SA-TIED
Southern Africa – Towards Inclusive Economic Development

Measuring labour earnings inequality in post-apartheid South Africa

Serena Merrino

SA-TIED Working Paper #100 | March 2020

UNITED NATIONS UNIVERSITY
UNU-WIDER

national treasury
Department:
National Treasury
REPUBLIC OF SOUTH AFRICA

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI

& evaluation
Department:
Planning, Monitoring and Evaluatio
REPUBLIC OF SOUTH AFRICA

the dti
Department:
Trade and Industry
REPUBLIC OF SOUTH AFRICA

SARS

TIPS

About the programme

Southern Africa –Towards Inclusive Economic Development (SA-TIED)

SA-TIED is a unique collaboration between local and international research institutes and the government of South Africa. Its primary goal is to improve the interface between research and policy by producing cutting-edge research for inclusive growth and economic transformation in the southern African region. It is hoped that the SA-TIED programme will lead to greater institutional and individual capacities, improve database management and data analysis, and provide research outputs that assist in the formulation of evidence-based economic policy.

The collaboration is between the United Nations University World Institute for Development Economics Research (UNU-WIDER), the National Treasury of South Africa, the International Food Policy Research Institute (IFPRI), the Department of Monitoring, Planning, and Evaluation, the Department of Trade and Industry, South African Revenue Services, Trade and Industrial Policy Strategies, and other universities and institutes. It is funded by the National Treasury of South Africa, the Department of Trade and Industry of South Africa, the Delegation of the European Union to South Africa, IFPRI, and UNU-WIDER through the Institute's contributions from Finland, Sweden, and the United Kingdom to its research programme.

Corresponding author: sm147@soas.ac.uk

The views expressed in SA-TIED working papers are those of the author(s) and do not necessarily represent the views of the programme partners or its donors.

UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2020/32

Measuring labour earnings inequality in post-apartheid South Africa

Serena Merrino*

March 2020

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

Abstract: This paper investigates the validity of household survey data published by Statistics South Africa since 1993 and later integrated into the Post-Apartheid Labour Market Series (PALMS). A series of statistical adjustments are proposed, compared, and applied to primary data with the purpose of generating time-comparable, unbiased estimates, and accurate standard errors of labour earnings inequality coefficients. In particular, corrections deal with outliers and implausible data records, missing observations, bracket responses, breaks in the series, under-reporting of high incomes, and quarterly frequency. This work lays the ground for future research on the redistributive dynamics of economic policy in South Africa, which notably suffers from the presence of spurious shifts in repeated cross-sections.

Key words: income inequality, distribution, heterogeneity, survey data, imputation

JEL classification: C31, D31, O15, R20

Acknowledgements: I am grateful to Professor Laurence Harris for his mentorship, and to the South African Reserve Bank, in the persons of Dr Chris Loewald and Dr Konstantin Makrelov, for hosting my field research.

* UNU-WIDER, Helsinki, Finland, and South African Reserve Bank, Pretoria, South Africa; sm147@soas.ac.uk

This study has been prepared within the UNU-WIDER project Southern Africa – Towards Inclusive Economic Development (SA-TIED).

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9256-789-7

https://doi.org/10.35188/UNU-WIDER/2020/789-7

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 The Post-Apartheid Labour Market Series

Despite there being a rich literature examining cross-sectional inequality in South Africa, no consensus has been reached on the quality of long-run time series. In effect, multiple generations of household surveys have been produced since the end of the apartheid regime by local statistical and research agencies—first and foremost the parastatal Statistics South Africa (Stats SA)—which provide nationally representative micro-level information on the labour market.[1] Although today these resources constitute an abundant pool of information, they were not originally designed for dynamic analysis and do not allow for straightforward comparability and immediate use in longitudinal studies. In other words, the nature of the data collected differs more or less substantially in each survey wave because of differences in, for example, the sample design instrument and definitions.

As a response to rising concerns over the validity of using distributional data to undertake time-comparative exercises, the University of Cape Town's DataFirst initiated a study of successive labour market cross-sections and integrated them into a single longitudinal data set. This project produced the so-called Post-Apartheid Labour Market Series (PALMS): a stacked cross-section consisting of a harmonized compilation of four household surveys[2] conducted after 1993 and focused on socioeconomic topics (Kerr et al. 2013). Specifically, PALMS consists of:

- The 1993 Project for Statistics on Living Standards and Development (PSLSD); Southern Africa Labour and Development Research Unit (SALDRU UCT); annual.
- The 1994–99 October Household Surveys (OHS); Stats SA; annual.
- The 2000–07 Labour Force Surveys (LFS); Stats SA; biannual (March and September).
- The 2008–18 Quarterly Labour Force Surveys (QLFS); Stats SA; quarterly. QFLS earnings data are released separately in Labour Market Dynamics (LMDSA).

Notably, the major advantage related to the latest release (PALMS version 3.3) is that it exhibits a labour income variable at individual level that is consistent from 1993 to 2017.[3] This is labelled '*realearnings*' and reports monthly earnings per capita before taxes and at constant prices as for December 2015. The full description given in Kerr and Wittenberg (2019b: 16) is as follows:

> Monthly REAL earnings variable generated from the earnings amount data (not bracket information) across all waves where earnings amounts were asked and data have been released (all waves except OHS 1996 and QFLS waves 2008, 2009 and 2012). This is the earnings variable deflated to 2015 Rands using CPI.

For this reason, PALMS has generated a new strand of academic literature that explores the short- and long-term dynamics of wage inequality in post-transition South Africa, as well as a vibrant discussion on the need for higher-quality time-consistent and more frequent microeconomic data. Although PALMS yields significant improvements in the treatment of labour data in South Africa,

[1] According to Devereux (1983), until the 1980s, government censuses ignored the personal incomes of black people, which had to be calculated as a residual of national accounts. For this and other reasons, this paper refers only to the post-apartheid period.

[2] For a detailed description of primary sources available, see Kerr and Wittenberg (2019a).

[3] PALMS version 3.3 includes the 2017 LMDSA data on earnings in quarters 3 and 4.

it still preserves a number of incongruities inherited from primary sources. To date, the South African literature that assesses the sensitivity to economic policy shocks of distributional trends is almost non-existent precisely because dynamic analyses would suffer from the presence of methodological shortcomings: spurious shifts among repeated cross-sections are inevitably related to real changes in the variables of interest. It is nonetheless necessary to use available resources to identify time trends and changes such that a more granular picture can shed light beyond stylized facts.

This paper investigates the features inherent in PALMS,[4] thoroughly reviews the literature addressing issues in South African labour data, and complements earlier studies by constructing a complete and robust time series of inequality to be used for dynamic economic policy analysis. The ultimate purpose of the paper is to improve longitudinal analysis on inequality in post-apartheid South Africa by generating unbiased estimates and accurate standard errors of inequality coefficients that can be better compared over time with quarterly-frequency data. It lays the ground for a second paper analysing the impact of monetary policy on labour income inequality in South Africa.

The paper is structured as follows. Section 2 offers a selective review of the literature that makes use of South African income and earnings disaggregated data. Then, in Section 3, the data underpinning this work is carefully analysed and different methods of adjustment proposed, compared, and implemented in defiance of data quality issues. In Section 4, I discuss trends of inequality through distinct measures based on the moments of the earnings distribution. While it is not feasible to fully address all problems pertaining to primary data collection, the final remarks discuss what assumptions are needed in order to make defensible comparisons over time. The final set of complete data on household-level pre-tax wage income at constant prices, along with the Stata code that was applied to the raw data, is available from the author on request.

2 Labour income in post-apartheid South Africa: a literature review

A number of attempts to quantify inequality dynamics since the advent of democracy in South Africa explore the quality of surveys and censuses available in the country and eventually comment on the comparability of relevant variables over time. Cichello et al. (2005) compare 1993 and 1998 earnings in the KwaZulu Natal Income Dynamics Study and reach different results when using the data as a panel and as a cross-section. Using the cross-sectional data by overlooking specific workers' dynamics shows that formal sector workers were better off in 1998. By contrast, the panel data indicate that workers who were already employed in the formal sector in 1993 experienced a fall in earnings, while informal workers started at a much lower average earnings point but experienced a rise due to mobility towards formal employment. Casale et al. (2004) use only the OHS 1995 and the LFS 2001:2 to analyse the position of women and ethnic groups in the labour market. In that paper, the authors make no data transformation and assert that 'these are the years in which the earnings data are most comparable' (Casale et al. 2004: 6). Despite data concerns, they observe that both mean and median earnings declined over the period. Burger and Yu (2007) compare the OHS and the LFS from 1995 to 2005 by excluding the outliers, the self-employed, and informal workers. They find that average earnings started to increase and their distribution to improve after 1998. Following Casale et al. (2004), their figures confirm no improvements in the relative earnings position of women, non-white population groups, or unskilled and semi-skilled

[4] The relatively long span of data necessary to implement this analysis precludes the use of administrative data recently released by the South African Revenue Service (SARS), which starts in 2011.

