UNIVERSITY OF CAPE TOWN

Measuring Wages and Inequality in South Africa Using Two Nationally Representative Data Series

Author: Bruce McDougall
Supervisor: Martin Wittenberg
19/02/2018

Abstract

The National Income Dynamics Study (NIDS) and the Post-Apartheid Labour Market Series (PALMS) are two data sources frequently relied upon for research into earnings in South Africa. This paper contributes to the literature in three ways. Firstly, I show how NIDS data can be adjusted to account for item non-response using a bracket reweighting technique and the effects thereof. Secondly, I consider how estimates of the wage distribution differ between NIDS and PALMS when using the most comparable estimation methods available. Finally, I discuss what the data reveal about the evolution of inequality in South African wages between 2008 and 2014.

Thesis paper for the master's programme in applied economics, February 2018.

1. Introduction

Inequality in the distribution of resources has taken centre-stage in economic discourse following the works of prominent economists such as Joseph Stiglitz (2012) and Thomas Piketty (2013). Their efforts served to undermine the preceding orthodox view[1] that inequality is an unavoidable and somewhat inconsequent by-product of economic growth. Similarly, economic inequality is now perhaps the most pivotal element of the political landscape, with factions around the world divided around questions of its tolerable levels, causes and cures.

The placement of inequality at centre-stage in economic discourse has led to a scramble amongst researchers to collect and analyse data on the distribution of wealth. Nowhere in the world is this research more pertinent than post-apartheid South Africa, which remains the poster-child for economic and societal inequality. Understanding this inequality crucially depends on an investigation of the distribution of earnings, as earnings have historically contributed the largest portion to individuals' total wealth. As such, this paper aims to add to the discussion by analysing the earnings distribution and how it has changed over time.

In order to provide a more comprehensive study, I analyse earnings using two different sources while comparing the results. These are the National Income Dynamics Study (NIDS) and the Post-Apartheid Labour Market Series (PALMS), two long-running datasets based on household survey data. Because both are designed to be nationally representative, provided the same estimation methods are used, NIDS and PALMS should theoretically produce comparable results in terms of earnings and inequality. However, applying the same methods in both datasets requires a significant re-working of the NIDS earnings variable, and a change from an imputation to a bracket reweighting (BRW) strategy for item non-response.

As such, the paper answers three primary research questions (RQs):

1) What are the necessary data quality adjustments in NIDS and what are their effect?

2) Are the distributions in NIDS and PALMS similar when using comparable estimation methods?

3) What do the data tell us about the evolution of inequality between 2008 and 2014?

The paper describes several data quality adjustments appropriate in NIDS, including the BRW technique which results in marginally increased measures of inequality. The comparison between NIDS and PALMS reveals an encouraging story; both datasets seem to be measuring the same underlying earnings distribution. Lastly, the results suggest that inequality has worsened at the top of the earnings distribution, while some improvement has probably occurred among the lower half. The effect on overall inequality is therefore ambiguous.

---

[1] The more orthodox view is perhaps most famously captured in Robert Lucas' quote of 2004: "Of the tendencies that are harmful to sound economics, the most seductive, and in my opinion most poisonous, is to focus on questions of distribution".

The layout of the rest of the paper is as follows. Section 2 provides a review of the existing literature. Section 3 introduces the methodology while Section 4 describes the data sources used. Sections 5, 6 and 7 answer the three research questions above. Section 8 then concludes.

2. Literature Review

2.1. Comparing Earnings in Different Datasets

A major advantage of this study is the ability to compare contemporaneous results using two different datasets that aim to measure the same underlying construct (earnings). Data collected and processed for release in any given dataset are invariably exposed to error and one can never have full confidence. Comparing results between two different datasets that aim to measure the same thing is useful as it provides a check on the consistency between the sources and informs the confidence by which one can draw conclusions. In this section I describe efforts that have previously been made to compare earnings-related[2] variables in South Africa using different data sources.

2.1.1 Using any data sources to make comparisons

While there are many data sources measuring earnings in RSA, comparisons between them are not straightforward and works that do so tend to be rare. Wittenberg (2014, 2017) provides two papers that make such comparisons. In the first case he compares earnings in Quarterly Labour Force Survey (QLFS) household data with Quarterly Employment Statistics (QES) firm data. He finds that average wages reported in firm surveys are irreconcilably higher than those reported in the QLFS (*ibid:* 41). Like others Wittenberg (2014) posits that the discrepancy arises due to underreporting of earnings in household survey data. He later (2017) compares earnings in QLFS survey data to South African Revenue Service (SARS) tax assessment data. He finds that under-reporting exists in both data sources, particularly amongst the wealthy and the self-employed. However, this problem is more pronounced in QLFS household data, as respondents have lower incentives to respond and tend to omit benefits such as medical aid and bonuses. As a result, Wittenberg (2017:15) concludes that QLFS data understate earnings on average by around 40%, with bigger gaps at the top of the distribution.

2.1.2 Comparing household surveys

The discrepancies mentioned in the papers above arise in part from the fact that the data are essentially of different types, being drawn from different sources (Wittenberg, 2014: 2 & 2017:6-7). This problem is mitigated when one specifically compares two household surveys. There are a number of household survey datasets available in South Africa, including the Project for

[2] The reader is assumed to be familiar with the difference between the variables related to wealth, such as earnings, wages, income and expenditure. However, for the purpose of clarity, a brief explanation of these terms is included in the Appendix A. Note that earnings in this paper exclude earnings from self-employment.

Statistics on Living Standards and Development (PSLSD), the October Household Survey (OHS), the Income and Expenditure Survey (IES), the Labour Force Survey (LFS), the Quarterly Labour Force Survey (QLFS), the National Income Dynamics Study (NIDS), and others. However, despite the number of data sources there are very few papers comparing results between them.

Leibbrandt et al (2010:25) compare household income in different years between the PSLSD (1993), IES (2000) and NIDS (2008). Unfortunately, the use of different time periods makes it difficult to draw direct inferences about consistency between the surveys. This is less of a problem in Finn, Leibbrandt & Woolard (2009:3) who in their paper compare the household income and expenditure per capita measure in NIDS 2008 to the IES 2006/2007. However, they note that the comparison is dubious without further work due to differences in the construction of the variables in NIDS as opposed to the IES.

Perhaps the most thorough comparison of different household surveys comes from the creation of the PALMS series itself. PALMS is a collection of consecutive surveys and comprises the PSLSD, OHS, LFS and the QLFS stacked in order as explained by Kerr & Wittenberg (2017). While much consideration went into comparing these surveys when they were combined, the nature of this comparison is different as the surveys lie side by side (chronologically) and as such are not expected to produce the same results as one would expect from contemporaneous cross-sections (excluding the effect of sampling variation).

As such, barring those few instances mentioned above, there is to my knowledge no other work comparing earnings in any given survey from within PALMS to an external dataset such as NIDS. Similarly, earnings in NIDS as a national cross section have not been compared to another cross section. As will be explained, because NIDS and PALMS aim to provide nationally representative cross-sections of earnings using weights, they should theoretically produce the same results (provided the analytical techniques are comparable). The theoretical justification for this is expanded upon in subsection 6.1.

2.2. Measurements of Inequality

2.2.1 Inequality in the literature

Inequality has long been a central issue in South Africa and the literature is well-developed. Of course, there are several dimensions to inequality and many ways to understand and analyse it. The first question to answer is: inequality of what?

There are as many potential measures of inequality as there are types of wealth. The literature discusses several variables within which inequality can arise, typically at the household or individual level. Woolard & Mwebe (2016) discuss household net worth, being the difference between gross wealth and gross liabilities in the household. Finn, Leibbrandt & Woolard (2009) describe monthly household income and expenditure per capita. Other papers focus on broader measures of wealth, such as social cohesion (Njozela et al, 2016) or subjective well-being (Kannemeyer, 2016). At the individual level, wages, self-employed income, social transfers, remittances and rental income are commonly discussed (Van Der Berg, 2010). In some analysis individuals' wealth or income is aggregated within a household to provide a total measure for

the household, or the opposite is done by dividing a household value by the number of household members to get a per-capita figure.

Apart from simply measuring inequality in these variables, authors examine its determinants. Existing work considers the roles that race (Leibbrandt et al, 2009; Van Der Berg, 2010), geotype[3] (Leibbrandt et al, 2010), education (Keswell & Poswell, 2004; Seekings, 2007), the wealth of one's parents (Girdwood & Leibbrandt, 2009), employment (Tregenna, 2011), gender (Bhorat & Goga, 2012), remittances (Biyase & Tregenna, 2016), social transfers (Posel, 2016), access to services (Leibbrandt et al, 2010) and a host of other variables have in explaining inequality.

Surveying the full extent of this literature is clearly a large undertaking and is not the purpose of this text.[4] However, two lessons can be drawn that are commonly agreed upon, which inform the subject of this paper. Firstly, there is widespread agreement that inequality has not declined following the demise of apartheid; in fact, several papers contend that it continues to rise (Woolard & Mbewe, 2016; Leibbrandt et al, 2012). Secondly, it is generally agreed that the labour market and earnings therein are paramount in explaining this inequality (Van Der Berg, 2010; Leibbrandt et al, 2010:19). These lessons reveal the importance of earnings inequality to inequality research in South Africa.

2.2.2 Earnings inequality in the literature

Authors agree that earnings in the labour market remain the largest part of total earnings, and therefore the most important contributor to overall inequality. For example, Van Der Berg (2010:15) suggests that earnings made up 63% of total earnings in 2005 - Leibbrandt et al (2010:23) contend that the figure is closer to 70%. Between 1993 and 2008, Lebbrandt et al (2010) find that earnings' contribution lay between 70%-90% of total inequality as given by the Gini coefficient.

The importance of earnings has motivated previous statistical analyses into the distribution of earnings. Wittenberg (2016) provides a useful starting point. He finds that while mean earnings have risen over the post-apartheid period (between 1993 and 2014), the gains have not been evenly distributed and the median earner has fallen behind. In terms of common percentiles (as explained in section 3.6.2), he notes that both the higher (p90 and p75) and the lower (p10 and p25) percentiles have gained relative to the median earner in the post-apartheid period.

Wittenberg (*ibid)* posits that a summary measure of inequality such as a Gini coefficient would likely reflect a worsening of income inequality given these trends. However, he warns that such measures which summarize trends are problematic as they conceal relevant information and will reach different conclusions depending on how they aggregate the underlying phenomena. In the same paper he provides an analysis of self-employed income but finds that including the

[3] In Leibbrandt's usage geotype refers to the distinction between a 'rural' or 'urban' setting.
[4] For an overview of the literature see, for example, Leibbrandt et al (2010).

self-employed did not have a large impact on inequality measures such as p-ratios. Lastly, he notes that the relative gains in the bottom of the distribution began to reverse around 2012.

Du Toit & Wittenberg (2016) reach similar conclusions. They agree that over the 1993-2014 period, inequality has widened at the top of the distribution, whilst narrowing at the bottom. Again, they stress that a summary figure (such as a Gini coefficient or an overall variance measure) would not be useful alone as these aggregate over contradictory effects. Controlling for compositional effects, they argue that gains experienced by the wealthy and the poor have arisen from increasing returns to education and experience, rather than increasing endowments of these factors (du Toit & Wittenberg, 2016:18). They also suggest that minimum wage laws had a role to play in reducing inequality in the bottom half of the distribution.

2.2.3 A shortfall in looking at earnings inequality

A downside of analysing earnings inequality rather than a broader measure is that it omits certain other forms of income and wealth that are relevant contributors to overall inequality. Authors such as Van Der Berg (2010:15) note the rising role of interest, dividends and rental income in total income. Corroborating this, some sources suggest that the share of earnings in total income has been declining: Van Der Berg (*ibid)* reports a share of earnings in total income at 81% in 1976. Du Toit & Wittenberg (2016:2) put this share at 69% in 1993; by 2005, the figure dropped to 63%.

Wealth is even broader than income or earnings. Discussing wealth inequality, Woolard and Mbewe (2016) find that inequality in wealth is extreme and rising faster than income inequality. As such, it seems that inequality in wealth, and in the returns to wealth are both increasing. These observations are in accord with Piketty's line of argument that globally the wealthy are increasingly benefitting from an unequal distribution of and increasing returns to capital. The implication for economists is that inequality research that focuses only on earnings will capture less of the overall story.

3. Methodology

3.1. A Note on Data Quality vs. Comparability

Because NIDS and PALMS differ in some regards, at times the best treatment of the data in one dataset is not the same as in the other. In particular, NIDS often offers more data than PALMS, allowing more "room for manoeuvre" in terms of adjusting the data to improve the quality. A tension therefore arises between the goal of improving data quality in NIDS as far as possible versus the goal of comparing NIDS to PALMS. In some cases, such as the decision I take to ignore the secondary job information available in NIDS, I make a slight sacrifice in NIDS in order to preserve comparability to PALMS. In the case of the preferred bracket reweighting technique (explained in 4.3.2), because it has a larger negative effect on comparability, I rather opt to provide sets of results with and without the technique applied, so that one can compare NIDS and PALMS before the adjustment.

3.2. The Instrument: Gross Wages

This subsection explains the chosen variable of interest for the paper, which is real gross earnings from work excluding self-employment. Choosing to consider earnings has certain

benefits as well as shortfalls. As discussed, narrower concepts unfortunately omit certain factors which contribute to inequality, thus they are likely to under-represent it.

Secondly, excluding earnings from self-employment is undesirable as these individuals are a non-negligible portion of the population. However, as Wittenberg (2016:2) notes, including them may confound the data these earnings are more prone to data quality issues: inconsistency between surveys (especially a longer series like PALMS), coverage problems and missing values are notably worse in the case of self-employment. Adjusting for these data quality problems is difficult although not impossible.

Including self-employed earnings would complicate the analysis for two further reasons. Firstly, as Wittenberg (2016:2) notes, the concept of gross earnings is not clearly defined for the self-employed– which (if any) deductions should be allowed? Secondly, including self-earnings information complicates the bracket-reweighting procedure which is preferred for reasons discussed in 3.4.2. There are cases where individuals report primary earnings in brackets and self-employed earnings as a point value, and vice-versa. How does one classify these responses?

Although the choice of excluding self-employed earnings is limiting and may lead to misestimates of inequality, it has the advantage of avoiding these analytical and data quality issues. This is why employed work has often been the focus of analysis, internationally and in South Africa (Wittenberg, 2016: 12). Using a narrower construct is also useful when making comparisons as there is greater certainty that one is comparing like with like, both between the surveys in RSA or internationally. Therefore, I opt to consider earnings from non-self-employed work, which I refer to as "employed" work, which provides "wages" or "earnings".

It is analytically preferable to make certain data transformations. The nominal amounts provided in NIDS need to be adjusted to real values to remove the effect of inflation, making intertemporal comparison more meaningful. The data are deflated to June 2015 prices to match the real earnings variable provided in PALMS v3.2. The CPI data used are nationally aggregated annual figures taken from Stats SA.[5] In the literature earnings and real earnings are commonly assumed to follow a somewhat log-normal distribution. The log transformation is essential for examining and comparing densities as it consistently scales the data into a narrower band.

Finally, there is a distinction to be drawn between gross and net wages. As Wittenberg (2016:2) discusses, the construct of interest in the surveys that comprise PALMS has almost universally been gross monthly earnings. Conversely, while NIDS provides both net and gross monthly amounts, the net measure is more comprehensive: NIDS collects bracket data for net earnings but does not allow respondents to do so for gross earnings. The bracket information is necessary for the bracket reweighting procedure performed on the data. PALMS v3.2 provides pre-calculated bracket weights based on gross earnings (see Kerr & Wittenberg, 2017).

To summarize, the available variables are presented in table 1 below (next page):

---

[5] Available in Appendix B-II and at http://www.Stats SA.gov.za/publications/P0141/CPIHistory.pdf

	NIDS Net	NIDS Gross	PALMS Gross	PALMS Net
Point Data	Yes	Yes	Yes	No
Bracket Data	Yes	No	Yes	-
Bracket Weights	Created in this paper	Not available due to no bracket data	Released in PALMS	-

Table 1: Earnings variables available in NIDS and PALMS.

Seeing as I would like to compare like with like, and that there are only gross wages in PALMS, gross monthly wages are the construct of choice. Therefore, for the remainder of the paper, any use of the words *earnings* or *wages* refers to real logged monthly values unless stipulated otherwise. Certain transformations of the NIDS data are necessary to create a comparable gross variable, as discussed in section 5.

3.3. Unit Non-Response and Post-Stratification

Both NIDS and PALMS release deign weights which theoretically make the sample realized representative of the population by weighting observed household units upwards by the inverse of the probability of selection. However, even when these are adjusted ex-post for household non-response, the sample may still be non-representative in a meaningful way. One reason for this is due to individual non-response, which will skew the sample if the missing individuals are missing in a non-random fashion. For example, if working-age individuals are less likely to be home at the time of survey, the sample will under-represent them. Analysis based on variables correlated with age would then be less accurate.

Fortunately, post-stratification can be used to adjust the design weights ex-post to make the weighted sample more reflective of the population in certain respects (geographic and demographic). This technique will also improve on mis-representation that may arise from other sources, such as sampling variation, fieldwork errors or the use of an outdated sampling frame (Branson & Wittenberg, 2011:5).

NIDS and PALMS both release post-stratified weights that provide (weighted) totals comparable to population data in terms of age, sex, race, and provincial totals as well as the total population estimate (NIDS, 2009a). The implicit assumption when performing analysis using these weights is that the nonresponse is missing completely at random (MCAR) with respect to the relevant variables of analysis within each age-sex-race-provincial category (NIDS, 2009b:5; NIDS, 2009a:6). The post-stratified weights released in NIDS and PALMS are not created in the exact same way however, a point I return to in section 6.1.

3.4. Item Non-Response and Bracket-Reweighting

The issue of how to deal with *item* non-response is more highly debated, and various approaches exist. Take the case of wages that are missing when someone has reported having a wage-paying job. The first approach is to assume the missing wage points are MCAR and simply ignore the missing information, looking at the picture when considering only those wage points (actual rand amounts) that were reported. However, this results in fewer observations, and more importantly will introduce bias if the item is not MCAR. An upside of the point

data and this approach is that the data are directly from the survey and unaffected by any errors that may arise when trying to adjustment them.

3.4.1 Problems with imputation

Another approach is to try regaining the data by means of imputation, which assumes that the data are missing at random, with response depending on an observed characteristic (NIDS, 2009b). This is the approach used by NIDS, which uses a simple wage imputation procedure in cases where there are at least 100 observations and the proportion of missing information is not greater than 0.6. A downside to such a strategy is that it involves creating new data based on existing data, which runs the risk of simply repeating existing patterns or relationships in the data according to the imputation algorithm. If the data are then used to estimate statistical significance, this will artificially reduce standard errors. NIDS notes this downside in the third technical appendix and stresses that additional treatment is preferable for the income variables (NIDS, 2009b:8).

Randomized imputation tries to mitigate this problem but unfortunately creates additional noise by adding a random error which can "dilute" the presence of true relationships or patterns in the data (Wittenberg, 2008). Additionally, because values output by an imputation process depend on the algorithm used and how it is operationalized, there is room for variation between datasets and researchers.

Another common strategy to deal with item non-response in questionnaires is to allow respondents to disclose the bracket (range) within with a variable lies rather than the value itself. This is preferred by respondents who consider the information sensitive. This improves the amount of information available but creates the question of what to do with the bracket data. Imputation using this information is again possible and indeed a popular deterministic imputation involves replacing the bracket observations with the midpoints of the brackets and some fixed multiple of the top bracket. This approach is taken by NIDS but has its own downsides. Imputing many observations with a single value can bias estimates of statistics such as percentiles or variance. The integrity of these statistics will be critical to the analysis of inequality. Mid-point imputation also causes issues in the context of nonparametric density plots as it creates spikes in the data which kernel functions struggle to smooth over (Wittenberg 2008).

