have tried to document each, focusing on sources that allow researchers to put together a picture of the evolution of earnings and employment in the post-Apartheid period, but pointing out that none are perfect and some may be better than others.

References

Arrow (2018). 'The Sudden Jump in the Unemployment Rate in 2015: Is There a Break in the QLFS Data?' Econ3×3, 16 May. Available at: http://www.econ3x3.org/article/sudden-jump-unemployment-rate-2015-there-break-qlfs-data (accessed 13 February 2019).

Bhorat, H., and M. Oosthuizen (2006). 'Evolution of the Labour Market: 1995–2002'. In H. Bhorat and R. Kanbur (eds), *Poverty and Policy in Post-Apartheid South Africa*. Cape Town: Human Sciences Research Council.

Bhorat, H., S. Goga, and D. Tseng (2013). 'Unemployment Insurance in South Africa. A Descriptive Overview of Claimants and Claims'. DPRU Working Paper 13/160. Cape Town: Development Policy Research Unit. Available at www.dpru.uct.ac.za/sites/default/files/image_tool/images/36/DPRU%20WP13-160.pdf (accessed 13 February 2019).

Branson, N., and M. Wittenberg (2007). 'The Measurement of Employment Status in South Africa using Cohort Analysis, 1994–2004'. *South African Journal of Economics*, 75(2): 313–26.

Branson, N., and M. Wittenberg (2014). 'Reweighting South African National Household Survey Data to Create a Consistent Series over Time: A Cross-Entropy Estimation Approach'. *South African Journal of Economics*, 82(1): 19–38.

Casale, D., M. Muller, and D. Posel (2004). '"Two Million Net New Jobs": A Reconsideration of the Rise in Employment in South Africa, 1995–2003'. *South African Journal of Economics*, 72(5): 978–1002.

Central Statistical Service (1998). 'Living in South Africa. Selected Findings from the 1995 October Household Survey'. Pretoria: CSS. Available at http://www.statssa.gov.za/publications/LivingInSA/LivingInSA.pdf (accessed 13 February 2019).

Fedderke, J., and W. Simbanegavi (2008). 'South African Manufacturing Industry Structure and Its Implications for Competition Policy'. ERSA Working Paper 111. Cape Town: University of Cape Town. Available at https://econrsa.org/papers/w_papers/wp111.pdf (accessed 13 February 2019).

Finn, A., and V. Ranchhod (2017). 'Short-Run Differences between Static and Dynamic Measures of Earnings Inequality in South Africa'. REDI3×3 Working Paper 35. Cape Town: University of Cape Town.

Flowerday, W., N. Rankin, and V. Schöer (2017) Continuity and Change: Shifts and Continuities in South African Regulation of Labor Market since 1994 and the Comparative Analysis of the Impact of Selected Labor Market Policies on Employment'. Paper presented at the 2017 CSAE Conference, University of Oxford. Available at https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=CSAE2017&paper_id=314 (accessed 13 February 2019).

Kerr, A. (2015a). 'A Guide to the Employers and the Self-Employed Series'. Cape Town: DataFirst, University of Cape Town. Available at https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/514/download/7060 (accessed 13 February 2019).

Kerr, A. (2015b). 'An Introduction to the Manpower Survey Data'. DataFirst Technical Paper 32. Cape Town: DataFirst, University of Cape Town.

Kerr, A. (2018). 'Job Flows, Worker Flows, and Churning in South Africa'. *South African Journal of Economics*, 86(S1): 141–66.

Kerr, A., and M. Wittenberg (2015). 'Sampling Methodology and Fieldwork Changes in the October Household Surveys and Labour Force Surveys'. *Development Southern Africa*, 32(5): 603–12.

Kerr, A., and M. Wittenberg (2017). 'Public Sector Wages and Employment in South Africa'. REDI3×3 Working Paper 42. Cape Town: University of Cape Town. Available at: http://www.redi3x3.org/paper/public-sector-wages-and-employment-south-africa (accessed 22 May 2019).

Kerr, A., and M. Wittenberg (2019). 'Statistical Uncertainty in the Statistics South Africa Quarterly Labour Force Surveys'. Unpublished Working Paper.

Kerr, A., D. Lam, and M. Wittenberg (2019). 'Post-Apartheid Labour Market Series [dataset]'. Version 3.3. Cape Town: DataFirst, University of Cape Town.

Kerr, A., M. Wittenberg, and J. Arrow (2014). 'Job Creation and Destruction in South Africa'. *South African Journal of Economics*, 82(1): 1–18.