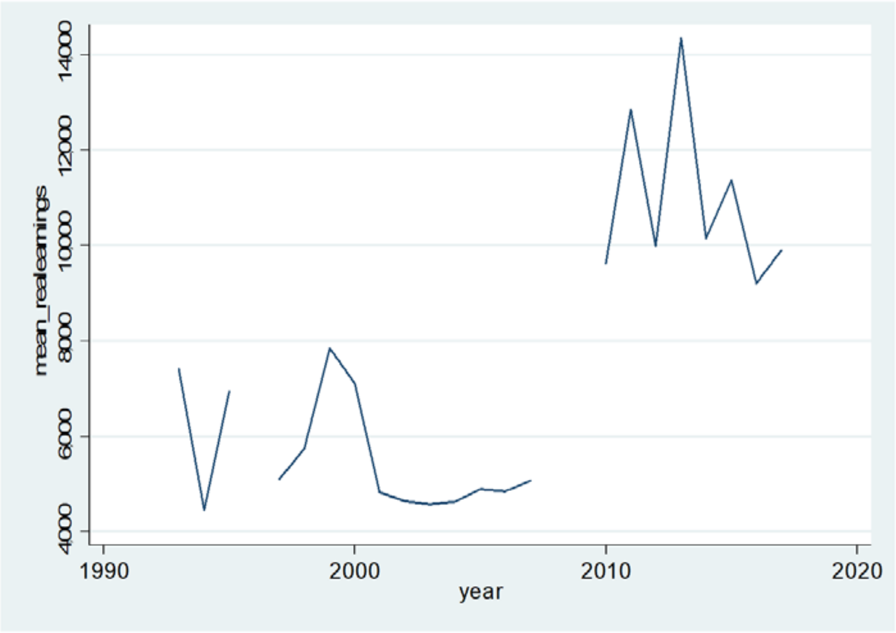
workers, but they show 'signs that there has been an decrease in between-group inequality in more recent years'. Bhorat et al. (2009) utilize the 1995 Income and Expenditure Survey (IES) and the 2005/06 Income and Expenditure Survey, looking at total income, and report increasing inequality over the period, from an income Gini coefficient of 0.64 in 1995 to 0.72 in 2005. Leibbrandt et al. (2010) include all forms of labour earnings from three comparable national household survey data sets: the PSLSD for 1993, the LFS and IES for 2000, and the National Income Dynamics Study (NIDS) for 2008. With no adjustment, they calculate the income Gini coefficient in South Africa and report that it rose from 0.66 in 1993 to 0.68 in 2000 and further to 0.70 in 2008. Finn et al. (2016) use the first four waves of NIDS from 2008 to 2014 and the 1993 PSLSD to investigate the shape of the association between parental and child earnings across the distribution.

While all previously mentioned authors rely on a few points in time, the most comprehensive study on long-run trends in labour income inequality in South Africa can be found in the work of University of Cape Town's Martin Wittenberg, which indeed serves as the basis for this discussion. Wittenberg and Pirouz (2013) use PALMSv2 to show the impact of different types of data quality adjustments (specifically they treat outliers, zero earnings, bracket responses, and missing observations) on the estimation of the average wage over the period 1994–2011. As already observed by Casale et al. (2004) and Leibbrandt et al. (2010), Wittenberg and Pirouz (2013) also evidence how the change in coverage between the OHSs and the successive LFSs generated a gap in the earnings series at the year 2000. Wittenberg and Pirouz conclude by arguing that it is possible to identify some real wage growth since 2000 despite the noise generated by these measurement changes. Wittenberg (2014b) builds on the previous paper to compare PALMS to firm-level data—namely the Survey of Employment and Earnings (SEE) and the Quarterly Employment Statistics (QES) surveys. He adds that the top tail of the earnings distribution has received larger gains than the 75th percentile; that both of them show significant real earnings growth; that the 10th percentile made real gains relative to the median, therefore experiencing a compression; and that among the self-employed there is no evidence for systematic shifts in the distribution over the post-apartheid period. Wittenberg (2017c) effects further adjustments to yield PALMSv2.1 and calculates wage inequality through the Gini coefficient. He argues that despite some noise in the estimates, the measurements made after the LFS 2007:1 are noticeably higher than those made from 2000 to 2006. Finn (2015) calculates the Gini wage inequality in PALMS using the same data-cleaning procedure suggested by Wittenberg (2014b): in contrast to Leibbrandt et al. (2010), who calculated overall income inequality, the Gini coefficient of real wages in 2003:1 (0.553) was almost identical in 2012:1 (0.554). By contrast, using the LFSs, Vermaak (2012) finds no trend that is robust to alternative coarse data adjustments—particularly the treatment of zero values and the choice of imputation methods.

3 Working with PALMS

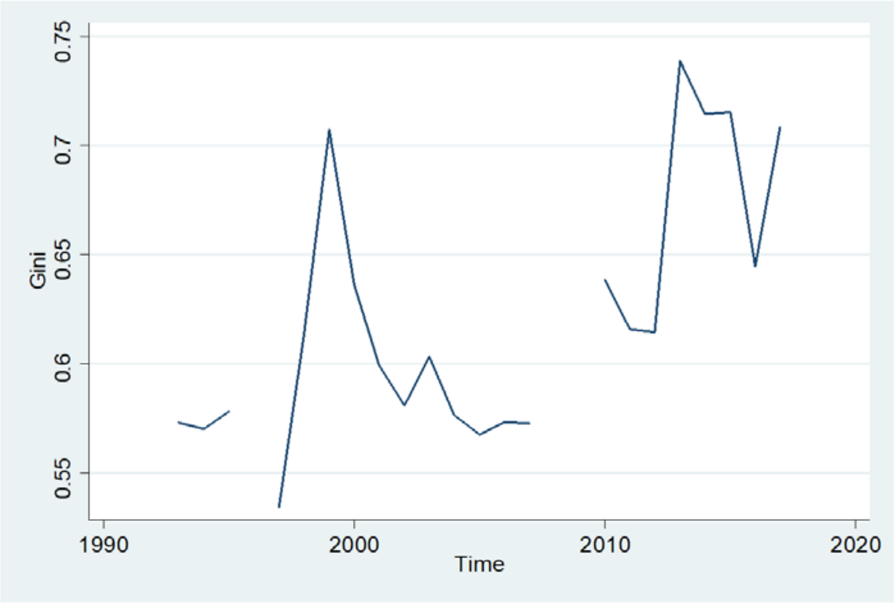
In PALMS, the variable reporting real earnings with no adjustment returns a mean of ZAR8,784 per month and a median of ZAR3,225. The number of observations, $N_{observed}$, in the original file is 963,492; this is higher than in any of the other approaches because every possible earner is included. However, in the original file more than 5 million real earnings observations are missing, including all individuals in years 1996, 2008, 2009, and 2018 and the first two quarters of 2019. Table A1 in the Appendix summarizes the main features of real earnings in PALMSv3.3 before any adjustment. It can be observed that for each wave the coefficient of variation of the random variable (standard deviation/mean) is significantly higher than 1: the high variance is due to the log-normal distribution of real earnings that is not centred on the mean and is positively skewed with long right tails.

Figure 1: Evolution of monthly average real earnings in PALMSv3.3 (no adjustment)



Source: author’s illustration based on PALMSv3.3.

Figure 2: Evolution of the Gini index in PALMSv3.3 (no adjustment)



Source: author’s illustration based on PALMSv3.3.

Plotting raw data against time also evidences the presence of issues. Figure 1 displays a clear trend of average real earnings and a puzzling volatility, with suspicious falls in 1994 and after 2000 and rises in 2012, among other things. It is evident that evolution over time is very similar in the two series shown in Figures 1 and 2: the Pearson coefficient of correlation is 0.76 and is statistically significant at 1 per cent. At first sight, it seems that when labour earnings increase, on average their distribution across individuals worsens (see Figure 2).

3.1 The benchmark data set

The first issue encountered in exploring the statistical properties of PALMSv3.3 is that unrealistic values are found with respect to the age category, with 708 respondents supposedly reporting as more than 100 years old (one individual is recorded as being 142 years old). For this reason, the sample is restricted to those typically assumed to be in the labour force—that is, to respondents in the age group 18–65 (see also Finn and Leibbrandt 2018).

Secondly, given that analysis is restricted to labour income, the unemployed, who can be assumed to receive zero earnings, are also excluded. This implies that the distribution may worsen when employment is created, if the newly employed receive lower-than-average earnings.

Lastly, PALMSv3.3 distinguishes between wage-earning employees and the self-employed through the dummy variable '*employerAll*'.[5] The latter was created to harmonize changes in the earnings question: respondents could report several jobs in the PSLSD, a maximum of two jobs in the OHSs, but only one main job afterwards. In the LFS, Wittenberg and Pirouz (2013: 6) note the impossibility of identifying 'those working for themselves (employers/self-employed) and those working for others'. Instead, they evidence the presence of a prior question which prevents individuals from reporting both types of income in the QLFS (Wittenberg and Pirouz 2013: 9). Because of the shift in recorded self-employment, Wittenberg (2017c) accounts for inequality across wage-earners only. By accounting only for waged employees, the mean is significantly decreased (from ZAR8,735 to ZAR7,260) given that the self-employed person earns more on average (ZAR17,868). At the same time, the median wage for public sector workers is significantly higher than the median pay of a self-employed person, especially if one considers unionized public employees only (Kerr and Teal 2012: 7). As a result of these considerations, excluding the self-employed is likely to decrease the inequality estimates.

The benchmark data set is therefore a constrained version of PALMSv3.3 (N = 1,250,645, of which $N_{observed}$ = 436,347) that accounts only for 18–65-year-old workers receiving income from a salary or wage.

3.2 Outliers

In PALMSv3.3 without adjustment, 164 individuals report real earnings of more than 1 million rand, and relatively higher numbers of millionaires are observed in the OHS 1999, LFS 2002:2, and QLFS from 2012 Q3 to 2013 Q4, 2015 Q3, 2015 Q4, 2017 Q1, and 2017 Q2 (see Table A1). The number of millionaires decreases to 28 after narrowing the scope down to wage employees, indicating that most millionaires were recorded among the self-employed category. Among them, there are some implausibly high values that could result from coding errors. For instance, in the OHS 1999 a 25-year-old woman from the Limpopo province who reports nine years of schooling (the basic education system consists of 12 grades) is coded as earning over 1 million rand a month while employed in private domestic services. However, extreme values exist in both tails of the real earnings distribution, not only at the top end. An example of a low outlier in the data is someone who reported working 48 hours a week in the formal sector and yet reported a monthly wage of ZAR4.80.