In light of the above and given the goals of discussing inequality and comparing wage distributions in NIDS and PALMS, this paper considers the use of imputed values less appropriate for three reasons: Firstly, the regression imputed data released in NIDS and PALMS were not created using the same imputation technique. NIDS uses deterministic regression imputation for completely missing items, while using midpoints for the brackets. PALMS by contrast uses multiple imputation (MI) for both completely missing and bracket earnings data. A strategy might be to perform multiple imputation on the NIDS data, but this technique cannot be operationalized in in PALMS without further information from Stats SA about the imputation released in PALMS. Secondly, and perhaps more importantly, regression imputation relies on the dubious assumption that wage point data being imputed is MCAR once the regression has controlled for the appropriate variables. This assumption is unlikely given the stylized fact that people tend to be less likely to disclose point-values of earnings at

higher income levels – the wealthy have tighter lips. In other words, the missingness depends on the value of the missing variable itself. Lastly, as discussed, bracket mid-point imputation is inappropriate for the goals of this paper.

3.4.2 Why bracket reweighting is preferable

In the face of this problem an alternative to imputation is the bracket reweighting technique detailed in Wittenberg (2008). The idea is to apply the logic of inverse probability weighting traditionally used for unit non-response to instead adjust a sample for *item* non-response in the context of bracket responses. The bracket and point information captured are used to estimate the likelihood (p) of a respondent providing a point value as opposed to a bracket depending on their earnings *category* (not the wage itself, which is missing). The reweighting approach then involves increasing the weight of the responses that are less likely to be observed in order that they better represent their true propensity in the population. This is done by multiplying the current weight of the observation by 1 over p.

Compared to imputation, an attractive feature of the bracket reweighting approach is that it does not impose strict rules or assumptions on the data, apart from the assumption that that within each bracket, wage non-response is MCAR. Another benefit is that it specifically accounts for the fact that richer people tend to be more hushed about their earnings. As such it is has the dual benefit of improving the accuracy of a density plot without potentially risking perverting it by creating spikes and possibly accentuating or creating false information. Wittenberg (2008) has argued that for these reasons the bracket reweighting technique can produce better results than imputation, especially in the context of nonparametric analysis.

Given the goal of this paper in improving data quality the BRW is an attractive technique. Unfortunately, however, the BRW cannot be applied to PALMS beyond wave 1 for reasons discussed in 4.3. To avoid the risk of conflating the comparison between NIDS and PALMS, I therefore provide a set of results before and after the BRW is applied. This allows one to examine the effect of the BRW where applicable while also providing an unaffected (pre-BRW) comparison. I further argue in 5.5.3 that it is not clear that the BRW worsens comparability even if it can only be applied fully to NIDS. A final benefit to performing the BRW is simply that it is a somewhat novel technique which warrants investigation; the BRW could be useful in other contexts or when applied to future releases of NIDS or PALMS. As such, the BRW is preferred in this context as it theoretically provides the best improvement to data quality in NIDS and PALMS (wave 1) without undermining the comparison.

A brief note on nomenclature may be helpful going forward. Where the term *full imputation* is used, this refers to imputation that accounts for both completely missing earnings data and those that provided bracket responses. For PALMS this would be the data created using MI techniques. In the case of NIDS, *full imputation* data refers to the regression imputation used for completely missing data *and* the midpoint imputations for brackets. In both cases, full imputation therefore refers to using all the imputed values released with the datasets. The term *imputed* or *imputed values* can refer to imputation for either type of missing item, but in the case where the data are imputed *using* bracket information, this is specifically made clear using the phrase *imputed from bracket responses* or similar.

3.5. Measurement and Processing Error

Apart from being missing, data can also be wrong. Wittenberg (2016:3) describes the following sources of measurement error: fieldworker fraud, honest fieldworker mistakes, respondent errors or respondent lies. Wittenberg (2014) has argued that respondents in household surveys may be deliberately reporting their net wages when the question expressly asks for gross amounts, to underrepresent their earnings. The use of proxy respondents may increase the likelihood of honest mistakes in the data. Processing error is also possible: mistakes can occur in the capture, storage or cleaning of the data, before it is released. Unfortunately, there are few tools available to mitigate these types of error.

3.6. Outliers

There are several approaches to dealing with outliers. PALMS flags outliers based on studentized residuals following a Mincerian wage regression. A blunter approach is to simply drop observations that are above a certain value. This is somewhat arbitrary however and additionally will not travel well over time due to the long-run growth of real variables and the inflation of nominal variables. NIDS does not release a flag for outliers and as such leaves it to the analyst. That said, examining the programme files reveals that a cut-off is implemented above which observations captured from survey are not included in the public release. The level is at R500 000 per month (admittedly very high) which occurs in wave 2-4, on the net but not the gross variable.

An alternative to implementing a fixed cut-off is to use a formula to set the threshold; for example, one could use a fixed number of standard deviations from the mean. However, this still has the detriment of imposing a somewhat arbitrary threshold on the data. Additionally, one should be cautious of accidentally removing valid observations in the highest brackets as these data points are relatively rare to begin with. Another problem with both options above is that they do not treat observations that are unusually low.

By contrast a regression approach is less arbitrary as it considers what the expected value should be and then flags large differences in the observed value. Following PALMS as explained in Wittenberg (2016:7), the regression approach is operationalized as follows: A Mincerian-style regression is run on (log) real wages with population group, gender, age, age as a quadratic and a quartic, education, and occupation as controls. I do not include a measure of experience as the common formula for its calculation (age minus years of education minus six) would make it collinear.

The regression is run simultaneously over all four waves and marks observations with studentized residuals greater than five. Five differs from the traditional value of three and rather follows the precedent set by PALMS. This choice is for the sake of being consistent and produces a satisfactory flag. 0.21% of real net and 0.12% of real gross values are flagged as outliers, which is comparable to the 0.14% of real gross data flagged in PALMS. Flagged observations include values that were far higher or far lower than the fitted value. These observations can either be imputed for or dropped from the analysis. Seeing as I opt to avoid regression imputation, I prefer to drop these observations from the analysis.

3.7. Analysis

This section introduces the techniques used to analyse wage distributions and inequality.

3.7.1 Summary statistics

Perhaps the simplest way to discuss and compare wage distributions is in terms of descriptive statistics. I estimate the weighted mean of gross wages in nominal and real terms, including a 95% confidence interval and standard errors. The calculation takes account of the complex survey design using the *svyset* command to produce correct standard errors (calculated using standard Taylor-linearized variance estimation). I also discuss the distributions in terms of the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles.

3.7.2 P-ratios

P-ratios were mentioned in 2.3.2 and these are popular way to examine inequality. A p-ratio is the ratio of one percentile to another, for example the 90^{th} to the 50^{th}. P-ratios measure the difference in magnitude between percentiles which is useful for comparison over time.

3.7.3 Nonparametric density plots

Density plots summarize distributions by representing the frequency of observations within particular ranges (bins). Histograms use fixed bins which are inappropriate for making comparisons as the appearance of the distribution depends significantly on the relationship of observed data points to the bin boundaries and the number of bins. Kernel function do not use fixed bins but rather count observations within "sliding" bins (or "windows") centred around each observation. By down-weighting observations near the ends of the bin, kernels can avoid spiking in the distribution. Kernel estimation is thus preferred as it provides a consistent estimate of the density while avoiding spiking. I use the Epanechnikov function with the Silverman plug-in bandwidth for window size. When comparing several densities, this bandwidth is calculated for the first of them, and then applied to the rest to provide the same degree of smoothing.

3.7.4 Nonparametric variance estimation

A second measure of inequality is simply to estimate the variance of log wages over the sample of interest. I compare these between NIDS and PALMS over time. Like a Gini coefficient, this measure runs the risk of over-summarizing the underlying trends as discussed in 2.3.2.

4. The Data

4.1. Data Sources

The first data source is the PALMS V3.2 dataset.[6] As mentioned PALMS is a series of consecutive surveys that have been stacked together as explained by Kerr & Wittenberg (2017). These are listed below:

1993	Project for Statistics on Living Standards and Development
1994-1999	October Household Surveys
2000:1-2007:2	Labour Force Surveys (Biannual)
2008:0-2015:4	Quarterly Labour Force Surveys

The second data source is the National Income Dynamics Study[7], Waves 1-4. These cover the following periods:

2008	Wave 1
2010-2011	Wave 2
2012	Wave 3
2014-2015	Wave 4

For each wave of NIDS the datafiles must be merged before the four waves can be appended. This process and some of the differences between waves are explained below; the remainder is in programme files available from the author should the reader wish to replicate this work.

4.2. Sample

Part-time workers are qualitatively different to full-time workers and including them will influence measures of inequality. This is especially true if part-time work is not fully subject to minimum wage legislation as workers could be paid very low wages. While this is an important facet of inequality, their inclusion might conflate measurements of inequality. This is because part time work was not measured consistently throughout PALMS (Wittenberg & Du Toit, 2016) and the manner and extent to which they were surveyed in NIDS may differ to PALMS. For the sake of creating consistent and comparable measures, I opt to exclude part-time workers, a restriction commonly applied in the literature (Wittenberg, 2016; Du Toit & Wittenberg, 2016:12).

Following du Toit & Wittenberg (2016), the sample is restricted to individuals between the ages 20 and 60 (inclusive) who work for more than 35 hours per week. Making this hour restriction unfortunately shrinks the sample, a reduction which results in there being only two

---

[6] Kerr, Andrew, David Lam and Martin Wittenberg (2017), Post-Apartheid Labour Market Series [dataset]. Version 3.2. Cape Town: DataFirst [producer and distributor], 2017.

[7] Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2008, 2010/2011, 2012, 2014/2015, Waves 1-4 [datasets]. Versions 6.1, 3.1, 2.1, and 1.1 respectively. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2016. Cape Town: DataFirst [distributor], 2016. Pretoria: Department of Planning Monitoring and Evaluation [commissioner], 2014.

observations in the top income brackets in 2008 for example. Reducing the number of brackets could help but this would cause a loss of precision seeing as there are precipitous drops in the response rate in the very highest brackets.

As discussed, I exclude self-employed earnings. This is possible in NIDS as main job earnings (also referred to as "primary job" earnings) are separated from self-employment earnings in the questionnaire. Main job earnings in NIDS exclude all self-employment earnings and only contain earnings from employed work for someone else.

However, there is still a question of whether individuals who report earnings from such a job (working for someone else) but *also* report self-employment earnings -possible in NIDS- should remain in the sample. This situation is different in PALMS v3.2, where earnings collected from the underlying surveys were *either* from self-employment or from an employed job. The documentation isn't explicit, but I will presume that PALMS chooses between self-employed and employed earnings according to which is the primary (largest) source of earnings. Therefore, my assumption is, if someone's primary source of income was employed work, PALMS recorded these earnings, the person would stay in the sample, and the self-employed earnings would be disregarded.

Thus, to replicate the above in NIDS, my approach is to categorize an individual according to whether their self-employed or employed work is the primary source of earnings. If their self-employment earnings exceed their main job earnings, I classify them as self-employed and drop them from the sample. This is hopefully comparable to dropping observations flagged as self-employed in PALMS.[8] I describe this distinction as "earner type". In PALMS the BRW is performed on earnings with the self-employed included. Therefore, I include self-employed individuals (but not the amounts) when performing the BRW before dropping them for analysis.

Henceforth "the sample" therefore refers to individuals primarily employed, working for more than 35 hours per week, between the ages of 20 and 60 years.

4.3. Known Data Quality Issues in PALMS

Wittenberg (2016:8) provides a succinct summary of the known data quality issues in the PALMS series and how they relate to wages and measures of inequality. Most important to the present analysis are the issues with PALMS earnings post-2010. From 2010-2012:2 (inclusive) missing values and bracket responses were imputed by Stats SA with no flags or bracket information from which one could separate the survey data from the imputation (Kerr & Wittenberg, 2017:3). Following 2012:2 Stats SA did not impute in this fashion, although there are a suspicious number of observations of exactly R400 000 (*ibid).* For these reasons, the PALMS release does not perform the BRW after 2010; the bracket weight is the exact same as the cross-entropy (post-stratified) weights in these waves. This unfortunately means one cannot compare the BRW in NIDS to the BRW in these waves of PALMS.

[8] The relevant variable in PALMS is employerAll. In NIDS (after data processing) the variable is "selfp".

5. RQ1: Adjusting the NIDS Data

Performing the BRW on the NIDS data is interesting as it can provide a better representation of the earnings distribution. This is the subject matter of this section, which addresses research question 1.

Performing the BRW in NIDS is reasonably complex. As mentioned, PALMS only collects gross earnings, which means one has to work with gross in both datasets. Unfortunately, net income is the variable of focus in the NIDS survey – it has a higher rate of observation and bracket information is collected for it. However, although there is no bracket data collected for gross, one can apply the bracket weights constructed for net to the gross variable *provided it is applied over the same set of observations.*

Creating bracket weights for net requires an appropriate net earnings variable. The net released in NIDS flags bracket responses as survey data which is inappropriate as it is necessary to separate bracket from point data. While undoing this and creating a new flag for the variable it turns out there are several other minor improvements that can be made along the way. Once an appropriate net variable is constructed, I create a gross variable with observations for the same individuals and then perform the BRW. The steps in this section are therefore:

5.1 Create the best possible net wages variable over a set of individuals (called them M)
5.2 Create the best possible gross wages variable with observations over M
5.3 Perform the BRW
5.4 Results

5.1. Creating the Best Possible Net Wage Variable

5.1.1 Available data

Creating the best possible net wage variable requires understanding the data available from the surveys. The situation is complex as NIDS collects primary, secondary, gross, net, and bracket data. Specifically, NIDS collects primary net wages which allow bracket response, and primary gross wages without bracket responses. The exact same is true of secondary wages. However, secondary brackets were few and drop off entirely after wave 2. Altogether, the relevant "raw" wage variables from survey data are available in table 2 below.

			Wave 1		Wave 2		Wave 3		Wave 4	
		Var Name	Code	Count	Code	Count	Code	Count	Code	Count
Gross	Primary Point	em1inc	E8	3019	E10	3485	E10	4364	Eb9	6027
Net	Primary Point	em1pay	E9	3088	E11	3511	E11	4416	Eb10	6145
Net	Primary Bracket	em1inc_cat*	E10	481	E12	507	E12	914	Eb11	547
Gross	Secondary Point	em2inc	E24	45	E29	37	E29	34	Eb22	53
Net	Secondary Point	em2pay	E25	42	E30	41	E30	37	Eb23	55
Net	Secondary Bracket	em2inc_cat*	E26	4	E31	2	E31	0	Eb24	0

Table 2: NIDS Wage Varibles Waves 1-4. *In wave 1 these are called em1inc_sh and em2inc_sh. The 'inc' in bracket info refers to net which is misleading. Observations less than zero excluded.

Individuals may report or fail to report the above six variables in several combinations- in the data 13 possible scenarios were realized. How to process these and arrive at a single net variable requires two choices that lead to three potential options for the net variable.

5.1.2 Choice 1: whether or not to combine primary and secondary

Combining the secondary net information with the primary is useful as it provides more observations and a truer reflection of the distribution of earnings in NIDS. Seeing as bracket response is available for the secondary question the variable is consistent with the BRW approach. A downside of combining the two is that the resulting variable will be less comparable to PALMS, as gross wages in PALMS only represents earnings from a primary job. For this reason, the first option (Option A) is to keep the primary only variable (called net1) and use it as a benchmark for comparison to the case when primary and secondary are combined. Option B is to combine primary and secondary net information (this variable is called (net1_2).

5.1.3 Choice 2: whether or not to impute from gross

Next one needs to decide what to do with the gross data. Considering the plan to perform the BRW, it is useful to conceptualize the response of a working individual as reflecting a point response, a bracket response, or a non-response (missing). So how does one categorize a person who provides no net information, or only a bracket response for net, but then later provides a point response for gross?

There are scenarios where it seems such a gross figure may reflect a valid point response that can be imputed for net and flagged as a point response. For example, some people hire others to file their taxes and deductions and may only be aware of their gross pay. Others might have forgotten their net pay while remembering the gross amount. This could lead them to either report gross point without net, or possibly gross point but net only in brackets. Imputing from gross to net for these cases is essentially going to capture more of these authentic point responses. If one considers gross when net is missing a response, it follows that gross when net is provided as bracket data is also a response (and not a refusal). As such, imputing when gross is available, and net is either undesirable type (missing or bracket) is consistent with the conceptualization of a bracket response as reflecting a *refusal* to the earnings question.

However, a downside of imputing from gross is that it will make the method less comparable to PALMS. This is because PALMS only allows respondents to answer one type earnings (gross). If they cannot remember the point value, they will either provide a bracket for gross or not respond at all. Allowing NIDS respondents two variables within which to respond when memory is hazy but not PALMS means that the response rates will be theoretically different, being based on a different response construct. Therefore, choosing to ignore gross and not impute is a choice more consistent with the BRW method as applied to PALMS.

To summarize, the decision involves choosing between the competing desiderata of being more conceptually consistent (impute from gross point) or being more comparable to PALMS (only consider net). Going forward my approach is to consider both. I perform the imputation (Option C) and compare this new variable (called net1_2gimp) with the other two options from before. Below I discuss the transformations needed to create the three options.

5.1.4 Option A: Net primary only (net1)

This is simply a cleaned version of the *em1pay* variable for each wave.

5.1.5 Option B: Combined net without imputation (net1_2)

Combining the primary and secondary cases is relatively straightforward if there is no imputation to be performed. The only complication is dealing with the bracket data. As discussed, I prefer to avoid imputation by bracket midpoint and rather flag the observations as bracket data for the BRW. However, there are rare cases when combining primary and secondary data where midpoint imputation is useful and will not cause spiking issues. My approach is as follows:

Scenario		Treatment
Primary	Secondary	
Bracket	Bracket	Check if summing midpoints would result in a higher bracket (as explained below). Flag as bracket.
Bracket	Point	The same as above. Flag as bracket.
Point	Bracket	Add the bracket midpoint to primary. Flag as point.

Table 3: Dealing with brackets when constructing net without imputation.

In row 1, primary bracket data is being combined with secondary bracket information. Permuting the different bracket combinations that are possible would produce a finite and exhaustive set of brackets into which each combination must fall. However, the number of distinct brackets would be unfeasibly large, resulting in low response rates which would be problematic for the BRW. A better heuristic is the one mentioned: to calculate whether the individual would be in a higher bracket (than the primary one recorded) if the midpoint of the secondary bracket was added - and adjust accordingly if this is the case. This is the same approach taken in row 2, when there is bracket primary and point secondary. In row 3 I simply add the secondary bracket midpoint to the point primary amount. This is not problematic in terms of spiking due to the variation in the primary point data.

5.1.6 Option C: Constructing net with imputation from gross (net1_2gimp)

The situation is more complicated if one wants to impute from gross– this subsection will explain how one can improve on the imputation from gross that NIDS does.

The NIDS variable with imputation is inappropriate for our purposes. NIDS' approach is relatively simple once the programme files are understood. NIDS combines primary (em1pay) and secondary net (em2pay), and primary and secondary bracket midpoints (fwag_ib), and if this new variable (called fwag) is non-missing, no imputation is performed. When brackets are present the midpoint is added to the fwag variable, which is problematic in the context of the BRW as the bracket information is lost. Further, because NIDS only imputes when net is completely missing, the approach sometimes misses available gross data. For example, if there is a net figure for the secondary job, but there is also gross primary available, NIDS would ignore the primary gross, and the secondary net figure would be the final wage amount. NIDS will also impute for item non-response of a working individual when there is no earnings data whatsoever. As discussed this is not the technique of choice.

To improve on this, I perform a more rigorous approach that considers every possible outcome for the collection of these primary and secondary figures. There were thirteen scenarios realized in the data (all waves), each requiring a different treatment. The conditions that define the scenario and the number of observations in each wave are captured in table 4 below.[9]

	Scenario				Observations			
	Primary		Secondary					
	Net	Gross	Net	Gross	Wave1	Wave2	Wave3	Wave4
a)	bracket	.	bracket	.	1	.	.	.
b)	bracket	.	point	.	.	1	.	.
c)	bracket	.	.	.	428	452	846	503
d)	point	.	bracket	.	2	2	.	.
e)	point	.	point	.	40	35	36	53
f)	point	.	.	point	3	.	.	.
g)	point	.	bracket	point	1	.	.	.
h)	point	.	.	.	3042	3474	4380	6092
i)	.	point	point	.	.	3	.	2
j)	.	point	.	.	54	20	11	14
k)	bracket	point	point	.	.	1	.	.
l)	bracket	point	.	.	23	36	7	12
m)	.	.	point	.	2	1	1	.

