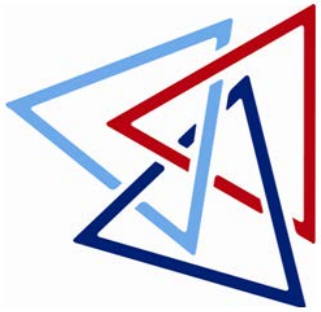


Wages and wage inequality in South Africa 1994-2011: The evidence from household survey data

Martin Wittenberg

Abstract

We analyse the long-term trends in wages in South Africa, using the data from the October Household Surveys, Labour Force Surveys and Quarterly Labour Force Surveys. We show that outliers and missing data need to be taken into consideration when working with these data. Our results show that overall mean real earnings among employees has risen over this period. Median real earnings, by contrast, have lagged. We show that the top end of the earnings distribution has moved away from the median, while there seems to have been a relative compression of the distribution right at the bottom.

Keywords: South Africa, inequality, earnings, data quality, multiple imputation

This is a REDI3x3 *affiliated paper* which, though not contracted within the Project, is closely related to the Project work; in addition, some of its original data work was funded by REDI3x3.

The **Research Project on Employment, Income Distribution and Inclusive Growth** is based at SALDRU at the University of Cape Town and supported by the National Treasury. Views expressed in REDI3x3 Working Papers are those of the authors and are not to be attributed to any of these institutions.

Wages and wage inequality in South Africa 1994-2011: The evidence from household survey data [1]

Martin Wittenberg

(DataFirst, SALDRU and School of Economics, University of Cape Town)

Inequality is again on the centre-stage in the international debate, with the run-away success of Piketty's book "Capital in the Twenty-First Century" (Piketty 2014). In many countries in the world inequality has widened over the last few decades, with gains disproportionately concentrated at the top end of the income distribution. This may be due to the weakening of redistributive policies in many OECD countries: tax rates on top earners and on corporations have come down, while union power has significantly weakened across the board.

South Africa is in some respects an anomaly: the government that came into power in 1994 was explicitly committed to redistribution and was in alliance with the trade union movement. Unlike in many other countries, union influence was actually extended through the labour relations system during this period. Nonetheless, as we will show later, inequality in labour earnings still widened during this period.

The South African case is therefore of considerable interest more generally. Indeed Piketty opens his analysis (2014, p. 39) with the Marikana massacre. That conflict was sparked by the suspicion that the bosses were benefitting more from growth than the workers, i.e. the impression of widening inequality is itself becoming a cause of labour instability.

Understanding wage trends in South Africa post-1994 is therefore important: it is interesting to South Africans in its own right, because the question of the distribution of the benefits of growth is central to the political debate; and it might illuminate the processes underpinning the growth of inequality more generally.

Analysing the trends should be fairly straightforward, given that South Africa has been collecting survey data on earnings since at least 1994. Nevertheless as noted in Wittenberg and Pirouz (2013) the series is bedevilled by breaks in the measurement process. So one of the

[1] This paper draws on and extends work reported on in two working papers, Wittenberg and Pirouz (2013) and Wittenberg (2014). Some of the original data work was done with the support of an infrastructure grant to DataFirst from the Redi3x3 project on "Employment/Unemployment, Income Distribution and Inclusive Growth". The Vice-Chancellor's Strategic Fund of the University of Cape Town paid for the initial construction of the PALMS dataset without which this research would not have been possible. The ILO commissioned some of the initial analyses of wage trends, which are reported in Wittenberg (2014). I would like to acknowledge the helpful comments and support from Patrick Belser and Kristen Sobeck in that work.

more significant general lessons to emerge from South Africa is the importance of careful measurement for the analysis of inequality.

In this paper we will present our best bet at what the long-run trends are, dealing with as many of the measurement issues as possible. In the process we will argue that confronting them is not a minor component of the research: it is absolutely central if one wishes to get the trends right.

The plan of the discussion is as follows: in the next section we review the literature. We then talk about the data underpinning our study and focus on three key measurement issues: changes in the instrument, outliers and missing data (including bracket responses). We show why these issues matter. In Section 3 we discuss the methods for analysing the data. In particular we discuss the algorithms for detecting outliers and multiple imputations. In Section 4 we show the impact of different types of data quality adjustments, while in section 5 we discuss wage trends and the evolution of wage inequality. We argue that there is compelling evidence that the top end of the earnings distribution has shifted away from the median, while inequality at the bottom end of the distribution has become reduced.

1. Literature review

The earnings information in South African household surveys has been used to many ends: to estimate discrimination (e.g. Rospabé 2002), investigate the union wage premium (e.g. Schultz and Mwabu 1998, Butcher and Rouse 2001), or to estimate returns to education (e.g. Keswell and Poswell 2004). Less has been written about the trends in earnings themselves. There is a literature looking at the long-run trends using the macroeconomic series based on firm surveys (e.g. Mazumdar and van Seventer 2002, Klein 2012), but that data series is subject to its own problems (see Wittenberg 2014).

Several papers have used survey evidence to discuss wage trends from 1994 to the early 2000s (see for example Casale 2004, Burger and Yu 2007). One of the issues which was debated was whether or not there had been a precipitous decline in real earnings over this period. Much of this was concentrated in the informal sector. Casale suggested that better enumeration of low wage work might explain some of the decline, but argued that

> "the fall in informal self-employment earnings between 1995 and 2001 is unlikely to be the result of improved data collection alone, as more and more people crowding into already low income-generating informal activities would be expected to depress average earnings even further." (Casale 2004, p.264).

Burger and Yu (2007), by contrast, suggest that the large drop in informal sector earnings is concentrated between October 1999 and February 2000, due to changes in the capturing of informal employment attendant on the introduction of the Labour Force Survey. They conclude that while there may have been a decline in earnings between 1995 and 1998 for for-

mal sector workers, the overall trend in earnings is likely to have been upward. They also highlight that extreme values ("millionaires") contaminate the trends observable overall.

Neither of these papers discuss how they deal with respondents who provided responses only in brackets. Nevertheless this issue is likely to be of some importance, since Burger and Yu attribute some excess "millionaires" in the October Household Surveys to this type of reporting:

> "This is due mainly to changes in the earnings intervals that individuals were allowed to specify without revealing their exact incomes, which permitted all workers in 1995 and the self-employed in 1996 to 1998 to answer in higher income brackets than were available to respondents in the subsequent years." (Burger and Yu 2007, p.6).

