BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 5 (2020) 39-49

BRILL OPEN
brill.com/rdj

The Post-Apartheid Labour Market Series

Social and Behavioural Sciences

Andrew Kerr
School of Economics & DataFirst, University of Cape Town (UCT), Cape Town, South Africa
andrew.kerr@uct.ac.za

Martin Wittenberg
DataFirst & School of Economics, UCT, Cape Town, South Africa
martin.wittenberg@uct.ac.za

Abstract

The Post-Apartheid Labour Market Series (PALMS) is a compilation of microdata from 69 household surveys conducted in South Africa. The dataset and the code used to create the data are publicly available from DataFirst, a data repository at the University of Cape Town (www.doi.org/10.25828/gtr1-8r20). To harmonise the data required understanding the differences across the surveys, which has generated new knowledge about the South African labour market.

Keywords

household survey data – South Africa – labour market – data harmonisation

– Related data set "Post-Apartheid Labour Market Series (PALMS)" with DOI www.doi.org/10.25828/gtr1-8r20 in repository "DataFirst"

1. Introduction

South Africa has conducted multiple nationally representative labour market-related household surveys since 1993. The Post-Apartheid Labour Market Series (PALMS) is a harmonised compilation of microdata from 69 of these surveys and was created by the authors as well as David Lam at the University of Michigan (Kerr, Lam & Wittenberg, 2019). The surveys included in PALMS are the October Household Surveys (1994–1999), the biannual Labour Force Surveys (LFS) (2000–2007) and the Quarterly Labour Force Surveys (QLFSS) (2008–2019), all conducted by Statistics South Africa, the National Statistics Office (NSO). PALMS also includes the 1993 Project for Statistics on Living Standards and Development (PSLSD) conducted by the Southern African Labour and Development Research Unit (SALDRU) at the University of Cape Town (UCT). PALMS is publicly available through DataFirst at UCT (see Figure 1). We have also released a guide to PALMS to help users understand the data as well as issues they might encounter when using it to undertake labour market analysis (Kerr & Wittenberg, 2019a). We discuss the most recent version (3.3) of the data, but we have released several prior versions, and we plan to continue to update it once a year, as new surveys are released.

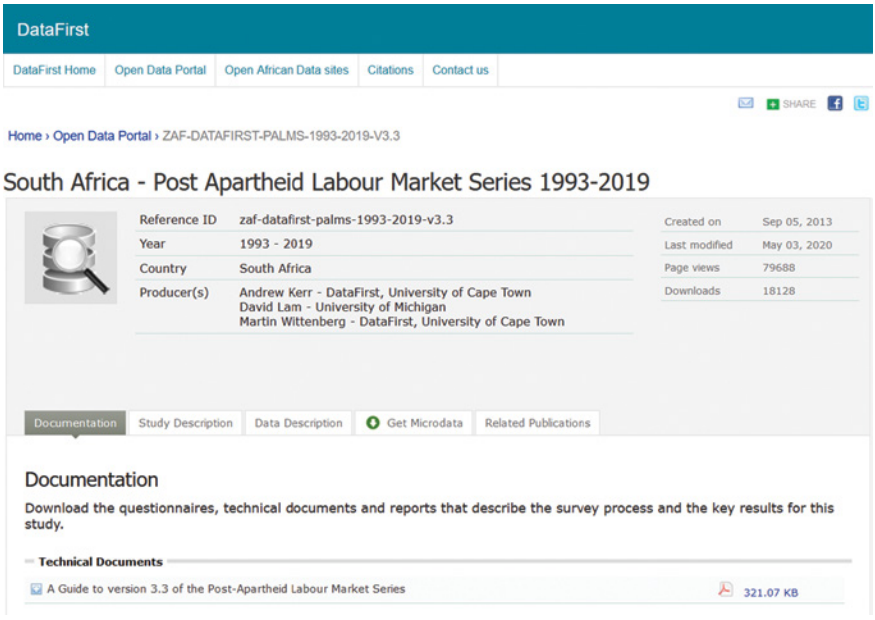

FIGURE 1 Screenshot of the PALMS startpage at Datafirst's Open Data Portal

2. Background

The surveys included in PALMS enable, in theory, research about inequality, unemployment, changes in employment structure and many other pressing labour market issues in South Africa. The microdata for all these surveys is publicly available. So any researcher could download each survey and assemble them to better understand the evolution of the South African labour market. Given the rapid change in the country since 1993, it was inevitable that the surveys differed from each other in a multitude of ways, as we discuss in more detail below.