Table 4: 13 Scenarios that are used to create net (with and without imputation).

The greyed rows indicate cases where imputation from gross is possible. As is clear, they are low in proportion to the bulk of the data. This raises the question of how the imputation is performed, the subject of the next subsection.

[9] Note that in this table I ignore gross (treat as missing) when there is net point available. For a more detailed account of how I treat these scenarios, see appendix B-III.

5.1.7 How to impute from gross

I consider three options for imputing from gross. The first two are based on the "elasticity" type regression NIDS uses. NIDS uses a univariate OLS regression to fit values for log(net) from log(gross). Seeing as the tax schedule is progressive, there is good reason to expect a non-linear relationship between gross and net. The second imputation technique is therefore to run the same elasticity-type regression, but using a multivariate regression including a quadratic, cubic and quartic term in gross wages.

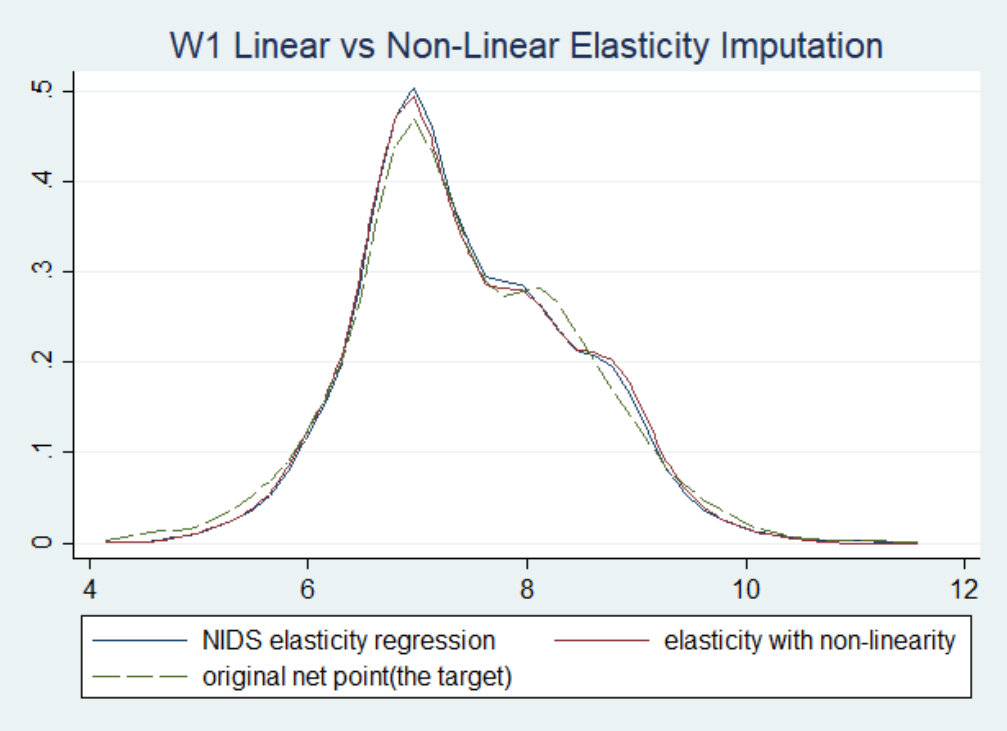


Figure 1a: Linear vs Non-linear elasticity for imputation from gross to net (Wave 1).

Figures 1a (above) and 1b (next page) compare the linear elasticity to the non-linear one for waves 1 and 2.

It appears that the non-linearity provides fitted values marginally closer to the observed figures in wave 1. In wave 2 the improvement is more noticeable, especially around the mode. For waves 3 and 4 the non-linear option provides an improvement of similar magnitude. The non-linear version is therefore preferable to linear elasticity imputation.

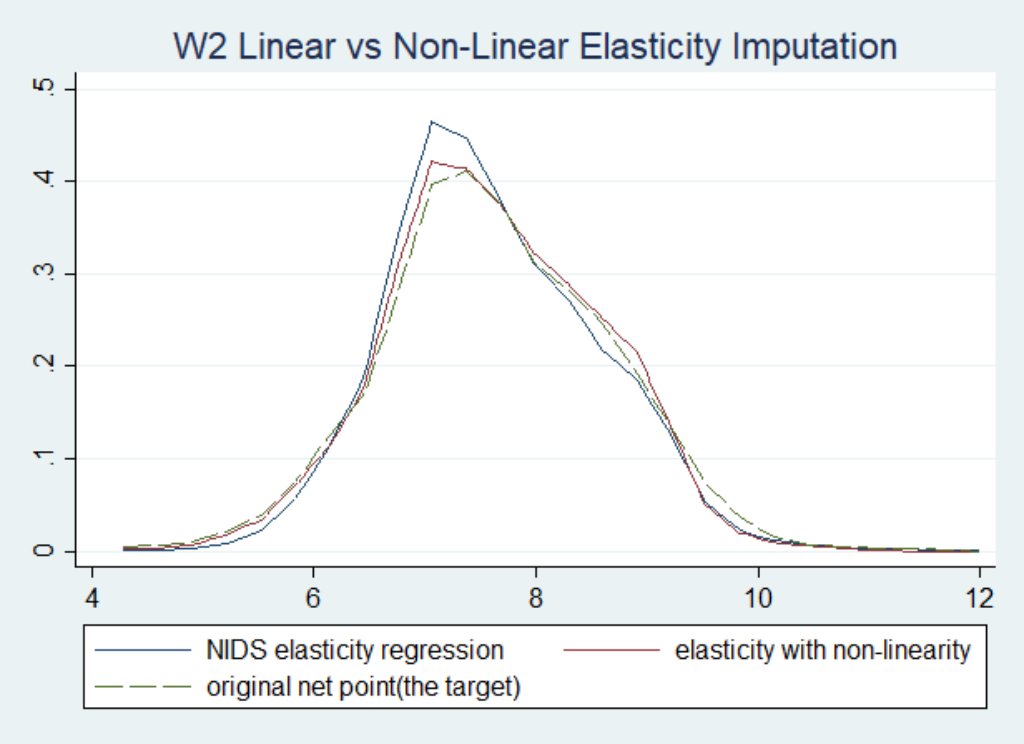


Figure 1b: Linear vs Non-linear elasticity for imputation from gross to net (Wave 2).

The third option is to convert the gross amounts to net figures by running them through the tax table of the appropriate year. The tax tables used are from SARS and are available in Appendix B-I.

Figure 2 (next page) compares the Tax Table (TT) imputation to the Non-Linear (NL) regression-based imputation and the target point values in all four waves. There is an interesting trend. At lower values, roughly anywhere below the mode in each wave, the TT under-predicts net earnings, whilst the NL regression consistently over-predicts it. As such the targeted values lie neatly in the middle. Above a certain point (indicated by the vertical lines) the pattern swaps, and the TT is overpredicting net wages and vice versa.

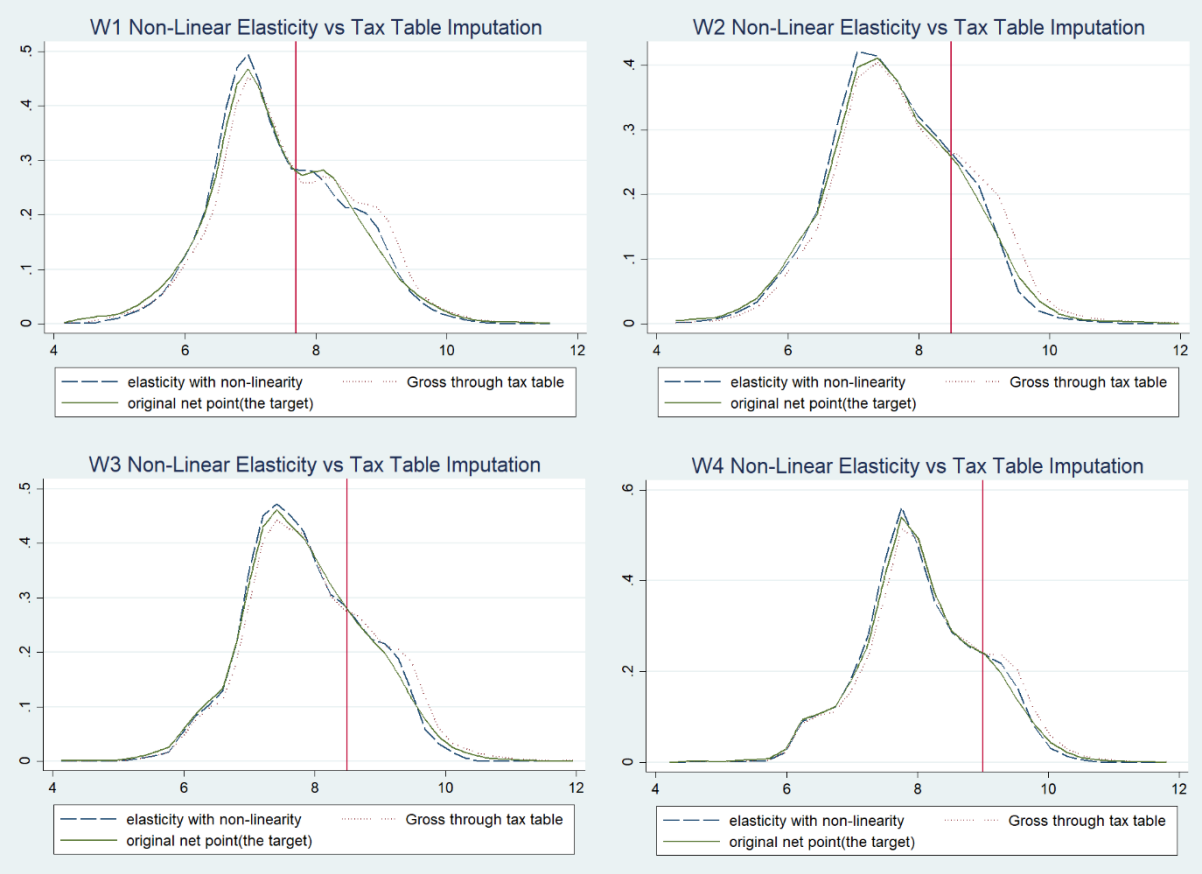


Figure 2: Non-linear 'elasticity' regression vs tax table imputation.

In order to better understand the TT conversion, I run the imputation backwards from net to gross. In this case, imputing backwards through the tax table produces *underestimates* of gross earnings at higher incomes (see Appendix B-V). It therefore seems the TT conversion does not capture the full extent of the gap between net and gross at higher incomes. This might be because non-tax deductions such as medical aid and pension are not being accounted for with this technique. It makes sense that this difference would be accentuated at higher incomes.

Wittenberg (2017) has posited that in PALMS individuals were reporting gross *after* deductions, which if true would contradict the above explanation, as both net and gross would be sans pension and medical aid earnings and the difference between them would theoretically be tax only.[10] However, providing there are enough individuals in NIDS providing honest gross and net figures, the TT would not fully account for the difference, and this explanation would then hold. Given this shortfall and the above discussion, my approach henceforth is to use the non-linear elasticity-type regression when performing imputations.

[10] In this discussion I am ignoring the case of the self-employed, who introduce other complications.

5.1.8 Choosing between the three options

This section compares the three net variables created as described above. Table 5 below provides counts, weighted counts, response rates and weighted response rates for the four waves of NIDS.

			Wave 1		Wave 2		Wave 3		Wave 4	
			Obs	w1_wgt	Obs	w2_wgt	Obs	w3_wgt	Obs	w4_wgt
Count and Weighted Count	net1	Point	3 088	6 510 335	3 511	7 900 104	4 416	9 547 019	6 145	11 718 240
		Bracket	481	1 083 409	507	1 311 252	914	1 841 818	547	1 150 398
		Total	3 569	7 593 744	4 018	9 211 356	5 330	11 388 837	6 692	12 868 638
	net1_2	Point	3 090	6 522 920	3 515	7 905 568	4 417	9 550 438	6 147	11 720 468
		Bracket	481	1 083 409	506	1 305 420	914	1 841 818	547	1 150 398
		Total	3 571	7 606 329	4 021	9 210 988	5 331	11 392 256	6 694	12 870 866
	net1_2 gimp	Point	3 167	6 676 444	3 572	8 054 038	4 435	9 593 690	6 173	11 763 458
		Bracket	458	1 029 595	469	1 193 949	907	1 813 933	535	1 131 971
		Total	3 625	7 706 039	4 041	9 247 987	5 342	11 407 623	6 708	12 895 429
Response Rates	net1	Point	0.865	0.857	0.874	0.858	0.829	0.838	0.918	0.911
		Bracket	0.135	0.143	0.126	0.142	0.171	0.162	0.082	0.089
		Total	1	1	1	1	1	1	1	1
	net1_2	Point	0.865	0.858	0.874	0.858	0.829	0.838	0.918	0.911
		Bracket	0.126	0.134	0.116	0.129	0.170	0.159	0.080	0.088
		Total	1	1	1	1	1	1	1	1
	net1_2 gimp	Point	0.874	0.866	0.884	0.871	0.830	0.841	0.920	0.912
		Bracket	0.126	0.134	0.116	0.129	0.170	0.159	0.080	0.088
		Total	1	1	1	1	1	1	1	1

Table 5: Comparing counts and response rates between three versions of net earnings.

The rate of bracket response as a share of total responses for the net1 variable does not improve (increase) significantly following the implementation of unfolding brackets in wave 2. The share of bracket responses does rise significantly in wave 3, but then drops again in wave 4. Reading off the table, bracket responses are 13% of responses in waves 1 and 2, 17% of responses in wave 3, and only 9% in wave 4. A feature which may have convoluted the response rates between the waves is that the range of the brackets is inconsistent between waves. For example, the top bracket starts at R50 000 in Wave 1, R8000 in Wave 2, R18 000 in Wave 3 and R24 000 in Wave 4. The inconsistency in the bracket response rate between waves will influence the magnitude of the BRW effect in each wave.

As is clear, combing primary and secondary has very little effect. The reason is that there are very few observations with secondary job observations as evidenced in tables 2 and 4 from above. For instance, in Wave 1 there are 3088 point observations for primary, and only 42 for secondary. The number of these where there is secondary and no primary is even fewer – in wave 1 it is just 2. This makes sense as someone should not report a secondary job if they do not report a primary job. As such, the counts and weighted counts are almost exactly the same

as for net1. For this reason, the weighted and unweighted response rates are also very similar between net1 and net1_2.

Imputation improves the number of observations and the response rates, but only marginally. This is because there are very few gross observations where there was not satisfactory net data available in net1 or net1_2. Applying the post-stratified weights to the data has practically no effect on the point response rates. In some cases it increases and in others it decrease but the changes are marginal.

What remains is to select between the three versions. Little hinges on this choice as they are very similar. Seeing as little is gained by imputation from gross, and that it was previously concluded that net without imputation is more consistent with PALMS, it seems safe to drop the imputed version. In the same vein, seeing as so little is gained from adding the secondary information in NIDS, and that PALMS does not collect secondary job data, I opt to drop secondary earnings from NIDS. Therefore, I go forward with net1 for the BRW and drop the other two alternatives.

5.2. Constructing the best possible gross over M

5.2.1 Imputation in the other direction

As discussed, a set of observations M for which there are a gross and net values is necessary to apply the BRW. This requires imputing for gross point data when it is missing and net point (net1) is available. I therefore run a 'backwards' (net to gross) NL regression to fit the necessary gross values. This applies to close to 100 observations per wave. The result is a set of individuals who, if there are net or gross wages, have a value for both recorded.

5.2.2 A note on cases where gross is equal or very close to net

In NIDS there is a high proportion of cases where the net figure is exactly equal to, or very close to, the gross figure from the survey. This is largely because most of the observations (70% on average) are below the bottom tax bracket. However, even above the tax bracket, there are around 8% of cases where gross and net are very similar. I define similar to mean being less than 5% apart. Table 6 below (following page) captures the proportion of cases that have the exact same or similar net and gross figures, above and below the bottom tax threshold, for each wave.

		Wave 1	Wave 2	Wave 3	Wave 4
Below Tax Threshold	Total	2157	2457	3136	4227
	Same or Close	1368	1437	2011	2876
	-Proportion	0.63	0.58	0.64	0.68
Above Tax Threshold	Total	987	1078	1292	1934
	Same or Close	83	100	131	187
	-Proportion	0.08	0.08	0.09	0.09
Overall	Tax Threshold	3833	4750	5296	5891
	-Proportion	0.69	0.70	0.71	0.69

Total above and below	3144	3535	4428	6161

Table 6: Cases where gross is very similar to net in NIDS.

It does not make sense that figures above the threshold can be the same or very close - some form of error must have occurred to explain this. There are reasons to assume that when gross is recorded as roughly equal to net, they are both reflecting a true net value. This is justified by a few observations. Firstly, when gross and net are very similar and above the threshold, they average around R9107 monthly pay. This is closer to the average of net figures (R9300) than the average of gross figure (R18000) when they are above the threshold.[11] Secondly, NIDS collects net more carefully and it appears first in the questionnaire - a lazy respondent or surveyor may simply repeat the net figure for gross. Lastly, Wittenberg (2014) has proposed that respondents are giving net figure amounts when asked for gross. Therefore, it might provide a truer reflection if one adjusts the gross upwards in these cases. However, because the same adjustment cannot be made in PALMS (which only collects gross), I opt not to.

5.3. Performing the BRW on NIDS

The BRW technique was then performed on the net data to create bracket weights that are comparable to those used in PALMS. This section details how this is performed.

The wave 1 technique is the standard approach for fixed brackets as explain in Wittenberg (2008); the point data are weighted up by the inverse of the likelihood of a point response in each earnings bracket. This assigns the weight of the bracket responses to the point data, leaving the total (weighted) number of respondents the same. The bracket observations are then dropped from the analysis.

In wave 2 NIDS switches to the "unfolding brackets" approach. This created several "brackets" that were not a range but rather a single point value, where respondents said their earnings were "close to X" amount. Having discrete brackets could work hypothetically but the result in this case was to create large discrete changes in the response probability resulting in spikes in the density plots. Part of the appeal of BRW was supposed to be that one can avoid creating spikes in the first place. Their presence in the questionnaire also seems somewhat illogical, as it is unclear why people would refuse to give a point value but then when asked later for bracket response be happy to indicate a single-value bracket.

A somewhat novel approach[12] is thus necessary to deal with the unfolding brackets that were introduced to NIDS in wave 2 and the discrete single-value brackets. The approach has been to apportion the weight of these discrete "brackets" into the normal brackets on either side of them. The weight of normal (non-discrete) bracket responses is then assigned to the appropriate point responses according to their value in the normal BRW technique described above for wave 1. For further detail on this technique, the brackets used and response probabilities, see the Appendix B-IV.

[11] There is an additional programme file called "information for appendices" which reveals these results.
[12] Suggested to me by Martin Wittenberg.

5.4. Response Rates Among the Population

5.4.1 Overall response rates

To summarize the net variable of choice for the BRW is net1 (Option A) and there is a corresponding gross value for each individual in M. The interest is in examining them over the "primarily employed" population. To get a better sense of the observations and response rates, Table 7 below considers counts and weighted counts amongst the population.

Firstly, I provide breakdown of responses by response type and earner type. I also divide non-responders by earner type, but this is based on whether or not there was *any* self-employment (there is no point data to consider which is primary in this category). Response and non-response together total the working population, which added to the non-working population gives the total population estimate.