The issue of bracket responses is discussed explicitly by Posel and Casale (2006), who note that people who prefer to give only ranges for their income look different on many dimensions from people who give actual amounts. They go on to consider various ways of converting the bracket information to actual earnings. They suggest that mid-point imputations (likely to have been used by Casale 2004 and Burger and Yu, 2007) perform reasonably. We'll discuss the question of imputations in more detail below. Von Fintel (2007) considers what happens when these sorts of imputations appear on the right hand side of regression analyses. He concludes also that mid-point estimates do not produce dramatically misleading results.

This line of research has not continued, mainly because the earnings question was removed from the Quarterly Labour Force Surveys (QLFS) that replaced the Labour Force Survey in 2008. The reason for taking out the question was, *inter alia*, criticism from an IMF delegation that assessed labour market statistics. Their objection (cited in Statistics South Africa's official response) was that

> Data on earnings are collected each survey but considered to be poor quality, especially for the self-employed. Question has relatively high refusal/non-completion rates. Data are not published; hence the importance of the data is not appreciated by the survey officers. (Statistics South Africa 2008, Section 2.3.5, pp.7-8).

The earnings question did reappear in late 2009, but the microdata was not released with the public release of the QLFS. The earnings information from 2010 and 2011 has become available, although it was released separately (and considerably later) as the "Labour Market Dynamics in South Africa" study.

This work (see also Wittenberg 2014) therefore presents the first discussion of the longer-run trends. It is also the first contribution which deals with the bracket responses and missing data across the entire time run using a multiple imputation framework (Vermaak 2012 does so for the LFSs only). As we argue below, this approach is likely to be the best overall, because it allows us to create confidence intervals correctly.

2. The data and measurement issues

The original data for this study come from the October Household Surveys (OHSs), Labour Force Surveys (LFSs), and Quarterly Labour Force Surveys (QLFSs). The earnings information for the latter was retrieved from the "Labour Market Dynamics in South Africa" releases for 2010 and 2011. In order to facilitate comparative work over this period, we used the PALMS version of these datasets (Kerr, Lam and Wittenberg 2013), which harmonises variable definitions over time.

One additional issue that requires some attention is that changes in the assumptions of the demographic models that are released with the datasets can produce shifts in estimates that are unrelated to "real" shifts (Branson and Wittenberg 2014). Consequently we use the cross entropy weights (ceweights2) released with PALMS.

2.1. Changes in the instrument

The instrument, broadly conceived, encompasses the questionnaire as well as fieldwork implementation. Wittenberg and Pirouz (2013) discuss how the earnings questions evolved over the period 1994 to 2012. Daniels (2013) looks at the period 1997 to 2003 in more detail. Some of the big changes in the questionnaire are:

- The shift from capturing net income in 1994 to gross income thereafter;
- The shift from two earnings questions in the OHSs (for wage work and self-employed income) to only one question in the LFS and then back to two separate questions in the QLFS, but with a prior question which prevents individuals from reporting both types of income;
- The brackets used in the early OHSs (1994 and 1995) versus those used thereafter.

Arguably, however, the biggest measurement change in the period is produced by the huge increase in coverage of marginal workers (in particular in the informal sector) between the last OHS (October 1999) and the first LFS (February 2000). Much of this is among self-employed agricultural workers, but there are detectable shifts in most sectors. Wittenberg (2014) shows that many of the additionally enumerated workers record low hours and low earnings. As a result the earnings distribution for the self-employed shifts dramatically between these periods. Thus far it has proved difficult to deal sensibly with this discontinuity in the series. For this paper we will therefore restrict our analysis to the wage data of employees. This will present a reasonable picture of the evolution in the formal sector. Given that the informal sector is small by international standards (see Kingdon and Knight 2001) this is not a major distortion of trends in South Africa. Indeed for much of the political debate, wage trends among employees are the salient issue in any case.

One further problem with the change between the OHSs and the LFSs is that the OHSs seem to have undersampled small households (Kerr and Wittenberg 2013). It is unclear at this stage what impact that might have had on the wage distribution in the earlier period.

One last issue that needs to be mentioned is that there are also shifts in the post-fieldwork processing of the survey data over this period. The 1994 data is heavily imputed. We deal with this by reverse-engineering the imputations and separating the respondents giving point values originally from those who supplied bracket responses (see Wittenberg 2008a). The earnings information from the QLFS is also supplied with imputations. In this case, however, we cannot separate out genuine point values from point values created by the imputation process.

2.2. Extreme values

Burger and Yu (2007) already drew attention to the undue influence that "millionaires" play in some of the surveys. Table 1 shows the situation for employees. We show both the raw counts of sample "millionaires" (in real 2000 Rands) as well as an estimate of how many people these sampled cases represent in the population. It is evident that there are considerably more millionaires in 1999 and September 2000 than in any of the other surveys. 1998 and the second and third quarters of 2010 also look to be on the high side.

Table 1 Number of millionaires among employees (in constant year 2000 Rands) per survey

	unweighted		weighted			unweighted		weighted	
Survey	**n**	**prop**	**total**	**prop**	**Survey**	**n**	**prop**	**total**	**prop**
1994		0		0	**05:1**		0		0
1995	2	0.000097	1 865	0.000211	**05:2**	4	0.000262	3 052	0.00031
1997		0		0	**06:1**		0		0
1998	10	0.001089	8 990	0.001048	**06:2**		0		0
1999	43	0.003576	27 570	0.003235	**07:1**	2	0.000117	824	0.000079
00:1	1	0.000174	614	0.000071	**07:2**	2	0.000125	2 794	0.000259
00:2	20	0.001049	14 357	0.001526	**10:1**	6	0.000334	3 678	0.000318
01:1	1	0.000059	86	9.70E-06	**10:2**	10	0.00056	6 277	0.000548
01:2	4	0.000247	2 466	0.000276	**10:3**	11	0.000644	7 511	0.000664
02:1		0		0	**10:4**	6	0.000358	3 611	0.000315
02:2	1	0.000068	2 441	0.000276	**11:1**	1	0.000061	1 041	0.000091
03:1		0		0	**11:2**	4	0.000243	3 737	0.000327
03:2		0		0	**11:3**	3	0.000173	1 937	0.000166
04:1		0		0	**11:4**	6	0.000335	2 647	0.000224
04:2		0		0					

The central point, however, is that these cases (although they are few in number) have sufficiently high incomes that they have a marked influence on the overall mean (as we will show below). A plausible explanation for the year-on-year shifts in sampled "millionaires" is that

they reflect differences in post-fieldwork data cleaning protocols[2]. It is obviously not desirable for a handful of cases to swing the entire trend.

2.3. Bracket responses and missing values

Posel and Casale (2006) made the case that the people who respond in brackets are materially different from those that give point values. Indeed the fraction of respondents who use the bracket option increases almost monotonically with income, as shown in Table 2 (for the Labour Force Surveys). The exception (at least in the initial LFSs) is the open category, which is anomalous due to the small numbers in it. This response pattern implies that using only the point values would seriously underestimate mean wages and inequality.