[5] After Neyens and Wittenberg (2016) showed that the self-employed agricultural worker series in PALMS was extremely inconsistent over time, PALMS' authors included the variable '*employedpreferred*', which is a dummy taking a value equal to 1 if the individual is not self-employed in the agricultural sector.

The key problem associated with suspect observations is that they exert leverage on the trend and the moments of the distribution, despite being unrepresentative of the true data-generating process. Burger and Yu (2007: 4) find that 'especially the 1999 OHS and the September 2000 LFS shows evidence of high earning outliers', and they arbitrarily take millionaires out of the data set according to different thresholds to reconcile preceding and subsequent surveys. Excluding the 164 millionaires from the distribution decreases the mean to ZAR7,502 while the median remains almost unchanged, signalling that the millionaires represent a negligible albeit influential number of the total sample. This simple procedure potentially underestimates inequality by removing both genuine and spurious top earners while ignoring the lower tail of the distribution. Therefore, I follow Wittenberg (2017b) and compare three procedures that distinctly detect contaminating observations: the BACON algorithm (Billor et al. 2000), a robust regression through iteratively reweighted least squares, and a studentized residuals approach.

In the latter case, real earnings in logs are linearly regressed over gender, race, a quadratic in age, education, and occupation levels. Then, each *i*th residual ($\hat{e}_i$) of the regression is normalized against a standard deviation that is calculated from a sample in which that observation is left out ($\hat{\sigma}_{-i}$). As a result, studentized residuals (r_i) follow a standard normal distribution and can identify extreme points of high leverage, corresponding to studentized residuals with an absolute value greater than 5:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}_{-i}} > |5|$$

This approach is the one suggested by the PALMS authors, given that it excludes a relatively moderate number of observations, but also by Finn (2015) on the LMDSA 2014 data set and Finn and Leibbrandt (2018). In PALMSv3.3, 1,374 real earnings observations are flagged[6] through the 'outlier' indicator, set as missing, and then multiply imputed in PALMSv3.3miincomes.[7] This approach leaves 30 millionaires.

Replicating this procedure returns 1,299 outliers, most of which correspond to those already flagged in PALMSv3.3, and none of which has a zero value. After removal of outliers, the scatterplot of studentized residuals shows no presence of extreme values; mean earnings are now ZAR7,035 and the median is ZAR3,293, and the final number of millionaires is two out of a total 435,048 observations. However, Section 3.7 shows the procedure adopted for multiple imputation of flagged outliers. All ten imputations of the 1,299 observations flagged as outliers return a mean of ZAR11,761 and a median of ZAR4,783. These statistics confirm that this pool of respondents obtains a substantially higher and less dispersed wage than the rest of the sample.

3.3 Zero-earners

Zero-earners are workers who report null labour income, for various possible reasons: (i) the respondent earns a positive income but is lying; (ii) zero surplus at the end of the period is equated with zero income; (iii) the individual is receiving not monetary pay but experience, income in kind.[8]

[6] Kerr and Wittenberg (2019a) report 476 flagged outliers, but using an old version of the data (Kerr: personal communication).

[7] PALMSv3.3miincomes is an auxiliary data set produced by DataFirst that consists of multiple imputations of certain observations flagged in PALMSv3.3 (see Section 3.7).

[8] In-kind income is non-cash payment received in exchange for services rendered. It can come in the form of provisions such as free rent or free meals, or in the form of property or an exchange of service.

Based on the LFS, Vermaak (2012) shows the extent to which zero incomes affect the analysis of earnings.

According to Wittenberg and Pirouz (2013), zero-earners represent a problem only among the LFS' self-employed due to the simplification of the instrument and increased coverage of informal subsistence workers.[9] On these grounds, Wittenberg (2017b) excludes the self-employed and does not correct for the remaining zero-earners.

In PALMSv3.3 there are 3,417 zero earnings observations that are not imputed (3,346 when restricting consideration to the benchmark data set). Most of these zero values are concentrated in the OHS 1994 and the QLFSs, and among the wage-employed.

A second solution would be to exclude all individuals who report zero earnings. This is perhaps the most common approach used by researchers working with household survey data in South Africa. For example, Vermaak (2012) exploits the variety of information provided by zero-earners: she treats them as missing values and then imputes the monetary wage that these individuals would receive under the same data-generating process.

A third option is to remove and then impute only implausible zero earnings, while the rest are maintained as true information. Selection is undertaken by excluding a series of categories, supposedly subject to significant errors, from zero-earners:

1. Excluding all zero-earning self-employed brings zero values from 3,346 to 2,969;
2. Excluding zero-earning public employees brings zero values to 2,341;
3. Excluding zero-earners with a written contract brings zero values to 1,347 (1,139 in OHS 1994);
4. Excluding zero-earners with unemployment insurance fund (UIF) scheme sponsored by the employer brings zero values to 1,331;
5. Excluding full-time zero-earners (employed for more than 30 hours in the previous week) brings zero values to 174 (82 in OHS 1994);
6. Excluding zero-earners who went to university brings zero values to 148 (60 in OHS 1994);
7. Excluding zero-earners at highest professional levels brings zero values to 145 (58 in OHS 1994).

The remaining 145 zero earnings[10] are considered plausible values, i.e. the individual performed only unpaid tasks such as working in a household business or in subsistence agriculture. So, leaving the true value of zero monetary compensation is considered the best way to take into account the structure of the South African labour market. On the other hand, there are 2,824 zero earnings that are flagged as implausible and imputed (see Section 3.7): imputed zero monetary earnings are only slightly lower than observed values, indicating that workers with such characteristics would not earn no monthly wages if they worked in paid employment: in other words, they are implausible records.

[9] Since the LFS questionnaire does not prompt respondents to include the value of other non-monetary benefits received as a result of their employment such as in-kind income.

[10] 71 female and 74 male; 105 African/black, 18 coloured, 5 Indian, 17 white (of which 4 are male); 66 are in elementary or domestic occupations, 9 are machine operators, 26 are in crafts, 28 in services or sales, and 11 are clerks.

3.4 Sample weights

An important issue to be addressed if one is to obtain a consistent set of estimates over time is the harmonization of the sample weights released with each survey. The purpose of weights is to inflate the sample to represent the entire population at each point in time. While sample weights are usually designed as inverse inclusion probability, Stats SA implements instead a post-stratification adjustment based on auxiliary population totals to reflect race, gender, and age group distribution. Due to the cross-sectional nature of the data, post-stratification weighting corrects sampling errors (i.e. non-response rates and out-of-date sampling frame) given the external information available at the particular year in question. However, as the demographic information contained in new censuses changes, later surveys are calibrated to previous aggregates that have become obsolete (Branson and Wittenberg 2014; Casale et al. 2004).

Along these lines of thought, Branson (2010) first proposed a cross-entropy estimation approach to create a new set of individual weights—to be common within households—which inflates the sample to a time-consistent external total, while maintaining the post-stratification sampling correction applied by Stats SA. She based the cross-entropy weighting on the 2003 demographic model from the Actuarial Society of South Africa (ASSA)—see Branson and Wittenberg (2014) for a detailed explanation. In order to account for higher survival rates and a growing population, PALMSv3.3 updates cross-entropy weights using Stats SA population estimates for 2019. Since the earliest observation available corresponds to the year 2003, PALMSv3.3 complements the 2019 aggregate figures with the ASSA 2008 growth rates over the period 1993–2002, and then applies a simple exponential growth model to the new 2003 cross-entropy weight back to 1993 (Kerr and Wittenberg 2019a). In PALMSv3.3, the variable '*ceweight1*' is the recommended weight to use in conjunction with *realearnings*.

3.5 Bracket responses

Given that individuals may be reluctant to disclose the exact rand amount that they earn, in many surveys it is customary to offer respondents the option of providing their income information in bands (Juster and Smith 1997). These data are widely known as 'coarsened' data because econometricians are not always certain how to combine the information in the categorical variable with that given as point values.

A common, albeit rudimentary, practice is to impute the categorical information by placing all earnings within a certain bracket at the same point, such as the midpoint of the interval. In the case of the highest 'open' category, where there is no explicit midpoint, the procedure is to place observations at some deterministic multiple of the lower bracket boundary, because the distribution in the upper tail is approximately Pareto (von Fintel 2007). However, these practices will lead to artificial spikes in the earnings density distribution and will distort moments other than the mean. For example, when the midpoint method was followed, the large lowest category in OHS 1995 allowed too much weight to the upper portion of that bracket and overestimation of the mean.

Recently, several authors have examined the nature of interval responses and the sensitivity of the resulting earnings distribution to different methods of approximating the distribution within intervals. Von Fintel (2007) points to the importance of including categorical respondents in the analysis, and so he shows that coefficient differences between midpoint imputation and interval regression are virtually negligible. Ardington et al. (2005) claim that bracket incomes are usually higher than those that give point values and show that inequality levels are generally underestimated as a result of collecting income information in bands, although fortunately not by much. Posel and Casale (2005) make the case that people who respond in brackets do so because of privacy concerns or because they do not know 'exactly' how much they earn. Wittenberg (2017b: 284) deals with top-income under-reporting in QLFS 2011 and suggests that many people who earn high incomes may opt not to disclose information to the survey organization even if they agreed to be surveyed. He shows that the provision of a single value decreases almost monotonically with income, with the exception of the highest open category, where bracket responses constitute a tiny proportion.

This scenario suggests that simply omitting bracket responses would incorrectly overlook responses that overwhelmingly come from the top end of the distribution (relatively few individuals are included in the open category). Wittenberg (2008) recommends that the analyst adjust the lower point response rate of higher incomes such that individuals giving point values within a particular bracket are weighted by the inverse of the probability of giving a point value response:

$$w_i = 1/p_i$$

where p_i is the share of point responses in the bracket corresponding to individual i.