				Wave 1		Wave 2		Wave 3		Wave 4	
				count	w1_wgt	count	w2_wgt	count	w3_wgt	count	w4_wgt
Population	Working	Answering	Point: pri employed	3 082	6 495 160	3 505	7 879 684	4 412	9 542 505	6 124	11 689 804
			Point: pri self-employed	6	15 176	6	20 419	4	4 515	21	28 436
			bracket	481	1 083 409	506	1 305 420	914	1 841 818	547	1 150 398
			Total anwering	3 569	7 593 744	4 017	9 205 524	5 330	11 388 838	6 692	12 868 638
		No Answer	no self-employment	2 399	4 082 998	950	2 024 177	1 195	2 007 026	1 566	2 041 463
			any self-employed	808	1 675 348	570	1 379 123	711	1 518 466	1 011	2 041 463
			Total no answer	3 207	5 758 346	1 520	3 403 300	1 906	3 525 491	2 577	4 547 189
		Total Working		6 776	13 352 090	5 537	12 608 824	7 236	14 914 329	9 269	17 415 827
	Total Not Working			21 453	35 947 995	25 747	41 351 835	30 086	44 049 088	35 682	47 008 079
	Total Pop			28 229	49 300 085	31 284	53 960 659	37 322	58 963 417	44 951	64 423 906

Table 7: Comparing response-types in the population. Pri stands for primarily.

As is clear, very few individuals are flagged as primarily self-employed by my definition. This is likely because most of the self-employed have already been excluded from the sample, having not had or reported primary job earnings. Again, only those with self-employment earnings less their primary job earnings would remain in the "point response primarily employed" category (row 1). These indivduals are presumably working for themselves part-time to supplement their primary earnings.

A notable feature of the table is the jump in bracket responses in wave 3 (in grey). Seeing as there is no corresponding decline in a different category, it is unclear where these responses are coming from and why so many more individuals chose to provide bracket information in this wave. In wave 4 the bracket responses fall again, and there is instead a much higher number of point responses.

We can get a finer sense of the distribution between response types by considering them in terms of their relative proportions, provided in table 8 below. The rows of table 8 sum to 1 according to the logic of table 7 above. For reference: rows 1-3 (boxed) sum to the total answering population, while rows 5-6 (boxed) sum to the total non-answering population. In turn, rows 4 and 7 (thick underline), the answering and non-answer groups, sum to the total working population. Rows 8 and 9 (double bottom line), the working and non-working groups, sum to the total population.

					Wave 1		Wave 2		Wave 3		Wave 4	
					count	w1_wgt	count	w2_wgt	count	w3_wgt	count	w4_wgt
1	Population	Working	Answering	Point: pri employed	0.86	0.86	0.87	0.86	0.83	0.84	0.92	0.91
2				Point: pri self-employed	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3				bracket	0.13	0.14	0.13	0.14	0.17	0.16	0.08	0.09
4				Total anwering	0.53	0.57	0.73	0.73	0.74	0.76	0.72	0.74
5			No Answer	no self-employment	0.75	0.71	0.63	0.59	0.63	0.57	0.61	0.45
6				any self-employed	0.25	0.29	0.38	0.41	0.37	0.43	0.39	0.45
7				Total no answer	0.47	0.43	0.27	0.27	0.26	0.24	0.28	0.26
8		Total Working			0.24	0.27	0.18	0.23	0.19	0.25	0.21	0.27
9		Total Not Working			0.76	0.73	0.82	0.77	0.81	0.75	0.79	0.73
10	Total Pop				1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 8: Comparing response-types in the population by proportion. Pri stands for primarily.

The rate of bracket response (row 3) is very similar to the net1 and net1_2 variables discussed earlier in table 5, which is expected given the tiny proportion of self-employed that have been separated out. The jump in bracket response rates in wave 3 is highlighted again and the decline in wave 4 is still apparent. The proportions suggest that the bracket responses in wave 3 are mostly coming from a decline in the proportion of people providing point estimates, with some also coming from an overall increase in the relative proportion answering as opposed to refusing (from .73 in wave 2 to .76). Conversely, in wave 4 the point response rate jumps to .91 and this is not driven by a higher non-refusal rate. These proportions provide clarity on distribution between response types, but it remains unclear what is driving these changes.

A final noticeable feature of table 8 is the lower overall rate of response (row 4) in wave 1, at 57% of working individuals, compared to the subsequent waves which average around 74%. This is like due to a peculiarity from being the first wave. Wave 2 and beyond changed to the unfolding brackets technique, but from table 8 it does not seem that this elicited a higher bracket response rate. In the surveys wave 1 included non-residents as temporary survey members which subsequent waves did not. However, these were dropped in data preparation. As such it is unclear to me why the total rate of response is higher in subsequent waves.

5.4.2 Response rates by bracket

Table 9 below tabulates the factor by which the post-stratified weight will go up per income bracket when the BRW is applied to each wave (I call this the "rescale factor"). This is determined by the weighted response rates per bracket. For reference the straight count is also

provided.[13] Note that the rows represent an ordering of the brackets of each wave and not fixed intervals across all four waves. For example, the final bracket is the 11th bracket in wave 2, representing earnings above R8 000. The final bracket in wave 4 is the 13th bracket, representing earnings above R24 000. The ordering is a useful way to combine the disparate brackets seeing as it is the comparison of response rates that is of primary interest.

Following wave 2, every second row is highlighted to indicate the troublesome discrete "brackets" representing answers of "around X" amount. In these cases the rescale factor is always near to 1 as the adapted BRW technique deliberately takes the extra bracket weight and apportions it elsewhere.

The results show that there is a somewhat linear trend of decreasing point response rates, and therefore increasing rescale factors (colour-coded), as the bracket number increases. This is in line with the stylized fact that the wealthy are more likely to respond in brackets. Notably, the response rates are far lower in the highest brackets of wave 1 than elsewhere, resulting in a greater rescale factor and thus a bigger BRW effect. This might indicate that the highest brackets need to be nominally far above the rest before the wealthy begin to feel uncomfortable disclosing their point values. This would also explain why the later waves, which have lower nominal brackets (at wave 4 the top bracket is still only R24 000 as opposed to R50 000 eight years' prior), do not see a precipitous drop in response rates in the same way.

	Observations by Bracket								Response Rate and Rescale Factor by Bracket							
	Wave 1		Wave 2		Wave 3		Wave 4		Wave 1		Wave 2		Wave 3		Wave 4	
Bracket Number	Point	bracket	point	bracket	point	bracket	point	bracket	response rate	rescale factor	response rate	rescale factor	response rate	rescale factor	response rate	rescale factor
1	64	11	439	15	224	12	388	8	0.83	1.21	0.95	1.06	0.91	1.11	0.98	1.03
2	280	25	56	5	42	28	3	14	0.91	1.10	0.92	1.01	0.72	1.02	0.19	1.01
3	749	72	206	12	682	55	678	14	0.90	1.11	0.93	1.12	0.95	1.10	0.99	1.03
4	522	58	178	32	76	107	196	50	0.92	1.09	0.82	1.03	0.52	1.03	0.81	1.02
5	441	67	713	63	1672	139	1842	38	0.91	1.10	0.91	1.18	0.94	1.16	0.98	1.08
6	283	42	75	80	16	218	196	163	0.87	1.15	0.44	1.05	0.18	1.06	0.57	1.04
7	209	33	887	117	791	71	1324	42	0.90	1.11	0.89	1.20	0.93	1.24	0.96	1.12
8	192	37	66	45	9	99	100	74	0.85	1.18	0.68	1.03	0.04	1.09	0.58	1.03
9	142	36	528	49	582	62	896	35	0.87	1.15	0.92	1.14	0.92	1.24	0.96	1.11
10	91	29	47	34	31	51	63	49	0.75	1.34	0.66	1.03	0.40	1.06	0.62	1.03
11	65	36	316	54	195	30	376	24	0.66	1.52	0.79	1.29	0.85	1.30	0.92	1.17

[13] For full tables showing counts, weighted counts, response rates, and the normal and adapted BRW technique, see Appendix B-IV.

12	38	19			15	15	9	20	0.67	1.49			0.61	1.06	0.25	1.05
13	7	14			81	27	76	16	0.29	3.48			0.79	1.34	0.83	1.25
14	5	2							0.34	2.93						

Table 9:Response rates and rescale factor by bracket.

As a final check one can consider the increase in the total weight of the point data post-BRW vis-à-vis the total before the BRW. The new point weight total will be equal to the sum of the total bracket response and point response weights before the BRW.

Wave	Total weight of points (wgt)	Total weight of points (BRW)	Factor
1	6 510 335	7 593 744	1.17
2	7 942 567	9 247 987	1.16
3	9 547 020	11 388 838	1.19
4	11 718 239	12 868 637	1.10

Table 10: Total increase in point weights per wave.

5.5. Results

In this section I consider the effect of the BRW on means, percentiles and density plots over the sample.

5.5.1 Estimates of the mean

I estimate the mean of NIDS gross wages in nominal and real terms over the four waves, each including a 95% confidence interval and standard errors. The estimates take account of the complex survey design using the svyset command which produces the correct standard errors (calculated using standard Taylor-linearized variance estimation).

As quality checks I estimate the mean without outliers (calculated as explained in 3.5), as well as providing another control which is to exclude the top 10 gross values in the data (some of which may have been flagged as outliers and already excluded). This helps to reveal the impact of these high values in terms of the mean estimation, as well as the effect these "super-earners" have on the BRW.

		WGT				BRW			
		Wave 1	Wave 2	Wave 3	Wave 4	Wave 1	Wave 2	Wave 3	Wave 4
real gross wages	Mean	6078	15665	9484	12809	7163	16457	9967	13364
	Se	(491.4)	(4786.5)	(1144.8)	(1937.6)	(752.7)	(5179.2)	(1252.9)	(1998.5)
	min	5114	6274	7238	9007	5686	6296	7509	9443
	max	7042	25055	11730	16610	8639	26617	12425	17284
- w/o outliers	Mean	6078	8902	9489	11075	7163	9299	9972	11612
	se	(491.4)	(857.9)	(1145.3)	(1064.4)	(752.7)	(923.3)	(1253.4)	(1163.5)
	min	5114	7218	7241	8987	5686	7487	7512	9329
	max	7042	10584	11735	13163	8639	11110	12430	13894
- w/o outliers or top 10	Mean	5 884	8 533	8 407	10 094	6860	8894	8773	10545
	se	(482.84)	(798.32)	(695.59)	(872.82)	(737.7)	(857.7)	(751.6)	(965.5)
	min	4936	6966	7042	8382	5412	7210	7298	8650
	max	6831	10098	9772	11806	8307	10576	10247	12438

Table 11a:Estimation of mean real gross wages.

Table 11a above provides estimates for real gross wages. The left panel does so using the post-stratified weights (WGT). Over the wave 1 period (2008) the estimated mean is R6 078 and this rises to R12 809 by wave 4 (2014). Wave 2 is anomalous with a mean gross wage of R15 665 using the post-stratified weights. This is the result of an extreme observation of R12 million per month which is doubtless an erroneous value.

Removing the outliers drastically reduces the mean estimate in this case, from R15 665 to R8 902. In other waves the outlier regression flagged fewer observation – in wave 1 there were no outliers flagged and thus no change when outliers are removed. In wave 3 the mean estimate actually *increases,* which reflects the fact that the outlier regression flags both unexpectedly high *and* unexpectedly low values. The bottom third of the table considers the effect when the highest 10 observations are removed. In some waves this makes little difference as the outlier regression had already flagged these values. For example, in wave 2 most of the top 10 observations were picked up by the outlier regression, thus dropping the 10 top produces a similar effect. In wave 1 dropping the top 10 has a greater effect as the outlier regression found no outliers.

Table 11b below provides the same means and with these the relative changes produced by the quality adjustments. For example, in wave 2 the estimated mean is reduced by 45.5% when outliers are removed. The "a%c" row refers to the accumulated percentage change in mean from the original (top row) mean estimate to the case with outliers *and* the top 10 removed. On average, the accumulated change represents around a 20% reduction in the estimated mean, when outliers and the top 10 are dropped (but it varies significantly).

		WGT				BRW			
		Wave 1	Wave 2	Wave 3	Wave 4	Wave 1	Wave 2	Wave 3	Wave 4
gross wages	Mean	6078	15665	9484	12809	7163	16457	9967	13364
	BRW					17.9%	5.1%	5.1%	4.3%
- w/o outliers	Mean	6078	8902	9489	11075	7163	9299	9972	11612
	%c	0.0%	-43.2%	0.0%	-13.5%	0.0%	-43.5%	0.0%	-13.1%
	BRW					17.85%	4.46%	5.09%	4.85%
- w/o outliers or top 10	Mean	5884	8533	8407	10094	6860	8894	8773	10545
	%c	-3.2%	-4.1%	-11.4%	-8.9%	-4.2%	-4.4%	-12.0%	-9.2%
	a%c	-3.2%	-45.5%	-11.4%	-21.2%	-4.2%	-46.0%	-12.0%	-21.1%
	BRW					16.6%	4.2%	4.3%	4.5%

Table 11b: Changes in estimation of mean real gross wages under different data quality adjustments.

The right-hand panel of both 11a and 11b represents the means when the BRW weights are applied. As expected, the estimated mean increases when the technique is performed. The effect is far more pronounced in wave 1, however. The greyed rows in 11b capture the percentage increase in the mean estimate when the BRW is applied. The mean in Wave 1 increases by 17.85% from R5 884 to R6 860 with outliers removed. In the other three waves the estimate increases by around 5%. Dropping the top 10 tends to reduce the scale of the BRW, but only marginally. In wave 1 the change in the BRW effect when the top 10 observations are dropped is most pronounced, falling from a 17.85% increase in the mean to one of 16.5%. Overall the BRW is far strong in increasing the mean in wave 1.

Reflecting back to table 9, there were only 16 observations in the top two brackets in wave 1. Seeing as these two brackets had far lower response rates than the others, these few individuals were likely having a large impact on the (re-weighted) distribution and the estimated mean. As previously discussed, it is possible that these are valid responses, and that wave 1 was the only wave with appropriately high earnings brackets. If this is the case, it would mean the subsequent waves are under-estimating the mean even when the BRW is performed.

The real gross variable is simply the nominal amounts multiplied by one scalar "deflator" per wave.[14] As such the proportions in table 11b will remain unchanged looking at nominal values, while the means and the associated confidence intervals and standard errors will also be scaled down. For completeness table 11c below provides the nominal gross mean estimates.

		WGT				BRW			
		Wave 1	Wave 2	Wave 3	Wave 4	Wave 1	Wave 2	Wave 3	Wave 4
nominal gross wages	Mean	5502	14180	8585	11595	6484	14897	9023	12098
	se	(444.8)	(4332.8)	(1036.3)	(1754.0)	(681.3)	(4688.3)	(1134.2)	(1809.1)
	min	4629	5679	6552	8154	5147	5699	6797	8548
	max	6374	22680	10618	15036	7820	24094	11247	15646
- w/o outliers	Mean	5502	8058	8589	10026	6484	8417	9027	10512
	se	(444.8)	(776.6)	(1036.7)	(963.5)	(681.3)	(835.8)	(1134.6)	(1053.2)
	min	4629	6534	6555	8135	5147	6777	6800	8445
	max	6374	9581	1062	11915	7820	1005	11252	12577
- w/o outliers or top 10	Mean	5326	7724	7611	9138	6210	8051	7941	9545
	se	(437.1)	(722.7)	(629.7)	(790.1)	(667.8)	(776.5)	(680.4)	(874.0)
	min	4468	6306	6375	7587	4899	6527	6606	7830
	max	6183	9141	8845	1068	7520	9573	9276	11260

Table 11c: Estimation of mean nominal gross wages.

5.5.2 Percentiles of the distribution

Below I tabulate the distribution of the real gross variable according to common percentiles (5, 10, 25, 50, 75, 90). Again, the right-hand side represents the same data when the BRW weights are applied. The bottom right quadrant captures the percentage that the wage at percentile x rises when the BRW is applied.

	WGT				BRW			
	Wave 1	Wave 2	Wave 3	Wave 4	Wave 1	Wave 2	Wave 3	Wave 4
P5	663	884	1 105	1 326	663	972	1 215	1 326
P10	939	1 326	1 547	1 758	972	1 326	1 642	1 768

[14] Note that in this context because the data are dated before the base year (2016) the deflator is really "inflating" in the sense that real wages are higher than nominal wages.

P25	1 436	2 209	2 209	2 762	1 547	2 209	2 430	2 872
P50	3 093	4 971	4 419	5 292	3 314	4 971	4 872	5 524
P75	7 733	9 942	9 942	12 152	8 838	11 047	10 274	13 256
P90	14 361	18 780	19 885	23 199	16 571	19 885	20 437	24 303
P95	18 780	27 618	29 533	33 141	23 199	29 827	32 036	35 350
P5					0.00%	9.99%	10.00%	0.00%
P10					3.53%	0.00%	6.14%	0.56%
P25					7.69%	0.00%	10.00%	4.00%
P50					7.14%	0.00%	10.25%	4.38%
P75					14.29%	11.11%	3.33%	9.09%
P90					15.38%	5.88%	2.78%	4.76%
P95					23.53%	8.00%	8.48%	6.67%

Table 12:NIDS real gross wages by percentiles of the distribution.

The percentiles unadjusted show significant real wage gains over the four waves. The bottom three common percentiles (5, 10 & 25) each roughly double over the period. The median earner's wage increases to around 1.7x its starting value. The top three percentiles increase by around 1.6x to 1.7x their starting values. As such, the unadjusted percentiles suggest somewhat comparable growth throughout the distribution, with compression at the bottom.

The bottom-right hand quadrant considers the effect of the BRW in scaling up the respective percentiles. Only wave 1 shows a neat linear increase in the estimates as the percentile increases. Wave 3 is somewhat contrary to what one expects from the BRW, with the middle percentiles increasing more that the upper and lower ones. The BRW effect in waves 1, 2, and 4 will cause a widening of inequality in each wave (as opposed to the post-stratified weights) by increasing the top of the distribution by a greater factor than the bottom, as expected.

5.5.3 Density plots – comparing different data quality techniques

In this section I consider the effect of different data quality adjustments on the kernel density plot of *nominal* net wages. The usefulness of examining nominal net is that one can compare the new variable to the (nominal net) variable released by NIDS. Figure 3 (next page) compares the following four variables over the sample as discussed in 4.2:

1. Point data only (net1)
2. The same point data using bracket midpoint imputation
3. Full imputation as released by NIDS
4. BRW performed on 1

Cases 1,2 and 3 are weighted with the post-stratified weights.

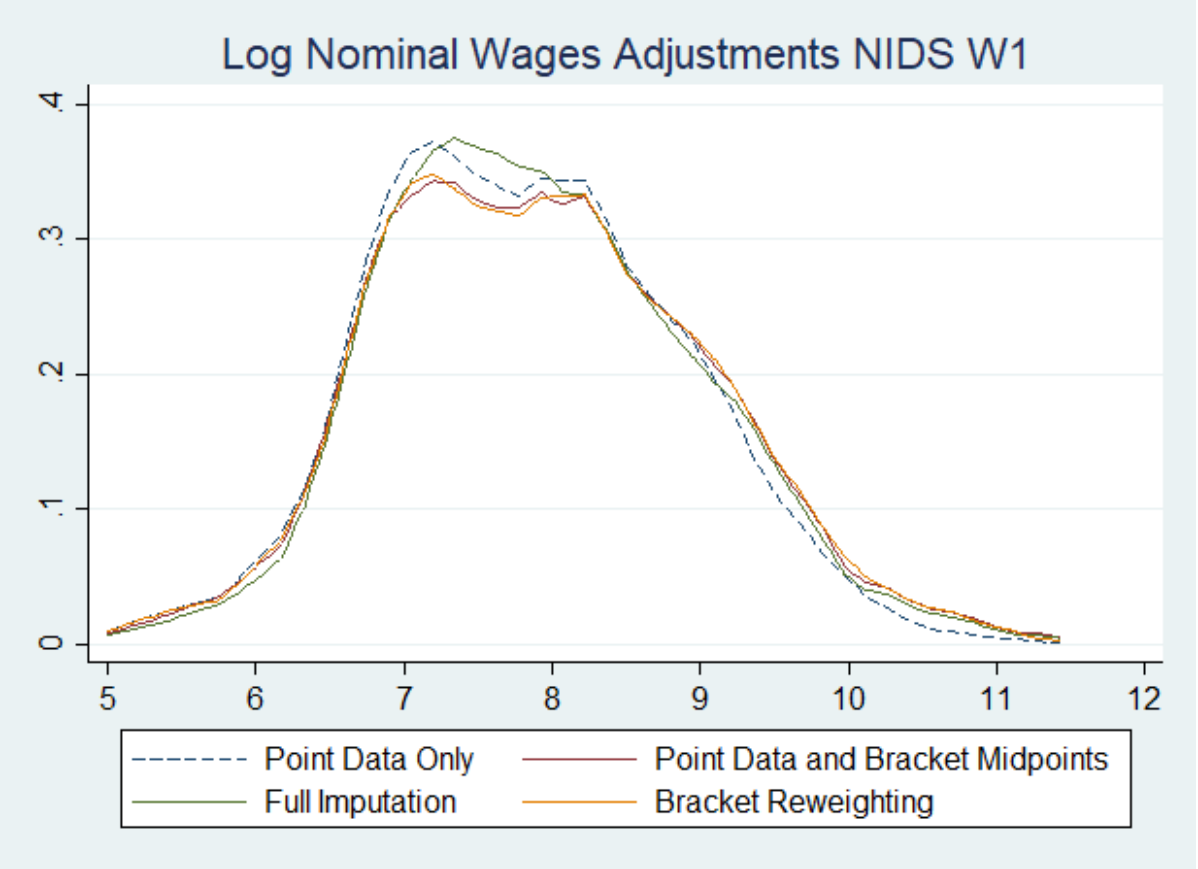


Figure 3: Effect on different data quality adjustment on nominal net wages wave 1.