PALMS is an important dataset for several reasons. Firstly, publicly available and harmonised microdata from these labour market-related surveys makes it much easier for researchers to better understand the South African labour market. Secondly, the availability of PALMS means that researchers do not have to duplicate work done by many others in creating the data. Thirdly, if results on important issues differ across researchers and these researchers are all using PALMS, then at least one can rule out as an explanation for these differences that the researchers created the data in different ways. Finally, many researchers have simply used two surveys to describe trends over time. But often the conclusions drawn depend on which two surveys were used. In South Africa, the 1995 OHS was used by many researchers to describe changes between 1995 and a later point but it turned out that 1995 was not an ideal anchor-point for such comparisons, for reasons that are not clear (Branson & Wittenberg, 2007). OHS 1995 found many more employed African men, many more orphans and a much smaller gender wage gap than the surrounding surveys (Wittenberg, 2014b). Having all the surveys together allows researchers to examine trends and to make sure their results are not an artefact of the two surveys they chose to compare.

3. Problem

There are many difficulties in constructing a consistent picture of earnings, employment and unemployment from South African household surveys. These include changes in questions on key-outcomes, sampling methods, weighting, fieldwork, data imputation and data processing by the data producer. These difficulties mean that depending on which surveys researchers use and what decisions they make about many aspects of the data processing and creation, they can reach very different conclusions. The point of PALMS is that a common dataset is created that researchers can build on, but also that

all the processing undertaken is made transparent by providing the code used to create the data, which allows others to criticise and/or replicate the data. Although we give examples of difficulties and issues from South African labour market-related surveys, these issues apply in many other contexts also, and we hope that this discussion is helpful for other researchers conducting similar exercises in different contexts.

4. Methods

We have documented many of the changes and inconsistencies across the surveys that PALMS included in prior research. We briefly discuss these here to demonstrate how the harmonisation of the various surveys in creating PALMS required a substantial research investment, generated a new research program in understanding the evolution of the South African labour market, as well as giving new insights into results and trends that were not well understood. The issues we discuss include changes in questions about earnings and processing of earnings data, earnings imputation methods, fieldwork changes and issues with weighting.

PALMS includes several labour market-related variables about education, employment status, employment and employer characteristics. It also includes a harmonised earnings variable. Several changes across the surveys make constructing even simple descriptions of changes in earnings over time very difficult. Wittenberg (2014a) provides a detailed discussion of the inconsistencies in earnings across the different surveys. These include very different questions to the self-employed about their income in different OHSs, changes in the way income bracket questions were asked, the extent of outliers in earnings across the surveys, as well as those reporting zero earnings. Probably the most valuable aspect of PALMS is the harmonising of all these earnings data into a single earnings variable. But PALMS also includes a separate earnings data file with the original variables from all the surveys, in case researchers need to investigate these further.

Between 10–15% of the employed in the surveys in PALMS are missing earnings data. There are also bracket responses and responses that are clearly outliers. These responses types are unlikely to occur randomly across individuals. Imputation is the usual solution to this type of problem. But single imputation will understate the true statistical uncertainty of any estimate. PALMS thus also includes a separate file with multiple imputed earnings data: earnings for refusals, bracket responses and outliers are all imputed for each individual to

allow analysts to understand the impact of missing earnings data and to generate unbiased standard errors. Schafer (1999) recommends imputing 5–10 times, and PALMS contains 10 imputations.

Data producers also undertake imputation of missing or non-sensical earnings responses. Unfortunately, and unlike the PALMS imputations, the methods used are often not carefully documented or explained. The 1994 OHS earnings data was heavily imputed without any documentation by the NSO. PALMS provides OHS 1994 earnings created from a process of reverse engineering based on Wittenberg (2008). The more recent QLFSs also have substantial imputations by the NSO, and again the methodology is not documented, and no imputation flags are provided in the data. We have used PALMS to show that there are two imputation regimes for two different periods of time and that the imputation is likely to be driving impossible changes in the Gini coefficient in earnings (Kerr & Wittenberg, 2019b), but, unfortunately, we cannot do any more without publicly available unimputed data from the NSO, which we have requested multiple times, unfortunately without success.