This strategy will outperform either imputation whenever the distribution of the variable in question is markedly different from the distributional assumptions implicit in the imputation strategy (earnings follow a log-normal distribution). Numerically, it will provide identical results to the imputation of means without altering any of the other moments, but it tends to inflate the standard errors.

PALMS' *realearnings* variable does not include bracket information, which is instead registered as missing values. Kerr and Wittenberg (2019a) use the *realearnings* variable in conjunction with '*bracketweight'* in order to take appropriate account of missing bracket responses. The variable *bracketweight* is the product of w_i (the inverse inclusion probability of a point value response in a particular bracket in a particular wave) and p_h(the cross-entropy weight for that particular individual created from the Stats SA 2019 demographic model). The *bracketweight* variable equals (i) the cross-entropy (CE) weight in a case where no bracket responses were given in a certain wave, as in all QLFSs, (ii) a value higher than the CE weight in the presence of bracket responses, or (iii) a missing value corresponding to missing earnings values. Thus, applying the bracket weight provided by PALMS recalibrates observations so as to represent the country's demographics (through the CE approach) and the bias implicit in the presence of bracket responses (inverse probability of giving a point value).

3.6 OHS 1996: a special case of bracket responses

The 1996 OHS (wave 3 in PALMSv3.3) was a much smaller survey (around 15,000 respondents),[11] given that it ran immediately after the census, and it displayed a much simpler instrument than usual since it captured no earnings amounts but only brackets. As a result, it records information in brackets only and the reweighting approach—which was based on the inverse inclusion probability, as illustrated in Section 3.5—is thus not applicable. In order to get around this issue, Wittenberg (2017b) proposes a variation of the two-sample two-stage least squares (TSTSLS) approach first introduced by Klevmarken (1982) and sometimes thought of as a 'cold deck' linear regression imputation, because an auxiliary sample is used to impute the missing variable in the

[11] In the original version of PALMSv3.3, wave 3 consists of 72,890 individuals, including non-respondents (see Table A1). This number decreases to 12,937 individual responses if one takes into account wage employees between 18 and 65 years old only.

main sample via the estimated conditional density.[12] The first step is to estimate a conventional Mincerian regression with individual *realearnings* observed in 1997 in logs as dependent variable ($y_{i,1997}$), and with the vector of individual characteristics $Z_{i,1997}$ to be used to control for unobserved time-invariant agent heterogeneity, which consists of: (i) gender, (ii) education, and (iii) a dummy for self-employment. In other words, labour income of individual i in 1997 ($Y_{i\ 1997}$) can be written as:

$$y_{i,1997} = \gamma Z_{i,1997} + v_i + u_{i,1997}$$

where v_i is the time-invariant error which is uncorrelated from the set of characteristics.

At this point, real earnings values in OHS 1996 are predicted through a random linear regression by using the estimated coefficients from the first-stage wage regressions in the 1997 data assumed to be constant over time ($\hat{\gamma}$), along with sample characteristics from the 1996 data that constitute the instrumental variables ($Z_{i,1996}$):

$$Y_{i\ 1996} = \hat{\gamma} Z_{i\ 1996} + \hat{v}_i + u_{i\ 1996}$$

This prediction is carried out through predictive mean matching (PMM) such that the imputed values fall within brackets (using the '*by*' and '*in'* qualifiers of '*mi impute pmm*'). The derivations presented above rely on the assumption that the main and auxiliary data sets are random samples from the same population and that the instrumental variables are independent and identically distributed across the two data sets.

Since *realearnings* is a continuous variable that is not normal standard, a linear regression may not return a distribution of predicted values that matches the observed values very well. Instead, being a semi-parametric procedure, PMM produces a distribution of imputed values that matches the observed non-normal distribution more closely. More specifically, PMM uses the linear prediction as a distance measure to form the set of nearest neighbours (possible donors) consisting of the complete values. Stata command *knn(#)* specifies the number of closest observations (*k* nearest neighbours) from which to draw the imputed value. In the above, Kerr et al. (2019) used five nearest neighbours to draw from.[13] The closest observation is the observation with the smallest absolute difference between the linear prediction for the missing value and that for the complete values. The process of PMM imputation is repeated *m* times to obtain *m* imputed data sets to be eventually analysed as though they were complete (Rubin 1987). In essence each realization of the stochastic process used in the imputation produces a different view of what the 'true' data might have been. Multiple imputation (MI) for missing data consists in imputing the missing values by using an appropriate model which incorporates unobserved variation ($u_{i\ 1997}$) drawn at random from a posterior predictive distribution that is conditioned on the observed values (in this case only Y_{1997}). This technique effectively adds some noise within the category without generating new data. In fact the two-stage linear imputation preserves relationships among variables involved in the imputation model but not variability around predicted values. It is important to keep in mind that the goal of MI is not to recover or replace the missing values; rather, it is to produce valid analytic results in the presence of missing data. Given the fact that we use a linear prediction, the

[12] Kerr and Teal (2012) and Finn et al. (2016) use the same methodology. Another way is to multiply draw the imputed value from some pre-specified random distribution (e.g. uniform or log-normal).

[13] Schenker and Taylor (1996) did simulations with three and ten *k*, finding small differences in performance, although with *k* = 3 there was less bias and more sampling variation. Based on their simulations, Morris et al. (2014) recommended *k* = 10 with large samples.

means of predicted log earnings ($\hat{Y}_{i\,1996}$) and actual auxiliary log earnings ($Y_{i\,1997}$) are identical while the most obvious difference between predicted and actual log earnings is the variance. The deflated figures are matched so that inflation is controlled for to some extent. The ten versions of these imputations for real labour incomes are contained in the '*imputed_real_v1'* to '*imputed_real_v10'* variables. Each round of imputations produces individual values of matrix $\hat{\mathrm{Y}}$ which contain no missing values and from which it is then possible to carry out further analysis, such as measuring inequality.

3.7 Multiple imputations over missing observations

Ardington et al. (2005) were the first to implement multiple imputations for missing income values in the South African data. They find that sequential multiple imputation methods produce higher income estimates than observed on average, because survey non-response rate was higher for white households in urban areas consistent with the view that survey response declines as income rises (Posel and Casale 2005). Vermaak (2012: 250) shows that the presence of zero earnings and the presence of missing earnings are likely to offset each other in the income distribution (although mean increases).

In PALMSv3.3, there are almost 5 million missing observations on real earnings (see Table A1); in the benchmark data set that maintains the wage-employed in the 18–65 age group exclusively, the number of missing observations is around 20 per cent of the observations of the *realearnings* variable (255,638 out of 1,217,213 in total). DataFirst provides another data set called PALMSv3.3miincomes, which multiply imputes (ten times) rand amounts for the following observations:

(i) the missing values of wage incomes in all waves, except waves in years 2008, 2009, 2018, and 2019;
(ii) 1,374 outliers flagged by Kerr and Wittenberg (2019a);
(iii) the OHS 1996 bracket responses.

Unlike imputation of (iii) in the previous section, the procedure for filling missing real earnings values in (i), as described in Wittenberg (2017b), requires a preliminary passage, which is: for each wave of the data set, an ordered logit model—with province, gender, education, race, a quadratic in age, and occupation as explanatory variables—is used to impute the brackets. The predicted brackets are then (along with covariates gender and education) used as independent variable in the linear regression to multiply impute rand amounts using PMM, exactly as in the second stage of the OHS 1996 imputation.

PALMSv3.3miincomes maintains some of the issues of PALMSv3.3, such as implausibly old workers and different extreme values, such that we cannot always make use of the multiple imputations already existing in PALMSv3.3miincomes. Given the non-replicability of MI,[14] I maintain the ten imputations of items (i) and (iii) already included in PALMSv3.3miincomes, but accounting for wage employees in the working-age group only. Then, given that outliers and implausible zero earnings flagged in Section 3.2 and 3.3 are different from PALMSv3.3, new imputations are run.

[14] MI estimates can be non-replicable in the sense that the estimates one person reports from a sample of *m* imputed data sets can differ substantially from the estimates that someone else would get if they re-imputed the data and obtained a different sample of *M* imputed data sets.

3.8 Breaks in the series

The South African labour income series is bedevilled by breaks. First, the earnings question was removed from the QLFS introduced in 2008, after the International Monetary Fund (IMF) delegation that assessed labour market statistics considered earnings data to be of poor quality, especially for the self-employed (Statistics South Africa 2019a: 2008, section 2.3.5, 7–8). The earnings question did reappear in late 2009, but data was released only from 2010, in the separate LMDSA (Wittenberg 2017b). Besides, even though the most recent version of PALMS contains surveys up to QLFS 2019 Q2, earnings data were published only up until 2017.

In order to fill the gaps in the stacked cross-section earnings observations in waves 23–30, 63–68 will be imputed using coefficient estimates extracted from auxiliary waves, following the two-stage procedure illustrated in Section 3.6 for the OHS 1996 series.[15] The proportion of missing values in each imputation wave does not exceed the 30 per cent of total observations.

Exploring the imputed values

One possible way of assessing the quality of the imputation is to compare real earnings in the auxiliary survey with earnings that have been predicted. Imputed values are considered reasonable only if the conditional distribution converges towards the observed distribution. A useful initial check can be done using graphical displays of the ten versions of the imputed data using:

(i) kernel density plot of the observed and imputed real earnings;
(ii) histogram of the observed and imputed series;
(iii) plot of the quantiles of the imputed series against quantiles of the observed series (quantile–quantile plot);
(iv) cumulative distribution plots of the observed and the imputed series.