Table 2 Proportions of respondents in each bracket giving point values, by LFS

Salary category	00:1	00:2	01:1	01:2	02:1	02:2	03:1	03:2
None	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R 1 - R 200	0.890	0.939	0.867	0.892	0.880	0.862	0.890	0.846
R 201 - R 500	0.877	0.922	0.889	0.855	0.864	0.872	0.873	0.857
R 501 - R 1 000	0.808	0.913	0.838	0.845	0.835	0.821	0.829	0.815
R 1 001 - R 1 500	0.703	0.845	0.765	0.733	0.717	0.710	0.737	0.680
R 1 501 - R 2 500	0.625	0.849	0.741	0.750	0.704	0.695	0.712	0.697
R 2 501 - R 3 500	0.526	0.849	0.662	0.655	0.594	0.600	0.609	0.577
R 3 501 - R 4 500	0.499	0.773	0.562	0.607	0.507	0.493	0.482	0.474
R 4 501 - R 6 000	0.513	0.777	0.580	0.611	0.518	0.523	0.492	0.455
R 6 001 - R 8 000	0.463	0.762	0.500	0.562	0.501	0.449	0.444	0.429
R 8 001 - R 11 000	0.473	0.661	0.464	0.448	0.398	0.383	0.372	0.336
R 11 001 - R 16 000	0.452	0.646	0.458	0.436	0.383	0.294	0.341	0.279
R 16 001 - R 30 000	0.336	0.668	0.398	0.338	0.401	0.272	0.303	0.297
R 30 000 or more	0.704	0.918	0.712	0.649	0.519	0.535	0.610	0.465
	04:1	**04:2**	**05:1**	**05:2**	**06:1**	**06:2**	**07:1**	**07:2**
None	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R 1 - R 200	0.870	0.868	0.858	0.870	0.886	0.863	0.881	0.843
R 201 - R 500	0.867	0.890	0.870	0.852	0.874	0.894	0.921	0.917
R 501 - R 1 000	0.847	0.881	0.854	0.858	0.877	0.888	0.897	0.905
R 1 001 - R 1 500	0.700	0.762	0.733	0.765	0.737	0.772	0.829	0.866
R 1 501 - R 2 500	0.697	0.703	0.715	0.734	0.751	0.758	0.805	0.825
R 2 501 - R 3 500	0.574	0.597	0.632	0.612	0.657	0.673	0.725	0.730
R 3 501 - R 4 500	0.461	0.521	0.549	0.540	0.567	0.598	0.635	0.651
R 4 501 - R 6 000	0.459	0.453	0.488	0.529	0.523	0.566	0.642	0.608
R 6 001 - R 8 000	0.441	0.446	0.464	0.505	0.505	0.489	0.581	0.563
R 8 001 - R 11 000	0.307	0.358	0.402	0.396	0.449	0.452	0.517	0.444
R 11 001 - R 16 000	0.303	0.361	0.352	0.398	0.419	0.406	0.440	0.479
R 16 001 - R 30 000	0.260	0.273	0.244	0.319	0.436	0.395	0.420	0.422
R 30 000 or more	0.373	0.356	0.325	0.362	0.401	0.389	0.409	0.345

[2] Since this table is based only on employees, the numbers cannot have anything to do with the brackets available for the self-employed, as conjectured by Burger and Yu (2007).

We obviously have no direct information on the earnings of individuals who give neither bracket nor point value. Posel and Casale (2006) suggest that their characteristics suggest that they are more likely to be higher income earners. Our results support that argument.

2.4. The case of zero earnings

Vermaak (2012) has argued that how one deals with individuals recording zero income can markedly affect one's analysis, particularly if one is investigating the "working poor". There are several mechanisms that might lead to a zero being recorded:

- The worker earns a positive income, but is lying or wants to signal that it is a pittance.
- The individual is self-employed and is not valuing consumption from own production or from inventories (in the case of traders). Zero surplus at the end of the period is equated with zero income.
- The individual works in a family enterprise and is not factoring in that they are receiving income in kind.
- The individual is self-employed and in fact made a loss, but can't report this.
- The individual is working as a volunteer (e.g. to gain experience).
- The individual is working on some deferrred compensation scheme. Own-account agriculture which may yield income only at harvest time may be an example of this also.

The majority of these cases are species of measurement error. They are also cases where it is somewhat doubtful whether the "data generating process" is directly comparable to those of individuals reporting positive earnings. Wittenberg and Pirouz (2013) suggest that zero earners are a big issue only in the LFSs (which are the surveys that Vermaak analysed), and then only among the self-employed, in particular subsistence agriculture workers. Since we are excluding the self-employed from this analysis we have chosen not to attempt to correct for the zero earners.

3. Methods for dealing with the measurement issues

3.1. Identifying extreme values

Burger and Yu (2007) flagged "millionaires" as extreme and showed that their removal affected the trend in mean incomes. While this procedure is simple and makes their point effectively, it is somewhat arbitrary and it risks removing some genuine high values along with the dubious ones. We used three procedures in order to flag suspicious observations: a) the BACON algorithm for outlier detection; b) extreme studentised regression residuals and c) robust regression.

3.1.1. Outlier detection

There are a number of outlier detection algorithms available (see Billor, Hadi and Velleman 2000 for a review). The BACON algorithm (Billor et al. 2000) has been implemented as a Stata routine (Weber 2010) and was used as the first approach.

The basic problem of every outlier detection algorithm is that the presence of the outliers can contaminate any statistics calculated to detect those outliers. The BACON algorithm begins with a small subset of observations assumed safe from contamination and then incrementally adds (in blocks) observations that are "close" to the existing safe set. The distance measure used to assess "closeness" is the Mahalanobis distance

$$\sqrt{(x_i - \bar{x})' S^{-1} (x_i - \bar{x})}$$

where $\bar{x}$ and S are the mean and covariance matrix calculated on the "safe" set and x_i is the vector under consideration.

One of the drawbacks of this procedure is that it works better on continuous data. Most of the covariates available to us, however, are discrete and this blunts the ability of the procedure to find anomalous data. It should be noted that these distance measures work better also with symmetrically distributed data, so the earnings data was logged before applying the routine. The BACON algorithm found only five clearly anomalous observations when using education categories as covariates. The flagged outliers were, indeed, all anomalous.

3.1.2. Studentised regression residuals

The BACON algorithm does not distinguish between a "dependent" and "independent" variables – all variables are treated symmetrically. Economists, however, tend to think of earnings as the outcome in which education and, perhaps, occupation, are the explanatory variables.