The results corroborate the discussion in section 3.4. The full imputation tends to repeat existing patterns, thus lowering the variance in the distribution and making it appear taller (around the mode) and slightly narrower. The bracket midpoint imputation adds some high-earnings, shifting the weight outwards and to the right, spreading the distribution and lowering the share that is near the mode. The BRW technique provides a very similar effect to midpoint imputation, which makes sense as it is based on the same underlying data. However, the BRW has a marginally stronger effect in widening the distribution.

As discussed in 4.3, the type of imputation done by Stats SA is unclear. If Stats SA imputed for brackets using midpoints, the BRW technique applied to NIDS might make NIDS results *more* comparable to (pre-imputed) PALMS given the similarity of their effects. As such, it prudent to compare the results using both the BRW and the post-stratified weights.

5.5.4 Density plots – effect on real net and gross

This section compares the total effects of data quality measures on the real net and gross variables.

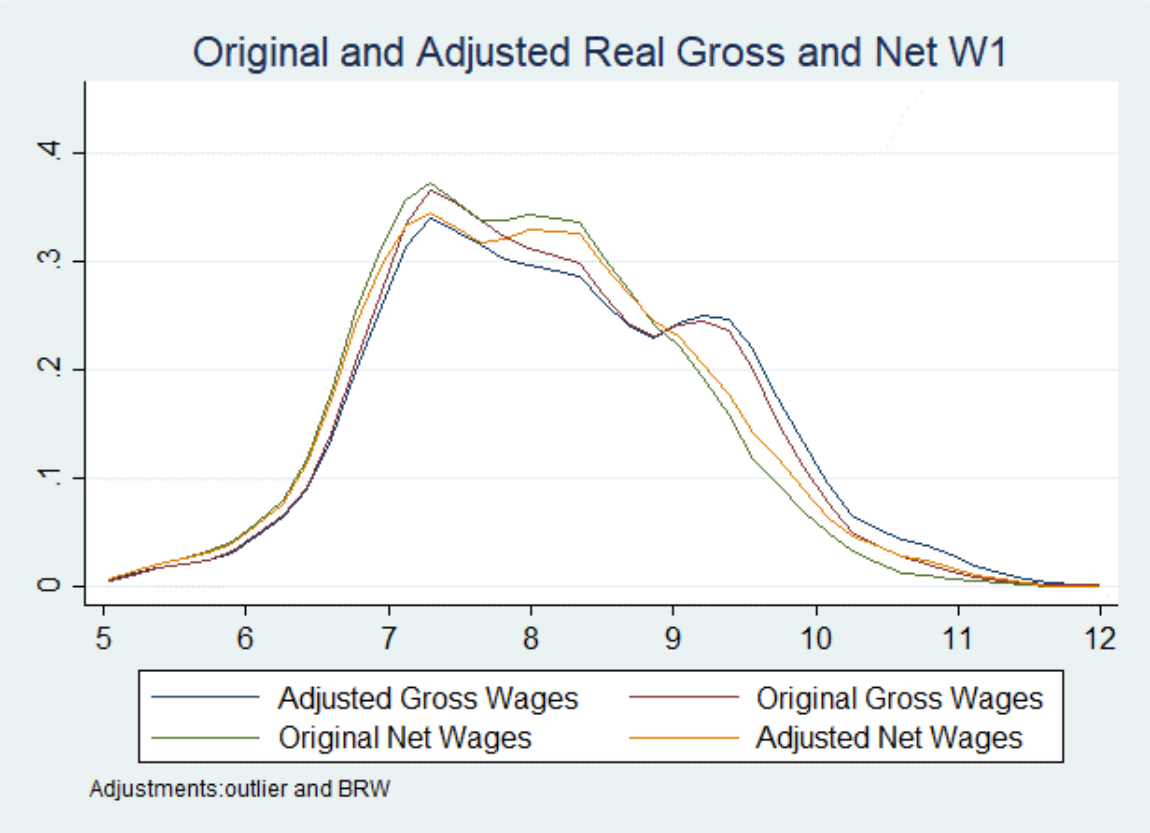


Figure 4: Wave 1 original and adjusted NIDS real gross and net wages.

Figure 4 compares the distributions of the wage data before and after adjustments, these being the new outlier procedures and the performing of the BRW.

The results are as follows. Adjusted and original gross is more spread out and higher in value than net as expected. For both gross and net, the adjustment process (which includes the bracket reweighting) creates additional variance in the distribution; the adjusted curves are more spread out than their counterparts. The effects of each adjustment can be considered separately but this reveals nothing surprising; the changed outlier flag makes little difference, while the BRW drives most of the change. Naturally, because the same BRW weights are used for gross and net, the effect of the BRW is the same on the gross and the net curves.

The picture above is much the same over the next three waves of NIDS, although the BRW technique had a smaller effect in waves 2, 3 and 4[15]. This is due to the far lower rates of bracket response as discussed in section 5.4.2.

The bracket reweighting has a similar effect in PALMS as evidenced in figure 5 (next page); higher values increase in weighted frequency whereas the lower values decrease. The effect is quite dramatic and far greater in magnitude than the change that happened in NIDS. This is partly to be expected given the higher rate of bracket response found in PALMS as opposed to NIDS. For example, the weighted bracket response rate in NIDS Wave 1 was 0.14 (table 9)

[15] These are included in appendix C-I.

whereas in PALMS it is 0.21. Note also that in certain other periods, such as 2010/2011, the BRW has no effect in PALMS. This is because the Stats SA released the data with bracket midpoint imputation as discussed in 4.3.

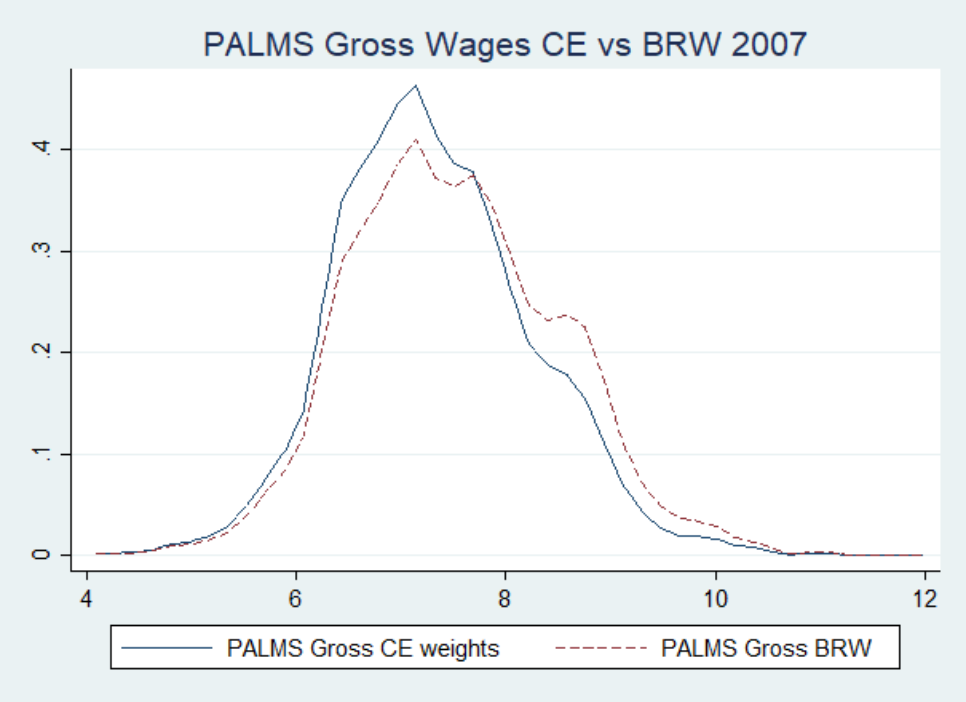


Figure 5: Effect of BRW on PALMS 2007.

6. RQ2: Comparing Wages in NIDS and PALMS

This section answers research two: are wage distributions in NIDS and PALMS similar when using comparable estimation methods? I begin with a discussion of the theoretical justification for this comparison and possible pitfalls, before turning to results.

6.1. Theoretical Justification Using Weights

There is an important distinction between NIDS and PALMS in that NIDS is a panel dataset that collects information on the *same* individuals over time whereas the surveys within PALMS make no effort to follow specific individuals. While this is true, both NIDS and PALMS release sets of weights that aim to make each cross section nationally representative by scaling the observations within the sample upwards by an appropriate amount. As such, both can be used to provide cross-sections of South African earnings data. Currently NIDS has released four waves of survey data: 2008, 2010-2011, 2012, 2014. The relevant earnings variable in PALMS covers 1993-2014 although it omits 2008 and 2009. One can thus compare earnings over the following cross-sections:

NIDS	PALMS (closest match)
2008	2007
2010/11	2010 and/or 2011
2012	2012
2014	2014

Using the most appropriate weights is critical. Both NIDS and PALMS release post-stratified weights estimated using a cross-entropy approach (NIDS Wave 1 Technical Paper 2; Kerr &

Wittenberg, 2017). The weights in PALMS are post-stratified and have additionally been adjusted to provide consistent trends from survey to survey in the series (as well as having other benefits).[16]

Post-stratified weights are most useful for the purposes of creating comparable cross-sections in NIDS and PALMS. Naturally, the quality of post-stratification depends on that of the external (auxiliary) population data used. In PALMS the cross entropy (CE) weights are post-stratified according to revised population totals using the ASSA 2008 model. The same is not true of NIDS, which relies on Stats SA census data. As such a discrepancy may arise due to differing population totals. Accounting for this discrepancy is beyond the scope of this paper. Therefore, the post-stratified weights (called WGT) are preferred and adopted.

Note that when I discuss the "waves" of PALMS it refers to a pooling of the available data corresponding to the given wave in NIDS. In PALSM waves 2-4 this results in a pooling of QLFS data. The QLFS does not select new individuals in each quarter but rather uses a rotating panel over the course of a year. As such, it is necessary to divide the weight of each observation by four, while clustering the standard errors by enumeration areas. This accounts for the correlation between repeated observations to provide more accurate errors that are not artificially deflated.

6.2. Results

6.2.1 Estimates of the mean

A useful starting point is to compare estimates of the mean in NIDS and PALMS, calculated in the same manner as before.

		WGT				BRW			
		Wave 1	Wave 2	Wave 3	Wave 4	Wave 1	Wave 2	Wave 3	Wave 4
NIDS real gross wages	Mean	6078	8902	9489	11075	7163	9299	9972	11612
	se	(491.4)	(857.9)	(1145.3)	(1064.4)	(752.7)	(923.3)	(1253.4)	(1163.5)
	min	5114	7218	7241	8987	5686	7487	7512	9329
	max	7042	10584	11735	13163	8639	11110	12430	13894
PALMS real gross wages	Mean	6236	8407	8687	9975	7971	8407	8687	9975
	se	(189.4)	(89.9)	(68.9)	(168.5)	(307.8)	(89.9)	(68.9)	(168.5)
	min	5864.23	8687.1	8552.02	9645.26	7367.95	8687.1	8552.02	9645.26
	max	6607	8584	8822	10306	8575	8584	8822	10306

Table 13: Mean estimates, standards errors and 95% confidence intervals NIDS vs PALMS. No outliers.

In table 13 above I compare means over the four periods, including standard errors and 95% confidence intervals as before. The means are encouragingly similar in wave 1: NIDS estimated

[16] See Branson & Wittenberg (2011) for a detailed explanation of what the PALMS post-stratified weights (called "CE weights" in the data) achieve and how these weights were constructed. For a detailed explanation of the data issues and inconsistencies between the surveys, and techniques used to minimize the effect on the analysis, see Wittenberg (2016).

mean is R6104 and the PALMS counterpart is R6236. Between 2008 and 2014 the NIDS mean rises faster than the PALMS estimation: by wave 4 it is R11 108 whist the PALMS counterpart is only R9 975.

The right-hand side panel provides the estimates using the bracket weights. NIDS means increase in the order of 5% for each wave as discussed in table 9. The PALMS estimate increases by as much as 28% when the BRW is applied over the wave 1 period. There is no change whatsoever in the PALMS means following wave 1 due to the data quality issues discussed in section 4.3. This exacerbates the difference between NIDS and PALMS on the right-hand side as PALMS has pre-adjusted values for item non-response, and not the preferred BRW technique.

It is unclear why the NIDS mean using post-stratified weights (or the entire distribution for that matter) appear to be rising faster than the corresponding PALMS measure over the same sample. One conceivable explanation is that there is a selection-bias mechanism occurring in NIDS due to the panel format that does not affect PALMS. If there was a variable that correlates positively with both income and the likelihood of answering the survey a second, third or fourth time, NIDS would produce higher estimates of income if this was not controlled for. For example, an unobserved variable such as patience might explain both success in the workplace and a higher rate of response. This would not affect PALMS in the same way as PALMS samples randomly each year (after taking into account survey design). An observed variable could also cause the problem if it was not accounted for by the post-stratification. For example, more educated individuals may be more likely to respond (less cognitive effort) and have higher incomes. Using the panel weights for attrition is not an option as it is not comparable to PALMS.

6.2.2 Percentiles of the distribution

The trends between NIDS and PALMS distributions are revealed more clearly by percentiles of the distribution as captured in table 14 below.

	BRW							
	Wave 1		Wave 2		Wave 3		Wave 4	
	NIDS	PALMS	NIDS	PALMS	NIDS	PALMS	NIDS	PALMS
P5	663	1 090	884	1 074	1 105	1 009	1 326	651
P10	939	1 453	1 326	1 417	1 547	1 366	1 758	1 074
P25	1 436	2 249	2 209	2 272	2 209	2 261	2 762	2 148
P50	3 093	4 495	4 971	4 266	4 419	4 434	5 292	3 749
P75	7 733	9 080	9 942	10 486	9 942	10 639	12 152	9 308
P90	14 361	17 303	18 780	18 487	19 885	19 949	23 199	20 285
P95	18 780	24 224	27 618	25 302	29 533	26 354	33 141	30 160

Table 14: Percentiles of the wage distribution NIDS vs PALMS (BRW applied where possible).

NIDS starts lower than PALMS in wave 1 at every percentile, but quickly increases to overtake PALMS. By wave 4 wages are higher at every percentile. Seeing as means and all the percentiles are increasing relative to PALMS, it is likely the whole distribution is growing relative to PALMS. The fact that the lower percentiles are increasing reflects that it is not

just the BRW that is causing gains in the distribution – even the very bottom has shifted upwards in NIDS. Again, it is unclear why NIDS is experiencing growth that PALMS is not.

Another notable feature is that wages at the 5th, 10th and 25th percentile in PALMS fell between wave 1 and wave 4, while the higher percentiles grew steadily. This implies that PALMS exhibited increasing inequality over the period, and also sheds light on why estimates of the mean grew faster in NIDS than their counterparts in PALMS. Conversely, wages rose at each percentile in NIDS and as such nothing can be directly inferred about inequality.

6.2.3 Comparing log real wage distributions in NIDS and PALMS

In this section I use kernel densities to compare net and gross wages in PALMS and NIDS.

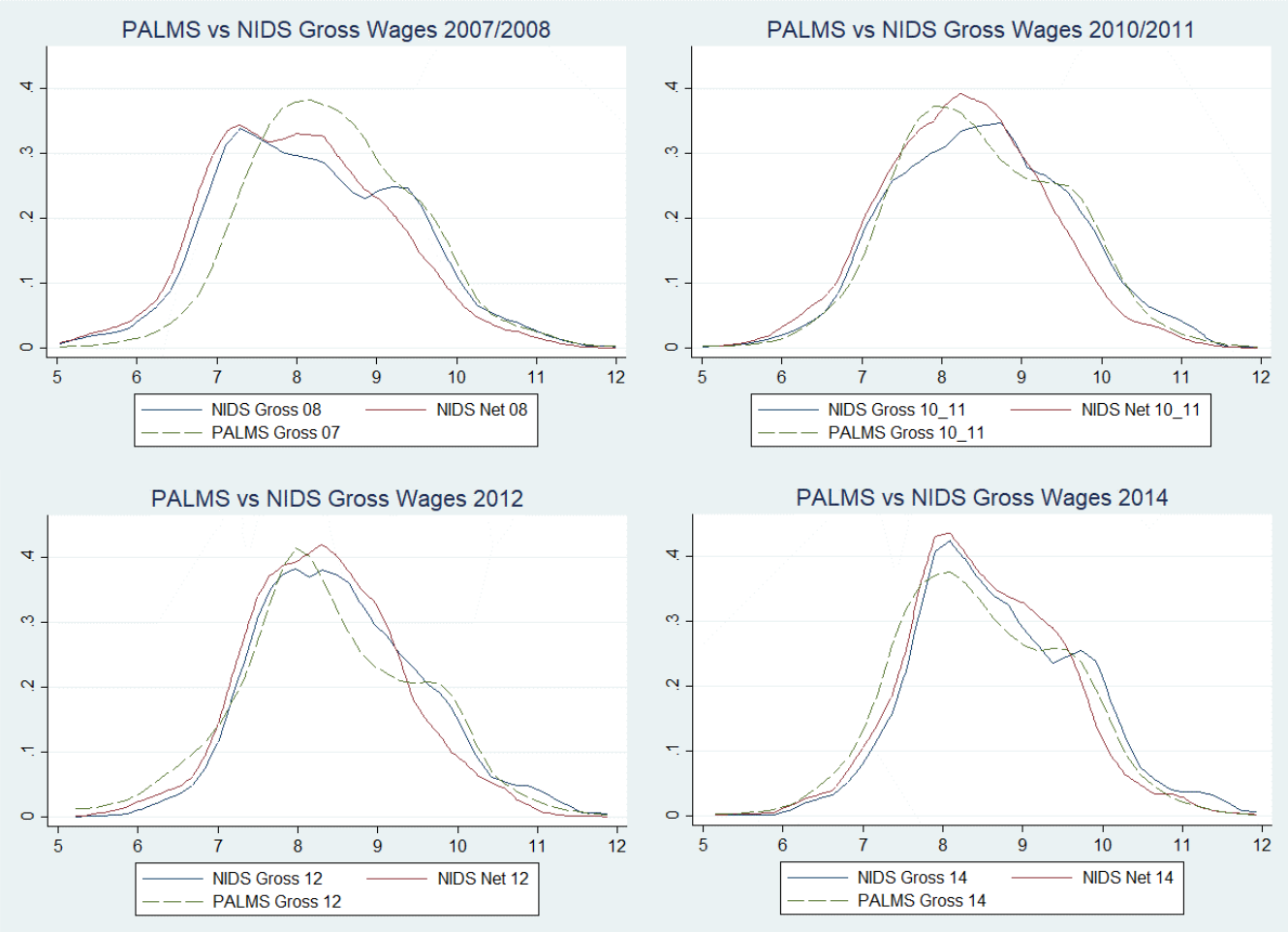


Figure 6: Density plots of real net and gross wages monthly, NIDS vs PALMS.