The imputed data is also checked numerically by generating descriptive statistics. In particular, for any individual, the imputed values may differ from round to round and hence the differences across the ten imputations performed capture the standard error in the estimate. Reporting both graphical and numerical checks for ten imputations and all waves is unfeasible—the .do file to replicate these steps remains available from the author on request.

3.9 Under-reporting

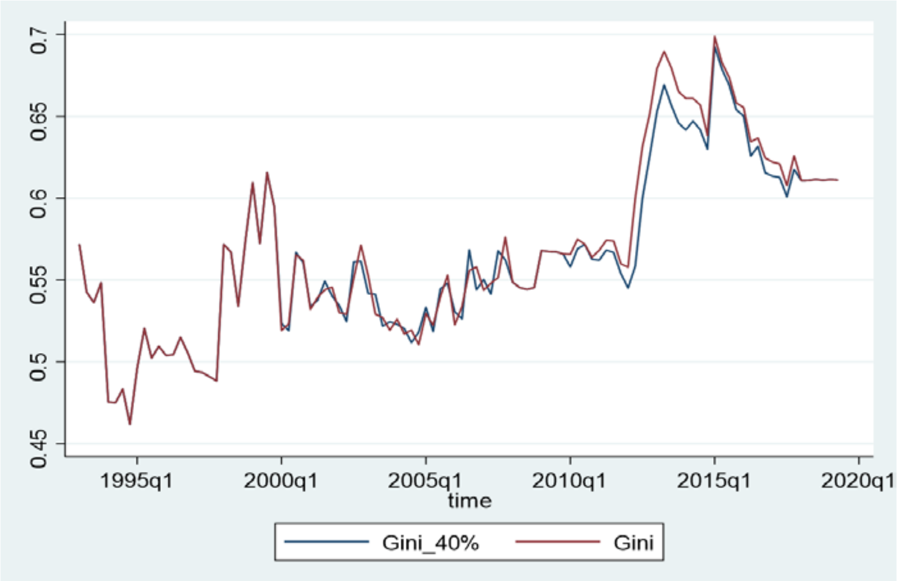
A number of studies compare the QLFS earnings data against other sources, particularly administrative data released by the South African Revenue Service since 2011, and suggest that it under-reports high incomes (Bassier and Woolard 2018; Seekings 2007; van der Berg et al. 2007; Wittenberg 2014a, 2017a). This mismatch can happen for many reasons: (i) the QLFS employs a nationally representative sample, and thus it includes bottom incomes and under-employed that the tax agencies usually overlook; (ii) the QLFS sample does not include a sufficient number of top earners because they make up a very small proportion of the population and are often reluctant to reveal their income information, 'while in the case of the tax authorities (SARS), they have much less choice since there are large penalties for non-cooperation or non-disclosure' (Wittenberg 2017a: 2); and (iii) the QLFS questionnaire only asks for income from one main employment, whereas the SARS data typically cover all sources of income (i.e., second jobs, benefits, and annual

[15] LFSs 2007 are used to impute all quarters of 2008 (all QLFSs 2008); the four quarters of QLFS 2010 are used to impute all quarters in 2009; the four quarters of QLFS 2017 are used to impute the four quarters of 2018 and the first two of 2019.

bonuses). Furthermore, when comparing the wage figures in the QLFS and the SARS data set, Wittenberg (2017b) notes that the gap is relatively uniform, at around 40 per cent, across different deciles. On these grounds, there exist two attempts to close the mean and the median earnings discrepancy: both Wittenberg (2014a) and Finn (2015) inflate the data to match firm-level information in the Quarterly Employment Survey (QES). These considerations necessarily imply that the estimate of the Gini coefficient through PALMS in the years 2000–19 will be lower than actually observed, yet higher than estimated through alternative data sources that completely ignore the lower deciles.

Although it could only be quantified once alternative data became publicly available, the problem of under-reporting earnings is inherent to the LFS waves too. It is widely acknowledged that between the last OHS (October 1999) and the first LFS (February 2000), there was an increase in coverage of marginal workers, and a consequent decline in earnings (Kerr and Wittenberg 2019a). Following previous instances, Figure 3 exhibits the evolution of the Gini index of wage inequality with and without 40 per cent inflation adjustment to the LFS and QLFS earnings data. As expected, the Gini coefficient goes up by a few percentage points.

Figure 3: Evolution of the Gini index accounting for under-reporting (1993 Q1 to 2019 Q2)



Source: author’s illustration based on PALMSv3.3 after adjustment.

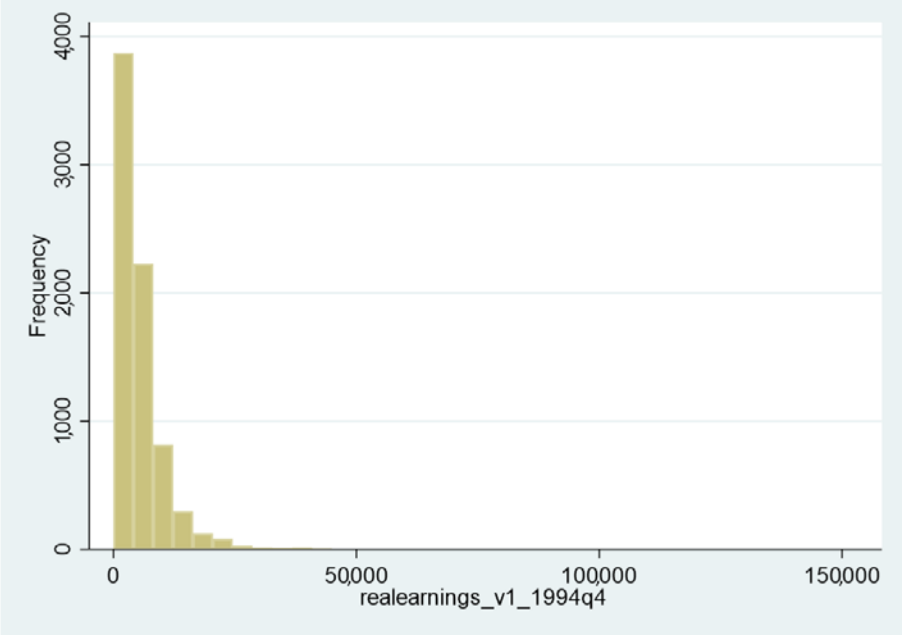
3.10 Quarter frequency (1993–2007)

Given the relationship of this paper to subsequent research on the relationship between monetary policy and income inequality, and considering the short timeframe in which monetary policy shocks propagate through the economy, it is necessary to derive sub-annual frequencies (i.e. quarterly) from annual or biannual surveys. Following Coibion et al. (2017), respondents are randomly assigned to four groups in annual waves (1993–99) or two groups in biannual waves (2000–08). Each group will then represent a quarter of the year. This approach is both elementary and simplistic, given that it negates any real shift among quarters.

4 Measuring wage inequality

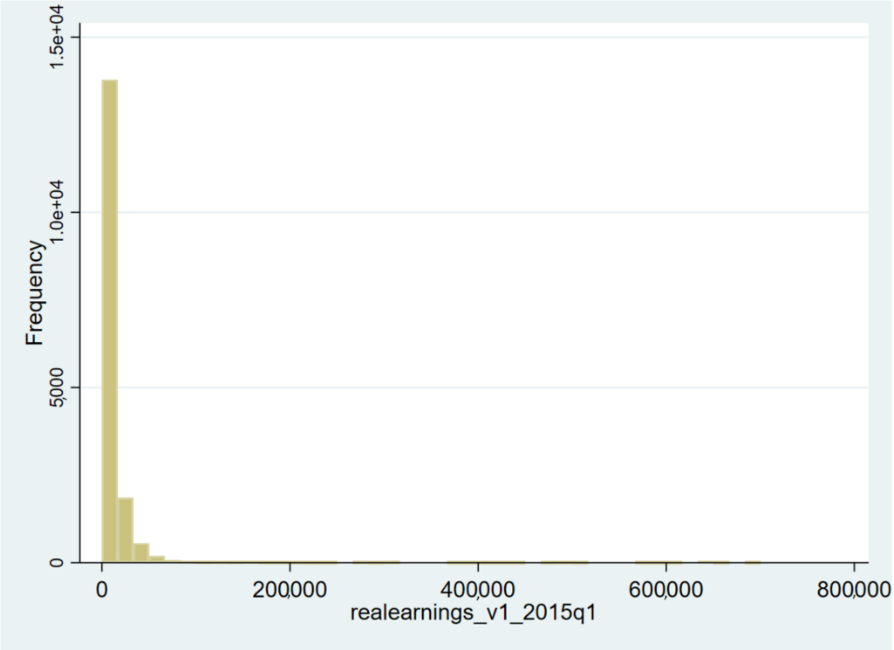
The final step of multiple imputations for missing data is to perform the desired analysis on each *m*th complete data set, then combine the results of the *m* analyses from every round, and finally average over the *m* estimates to obtain a point value with associated standard errors. Figures 4 and 5 show the frequency distribution of real wages across workers with their respective moments, in two distinct points in time: real wages are more evenly distributed in last quarter of 1994 than in the first quarter of 2015. This is confirmed by the Gini index plotted in Figure 6a.

Figure 4: Real monthly wage frequency distribution (1994 Q4)



Source: author’s illustration based on PALMSv3.3 after adjustment.