Another way of looking for extreme values is therefore to estimate a Mincerian style regression and to look for large residuals. We ran this regression pooling over all surveys, but using survey specific intercepts, and as additional explanatory variables gender, race (both interacted with survey), a quadratic in age, education categories and occupation categories.

One issue that has to be addressed when trying to identify "extreme" residuals, is that points of high leverage will tend to be associated with smaller residuals. The "Studentised residual" corrects for that. It is defined as

$$r_i = \frac{e_i}{\sqrt{s^2_{(i)}(1 - h_i)}}$$

where e_i is the unstandardized residual, $s^2_{(i)}$ is an estimate of the residual variance with the *i*-th observation removed and h_i is the leverage.

> "Studentized residuals can be interpreted as the t statistic for testing the significance of a dummy variable equal to 1 in the observation in question and 0 elsewhere (Belsley, Kuh, and Welsch 1980). Such a dummy variable would effectively absorb the observation and so remove its influence in determining the other coefficients in the model." (StataCorp 2013a, p.1877).

Studentised residuals with absolute values greater than five were flagged as extreme. The probability of being flagged if the regression errors are normally distributed is $5.735*10^{-7}$. On a sample of around 500 000, one would expect on average to see 0.3 observations this extreme. In practice the procedure flagged 476 observations, including a number deemed to be implausibly low.[3]

3.1.3. Robust regression

One problem with using the "Studentised residual" approach is that the presence of a group of outliers will again contaminate the data, making it harder to identify the bad observations. Robust regression attempts to deal with this issue recursively (similar to the BACON approach): progressively downweighting observations that appear to be problematic until the regression results and the weights are in agreement. Running this procedure led to over 1000 observations being weighted right down to zero. Observations with zero weights are identified as not belonging in the regression, i.e. they are outliers. Every one of the "extreme" values identified through the studentised residuals was also identified as problematic by the robust regression.

To err on the side of caution we used the second of our three approaches, i.e. we flagged only the subset of observations with extreme studentised residuals and not all of the others identified by the robust regression.

3.2. Dealing with brackets and missing values

There are two broad approaches to dealing with missing values: one can reweight the observed values to account for the missing ones, or one can impute for the missing data. There are several ways of doing the latter, as we'll discuss below.

3.2.1. Reweighting

Reweighting approaches are typically used to deal with unit non-response, but the same logic can be used for item non-response. Wittenberg (2008b) discusses how to apply this approach in the context of bracket response and income data. Individuals giving point values within a particular bracket are weighted up by the inverse of the probability of giving a point value response. Consider, for instance, the first column of Table 2 (for the LFS 2000:1). We see that individuals within the bracket R 1-200 gave actual Rand responses 89% of the time. These individuals would therefore get revised weights of $w_i/0.89$ while individuals within the bracket R 16 001-R 30 000 will get weights of $w_i/0.336$, i.e. they will be weighted up relative to individuals in the lowest bracket. Underpinning this approach is the idea that once we control for the bracket, the information is "missing at random". This approach is, in fact, also adopted by the imputation approaches considered below[4]. If this assumption is unwar-

[3] It should be noted that observations with missing information on the covariates would not get flagged as outliers.

[4] Posel and Casale (2006) argue that this assumption is dubious and that the estimation of a Heckman selection model gets around this. Unfortunately if income is really a determinant of who responds ***within*** a bracket, then

ranted, then some form of the EM algorithm discussed by Wittenberg (2008b) would become necessary.

3.2.2. Deterministic imputation: means, midpoints and conditional means

All of the imputation procedures discussed by Posel and Casale (2006) fall into this category. One of the simpler methods is to assign the category means to individuals giving bracket responses. The global mean with this procedure is, in fact, identical to the one calculated by the reweighting method. Other distributional statistics, however, will be different. Indeed one of the major drawbacks of both mean imputation and midpoint imputation is that it produces artificial spikes in the data at the imputation values, which would affect, for example, the percentiles.

Another issue that midpoint imputation confronts is what to do about the "open" category. The typical procedure is to take some multiple of the lower bracket boundary. Both von Fintel (2007, p.297) and Yu (2011, p.14) suggest that the factor that should be used is 1.1, which is implausibly conservative. Simkins (personal communication) by contrast suggests that the value of the lower bracket boundary should be doubled, because the distribution in the upper tail is approximately Pareto with a coefficient of around two.

One way of avoiding excessive spiking is to use the predicted values from a Mincerian wage regression. Posel and Casale (2006) discuss various ways of specifying these, ranging from OLS estimated on the point values, to the estimation of a Heckman selection model, to running separate OLS regressions within categories. One of the difficulties that some of these procedures encounter is that they can lead to predictions outside the range of the bracket itself. This is obviously problematic.

There is a second problem. Assume that the predictions do, in fact, return the "true" conditional mean $\mu_{(y|x)}$. The actual value will be the mean plus an idiosyncratic error, i.e. $y = \mu_{(y|x)} + \varepsilon$. Even if the conditional means are not all located at the same point, the variance of these imputed values will be lower than the true variance, since the error term is omitted. The imputed values will still distort some of the higher order moments. This will affect, in particular, inequality measures calculated on the data.

3.2.3. Stochastic imputation: parametric and nonparametric

One way of incorporating the missing "noise" in the process of imputation is to explicitly add it to the imputation. There are broadly two ways in which this can be done: the error can be drawn from some distribution specified *a priori* (e.g. a normal, log-normal or uniform distri-

---

income belongs in the "selection" equation of the Heckman model and omitting it will produce inconsistent estimates of the probabilities. The standard Heckman model does not work if the dependent variable in the "main regression" is an explanatory variable in the selection equation.

bution) or it can be drawn from some empirical distribution (e.g. the actual point values within a bracket, or regression residuals observed within the dataset).

If the actual point values are used, the procedure is referred to as a "hot deck". This is one of the most popular ways in which imputation is performed by national statistics offices. Andridge and Little (2010) discuss some of the different ways in which this can be implemented. Typically the missing information is copied from an observation that looks similar on some relevant co-variates (e.g. gender, age, race or location). One of the key questions, however, is how to measure this similarity. If only one or two variables are used (e.g. gender and race), one risks copying information from individuals that are otherwise quite dissimilar (e.g. on education or union status). The more variables are used, however, the smaller the "donor pool" and this can lead to the same value being copied to many other observations.

One attractive option is not to match directly on the characteristics, but "predictive mean matching". This involves regressing the outcome (in this case earnings) on the available co-variates (in this case dummies for earnings brackets, gender, education, and whether it was an employee or somebody self-employed). Observations are then matched on closeness on the ***predicted*** outcomes. Notice that this will be defined even for individuals with missing earnings (provided the explanatory variables are not missing also). Unlike with the imputations of the predicted values discussed in the previous section, it is not the predicted value that is imputed but the actual value from an observation with similar predicted outcomes[5].