Makgetla, N. (2016). 'The Jobs Bloodbath that Wasn't: What Happened to Employment in the First Quarter of 2016?' Policy Brief 6/2016. Pretoria: Trade & Industrial Policy Strategies. Available at: https://www.tips.org.za/images/Policy_Brief_What_happened_to_employment_in_first_quarter_2016_-_May_2016_final.pdf (accessed 13 February 2019).

McDougall, B (2018). 'Measuring Wages and Inequality in South Africa Using Two Nationally Representative Data Series'. Unpublished master's dissertation. Cape Town: University of Cape Town.

Naughtin, T. (2016). 'Firm Productivity, International Trade and Competition: Using Micro Data to Examine the Dynamics of South African Firms'. Unpublished PhD thesis. Stellenbosch: University of Stellenbosch. Available at http://scholar.sun.ac.za/handle/10019.1/100085 (accessed 13 February 2019).

Pieterse, D., E. Gavin, and F. Krueser (2018). 'Introduction to the South African Revenue Service and National Treasury Firm-Level Panel'. *South African Journal of Economics*, 86(S1): 6–39.

Seekings, J., and N. Nattrass, N. (2015). *Policy, Politics and Poverty in South Africa*. Houndmills, Basingstoke: Palgrave Macmillan.

Statistics South Africa (2000). 'Statistical Release P0317. October Household Survey 1999'. Pretoria: Statistics South Africa.

Statistics South Africa (2001). 'Discussion Paper 1. Comparative Labour Statistics. Labour Force Survey: First Round Pilot, February 2000'. Pretoria: Statistics South Africa.

Statistics South Africa (2008). 'Guide to the Quarterly Labour Force Survey, August 2008'. Pretoria: Statistics South Africa. Available at: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/498/download/6619 (accessed 13 February 2019).

Statistics South Africa, (2016). 'Statistical Release P0211. Quarterly Labour Force Survey Quarter 1: 2016'. Pretoria: Statistics South Africa. Available at: http://www.statssa.gov.za/publications/P0211/P02111stQuarter2016.pdf (accessed 13 February 2019).

UNU-WIDER (no date). 'Firm Data Inventory for South Africa'. Helsinki: UNU-WIDER. Available at: https://www3.wider.unu.edu/sites/default/files/Opportunities/PDF/firm-data-inventory-for-south-africa.pdf (accessed 13 February 2019).

Vilhuber, L. (2013). 'Methods for Protecting the Confidentiality of FirmLevel Data: Issues and Solutions'. Centers, Institutes, Programs 3-2013. Ithaca, NY: Labor Dynamics Institute, Cornell University. Available at: https://digitalcommons.ilr.cornell.edu/cgi/

viewcontent.cgi?referer=&httpsredir=1&article=1018&context=ldi (accessed 13 February 2019).

Wittenberg, M. 2008. October Household Survey 1994. Mellon Data Quality Project Technical Papers.

Wittenberg, M. (2014a). 'Analysis of Employment, Real Wage, and Productivity Trends in South Africa since 1994'. Conditions of Work and employment Series 45. Geneva: International Labour Office. Available at: http://www.ilo.org/wcmsp5/groups/public/@ed_protect/@protrav/@travail/documents/publication/wcms_237808.pdf (accessed 13 February 2019).

Wittenberg, M. (2014b). 'Data Issues in South Africa'. In H. Bhorat, A. Hirsch, R. Kanbur, and M. Ncube (eds), *Oxford Companion to the Economics of South Africa*. Oxford: Oxford University Press.

Wittenberg, M. (2016). 'Trends in Earnings and Earnings Inequality in South Africa: 1993–2014'. Unpublished working paper.

Wittenberg, M. (2017a). 'Measurement of Earnings: Comparing South African Tax and Survey Data'. REDI3×3 Working paper 41. Cape Town: University of Cape Town.

Wittenberg, M. (2017b). 'Wages and Wage Inequality in South Africa 1994–2011. Part 1: Wage Measurement and Trends'. *South African Journal of Economics*, 85(2): 279–97.

World Bank (2018). 'Overcoming Poverty and Inequality in South Africa. An Assessment of Drivers, Constraints and Opportunities'. Washington, DC: World Bank. Available at: http://documents.worldbank.org/curated/en/530481521735906534/pdf/124521-REV-OUO-South-Africa-Poverty-and-Inequality-Assessment-Report-2018-FINAL-WEB.pdf (accessed 13 February 2019).