Kerr and Wittenberg (2019b) discuss several fieldwork-related issues in the surveys in PALMS. These include the vast overestimate of both subsistence agricultural workers in the two LFSs from 2000 and of informally self-employed in February 2001, which together resulted in an impossibly large increase in the employment rate and labour force participation rate at the start of the new millennium. Since the QLFSs began in 2008, Statistics South Africa has excluded subsistence agriculture from the definition of employment. We cannot undo these kinds of issues post-fieldwork, but some partial solutions are possible. For example, PALMS includes an employment dummy variable that excludes those employed in subsistence agriculture, given the inconsistencies in their treatment over the surveys. The PALMS guide also explains some of these issues, so that researchers are made aware of them (Kerr & Wittenberg, 2019a).

Many household surveys are calibrated to a demographic model of the population they cover. This means that the sample design weights adjusted for non-response are calibrated, so the weighted totals match the best estimates from the demographic model on a few key characteristics, usually sex, age groups, province and (in South Africa) self-declared race. The difficulty is that the demographic models may be updated and improved, and thus earlier population estimates may be incorrect. This leads to large jumps in totals whenever major adjustments are made to these demographic models (Branson & Wittenberg, 2014). To ameliorate this issue, PALMS includes weights constructed using a consistent demographic model for the entire period, using cross-entropy weighting to calibrate the weights (Wittenberg, 2010).

5. Data

– **Post-Apartheid Labour Market Series (PALMS) deposited at DataFirst – doi:**www.doi.org/10.25828/gtr1-8r20
– **Temporal coverage:** 1993–2019 and onward

There are 69 surveys included in PALMS version 3.3. The earliest survey is the 1993 PSLSD. Before this, the Apartheid state did not collect nationally representative data, as it did not want to publicise the dire state of the living standards of many South Africans. The data that it did collect was also not publicly available. The PSLSD is also the only survey in PALMS not conducted by the NSO. It was included because it is probably the most well-used household survey since it was undertaken in 1993 and contains valuable information about the state of the labour market just before the advent of democracy. The OHSs were run annually between 1994 and 1999, the 1994 survey being run 6 months after the first democratic elections. The OHSs (and the PSLSD) included a much broader set of questions than just those about the labour market, but we have not included much of this data in PALMS since it is a labour-focused dataset. The LFSs were run biannually and focused on labour market-related issues. The QLFSs have run every quarter since 2008 and are similar to the LFSs. Table 1 summarises the data.

Several types of variables are included in PALMS. These include the household and person identifiers, the survey design variables, basic demographic and location information on each individual, educational attainment, labour force status and numerous variables pertaining to the individual's employment and employer. Table 2 shows a summary of the variable types in PALMS, with some examples. Since the surveys differed, not all surveys have the same set of variables, but there is a core of common variables. PALMS includes the original household and person identifiers so that researchers can merge in

TABLE 1 Summary of surveys included in PALMS

Year	Survey	Freq	Number of Surveys	Sample size (households)
1993	PSLSD	1 per year	1	9000
1994–1999	OHS	1 per year	6	16000–30000
2000–2007	LFS	2 per year	16	30000
2008–2019	QLFS	4 per year	46	30000–33000

TABLE 2 Summary of variable types in PALMS, with examples

Variable type	Variable Examples
Identifiers	Household and person id
Complex survey information	Weight, Enumerator Area, Stratum, survey
Location data	Province, urban/rural
Demographic characteristics	Age, sex, population group
Educational data	Years of education, enrolment status
Job characteristics	Occupation, contract type, hours worked, job start year
Employer characteristics	Firm size, industry, employer type
Earnings data	Monthly real income, outlier indicator

other data from the surveys that are not in PALMS. All the surveys in PALMS collected data on all resident members. We have included children and the elderly in the data even though they have no labour market data because many labour related research questions involve these groups.

All the surveys included in PALMS were two-stage cluster samples with stratification, although the number of households sampled per cluster and the strata have varied substantially over the surveys. One of the features of PALMS is the inclusion of the correct strata and cluster variables, which were often incorrectly released in the original versions released by Statistics South Africa. The inclusion of these variables allows the user to specify the correct sample design in the statistical software used to conduct any analysis (along with the cross-entropy weights discussed earlier). The PSLSD sample size was 9000 households. The sample sizes of the OHS varied between 16 000 and 30 000 households, partly as a result of how much funding was available to conduct the surveys. The sample size for the LFSS and QLFSS until 2014 was 30 000 households and 33 000 households from 2015 onwards. The realised samples are lower than this for all the surveys as a result of refusals, non-contact, vacant dwellings etc.