Figure 6 above plots adjusted net and gross wages from NIDS and the gross wages released in PALMS. I include net wages from NIDS as a robustness check (all real variables). The results are somewhat encouraging. All variables in both data sources lie in a similar position and take on the familiar bell-curve shape associated with wage distributions.

At this point one can comment on the spread of the distribution which provides a rough gauge of the variance and inequality. The NIDS curves suggest a decrease in the variance of wages over the period, with the density plot becoming 'narrower and taller' over time. The spread of the PALMS curve is roughly unchanged over the four waves, while the density becomes increasingly bi-modal. The "second hump" in PALMS beyond wave 1 at around R18 000 (or

9.8 in the log scale) might represent a bracket imputation done by Stats SA for a single bracket or a set of adjacent brackets containing a high proportion of observations. As discussed in 4.3 it is impossible to know how this imputation was done.

The NIDS and PALMS distribution are most different in Wave 1, with NIDS real gross wages significantly lower than PALMS despite being a year later. As discussed in 6.2.1 and 6.2.2 above, NIDS then catches up with PALMS and by 2014 wages overall seem slightly higher in NIDS. Part of the reason is that the bottom portion of NIDS has experiences gains that PALMS has not – this is reflected in left hand siding shifting inward (rightward), making the weight greater (taller) around the mode. Above I proposed that part of the overall difference between PALMS and NIDS might be due to the survey design creating an upward bias in NIDS vis-à-vis PALMS.

7. RQ3: Inequality in NIDS and PALMS

From above I have constructed a comparable wage measure in NIDS and found that NIDS and PALMS seem to be measuring the same underlying construct with reasonable degree of congruence. In this section I turn to consider what each dataset suggests about the evolution of inequality between 2008 and 2014.

7.1. Results

7.1.1 Effect of Data Quality Adjustment in NIDS

A simple measure of wage inequality is to calculate the weighted variance of the wage distribution. Before considering both PALMS and NIDS, it is useful to decompose the effects of the data quality adjustments performed on NIDS. This provides a check on what is driving the final results. Figure 7 (following page) considers how the variance of wages changes as one moves from NIDS net as in the public release to the final gross wages variable constructed in this paper. The adjusted net figures (red) which exclude imputations and bracket midpoint (unlike blue) show around a 0.05 increase in variance in all waves.

Gross wages naturally have higher variance which is expected seeing as the purpose of the progressive tax schedule in South Africa specifically aims to compress the distribution of wages. The gross variances measures (adjusted and unadjusted) both lie at minimum a point above their net counterparts.

The adjustment effect is largest in wave 1, which is expected given the discussion of response rates raised earlier. The effect is smaller in gross wages, excluding wave 1 in which it is large. Part of the reason the effect on net is less than on gross is that gross unadjusted did not contain any imputation. The net unadjusted figure is the net *fwag* variable released by NIDS, which includes NIDS' imputation. As discussed, imputation likely reduced the variance. Thus in the net case the adjustment is undoing the imputation as well as applying the BRW which is not true of the gross variable.

The figure suggests that in both variables, the variance is highest in wave 1, decreasing quite drastically over waves 2 and 3, before levelling out by wave 4.

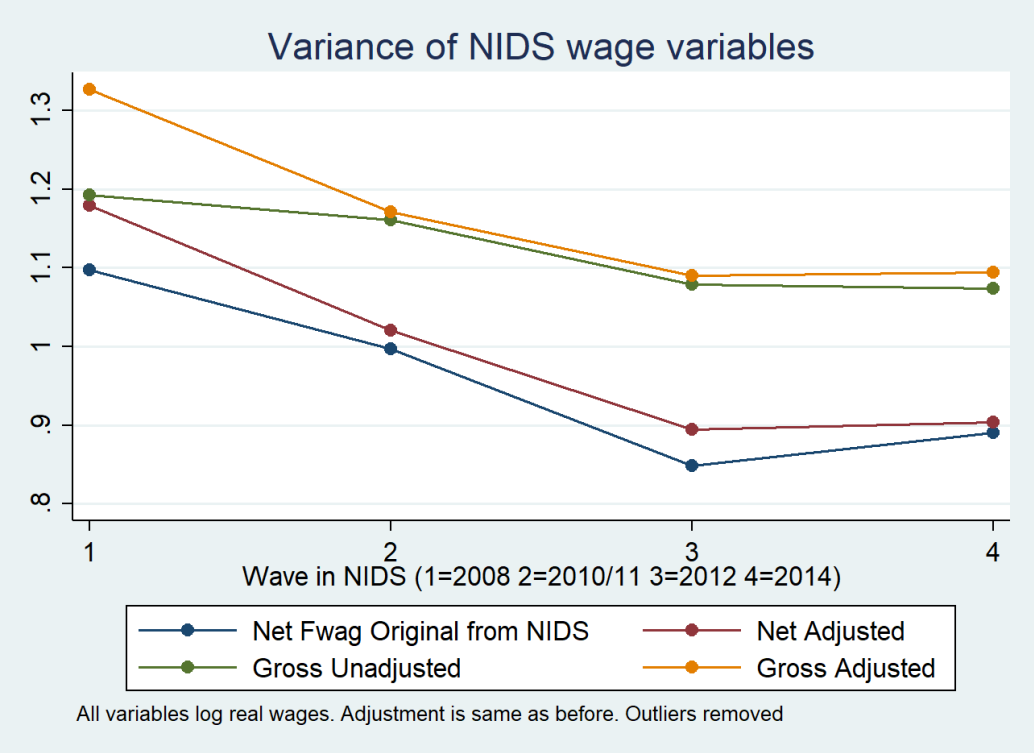


Figure 7: Effect of Data Quality Adjustments on the variance of NIDS wages.

7.1.2 NIDS and PALMS: A look at the Percentiles

Before comparing p-ratios it is useful to get an over-view of the evolution of the different percentiles.

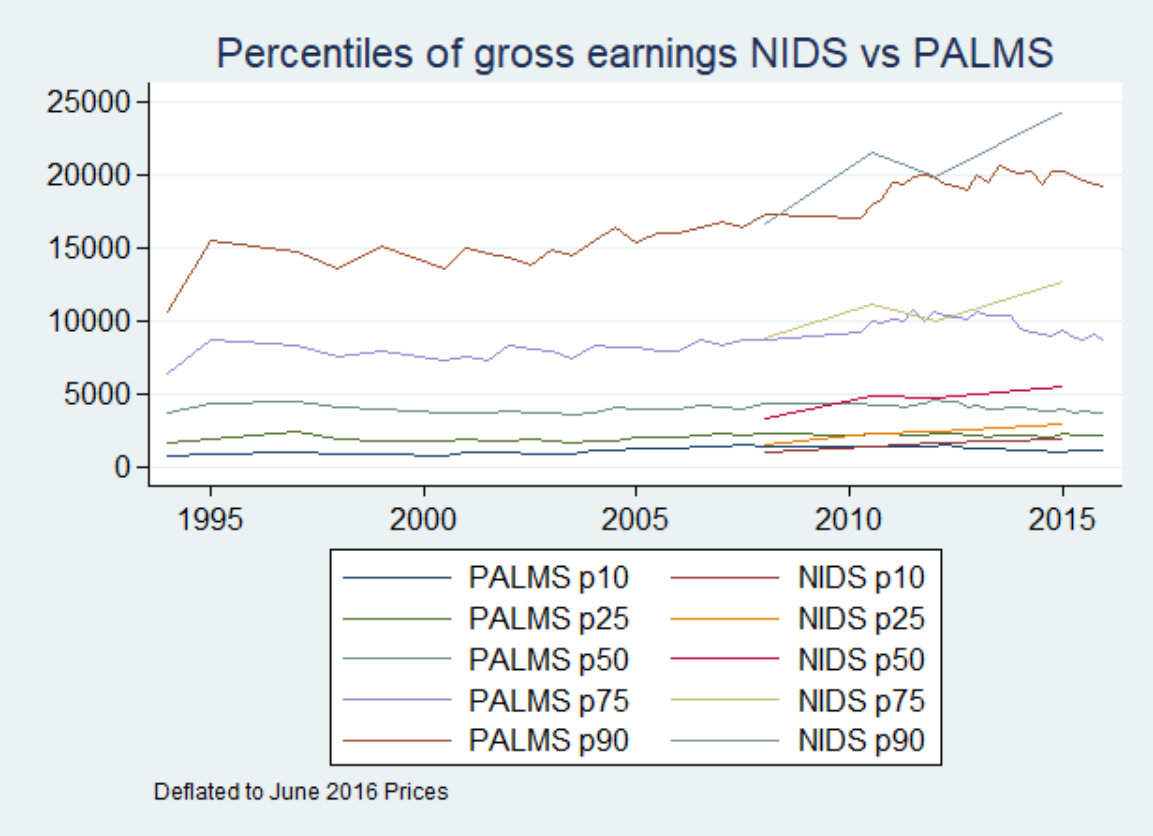


Figure 8: p10, p25, p50, p75, p90 in NIDS and PALMS 1995-2015.

Figure 8 above provides plots connecting common percentiles over all periods for which there are data. The NIDS percentile lines (appearing on the right) track their PALMS counterparts quite well. Again, the increase in the NIDS distribution relative to PALMS as time progresses is clear. This is most apparent in the p10 and p25, which increase for NIDS post 2010, while remaining somewhat unchanged in the case of PALMS.

The overall picture suggests a common story in the literature: the wages of the wealthiest 10% dwarf those of the majority, and this difference has remained if not grown since apartheid.

7.1.3 P-ratios as Inequality

The disparity between high and lower earners given by percentiles is more precisely measured by taking their ratio. I discuss p-ratios in terms of the "upper-half" (everyone at or above p50) and the "lower half" (everyone at or below the p50) or the distribution. Thus, the P90/p50 and p75/p50 ratios are above, while the p25/p50 and p10/p50 ratios are below.

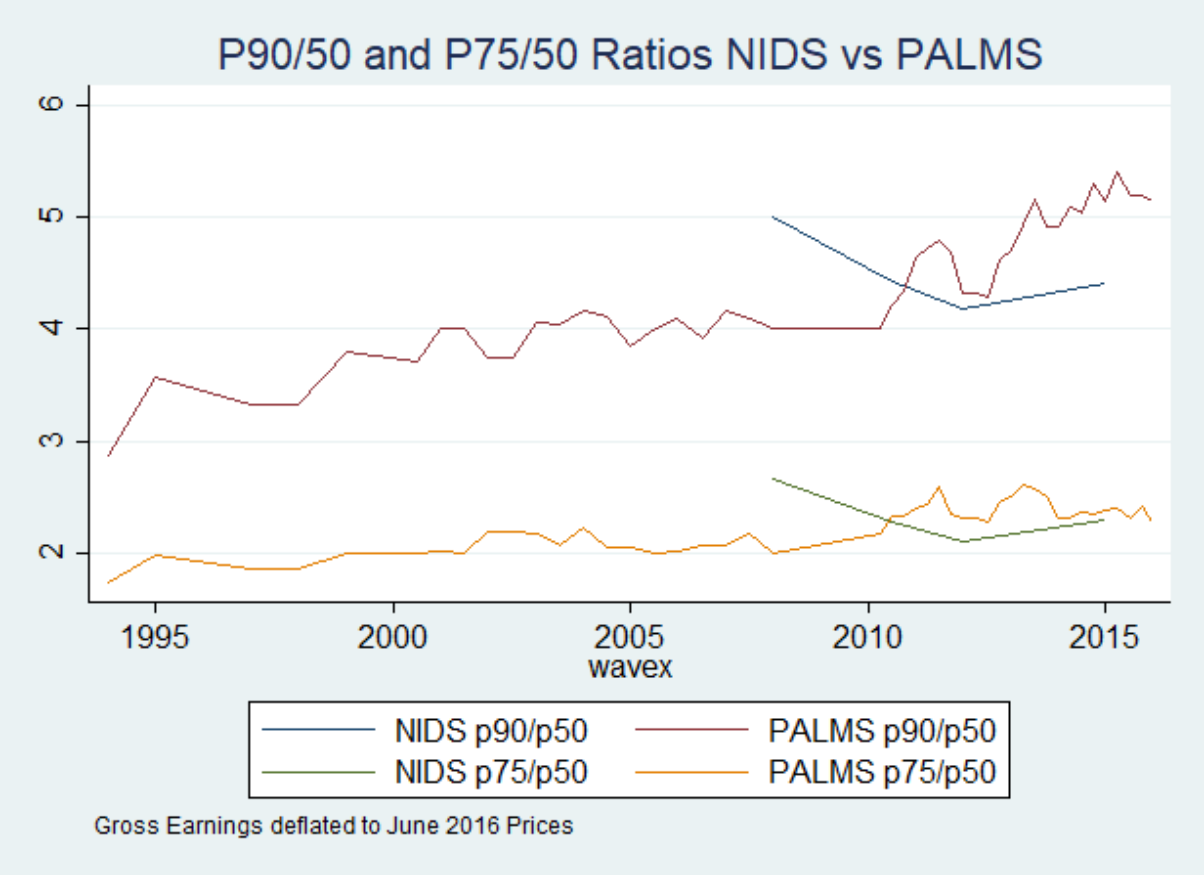


Figure 9: p-ratios above the median.

Figure 9 above tracks the p90/p50 and p75/p50 ratios. Over the 20-year period PALMS data shows steady gains of the wealthiest 10% relative to the median. By 2015 the p90 stands at 5 times the p50, meaning that the richest 10% earn at least quintuple the wage of the median earner. Conversely, the p75/p50 ratio seems somewhat stable over the period, suggesting a more constant relative distribution of earnings in this quartile. The NIDS ratios are similar in magnitude to PALMS (least so in wave 1), but neither shows a clear trend over the 2008-2014.

Overall, the PALMS "above" curves together suggest a moderate expansion of inequality post 2010, whilst the NIDS curves are ambiguous on this front.

Figure 10 below (next page) looks at the bottom half of the distribution using the p10/p50 and p25/p50 ratios.

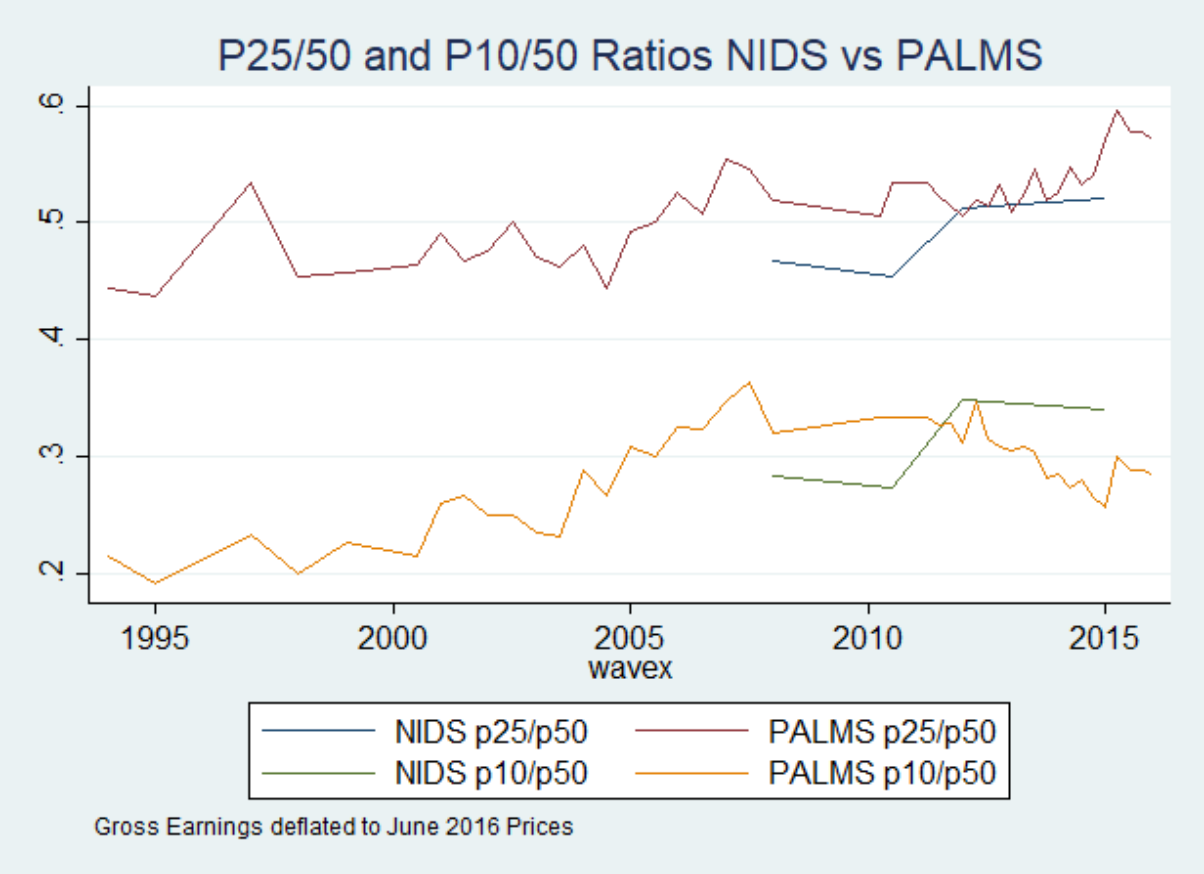


Figure 10: p-ratios below the median.

Figure 10 shows a gradual compression in wages in both NIDS and PALMS, as the 10th and 25th percentile move closer to the median. This suggests that the wages of the bottom half of the distribution are moving towards that of the median earner. PALMS shows a long-term trend of modest relative gains of both the lowest decile and the lowest quintile. However, beyond 2010 the upward trend in the p10/p50 reverses with the median moving away from the p10. The NIDS p25/p50 increases over the four waves, mirroring the p25/p50 in PALMS. Beyond 2012 the p10/p50 declines rapidly in PALMS while it remains stable in NIDS.

Overall figure 10 shows modest relative gains for the poor as compared to the median over the post-apartheid period, which translates to a reduction in inequality. This result is quite robust given the two different datasets being used, and the similarily between the results. However, there is some evidence of a reversal at the bottom in the post 2010 period, particularly in PALMS. Conversely, NIDS showed more sustained compression in the bottom half between 2008 and 2014.

Taking figures 9 and 10 together, PALMS shows expansion at the top and compression at the bottom between 1995 and 2009, providing an ambiguous effect on overall inequality. Post 2010, PALMS shows expansion at the top, and ambiguous effects at the bottom; together these imply a weak increase in inequality. In the case of NIDS, the top was ambiguous, with the bottom showing compression. This implies an overall reduction in inequality. Therefore, in

1995-2009, PALMS suggests ambiguous trends in overall inequality. In 2008-2014, PALMS suggests moderate increases, whilst NIDS suggests decreases in inequality.

7.1.4 Variance as Inequality

Given the discussion immediately above it is interesting to consider how a measure of overall inequality would reflect these underlying changes. In this section I use the variances of wages as a single "all-encompassing" measure of wage inequality. As above I consider both NIDS and PALMS over the post-apartheid period. As a robustness check, and in line with the literature, I consider both hourly and monthly wages, and I also consider the effect of weighting the prior measure by hours worked. As it turns out, the story is not affected by these adjustments.