The advantage of drawing from an empirical distribution is that no new data is created. However this approach can obviously only be used where there is an empirical distribution from which data can be drawn. In the case of the 1996 October Household Survey information was **only** collected in brackets, so there is no empirical distribution of point values. In order to get around this issue, we drew from the 1997 OHS empirical distribution, deflated by the CPI, to account for price changes between the surveys[6].

3.2.4. Multiple imputations

One of the big problems with all of the imputation processes discussed so far is that the resulting value does not signal any of the uncertainties implicit in the process of producing the imputation. The imputed value is not a real measurement, although the estimation techniques will treat it as such; it is the true value plus some measurement error, but the error has been rendered invisible.

5 It is actually preferable to take the draw not from the predicted mean outcomes, but from somewhere in the "posterior distribution". For a fuller discussion see StataCorp 2013b.

6 Observe that this does not force the 1996 distribution to be the same as the 1997 one: to the extent to which there are fewer observations in higher income brackets, there will be fewer draws in those data regions than there would be for the 1997 missing values. It does assume that within brackets the shape of the distribution is not altered by inflation.

The theoretical solution in the case of stochastic imputation is to do the imputation multiple times and perform any statistical analyses (e.g. calculation of summary statistics) on all of the resulting datasets (Rubin 1987, StataCorp 2013b). In essence each realisation of the stochastic process used in the imputation produces a different view of what the "true" data might have been. By taking into consideration the differences in estimates **between** analyses run on different versions of the data, as well as by using the standard tools to estimate the variance of the estimators **within** any of the complete versions of the data, one can obtain appropriate point estimates and standard errors. To state this more precisely, assume that $\hat{\beta}_j$ is the estimate from the *j*-th complete dataset and that $\hat{U}_j$ is the corresponding estimator of the covariance-matrix of $\hat{\beta}_j$, then Rubin's multiple imputation estimate of **β** will be given by

$$\hat{\beta}_{MI} = \sum_{j=1}^{M} \hat{\beta}_j ,$$

where M is the number of multiple imputations (in our case 10). The estimate of the covariance matrix of $\hat{\beta}_{MI}$ will be given by

$$\hat{V}_{MI} = \bar{U} + \left(1 + \frac{1}{M}\right) B$$

where

$$\bar{U} = \sum_{j=1}^{M} \hat{U}_j$$

is the average of the "within" dataset estimates of the covariance and

$$B = \sum_{j=1}^{M} \frac{1}{M-1} (\hat{\beta}_j - \hat{\beta}_{MI})(\hat{\beta}_j - \hat{\beta}_{MI})'$$

is an estimate of the "between" dataset covariance (StataCorp 2013b, pp.64-65).

3.2.5. Imputing for observations where income is completely missing

In cases where we had neither bracket nor point value information (and this included the outliers flagged in the prior step) we first multiply imputed the brackets using an ordered logit model with province, gender, education, race, a quadratic in age and occupation as explanatory variables. The imputed brackets were then (along with gender and education) used to multiply impute Rand amounts, using predictive mean matching.

3.3. Estimating the standard errors

We use multiple imputation in order to get reasonable estimates of the standard errors of our coefficients. Nevertheless as Rubin's formula (in section 3.2.4) shows, we need an initial estimate of the covariance matrix to begin the process. We use the "linearised" variance estimator for our estimates of the mean, but there is no simple equivalent appropriate for the weighted percentiles and their ratios that we also report below. We estimate those

standard errors by means of a clustered bootstrap. That is not entirely satisfactory, because bootstrap methods are not guaranteed to work with data extracted by disproportional sampling. The confidence intervals reported for those estimates should therefore be interpreted as indicative rather than definitive.

4. Results: the impacts of the different data adjustments

The results of different approaches to correcting for outliers, brackets and missing values is shown in Table 3. We have only provided information up to 2002 in order to make the methodological points. The substantive discussion with our preferred estimation method will be deferred to the next section.

Table 3 Estimates of mean wage according different data quality adjustments

	Point values only		Reweighted		Imputations (no outliers)			
	outliers	removed	outliers	removed	mean	midpt	hotdeck	multiple
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1994	1725.6	1726.2	2123.0	2123.7	2121.7	2556.6	2099.8	2374.9
	(63.08)	(63.13)	(76.77)	(76.81)	(61.68)	(85.82)	(57.69)	(79.95)
1995	2620	2620.3	2793.6	2793.9	2793.9	2880.3	2815.6	3028.1
	(54.73)	(54.74)	(59.33)	(59.34)	(53.15)	(57.47)	(54.32)	(66.63)
1996						2864.6	2650.6	2838.1
						(84.72)	(75.42)	(100.7)
1997	2049.2	2050.1	2660.0	2660.9	2660.8	2653.7	2664.1	2867.5
	(42.5)	(42.51)	(95.37)	(95.39)	(52.77)	(60.29)	(55.41)	(70.15)
1998	2174.5	2044.8	2826.8	2667.8	2684.8	2575	2675.3	2817.9
	(90)	(75.37)	(111.01)	(96.57)	(68.33)	(67.95)	(72.03)	(79.7)
1999	3150.7	1984.3	3614.0	2663.2	2698.7	2747.2	2689.6	3093.7
	(327.01)	(77.62)	(259.53)	(84.85)	(66.26)	(74.57)	(68.73)	(111.25)
2000:1	1904.3	1878	2355.7	2332.2	2331.8	2391.5	2474.2	2446.7
	(80.22)	(73.01)	(90.96)	(85.78)	(69.45)	(84.94)	(74.63)	(72.67)
2000:2	5095.1	2400.8	5105.1	2593.6	2594	2748.1	2640.1	2699.1
	(1062.69)	(74.85)	(990.97)	(78.26)	(72.71)	(85.54)	(74.65)	(79.74)
2001:1	1989.7	1980.1	2451.0	2442	2442	2461.7	2538.9	2513.6
	(43.67)	(42.25)	(61.42)	(60.53)	(51.24)	(55.77)	(54.46)	(61.7)
2001:2	2137.3	2101.4	2586.0	2543.7	2544.5	2624	2625	2683.8
	(59.3)	(50.3)	(77.94)	(69.3)	(55.21)	(65.37)	(57.25)	(60.77)
2002:1	1937.5	1937.5	2466.8	2466.8	2466.8	2550.8	2562.8	2688.9
	(43.24)	(43.24)	(63.19)	(63.19)	(51.15)	(63.58)	(53.05)	(60.55)
2002:2	1886.6	1886.6	2538.9	2538.9	2538.9	2640.9	2734.3	2819.7
	(84.74)	(84.74)	(116.87)	(116.87)	(78.69)	(94.18)	(100.92)	(100.01)

Estimated standard errors in parentheses, correcting for clustering, but not correcting for imputations (except in the multiple imputations case)

The first issue is the impact of extreme values. The relevant contrast is between columns 1 and 2 and between 3 and 4. The impact of extreme values is pronounced in two surveys, viz. 1999 and September 2000, although even in 1998 they increase average wages by more than 5%. The offending entries have been highlighted in grey. It is worthwhile pointing out that in some cases average wages went up (slightly) with the removal of extreme values. The

reason for this is that our approach to outlier detection also removed observations deemed to have implausibly low earnings.