6. Palms Use Case

Having explained the PALMS data, we now show two brief examples of how PALMS can be used to shed light on aspects of the South African labour market. Figure 2 shows the employment rate, non-participation rate and broad and

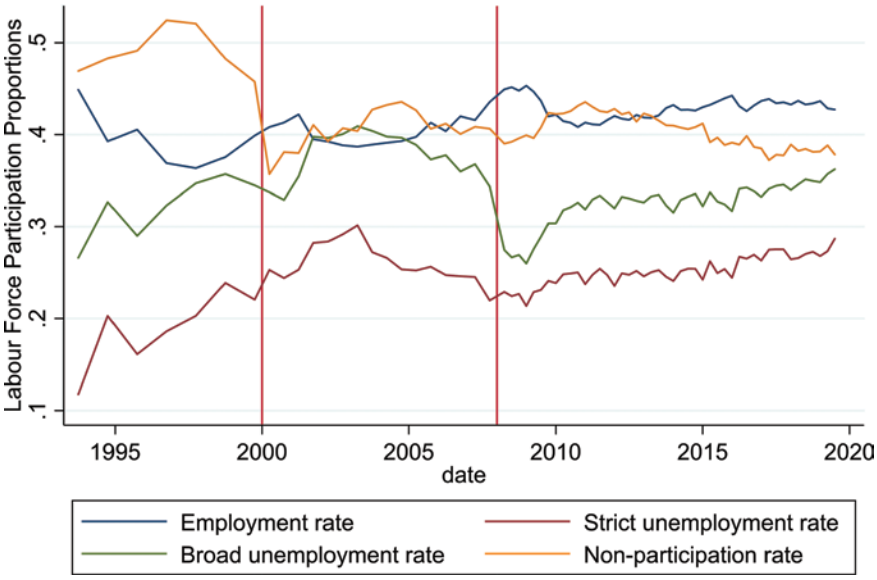

FIGURE 2 Post-Apartheid Labour Force Participation
OWN CALCULATIONS FROM PALMS V3.3

strict unemployment rates. It shows the well-known fact that unemployment has grown dramatically in the post-Apartheid period and is very high, whether one uses the strict or broad rates (broad unemployment includes those who have not looked for work but who want work). But the figure also shows the much less well-known fact that the employment rate has been roughly constant over the post-Apartheid period. These two facts are possible because of the big decline in the non-participation rate. The two red lines show the change between the OHSS and LFSS (in 2000) and the LFSS and the QLFS (in 2008). Clearly, there have been changes in definitions between these surveys that impact the measurement of labour force participation.

Figure 3 shows various percentiles of the earnings distribution, extending the work of Wittenberg (2017a, 2017b). Median earnings has declined since 1993, whilst the 75th and 90th percentile have increased, the 90th percentile very substantially. The 25th and 10th percentiles have actually increased.

Table 3 shows the 95% confidence intervals for these percentiles in 1993 and 2017, accounting for the complex survey design. These changes mean that inequality within the bottom half of the distribution has decreased, whilst inequality in the top half has increased substantially. It seems that earnings