Figure 5: Real monthly wage frequency distribution (2015 Q1)



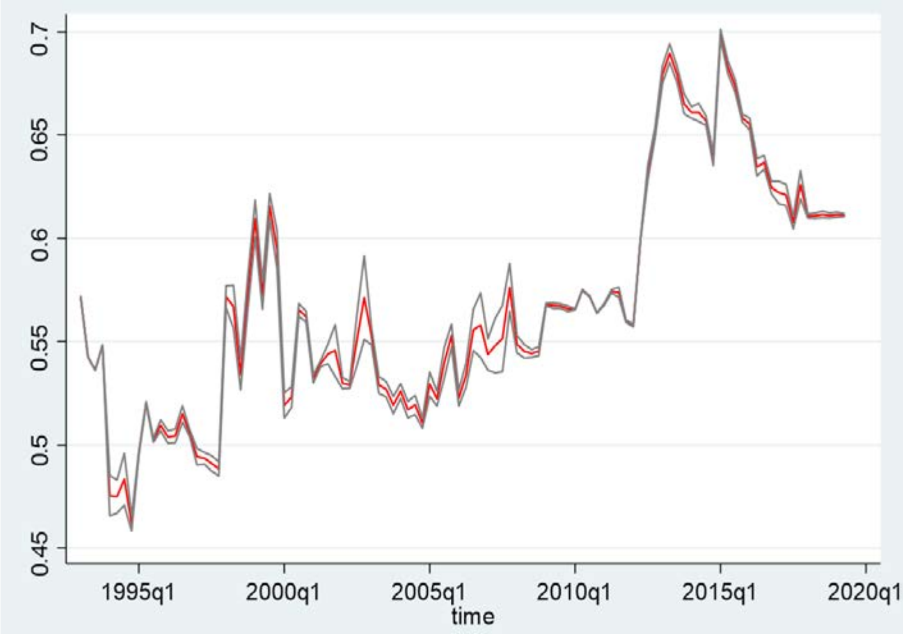
Source: author’s illustration based on PALMSv3.3 after adjustment.

In what follows I shall describe the changes in inequality over time and across multiple measures. These figures are based on individual monthly wage income (excluding the self-employed and the unemployed) at gross level (pre-tax).

4.1 The Gini index

The Gini index, the most widely cited measure of inequality, measures the extent to which the Lorenz curve for the actual wage distribution across employees deviates from a perfectly equal distribution (45-degree line). More precisely, it is computed as the ratio of the area between the two curves and so it lies in the interval between 0 (perfect equality) and 1 (perfect inequality). The Gini index satisfies most of the criteria that make a measure of inequality robust. It is independent were all incomes to be multiplied by the same number or were the population size to change. It is independent of income exchanges across workers, but it is sensitive to income transfers from one tail of the distribution to the other. Figures 6a and 6b show the evolution of the Gini index as a measure of wage inequality on an individual and a household basis, respectively.

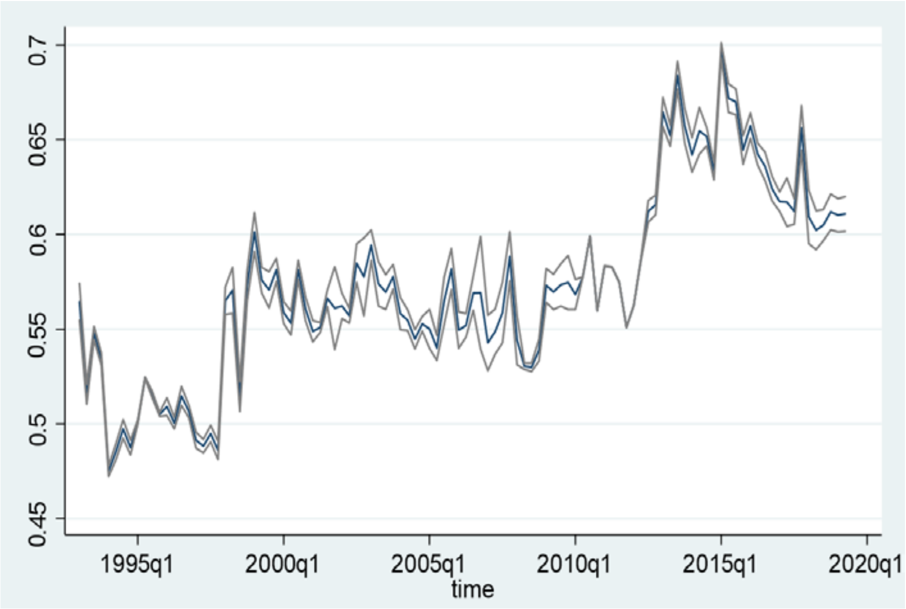
Figure 6a: Evolution of the Gini index, individual basis (1993 Q1 to 2019 Q2)



Source: author's illustration based on PALMSv3.3 after adjustment.

Measurement of earnings can be on a per capita or household basis, which averages the incomes of all the people sharing a particular dwelling. The latter is widely recognized to be better at capturing average standards of living. 'The inequality in the labour market translates into even higher inequality in society given that high earners tend to live together with other high earners while low wage individuals often end up sharing their incomes with the unemployed' (Wittenberg 2017b: 279). Although available data are surveyed on an individual basis, households are tracked too through the id variable '*unqr*', making it possible to average individual records across households.

Figure 6b: Evolution of the Gini index, household basis (1993 Q1 to 2019 Q2)

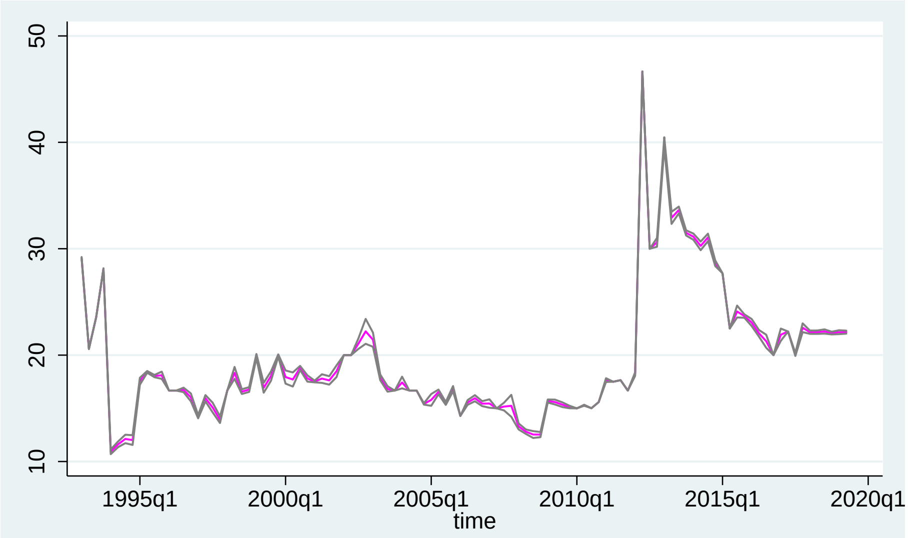


Source: author's illustration based on PALMSv3.3 after adjustment.

4.2 The P90/P10 dispersion ratio

The ratio of the wage share of the richest 10 per cent of workers to the share of the poorest 10 per cent ignores information about incomes in the middle of the income distribution. It is found to equal 19.6 on average over the period 1993–2019, showing an alarming peak in 2012 Q2 that is sustained until 2014.

Figure 7: Evolution of the P90/P10 dispersion ratio (1993 Q1 to 2019 Q2)

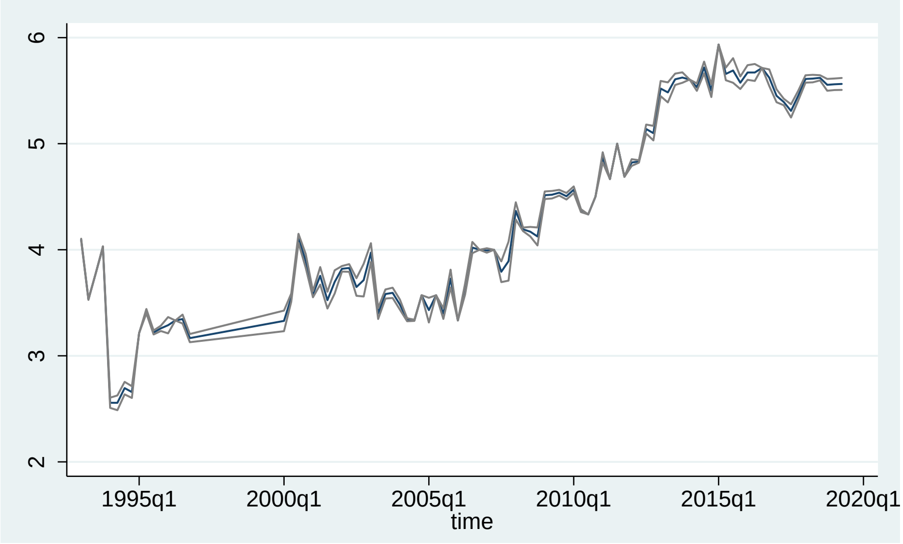


Source: author's illustration based on PALMSv3.3 after adjustment.

4.3 The P90/P50 dispersion ratio

A similar explanation can be applied to the P90/P50 ratio, the income share of the richest 10 per cent with respect to the lower 50 per cent of the wage distribution. The average ratio is 4.7, which implies that the richest receive five times more income than the poorest. Figure 8 shows a well-defined, positive trend that peaked in the first quarter of 2015. This evidence suggests that the wage differential between the ninth and the fifth decile of the wage distribution has been increasing over time: while the richest have become richer, the wage of the poorest 50 per cent has not increased proportionally. Yet, it seems that P50 changes are more closely related to P90 than P10.