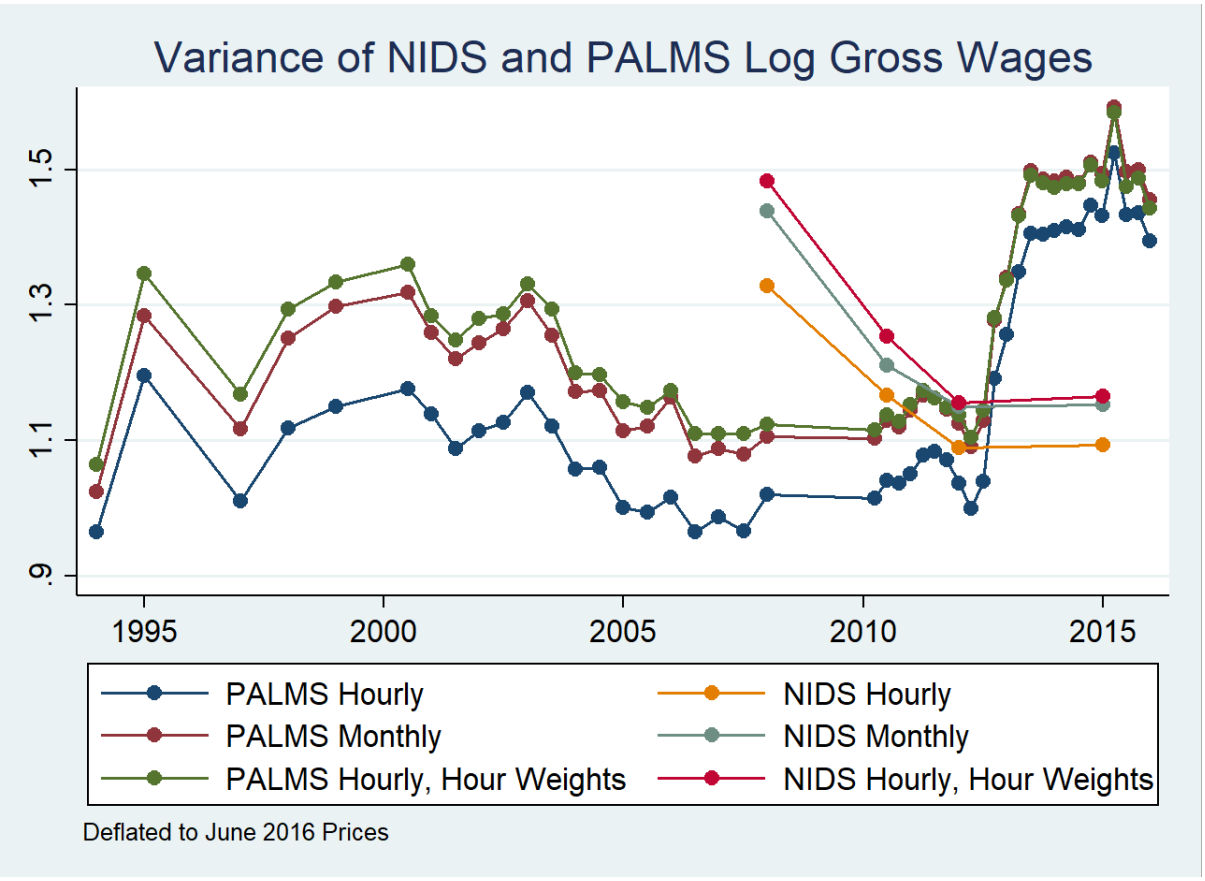


Figure 11: Variance of wages in NIDS and PALMS.

The PALMS portion of figure 11 above reproduces figure 1 of du Toit & Wittenberg (2016), whilst extending the data to 2015. The results are pleasing in the sense that the picture reflects the same trends observed there. Notably, the variance of PALMS wages hikes drastically to unprecedented levels after 2011, a result not reflected in du Toit & Wittenberg (*ibid)* as the data were not yet available to them. Part of this hike in inequality is expected given the declining p10/p50 and increasing p90/p50 observed in PALMS over this period. However, these changes in the p-ratios were modest compared to the drastic rise in figure 11 above, and there is likely another phenomenon at play. In 4.3 I noted significant data quality issues in this era of PALMS - it is possible that much of the dramatic rise in variance (and the expansion in the p90/50) is due to spurious data quality issues rather than an underlying change in the population. This would also explain why the corresponding measure in NIDS does not spike in

this period. Investigating what is causing this somewhat anomalous set of results in PALMS would require a thorough examination of StataSA data that is beyond the scope of this paper.

The choice between hourly and monthly wages makes no difference to the trends in NIDS or PALMS. This is not surprising as differences between them can only arise from changes in average hours worked, something there is no a priori reason to expect. Weighting the data by hours worked increases the variance marginally in NIDS and PALMS. As such the order of the curves and the differences between them are much the same in NIDS and PALMS, with hourly wages at the bottom and "hourly with hour weights" on top.

The long-term trend in PALMS before 2011 suggests that the variance in wages has been somewhat constant, at least since 2005. This is consistent with the discussion of p-ratios. NIDS comes into the picture showing a far greater measure of inequality, but this quickly returns to be in line with PALMS. The NIDS decline is consistent with the discussion about p-ratios, where NIDS showed a compression of the bottom half of the distribution, and ambiguous effects in the top half. As mentioned, these together imply a (moderate) reduction in variance, which is confirmed in figure 11 above.

That said, NIDS has notably lower BRW effects post 2010, which to some extent must be lowering the variance estimates (as the BRW tends to increase variance). NIDS had a moderate BRW effect due to low rates of bracket response. In 5.4.2 I posited that the brackets may have been nominally too low in their values to properly capture the non-response of the wealthy. So, part of this decline is possibly due to data quality issues and adjustments and not from underlying changes in the construct being measured.

Overall there is no clear message from figure 11. To be sure, much of this is because the variance measure is aggregating too much information as highlighted in the literature– there are different and opposing phenomena in the bottom and top halves of the distribution. Data quality issues also plague the results. The quality of PALMS earnings data post-2010 has been called into question, and this coincides suspiciously with the radical change in variance estimates. The NIDS results seem more reliable: they corroborate the discussion of p-ratios and do not suffer the same shortfalls. However, as I have proposed, they may underestimate inequality due to inaccurately high point response rates.

8. Conclusion

The major contribution of this paper has been to show how best one can adjust NIDS data to improve the measurement of wages whilst also creating a variable that is comparable to PALMS. The net wage variable released by NIDS was adjusted to remove imputation, improve outlier flagging, and account for item non-response using the BRW. The new weights could then be applied to gross wages, creating a variable fit for comparison to PALMS.

The first research question addressed the effects of these improvements. I found that the BRW acted in a similar fashion as in PALMS, to increase the weight of the higher earners, thus widening the distribution and increasing estimates of the mean. The effect of the BRW was lower than might be expected, due to high rates of point response in all income brackets post-

2008. However, it still appeared to have a marginally stronger effect on the overall distribution than midpoint imputation.

The second research question investigated the similarity between wage distributions in NIDS and PALMS. Section 6 found that the distributions were overall encouragingly alike; NIDS and PALMS appear to reflect a similar data generating process. This finding supports their theoretical capacity to provide nationally representative cross-sections. There were some troubling disparities, however. Overall, real gross wages appear to be growing at a faster rate when measured by NIDS as opposed to PALMS. Another disparity was that NIDS data showed gains at the bottom of the distribution, while PALMS data suggested that the wealth of the very poorest has stagnated relative to the median since 2010.

The third research question sought to draw conclusions about the evolution of inequality. Between 1995 and 2010 PALMS data confirmed findings elsewhere in the literature; overall PALMS shows expansion at the top of the wage distribution and compression at the bottom, with ambiguous total effects over the period. The post-2010 PALMS data suffer from serious data quality issues introduced by Stats SA; it is not clear that reliable conclusion can be drawn from them. NIDS data seem more reliable over the 2008-2014; they suggest an overall decline in wage inequality, due to ambiguous effects at the top and compression at the bottom. As noted, this may partly be due to unrealistically high point response rates.

The paper also revealed the benefits of the BRW technique. The BRW approach is a feasible way of regaining some of the lost information due to item non-response, with low risks of perverting the data. The BRW can theoretically provide a more accurate representation of the variance of wages than alternative imputation techniques. The paper found that variance is higher under than BRW than under bracket midpoint imputation, and far higher than the case with simple imputation, which actually reduced the variance relative to the raw point data.

The paper leaves much for future study. I have only scratched the surface of the various techniques and measures used to examine inequality in the literature. Other variables can be brought into the analysis, allowing, for example, inequality by demographic groups, analysis of residual inequality, or controlling for compositional shifts in the distribution. The sample could also be extended to include the self-employed and possibly part-time workers. Lastly, performing multiple imputation on NIDS would offer another way of comparing NIDS and PALMS, potentially improving the representation of wage distributions.

Appendix A: Clarifying Wealth, Income, Earnings and Wages[17]

Wealth can be thought of as a stock and in a broad conception involves both material and non-material elements. For example, psychological wellbeing or social status are non-material but contribute to a person's wealth. Economists for the most part avoid these elements however, as they are abstract and difficult to measure. By contrast, measurable elements including those that can be converted into a monetary value are often the subject of analysis; a common measure of wealth is *net worth*, defined as gross assets minus gross liabilities.

As opposed to wealth, income is a flow variable representing a change in wealth of a particular rate. Economists will very seldom consider the nonmaterial elements of income because they are very hard to conceptualize and measure. Income in turn can be broken down depending on its source, for example income from a primary job, secondary job, interest income, rental income, income from self-employment, and so on. When one considers the elements of income that accrue from paid work, the term used is 'earnings'. Earnings can come from self-employment or wage/salaried work; wages however can only come from working for someone else. However, in this paper earnings *excluding* self-employment earnings are still referred to as earnings for simplicity, and because wages might be a misleading term to use as I do not intend to exclude salaried pay. Thus 'wages' and 'earnings' refer in this work to any pay (a salary or a wage) accruing from working as an employee.

[17] This section is based on a typical categorization of wealth and income, such as in Woolard & Mbewe (2016).

Appendix B: Additional Tables

I. Tax Tables – From SARS budget pocket guides

2008

Individuals and special trusts

Taxable Income (R)	Rates of Tax (R)
0 - 122 000	18% of each R1
122 001 - 195 000	21 960 + 25% of the amount above 122 000
195 001 - 270 000	40 210 + 30% of the amount above 195 000
270 001 - 380 000	62 710 + 35% of the amount above 270 000
380 001 - 490 000	101 210 + 38% of the amount above 380 000
490 001 and above	143 010 + 40% of the amount above 490 000

Trusts other than special trusts	
Rate of Tax	40%

Tax Rebates

Rebates	R
Primary	8 280
Additional (Persons 65 and older)	5 040

Tax Thresholds

Age	Tax Threshold (R)
Below age 65	46 000
Age 65 and over	74 000

2010/2011

Individuals and special trusts

TAXABLE INCOME (R)		RATE OF TAX (R)
0 – 140 000		18% of taxable income
140 001 – 221 000	25 200 +	25% of taxable income above 140 000
221 001 – 305 000	45 450 +	30% of taxable income above 221 000
305 001 – 431 000	70 650 +	35% of taxable income above 305 000
431 001 – 552 000	114 750 +	38% of taxable income above 431 000
552 001 and above	160 730 +	40% of taxable income above 552 000

Trusts other than special trusts Rate of Tax - 40%

Tax Rebates

REBATES	
Primary	R10 260
Additional (Persons 65 and older)	R 5 675

Tax Thresholds

AGE	TAX THRESHOLD
Below age 65	R57 000
Age 65 and over	R88 528

2012

INCOME TAX: INDIVIDUALS AND TRUSTS

Tax rates (year of assessment ending 28 February 2013)

Individuals and special trusts

Taxable Income (R)	Rate of Tax (R)
0 - 160 000	18% of taxable income
160 001 - 250 000	28 800 + 25% of taxable income above 160 000
250 001 - 346 000	51 300 + 30% of taxable income above 250 000
346 001 - 484 000	80 100 + 35% of taxable income above 346 000
484 001 - 617 000	128 400 + 38% of taxable income above 484 000
617 001 and above	178 940 + 40% of taxable income above 617 000

Trusts other than special trusts Rate of Tax - 40%

Tax Rebates

Rebates	
Primary	R11 440
Secondary (Persons 65 and older)	R6 390
Tertiary (Persons 75 and older)	R2 130

Tax Thresholds

Age	Tax Threshold
Below age 65	R63 556
Age 65 to below 75	R99 056
Age 75 and over	R110 889

2014

INCOME TAX: INDIVIDUALS AND TRUSTS

Tax rates (year of assessment ending 28 February 2015)

Individuals and special trusts

Taxable Income (R)	Rate of Tax (R)
0 – 174 550	18% of taxable income
174 551 – 272 700	31 419 + 25% of taxable income above 174 550
272 701 – 377 450	55 957 + 30% of taxable income above 272 700
377 451 – 528 000	87 382 + 35% of taxable income above 377 450
528 001 – 673 100	140 074 + 38% of taxable income above 528 000
673 101 and above	195 212 + 40% of taxable income above 673 100

Trusts other than special trusts: Rate of Tax 40%

Tax Rebates and Tax Thresholds

Rebates	
Primary	R12 726
Secondary (Persons 65 and older)	R7 110
Tertiary (Persons 75 and older)	R2 367
Age	**Tax Threshold**
Below age 65	R70 700
Age 65 to below 75	R110 200
Age 75 and over	R123 350

II. Inflation Data

Yearly Index and Deflation Calculation		
Year	CPI at June	CPI 'Deflator'
2007	61.78	1.6186
2008	**69.27**	**1.4436**
2009	73.62	1.3584
2010	76.66	1.3045
Dec 2010	**77.63**	**1.2429**
2011	80.46	1.2429
2012	**84.91**	**1.1777**
2013	89.58	1.1164
2014	**95.55**	**1.0466**
2015	100.00	1.0000
2016	106.30	0.9408
2017	111.73	0.8950

Inflation information based on Stats SA data.

III. Detailed explanation for combining primary and secondary

01 =point data. Original NIDS flag is 1=survey, 2=imputed.

Scenario					Treatment					
	Primary		Secondary		New Variables				Vs. NIDS	
	Net	Gross	Net	Gross	net only	flag	net incl imputation from gross	flag	Same as NIDS?	NIDS flg
a)	bracket	.	bracket	.	If adding the midpoint of secondary bracket pushes primary into next bracket up, recode primary as such. Otherwise drop secondary bracket	0	Same as net_only	0	No	1
b)	bracket	.	point	.	If adding secondary point to primary midpoint pushes obs into next bracket, recode primary as such. Otherwise drop secondary point	0	Same as net_only	0	No	1
c)	bracket	.	.	.	Record as bracket	0	Same as net_only	0	No	1
d)	point	.	bracket	.	Add bracket secondary midpont to point primary	1	Same as net_only	1	Yes	1
e)	point	.	point	.	Sum the two	1	Same as net_only	1	Yes	1
f)	point	.	.	point	Only use primary net info	1	Impute secondary from gross and add to primary	1	Yes/No	1
g)	point	.	bracket	point	Add bracket secondary midpont to point primary	1	Impute secondary from gross and add to primary	1	Yes/No	1
h)	point	.	.	.	Only use primary net info	1	Same as net_only	1	Yes	1
i)	.	point	point	.	Store secondary bracket only. Flag as bracket.	0	Impute primary from gross and add to secondary	1	YES	1
j)	.	point	.	.	Flag as missing (no imputation)	.	Impute primary from gross	1	No/Yes	2
k)	bracket	point	point	.	If adding secondary point to primary midpoint pushes obs into next bracket, recode primary as such. Otherwise drop secondary point	0	Impute primary from gross, add to secondary	1	No/No	1
l)	bracket	point	.	.	Record net as bracket and flag as such	0	Impute primary from gross	1	No	1
m)	.	.	point	.	Use secondary point and flag as point	1	Same as net_only	1	Yes	1

IV. Detailed Bracket-Response Adjustments in NIDS

"Raw" Data from release with w2_wgt			Unadusted BRW would produce:		proportion of points that fall in the bracket immediately above, equal to X, and in the bracket immediately below:					Calculating New Weights for Entire Bracket (of Point Data grouped together)					End Result	
Brackets	Weight of brac	weight of points in brac	pr(point response)	rescale factor	bracket 2	bracket 4	bracket 6	bracket 8	bracket 10	Weight of brac	weight of points in brac	Extra from below	Extra from above	Extra from "around X"	total	Rescale Factor
1 0-699	31 262	638 403	0.95	1.05	0.63					**31262**	**638403**	**5480**			675 145	**1.058**
2 700	8 709	97 650	0.92	1.09	0.10						**97650**			**838**	98 488	**1.009**
3 701-999	22 145	278 518	0.93	1.08	0.27	0.16				**22145**	**278518**	**8787**	**2391**		311 841	**1.120**
4 1000	53 736	252 935	0.82	1.21		0.15					**252935**			**7980**	260 915	**1.032**
5 1001-1799	118 869	1 171 840	0.91	1.10		0.69	0.35			**118869**	**1171840**	**54858**	**36970**		1 382 537	**1.180**
6 1800	158 515	125 642	0.44	2.26			0.04				**125642**			**5882**	131 524	**1.047**
7 1801-3999	248 954	2 088 605	0.89	1.12			0.62	0.53		**248954**	**2088605**	**61838**	**97775**		2 497 172	**1.196**
8 4000	117 303	244 213	0.68	1.48				0.06			**244213**			**7230**	251 444	**1.030**
9 4001-7999	138 536	1 629 151	0.92	1.09				0.41	0.54	**138536**	**1629151**	**43171**	**48235**		1 859 092	**1.141**
10 8000	80 683	155 639	0.66	1.52					0.05		**155639**			**4124**	159 763	**1.026**
11 8000+	326 709	1 259 971	0.79	1.26					0.41	**326709**	**1259971**		**33388**		1 620 068	**1.286**

Wave 2 Example of Modified BRW technique.

As is clear, the 'old' BRW approach in columns 4 and 5 do not produce likelihoods that will work to provide a consistent smooth distribution. Rather they vary from bracket to bracket and do not decrease in an orderly fashion as assumed to be the case. The single-value brackets are particularly problematic, in all cases showing significantly lower likelihoods and therefore higher rescaling factors than the 'ranged' brackets on either side. This will cause spiking in the final distribution. The sudden changes between brackets will result in the kernel function becoming biased as the second derivative will be large.

An alternative approach, suggested to me by Martin Wittenberg, is to apportion the weight of these problematic 'discrete' brackets, being 2,4,6,8,10 in wave 2 above, to the brackets immediately above and below them. Columns 6-15 reveal how the total weight of these brackets was diluted into the brackets on either side, according to the proportion to which these were observed in the raw point data. The result is satisfactory. Firstly, the problem of discrete jumps in the rescaling factor is mitigated. Secondly, these discrete values will no longer experience a disproportionately large increase in their weight following the BRW.