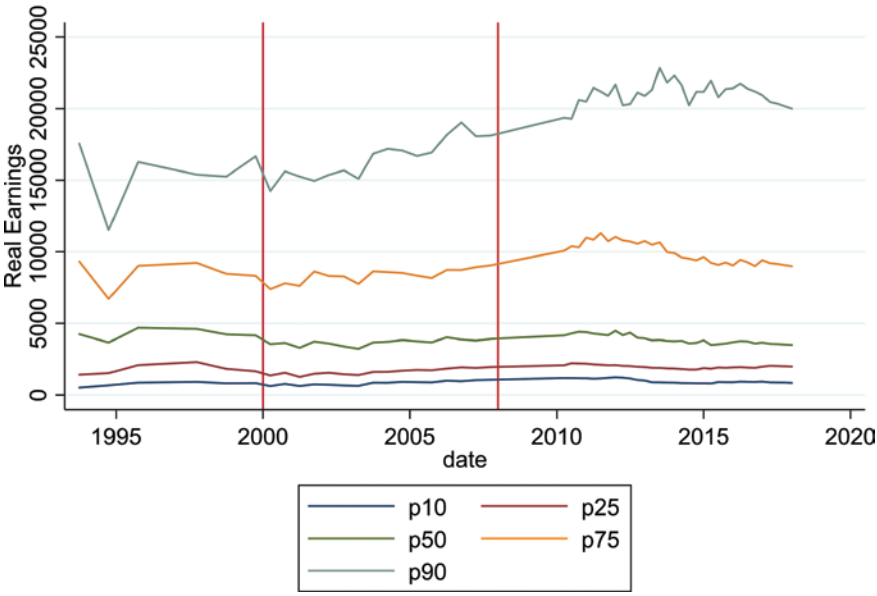

FIGURE 3 Earnings distribution in South Africa
OWN CALCULATIONS FROM PALMS V3.3. OUTLIERS ARE EXCLUDED AND BRACKET WEIGHTS ARE USED. REFUSALS ARE NOT IMPUTED. EARNINGS ARE DEFLATED TO DECEMBER 2017

TABLE 3 Changes in earnings percentiles

	1993			2017		
Percentile	**Rand value**	**CI Lower**	**CI Upper**	**Rand value**	**CI Lower**	**CI Upper**
10	519	432	649	881	819	916
25	1427	1073	2048	2000	1967	2017
50	4253	3200	5406	3582	3530	3700
75	9334	6974	12650	9078	8649	9700
90	17551	15468	19486	20350	20000	20469

Note: The values are deflated to December 2017 rands. CI upper and lower are the lower and upper limits of the 95 percent confidence interval. Confidence intervals account for clustering and weights.

across the distribution has been flat or declining since about 2012. Earnings information was collected in several different ways across the surveys in PALMS, and the harmonisation process has made the creation of the sorts of trends displayed in Figure 2 and Figure 3 much simpler.

7. Concluding Remarks

The Post-Apartheid Labour Market Series (PALMS) is a publicly available source of harmonised microdata focusing mainly on the South African labour market and has been created from 69 household surveys conducted between 1993 and 2019. All the code used to create the data is also publicly available. PALMS is South Africa specific but provides an example that could be followed by researchers in other contexts. Since such a data source is a public good, similar projects may have to be undertaken by NSOs or funded by international donors. The creation of PALMS has not just been an exercise in data production. It has led to a substantive research program that has improved the quality of the data and shed new light on several aspects of the South African labour market.

References

Branson, N., & Wittenberg, M. (2007). The measurement of employment status in South Africa using cohort analysis, 1994–2004. *South African Journal of Economics*, *75*(2), 313–326.

Branson, N., & Wittenberg, M. (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.

Kerr, A., Lam, D., & Wittenberg, M. (2019). *Post-Apartheid Labour Market Series 1993–2019* (Version 3.3) [Dataset]. DataFirst [producer and distributor].

Kerr, A., & Wittenberg, M. (2019a). *A guide to version 3.3 of the Post-Apartheid Labour Market Series*. https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/434/download/10286.

Kerr, A., & Wittenberg, M. (2019b). *Earnings and employment microdata in South Africa* (WIDER Working Paper 2019/47). https://www.wider.unu.edu/sites/default/files/Publications/Working-paper/PDF/wp-2019-47.pdf.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, *8*(1), 3–15.

Wittenberg, M. (2008). *October household survey 1994*. Mellon Data Quality Project Technical Paper.

Wittenberg, M. (2010). An introduction to maximum entropy and minimum cross-entropy estimation using Stata. *The Stata Journal*, *10*(3), 315–330.

Wittenberg, M. (2014a). *Analysis of employment, real wage, and productivity trends in South Africa since 1994* (Conditions of Work and Employment Series No. 45). International Labour Office. https://www.ilo.org/travail/whatwedo/publications/WCMS_237808/lang--en/index.htm.

Wittenberg, M. (2014b). Data issues in South Africa. In H. Bhorat, A. Hirsch, R. Kanbur & M. Ncube (Eds), *The Oxford companion to the economics of South Africa* (ch. 7, pp. 79–83). Oxford University Press.

Wittenberg, M. (2017a). Wages and wage inequality in South Africa 1994–2011: Part 1 – Wage measurement and trends. *South African Journal of Economics*, *85*(2), 279–297.

Wittenberg, M. (2017b). Wages and wage inequality in South Africa 1994–2011: Part 2 – Inequality measurement and trends. *South African Journal of Economics*, *85*(2), 298–318.