The second issue is the case of the brackets. The most relevant contrast is between columns 2 and 4. The former column uses only the point values, while the latter reweights those values to take into consideration the people providing brackets. It is important to note that in other respects both estimates use the same information – the same point values and the same samples. Correcting for brackets raises average incomes by 24% over the period.

We next consider how the "reweighting" approach compares to the imputations. We have four imputation approaches in the table: column 5 gives bracket responders the mean income of the point values recorded within the bracket, while column 6 gives the midpoint and twice the lower bound for the "open" category. Column 7 does a single "hot deck" round, correcting both for bracket responses and cases where the earnings information was completely missing. Column 8 is the multiple imputation version of what is done in column 7. Indeed column 7 is one of the imputation rounds calculated for the results in column 8.

Comparing first the reweighting approach to mean and midpoint imputations, we observe that they are generally very close. Indeed reweighting and mean imputation should lead to identical results, bar any rounding errors. There is a very small disagreement between the two sets of estimates in 1998 and 1999 which is due to marginal differences in the data cleaning steps. The results with the mid-point estimates are also reasonably close, although those are slightly higher, to the tune of 2%. Perhaps the adjustment for the open category was a bit too generous. We observe, however, that neither reweighting nor mean imputation is possible with the 1996 data, where we only have bracket information.

If the "reweighting" approach and these deterministic imputation methods give such similar results, why not simply run with these very simple data adjustments? There are two major reasons for not following this route. Firstly, the standard errors will be wrong (compare for instance the standard errors in columns 5 versus 4 or 8). Secondly, other moments of the distribution will, in fact, be affected. Table 4 shows estimates of several percentiles as well as of the p90/p10 ratio (a measure of inequality) according to the "reweighting" and "mean imputation" approaches. While the means of the two series are very close to each other, this cannot be said for the other percentiles. Differences in excess of 10% are not uncommon in Table 4. The problem, of course is that the mean imputation method creates spikes in the data and depending on where these are located they will lead to either an over- or underestimate of any particular percentile. The extent of the distortion will depend on the absolute numbers of observations at the spike. In short mean and midpoint imputation have little to recommend them, certainly when compared to the "reweighting" approach, which is also fairly easy to implement.

Correcting also for the individuals who refused to provide point values and brackets is done in columns 7 and 8. Comparing the final column to the "reweighting" approach in column 4 we see that average wages are around 7% higher when we bring the completely missing information also into consideration. This supports Posel and Casale's (2006) contention that the characteristics of outright refusers look more like the high bracket earners than the rest of the population.

Table 4 Selected percentiles estimated according to the reweighting and mean imputation approaches

	p10 mean	p10 rewt	%	p25 mean	p25 rewt	%	p50 mean	p50 rewt	%
1994	304.8	358.5	-15.0%	725.7	737.3	-1.6%	1816.8	1567.5	15.9%
1995	446.1	409.3	9.0%	955.0	955.0	0.0%	1835.9	1882.7	-2.5%
1997	441.9	403.5	9.5%	1032.6	930.2	11.0%	1744.2	1744.2	0.0%
1998	373.5	320.2	16.7%	870.8	747.1	16.6%	1667.4	1600.9	4.2%
1999	367.3	314.8	16.7%	810.3	682.1	18.8%	1460.8	1574.0	-7.2%
2000:1	310.6	310.6	0.0%	672.4	621.1	8.2%	1390.1	1434.4	-3.1%
2000:2	340.6	314.7	8.2%	688.3	639.1	7.7%	1474.9	1474.9	0.0%
2001:1	336.2	332.8	1.0%	691.6	665.5	3.9%	1440.9	1440.9	0.0%
2001:2	339.0	329.6	2.9%	713.5	659.1	8.3%	1506.6	1506.6	0.0%
2002:1	317.6	317.6	0.0%	687.6	635.2	8.2%	1375.2	1451.9	-5.3%
2002:2	330.0	296.4	11.3%	659.9	592.7	11.3%	1439.5	1439.5	0.0%
	p75 mean	**p75 rewt**	**%**	**p90 mean**	**p90 rewt**	**%**	**p90/p10 mean**	**p90/p10 rewt**	**%**
1994	2902.8	2612.5	11.1%	3936.1	4182.9	-5.9%	12.91	11.67	10.7%
1995	3618.0	3492.5	3.6%	6341.8	6139.2	3.3%	14.22	15.00	-5.2%
1997	3488.4	3398.8	2.6%	5814.0	5651.2	2.9%	13.16	14.01	-6.1%
1998	3428.2	3201.7	7.1%	5941.4	5336.2	11.3%	15.91	16.67	-4.6%
1999	3371.6	3148.0	7.1%	5676.0	5771.2	-1.7%	15.45	18.33	-15.7%
2000:1	3105.6	2898.6	7.1%	5683.4	5176.0	9.8%	18.30	16.67	9.8%
2000:2	2964.1	2949.9	0.5%	5899.7	5899.7	0.0%	17.32	18.75	-7.6%
2001:1	3002.0	2881.8	4.2%	5331.4	5763.7	-7.5%	15.86	17.32	-8.5%
2001:2	3057.9	3295.7	-7.2%	5649.7	5649.7	0.0%	16.67	17.14	-2.8%
2002:1	2994.6	3176.0	-5.7%	5444.6	5444.6	0.0%	17.14	17.14	0.0%
2002:2	3386.1	3210.0	5.5%	6173.2	5757.8	7.2%	18.71	19.43	-3.7%