Figure 8: Evolution of the P90/P50 dispersion ratio (1993 Q1 to 2019 Q2)

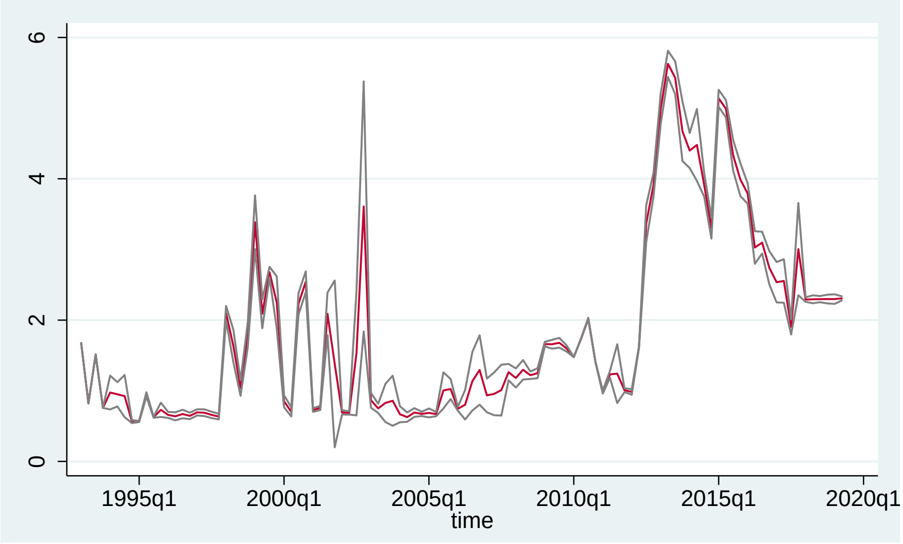


Source: author’s illustration based on PALMSv3.3 after adjustment.

4.4 The generalized entropy index

Measures from the generalized entropy (GE) class are sensitive to changes at the higher end of the distribution if the weight given to distances between incomes at different parts of the income distribution is high. The GE index calculated here employs a parameter equal to 2 such that the index is especially sensitive to the existence of large incomes. Figure 9 reveals the worrying presence of high incomes around 2000, while it confirms previous observations over the 2014–16 period.

Figure 9: Evolution of the generalized entropy index (1993 Q1 to 2019 Q2)

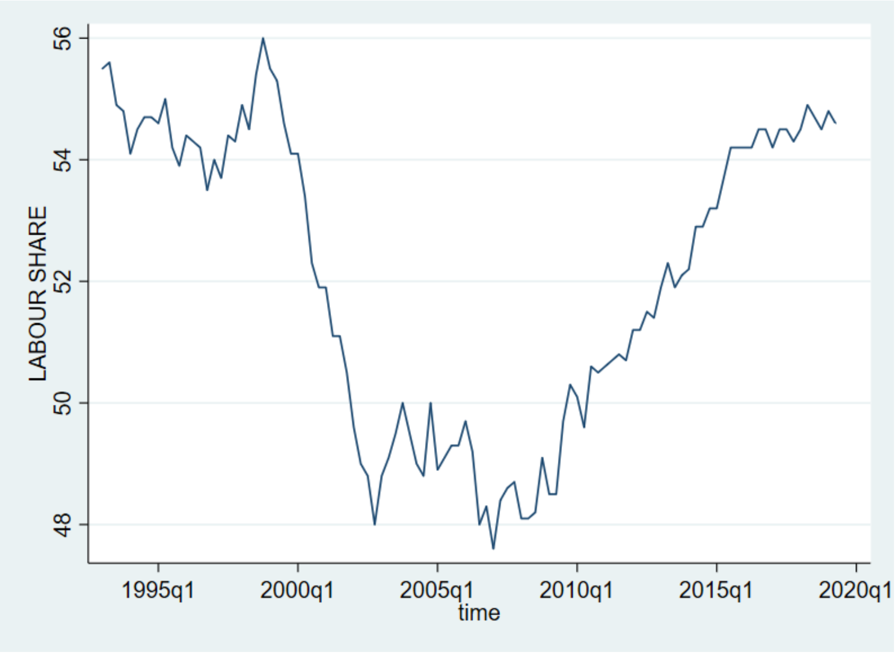


Source: author’s illustration based on PALMSv3.3 after adjustment.

4.5 Labour share of income

The wage or labour share is the part of national income allocated to workers in the form of monetary compensation as opposed to the part of value added going to the capital input. In advanced economies a declining labour income share constitutes a major factor in understanding rising inequality, since labour income is more equally distributed than capital income and represents a higher share of total income for lower- and middle-income groups. If these considerations hold in South Africa, the labour share is expected to be inversely related to inequality. In post-apartheid South Africa, the share of labour declined while that of capital increased until 2008 (see Figure 6a). Burger (2015) notes that this happened as a consequence of the widening gap between real wages and labour productivity. Comparing Figure 10 with previous inequality measures, the labour share moves together with inequality suggesting that, given that the lower end of the labour income distribution is structurally unemployed or economically inactive, increasing wages and employment opportunities affect higher incomes relatively more. The functional distribution of income does not seem to be a good proxy for inequality in South Africa.

Figure 10: Evolution of the labour share of income (1993–2019 Q2)



Notes: the labour income share is defined as the ratio of compensation of employees to gross value added at factor cost. It is calculated by dividing the compensation earned during a certain period by the economic output produced over the same period.

Source: author's illustration based on SARB (2019).

5 Conclusions

Despite substantial adjustments that were already implemented to Stats SA survey data, a number of problems have been inherited from the primary data used to compile PALMS in the first place, and as such they have no post-fieldwork solution. Inevitably, the time series plotted in Figures 6 to 9 may still feature characteristics that should be ascribed more to methodological than to real variation.[16] Nevertheless, this work contributes to the previous literature on South African disaggregated data by improving existing data quality, delivering a robust time series of labour income inequality among wage employees, and thus facilitating long-run dynamic policy analysis.

References

Ardington, C., D. Lam, M. Leibbrandt, and M. Welch (2005). 'The Sensitivity of Estimates of Post-Apartheid Changes in South African Poverty and Inequality to Key Data Imputations'. Working Paper 106. Cape Town: Centre for Social Science Research (CSSR), University of Cape Town.

Bassier, I., and Woolard, I. (2018). 'Exclusive Growth: Rapidly Increasing Top Incomes amidst Low National Growth in South Africa'. REDI 3X3 Working Paper 47. Cape Town: SALDRU, University of Cape Town.

Bhorat, H.I., C. van der Westhuizen, and J. Toughedah (2009). 'Income and Non-Income Inequality in Post-Apartheid South Africa: What Are the Drivers and Possible Policy Interventions?'. Working Paper 09/138. Cape Town: Development Policy Research Unit (DPRU), University of Cape Town.

Billor, N., A.S. Hadi, and P.F. Velleman (2000). 'BACON: Blocked Adaptive Computationally Efficient Outlier Nominators'. *Computational Statistics & Data Analysis*, 34(3): 279–98.

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.

Burger, P. (2014). 'Wages, Productivity, and Labour's Declining Income Share in Post-Apartheid South Africa'. Presidential Address, 2014 Annual General Meeting of the Economic Society of South Africa, University of the Witwatersrand, Johannesburg.

Burger, R., and D. Yu (2007). 'Wage Trends in Post-Apartheid South Africa: Constructing an Earnings Series from Household Survey Data'. Working Paper 10/06. Stellenbosch: University of Stellenbosch.

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

Cichello, P.L., G.S. Fields, and M. Leibbrandt (2005). 'Earnings and Employment Dynamics for Africans in Post-Apartheid South Africa: A Panel Study of Kwazulu-Natal'. *Journal of African Economies*, 14(2): 143–90.

Coibion, O., Y. Gorodnichenko, L. Kueng, and J. Silvia (2017). 'Innocent Bystanders? Monetary Policy and Inequality'. *Journal of Monetary Economics*, 88(C): 70–89.

Devereux, S. (1983). 'South African Income Distribution 1900–1980'. Working Paper 51. Cape Town: SALDRU, University of Cape Town.

[16] A detailed review of these issues lies outside the scope of this paper. For instance, Wittenberg and Pirouz (2013) and Wittenberg (2014a) both discuss how the earnings questions evolved across waves over the period 1994–2012 and how this impacted on data collection. For a comprehensive discussion see Kerr and Wittenberg (2019a).

Finn, A. (2015). 'A National Minimum Wage in the Context of the South African Labour Market'. Working Paper 153. Cape Town: SALDRU, University of Cape Town.

Finn, A., and M. Leibbrandt (2018). 'The Evolution and Determination of Earnings Inequality in Post-Apartheid South Africa'. WIDER Working Paper 2018/83. Helsinki: UNU-WIDER.

Finn, A., M. Leibbrandt, and V. Ranchhod (2016). 'Patterns of Persistence: Intergenerational Mobility and Education in South Africa'. Working Paper 175. Cape Town: SALDRU, University of Cape Town.

Juster, F.T., and J.P. Smith (1997). 'Improving the Quality of Economic Data: Lessons from the HRS and AHEAD'. *Journal of the American Statistical Association*, 92(440): 1268–78.

Kerr, A., and F. Teal (2012). 'The Determinants of Earnings Inequalities: Panel Data Evidence from South Africa'. IZA Discussion Paper 6,534. Bonn: IZA (Institute for the Study of Labor).

Kerr, A., and M. Wittenberg (2019a). 'Earnings and Employment Microdata in South Africa'. WIDER Working Paper 2019/47. Helsinki: UNU-WIDER.

Kerr, A., and M. Wittenberg (2019b). 'A Guide to Version 3.3 of the Post-Apartheid Labour Market Series'. Working Paper. University of Cape Town: DataFirst.

Klevmarken, W.A. (1982). 'Missing Variables and Two-Stage Least-Squares Estimation from More Than One Data Set'. Working Paper 62. Stockholm: Research Institute of Industrial Economics.

Leibbrandt, M., I. Woolard, A. Finn, and J. Argent (2010). 'Trends in South African Income Distribution and Poverty since the Fall of Apartheid'. Social, Employment and Migration Working Papers 101. Paris: Organisation for Economic Co-operation and Development (OECD).