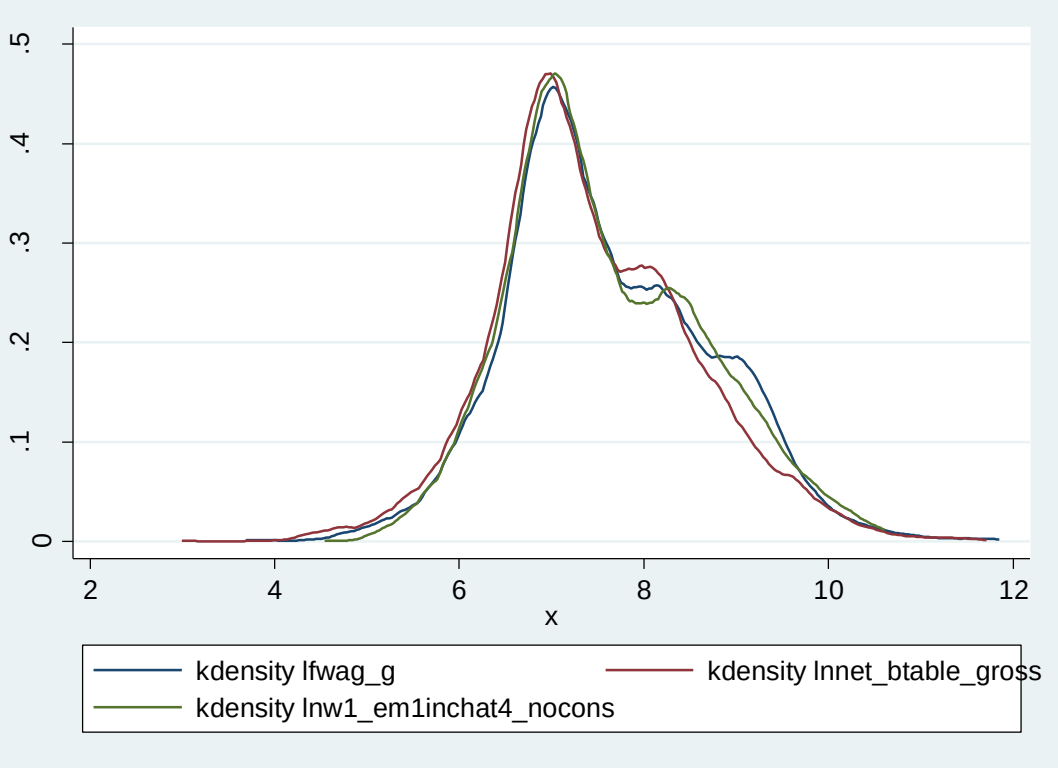
Wave 3

	"Raw" Data from release with w3_wgt			Unadusted BRW would produce:		proportion of points that fall in the bracket immediately above, equal to X, and in the bracket immediately below:						Calculating New Weights for Entire Bracket (of Point Data grouped together)					End Result	
	Brackets	Weight of brac	weight of points in brac	pr(point response)	rescale factor	bracket 2	bracket 4	bracket 6	bracket 8	bracket 10	bracket 12	Weight of brac	weight of points in brac	Extra from below	Extra from above	Extra from "around X"	total	Rescale Factor
1	0-599	31 750	331 757	0.91	1.10	0.25						**31 750**	**331 757**	**5 984**			369 491	**1.11**
2	600	24 077	63 265	0.72	1.38	0.05							**63 265**			**1 141**	64 406	**1.02**
3	600-1299	51 400	939 894	0.95	1.05	0.70	0.22					**51 400**	**939 894**	**29 193**	**16 952**		1 037 440	**1.10**
4	1300	130 275	146 899	0.53	1.89		0.04						**146 899**			**4 563**	151 462	**1.03**
5	1300-3099	194 355	3 107 450	0.94	1.06		0.74	0.60				**194 355**	**3 107 450**	**200 221**	**96 518**		3 598 545	**1.16**
6	3100	335 599	74 526	0.18	5.50			0.01					**74 526**			**4 802**	79 328	**1.06**
7	3101-5899	163 677	2 026 543	0.93	1.08			0.39	0.55			**163 677**	**2 026 543**	**192 075**	**130 576**		2 512 870	**1.24**
8	5900	352 086	14 103	0.04	25.96				0.00				**14 103**			**1 337**	15 440	**1.09**
9	5901-11000	137 619	1 674 143	0.92	1.08				0.45	0.73		**137 619**	**1 674 143**	**99 897**	**158 674**		2 070 333	**1.24**
10	11000	136 229	89 165	0.40	2.53					0.04			**89 165**			**5 321**	94 486	**1.06**
11	11001-17999	93 306	519 697	0.85	1.18					0.23	0.48	**93 306**	**519 697**	**32 418**	**31 011**		676 431	**1.30**
12	18000	67 324	104 405	0.61	1.64						0.10		**104 405**			**6 513**	110 918	**1.06**
13	18000+	124 124	455 174	0.79	1.27						0.42	**124 124**	**455 174**		**28 393**		607 690	**1.34**

Wave 4

"Raw" Data from release with w4_wgt			Unadusted BRW would produce:		proportion of points that fall in the bracket immediately above, equal to X, and in the bracket immediately below:						Calculating New Weights for Entire Bracket (of Point Data grouped together)					End Result	
Brackets	Weight of brac	weight of points in brac	pr(point response)	rescale factor	bracket 2	bracket 4	bracket 6	bracket 8	bracket 10	bracket 12	Weight of brac	weight of points in brac	Extra from below	Extra from above	Extra from "around X"	total	Rescale Factor
1 0-749	10 886	514 829	0.98		0.33						**10 886**	**514 829**	**2 572**			528 286	1.03
2 750	7 680	1 848	0.19		0.00							**1 848**			**9**	1 857	1.00
3 750-1499	11 021	1 020 806	0.99		0.66	0.25					**11 021**	**1 020 806**	**18 694**	**5 099**		1 055 621	1.03
4 1500	75 890	322 338	0.81			0.08						**322 338**			**5 903**	328 241	1.02
5 1501-2999	50 946	2 800 883	0.98			0.68	0.49				**50 946**	**2 800 883**	**121 381**	**51 293**		3 024 503	1.08
6 3000	248 285	325 693	0.57				0.06					**325 693**			**14 114**	339 807	1.04
7 3001-5999	109 058	2 602 643	0.96				0.45	0.52			**109 058**	**2 602 643**	**82 785**	**112 790**		2 907 276	1.12
8 6000	158 360	221 072	0.58					0.04				**221 072**			**7 032**	228 104	1.03
9 6001-11999	101 320	2 154 875	0.96					0.43	0.62		**101 320**	**2 154 875**	**71 079**	**68 543**		2 395 817	1.11
10 12000	113 973	183 270	0.62						0.05			**183 270**			**6 045**	189 316	1.03
11 12001-23999	96 923	1 117 122	0.92						0.32	0.71	**96 923**	**1 117 122**	**55 992**	**36 849**		1 306 885	1.17
12 24000	78 690	26 166	0.25							0.02		**26 166**			**1 311**	27 478	1.05
13 24000+	87 368	426 694	0.83							0.27	**87 368**	**426 694**		**21 386**		535 448	1.25

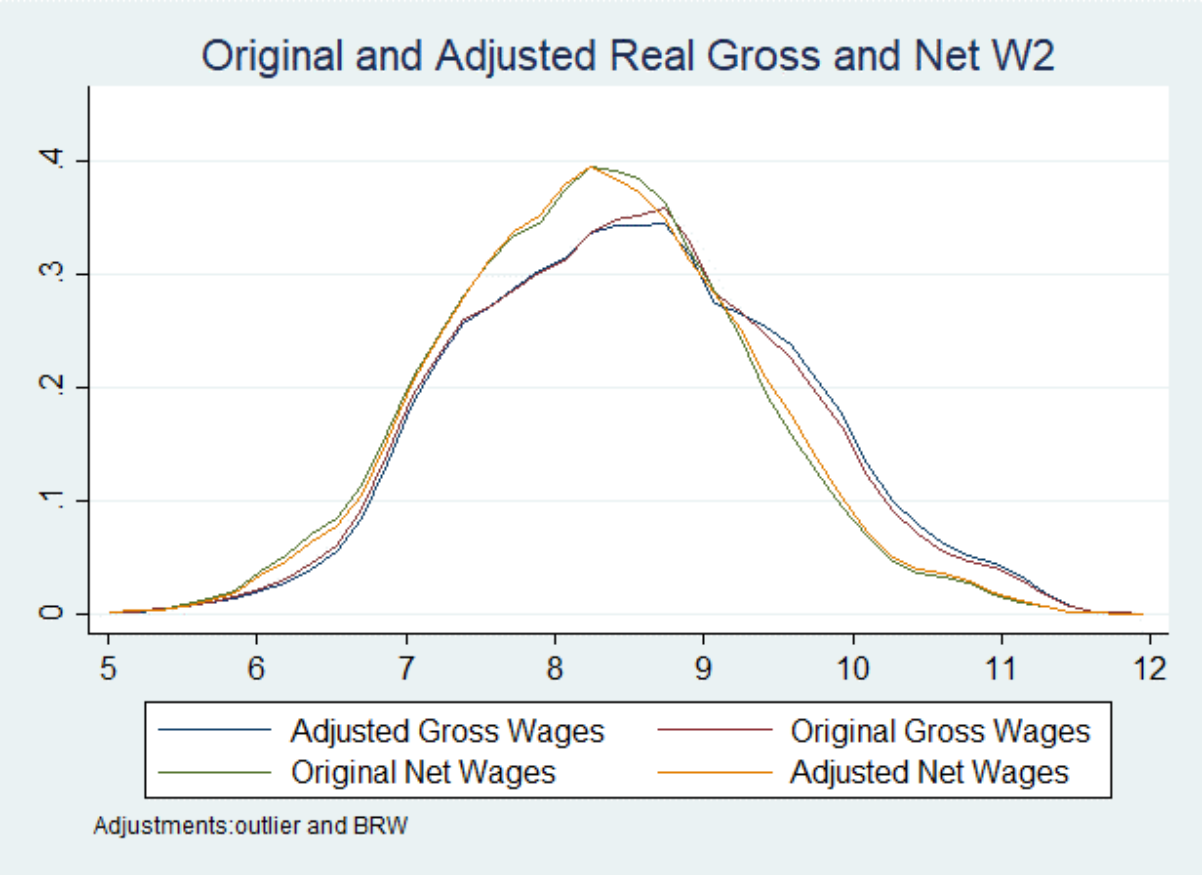
V. Reverse TT imputation

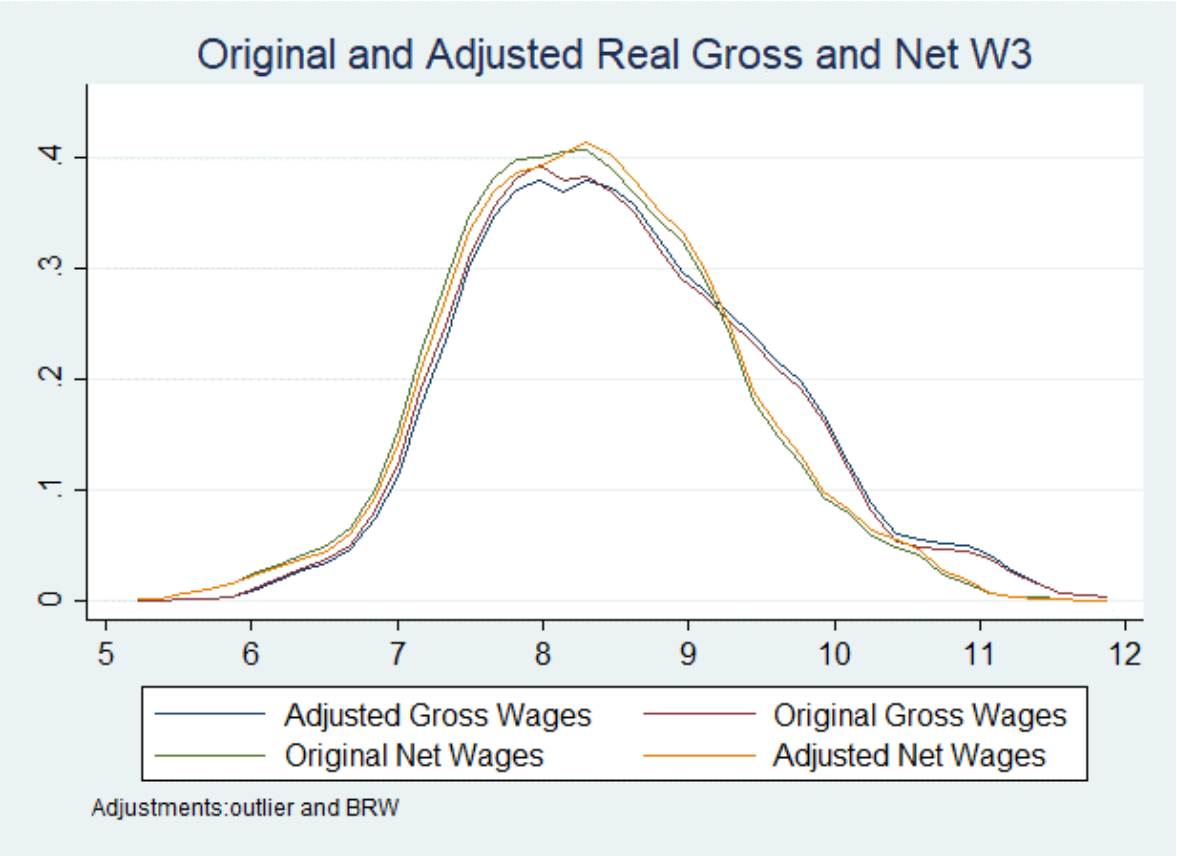


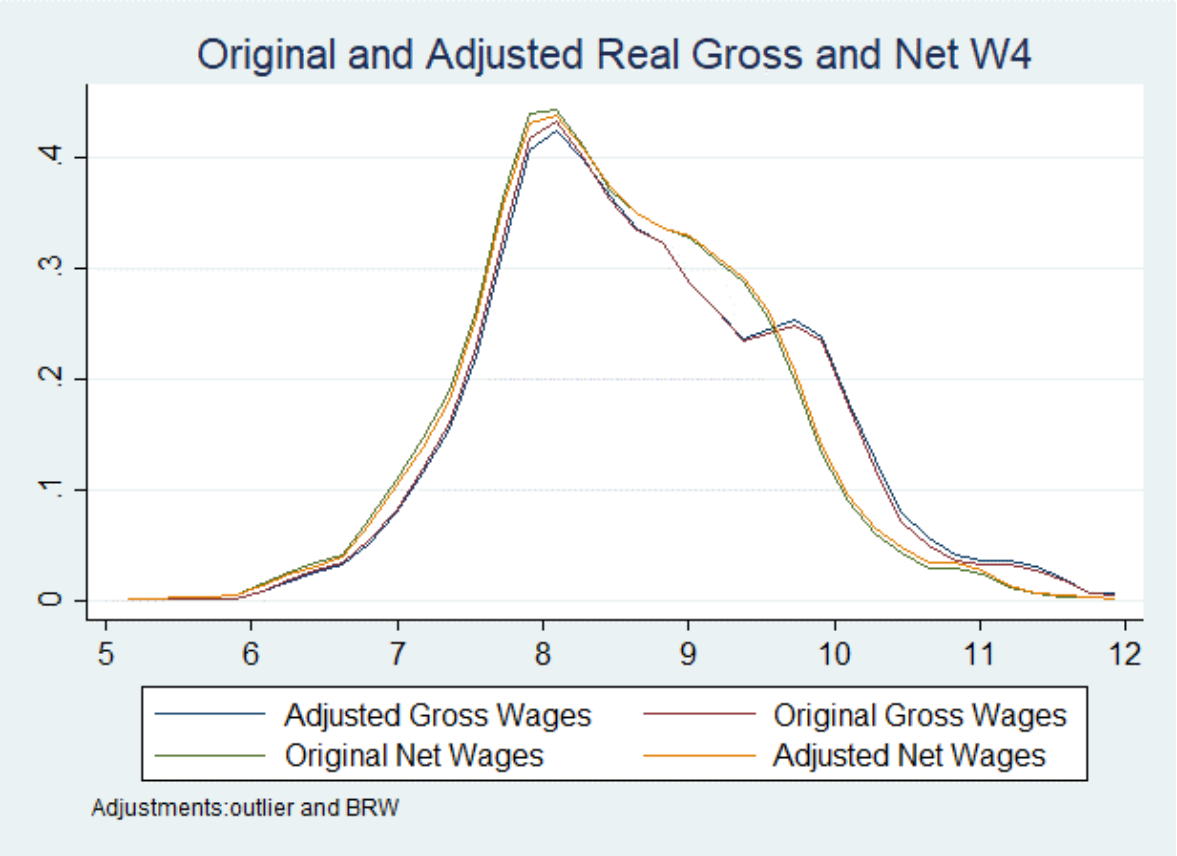
Fwag_g is the gross primary data only. Lnnet_btable_gross is gross imputed from net backwards through the TT. The third curve is the non-linear elasticity regression, which fits the observed fwag_g gross points more closely.

Appendix C: Additional Figures

I. Quality Adjustments Wave 2-4




Original and Adjusted Real Gross and Net W3
0
.1
.2
.3
.4
5
6
7
8
9
10
11
12
Adjusted Gross Wages
Original Gross Wages
Original Net Wages
Adjusted Net Wages
Adjustments:outlier and BRW


Original and Adjusted Real Gross and Net W4
0
.1
.2
.3
.4
5
6
7
8
9
10
11
12
Adjusted Gross Wages
Original Gross Wages
Original Net Wages
Adjusted Net Wages
Adjustments:outlier and BRW

Bibliography

Bhorat, H & Goga, S. 2012. The Gender Wage Gap in the Post-Apartheid South African Labour Market. *Development Policy Research Unit*: *Working Papers 12148*. University of Cape Town.

Biyase, M. & Tregenna, F. (2016). Determinants of remittances in South Africa. Cape Town: S*outhern Africa Labour and Development Research Unit: Working Paper No 176.* NIDS Discussion Paper 2016/3.

Branson, N. & Wittenberg, M. 2011. Re-weighting South African National Household Survey Data to create a consistent series over time: A cross entropy estimation approach. *Southern Africa Labour and Development Research Unit: Working Paper No 54.*

Di Tella, R & Rotemberg, J. J. 2016. Populism and the Return of the 'Paranoid Style': Some Evidence and a Simple Model of Demand for Incompetence as Insurance against Elite Betrayal. *Harvard Business School: Working Paper No 17-056.*

du Toit, A & Wittenberg, M. 2016. *Changes in South African earnings inequality 1994-2011: Markets, norms, compositional shifts and measurement.* Cape Town: Data First. (no paper number)

Finn, A., Leibbrandt, M. & Woolard, I. 2009. *Income & Expenditure Inequality: Analysis of the NIDS Wave 1 Dataset.* NIDS discussion paper 5.

Kannemeyer, C. 2016. Subjective well-being: Adult South Africans' Life Satisfaction (2008 - 2014). *SALDRU Working Paper Number 177.* NIDS Discussion Paper 2016/4.

Kerr, A & Wittenberg, M. 2017. A Guide to version 3.2 of the Post-Apartheid Labour Market Series (PALMS). Cape Town: DataFirst [producer and distributor], 2016.

Leibbrandt, M., Bhorat, H. & Woolard, I. 2001. Household inequality and the labour market In South Africa. *Contemporary Economic Policy.* 19(1):73-86.

Leibbrandt, M., Woolard, I., McEwen, H. & Koep, C. 2010. *Employment and Inequality Outcomes in South Africa. Employment and Inequality Outcomes in South Africa: What Role for Labour Market and Social Policies*? Southern Africa Labour and Development Research Unit, University of Cape Town.

Leibbrandt, M. & Nyaruwata, G. 2009. NIDS Wave 1 Discussion Paper 8: Personal Debt and Financial Access: Analysis of the NIDS Wave 1 Dataset.

Leibbrandt, M., Woolard, I., Finn, A., Argent, J. 2010. Trends in South African income distribution and poverty since the fall of Apartheid. *OECD social, employment and migration working papers* (no. 101).

Leibbrandt, M., Finn, A. and Woolard, I.: 2012. Describing and decomposing post-apartheid income inequality in South Africa. *Development Southern Africa.* 29(1):19-34.

NIDS, 2009a. Technical Paper no. 2: Weights: Report on NIDS Wave 1.

NIDS, 2009b. Technical Paper no. 3: Household Income: Report on NIDS Wave 1.

Njozela, L., Shaw, I. & Burns, J. 2016. Towards measuring social cohesion in South Africa. Cape Town: *SALDRU: Working Paper Number 187.* NIDS Discussion Paper 2016/14

Piketty, T. 2013. *Capital in the 21st Century.* Boston: Harvard University Press.

Posel, D. (2016). Inter-household transfers in South Africa: Prevalence, patterns and poverty. S*outhern Africa Labour and Development Research Unit: Working Paper No 180.* NIDS Discussion Paper 2016/7.

Ranchod, V. 2009. *Labour Market: Analysis of the NIDS Wave 1 Dataset.* NIDS Wave 1: Discussion Paper 12.

Seekings, J.2007. *Poverty and inequality in South Africa after apartheid.* Centre for Social Science Research: Working Paper 200. University of Cape Town.

Stiglitz, J. 2012. The Price of Inequality. United States: W. W. Norton & Company. 0393345068

Tregenna, F. 2011. Earnings inequality and unemployment in South Africa. *International Review of Applied Economics.* 25(5): 585-598.

Wittenberg, M. 2008. Nonparametric estimation when income is reported in bands and at points. Working Paper 94. *Economic Research Southern Africa.* Available Online: http://www.econrsa.org/papers/w_papers/wp94.pdf

Wittenberg, M. 2014. Analysis of employment, real wage, and productivity trends in South Africa since 1994. Geneva: International Labour Office. *Conditions of Work and Employment Series (45).*

Wittenberg, M. 2016. Trends in Earnings and Earnings Inequality in South Africa: 1993-2014. South African Labour Development Research Unit.

Woolard, I & Mbewe, S. 2016. Cross-Sectional Features of Wealth Inequality in South Africa: Evidence from The National Income Dynamics Study. *Southern Africa Labour and Development Research Unit: Working Paper No 185.*