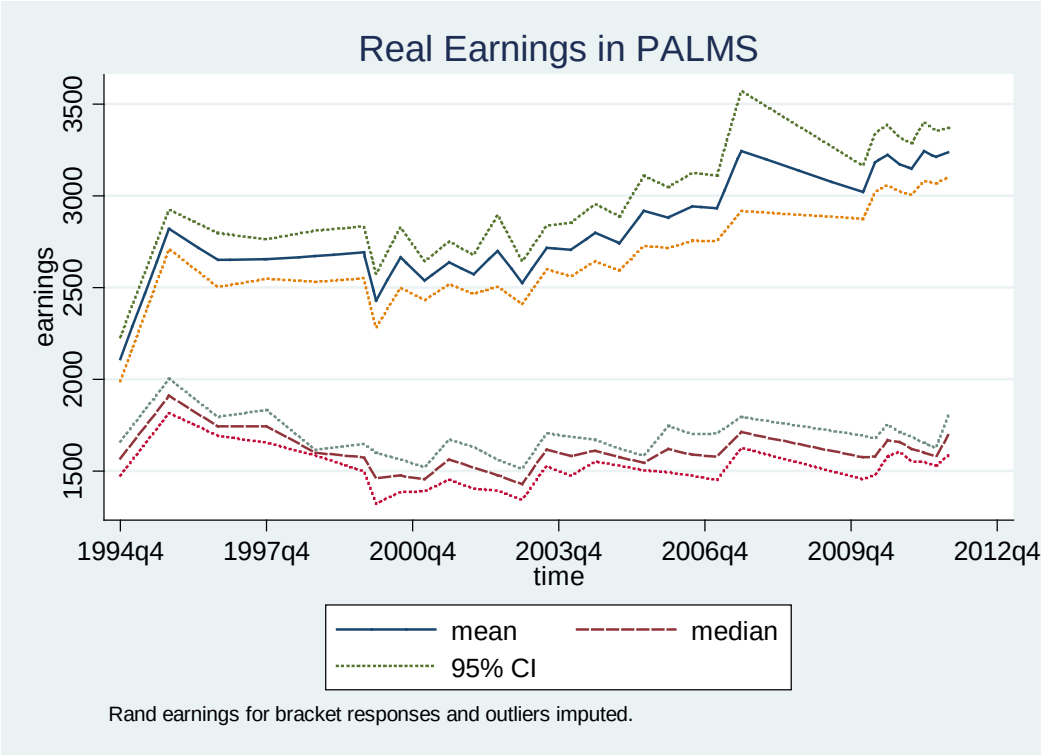
The final issue that we would like to address is the effect of doing the imputation multiple times, rather than just once. The key reason is to estimate the standard errors correctly. The impact of doing it can be seen by comparing the estimated standard errors in columns 7 and 8 of Table 3. On average across the surveys listed, the estimated standard error in column 7 is only 86% of the estimated standard error in column 8. As indicated by the theory, doing the imputation only once provides biased and inconsistent estimates of the covariance matrix.

5. Trends in wages and wage inequality

Our preferred estimation method is therefore the multiple imputation approach. The first set of results are shown in Figure 1. Several points stand out. The 1994 numbers are clearly

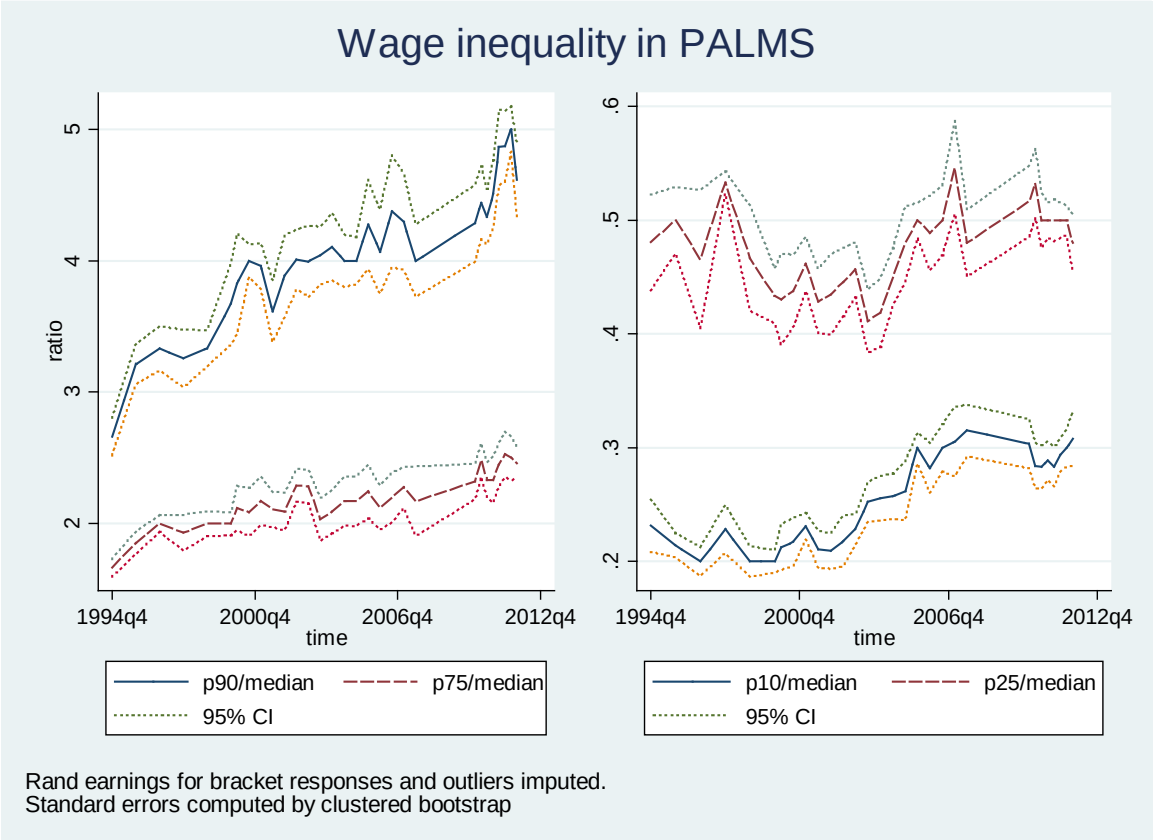
too low. As we pointed out above, this is the only period where the concept is net rather than gross earnings. Secondly the dip between October 1999 and February 2000 is clear in both the mean and the median earnings series. This is due to the better enumeration of more marginal forms of employment. Thirdly we see that the OHS earnings series (particularly when looking at median earnings) look a bit high. We suspect that this is due to the undersampling of small households (in particular hostels, backyard shacks and domestic workers). From February 2000 to the end of the series, however, both mean and median earnings are on an upward trajectory. The rate of increase, however, is stronger in the case of the mean than the median. The point estimates suggest that mean real earnings increased by 33% between 2000 and the end of 2011 while median real earnings increased by 16%. The figures convert to average annual increases of 2.45% and 1.26% respectively.