Morris, T.P., I.R. White, and P. Royston (2014). 'Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws'. BMC Medical Research Methodology, 14(75): online, open access.

Neyens, L., and M. Wittenberg (2016). 'Changes in Self-Employment in the Agricultural Sector, South Africa: 1994 -2012'. Working Paper 172. Cape Town: DataFirst.

Posel, D., and D. Casale (2005). 'The Continued Feminisation of the Labour Force in South Africa'. *South African Journal of Economics*, 70(1): 156–84.

Rubin, D.B. (1987). 'The Calculation of Posterior Distributions by Data Augmentation. Comment: a Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations when Fractions of Missing Information Are Modest: the SIR algorithm. Journal of the American Statistical Association, 82(398), 543-546.

Schenker, N., and J.M.G. Taylor (1996). 'Partially Parametric Techniques for Multiple Imputation'. *Computational Statistics & Data Analysis*, 22(4): 425–46.

Seekings, J. (2007), 'Poverty and Inequality after Apartheid'. CSSR Working Paper 200. Cape Town: CSSR, University of Cape Town.

Van der Berg, S., M. Louw, and R. Burger (2007). 'Post-Apartheid South Africa: Poverty and Distribution Trends in an Era of Globalization'. MPRA Paper 9,065. Munich: Munich Personal RePEc Archive (MPRA).

Vermaak, C. (2012). 'Tracking Poverty with Coarse Data: Evidence from South Africa'. *The Journal of Economic Inequality*, 10: 239–65.

Von Fintel, D. (2007). 'Earnings Bracket Obstacles in Household Surveys: How Sharp Are the Tools in the Shed?'. Stellenbosch Economic Working Papers 08/06. Stellenbosch: Stellenbosch University.

Wittenberg, M. (2008). 'Nonparametric Estimation when Income Is Reported in Bands and at Points'. Working Paper 94. Cape Town: University of Cape Town.

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 Organization.

Wittenberg, M. (2014b). 'Wages and Wage Inequality in South Africa 1994–2011: The Evidence from Household Survey Data'. Working Paper 135. Cape Town: SALDRU, University of Cape Town.

Wittenberg, M. (2017a). 'Are We Measuring Poverty and Inequality Correctly? Comparing Earnings Using Tax and Survey Data'. Econ3X3 website, 3 October..

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.

Wittenberg, M. (2017c). 'Wages and Wage Inequality in South Africa 1994–2011: Part 2—Inequality Measurement and Trends'. *South African Journal of Economics*, 85(2): 298–318.

Wittenberg, M., and F. Pirouz (2013). 'The Measurement of Earnings in the Post-Apartheid Period: an Overview'. Technical Paper 23. Cape Town: DataFirst.

Data sets

Branson, N. (2010). 'South Africa—OHS-LFS Consistent Series Weights 1994–2007'. Cape Town: DataFirst.

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

SARB (South Africa Reserve Bank) (2019). 'Compensation of Employees to GDP at Factor Cost 1993.1–2019.2'. Available at: https://www.resbank.co.za/Research/Statistics (accessed 15 December 2019).

South Africa Data Archive (SADA). Available at: http://sada-data.nrf.ac.za (accessed 15 October 2019).

Statistics South Africa (2010). 'October Household Surveys 1994–1999'. Pretoria: Statistics South Africa (producer). Cape Town: DataFirst (distributor).

Statistics South Africa (2012). 'Labour Force Surveys 2000.1–2007.2'. Pretoria: Statistics South Africa (producer). Cape Town: DataFirst (distributor).

Statistics South Africa (2019a). 'Quarterly Labour Force Surveys 2008.1–2011.4'. Pretoria: Statistics South Africa (producer). Cape Town: DataFirst (distributor).

Statistics South Africa (2019b). 'Labour Market Dynamics in South Africa 2010–2011'. Pretoria: Statistics South Africa (producer). Cape Town: DataFirst (distributor).

Appendix

Table A1: Summary statistics of *realearnings* by survey wave

Wave	$N_{observed}$	Mean	Median	σ	$N_{missing}$	N_{zero}		$N_{million}$
					E	SE	E	
PSLSD 93	8,592	7,400.6	3,892.2	15,618.9	31,552	0	27	0
OHS 94	16,007	4,450.1	2,431.6	12,639.8	116,462	7	1,175	1
OHS 95	24,432	6,930.8	3,971.9	16,479.0	106,355	10	0	1
OHS 96	72,890	-	-	-	72,890	-	-	-
OHS 97	17,972	5,094.2	3,182.2	10,966.5	122,043	28	0	0
OHS 98	10,030	5,740.3	2,824.7	22,269.6	72,223	45	0	2
OHS 99	13,498	7,843.1	2,407.9	77,701.5	93,152	2	0	11
LFS 00:1	6,767	4,910.3	2,372.7	20,045.3	31,762	0	0	1
LFS 00:2	22,392	9,307.5	2,697.8	173,586.9	82,978	7	15	13
LFS 01:1	21,615	4,671.3	2,495.9	20,051.3	86,111	0	0	1
LFS 01:2	18,921	4,964.3	2,241.1	12,159.0	87,518	0	1	1
LFS 02:1	19,134	4,766.7	2,241.6	7,830.9	90,276	0	0	0
LFS 02:2	16,806	4,501.1	2,376.7	10,970.3	85,674	0	0	0
LFS 03:1	16,882	4,439.7	2,449.1	8,070.1	83,952	0	0	0
LFS 03:2	15,864	4,700.2	2,308.7	11,029.2	82,903	0	0	0
LFS 04:1	16,191	4,662.2	2,336.0	7,099.9	82,065	0	0	0
LFS 04:2	17,369	4,578.9	2,452.6	10,128.1	92,519	0	0	0
LFS 05:1	17,782	4,518.3	2,420.8	6,628.2	92,889	0	0	0
LFS 05:2	18,418	5,266.4	2,522.5	80,012.2	90,661	1	1	1
LFS 06:1	19,167	4,860.9	2,852.1	8,262.9	89,178	0	0	0
LFS 06:2	18,948	4,823.0	2,717.4	9,001.6	87,952	0	0	0
LFS 07:1	19,918	5,078.5	2,852.1	9,387.4	89,633	0	0	0
LFS 07:2	18,881	5,050.8	2,717.4	8,822.7	87,105	0	0	0
QLFS 10:1	21,016	8,931.8	3,722.4	81,465.4	66,855	17	32	3
QLFS 10:2	20,965	8,312.1	3,857.3	18,691.7	65,653	14	25	0
QLFS 10:3	20,105	10,546.5	3,826.5	180,494.3	64,519	18	0	3
QLFS 10:4	19,827	10,674.9	3,952.5	180,648.7	63,530	4	15	3
QLFS 11:1	19,477	24,652.0	4,003.3	128,245.7	63,095	7	30	6
QLFS 11:2	19,450	9,379.9	3,978.0	30,919.5	62,687	19	11	6
QLFS 11:3	20,372	8,723.8	3,982.3	17,351.4	63,905	8	29	0
QLFS 11:4	20,939	8,613.9	4,139.0	13,853.9	64,014	13	25	0
QLFS 12:1	20,672	8,306.5	4,047.8	12,900.5	63,583	26	63	0
QLFS 12:2	20,493	9,128.3	4,062.5	28,876.1	64,238	15	185	2
QLFS 12:3	19,671	10,695.2	3,960.6	77,200.0	66,089	7	150	6
QLFS 12:4	19,089	11,773.3	3,916.8	166,311.1	66,023	8	128	8
QLFS 13:1	18,496	16,175.0	3,693.9	369,648.0	66,696	14	171	12
QLFS 13:2	19,012	14,918.1	3,679.2	391,168.6	67,903	19	121	6
QLFS 13:3	19,415	14,114.9	3,508.2	382,233.5	67,642	19	139	8
QLFS 13:4	19,985	12,125.6	3,488.2	103,485.3	67,538	18	127	12
QLFS 14:1	19,612	9,766.7	3,603.3	35,218.0	67,386	11	79	4
QLFS 14:2	19,024	11,575.0	3,570.5	229,259.2	65,722	19	83	5
QLFS 14:3	19,189	9,655.6	3,528.7	65,956.3	66,113	10	95	2
QLFS 14:4	18,980	9,600.4	3,528.7	86,665.7	65,294	4	72	2
QLFS 15:1	17,019	11,663.8	3,463.2	118,317.3	55,542	5	61	5
QLFS 15:2	17,192	11,027.2	3,408.9	117,790.4	54,815	0	23	3

QLFS 15:3	17,204	11,312.6	3,396.7	112,190.1	54,665	4	36	6
QLFS 15:4	16,717	11,451.3	3,390.3	138,546.1	53,628	0	19	7
QLFS 16:1	15,833	9,890.5	3,530.1	83,167.7	53,179	0	14	1
QLFS 16:2	14,936	8,790.6	3,421.7	25,324.7	52,718	0	6	0
QLFS 16:3	15,508	8,534.1	3,391.2	22,450.7	53,901	0	0	0
QLFS 16:4	15,675	9,572.2	3,559.8	87,172.4	53,556	0	2	1
QLFS 17:1	15,808	11,609.2	3,547.3	155,817.7	53,545	1	15	7
QLFS 17:2	15,551	10,749.4	3,526.6	133,777.6	54,362	0	14	6
QLFS 17:3	15,557	8,724.0	3,429.5	67,957.3	53,702	1	18	4
QLFS 17:4	15,117	8,485.4	3,466.0	48,709.5	53,234	0	15	4
PALMSv3.3	963,492	8,784	3,225	221,258	4,648,568	387	3,022	132

Note: suspect values are in red.

Source: author’s calculation based on PALMSv3.3 (no adjustment).