Figure 1 Real earnings among employees, 1994-2011



It is therefore clear that increases have gone disproportionately to higher income earners. This is shown in more detail in Figure 2. In this figure we graph various ratios: earnings of the 90th and the 75th percentiles relative to the median (left panel) and the 25th and 10th percentiles relative to the median (right panel). It is immediately apparent that earnings at the top end have increased strongly relative to the median – with much bigger gains at the 90th percentile than the 75th. Judging by the confidence intervals shown in those figures it is very unlikely that these trends reflect sampling noise.

Figure 2 Wage increases of employees at different percentiles relative to the median



The right-hand panel in that figure suggests that the gap between the 25th percentile and the median has stayed constant over time, while the 10th percentile has shifted up. This suggests some compression of the earnings distribution right at the bottom.

Conclusion

We have argued both a methodological and a substantive point. Methodologically it is clear that wage trends cannot be read off the raw data, without paying attention to the data quality issues enumerated above. Furthermore considerable care has to be taken to get the "right" point estimates, particularly if one wants to move beyond a simple concern with the mean. Our substantive results suggest that the mean obscures a number of important shifts in the underlying wage distribution. It is clear that inequality in earnings among employees has increased over the post-apartheid time period. Some of the mechanisms underpinning inequality increases in other societies – such as decreases in union power – should not have been operating in South Africa. Clearly more work needs to happen to unpick the mechanisms underlying these shifts. Such work must, however, start from a basis where the data issues are adequately dealt with. The multiply imputed incomes provided in PALMS (Kerr, Lam and Wittenberg 2013) are a good place from where to begin such analyses.

* * *

Datasets

Kerr, Andrew, David Lam and Martin Wittenberg (2013), Post-Apartheid Labour Market Series [dataset], Version 2.1, Cape Town: DataFirst [producer and distributor], 2013. [zaf-datafirst-palms-1994-2012-v2.1]

Statistics South Africa, Labour Force Surveys 2000.1–2007.2 [datasets], Pretoria: Statistics South Africa [producer], Cape Town: DataFirst [distributor], 2013.

Statistics South Africa, Labour Market Dynamics in South Africa 2010–2011 [datasets], Pretoria: Statistics South Africa [producer], Cape Town: DataFirst [distributor], 2012.

Statistics South Africa, October Household Surveys 1994–1999 [datasets], Pretoria: Statistics South Africa [producer], Cape Town: DataFirst [distributor], 2013.

Statistics South Africa, Quarterly Labour Force Surveys 2010.1–2011.4 [datasets], Pretoria: Statistics South Africa [producer], Cape Town: DataFirst [distributor], 2013.

References

Andridge, R. R. and R. J. A Little (2010), "A review of hot deck imputation for survey nonresponse," ***International Statistical Review***, 78(1), 40–64.

Billor, N., A.S. Hadi and P. F. Velleman, (2000) "BACON: blocked adaptive computationally efficient outlier nominators", ***Computational Statistics and Data Analysis,*** 34, 279–298.

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

Burger, Rulof and Derek Yu (2007), "Wage trends in post-Apartheid South Africa: Constructing an earnings series from household survey data", Working Paper 07/117**,** Development Policy Research Unit, University of Cape Town.

Butcher, Kristin and Cecilia Rouse (2001), "Wage Effects of Unions and Industrial Councils in South Africa", ***Industrial and Labor Relations Review***, 54 (2), 349-374.

Casale, Daniela (2004), "What has the feminisation of the labour market 'bought' women in South Africa? Trends in labour force participation, employment and earnings, 1995-2001", ***The Journal of Interdisciplinary Economics,*** 15, 251–275.

Daniels, Reza (2013), "Questionnaire design and response propensities for employee income microdata", Working Paper 89, SALDRU, University of Cape Town. Revised version. Retrieved from www.saldru.uct.ac.za on 18 April 2013.

Kerr, Andrew, and Martin Wittenberg, (2013), "Sampling methodology and field work changes in the October Household Surveys and Labour Force Surveys", DataFirst Technical Paper Number 21, Cape Town: DataFirst, University of Cape Town.

Keswell, Malcolm and Laura Poswell (2004), "Returns to Education in South Africa: A Retrospective Sensitivity Analysis of the Available Evidence", ***South African Journal of Economics***, 72 (4), 834–860.

Kingdon, Geeta and John Knight (2001), "Why high open unemployment and small informal sector in South Africa?" University of Oxford: Centre for the Study of African Economies.

Klein, Nir (2012), "Real Wage, Labor Productivity, and Employment Trends in South Africa: A Closer Look", IMF Working Paper, WP/12/92, International Monetary Fund.

Mazumdar, Dipak and Dirk van Seventer (2002), "A Decomposition of Growth of the Real Wage Rate for South Africa: 1970-2000", ***South African Journal of Economics***, 70(6), 1076–1102.

Piketty, Thomas (2014), ***Capital in the Twenty-First Century***, Cambridge, Mass.: Belknap Press of Harvard University Press.

Posel, Dorrit, and Daniela Casale (2006) "Who Replies in Brackets and what are the Implications for Earnings Estimates? An Analysis of Earnings Data from South Africa", Economic Research Southern Africa, Working Paper 7.

Rospabé, Sandrine (2002), "How Did Labour Market Racial Discrimination Evolve after the End of Apartheid? An Analysis of the Evolution of Employment, Occupational and Wage Discrimination in South Africa between 1993 and 1999", ***South African Journal of Economics***, 70 (1), 185–217.

Schultz, T. Paul and Germano Mwabu (1998), "Labor Unions and the Distribution of Wages and Employment in South Africa", ***Industrial and Labor Relations Review***, 51(4), 680–703.

StataCorp, (2013a), ***Stata 13 Base Reference Manual***, Stata Press, College Station, TX.

------- (2013b), ***Stata 13 Multiple Imputation Reference Manual***, Stata Press, College Station.

Statistics South Africa, (2008), "Report on the response by Statistics South Africa to recommendations of the International Monetary Fund on improvements to the Labour Force Survey", mimeo. Available at http://www.statssa.gov.za/qlfs/docs/improvements_to_lfs.pdf downloaded on 8 June 2013.

Vermaak, Clare (2012), "Tracking poverty with coarse data: evidence from South Africa", ***Journal of Economic Inequality***, 10:239–265.

von Fintel, Dieter (2007), "Dealing with earnings bracket responses in household surveys: How sharp are midpoint imputations?" ***South African Journal of Economics,*** 75(2), 293–312.

Weber, Sylvain (2010) "st0197. bacon: An effective way to detect outliers in multivariate data using Stata (and Mata)", ***Stata Journal,*** 10(3), 331–338.

Wittenberg, Martin (2008a), "Income in the October Household Survey 1994", DataFirst technical paper 7, University of Cape Town, available at http://datafirst.uct.ac.za/images/docs/DataFirst-TP08_07.pdf

---- (2008b), "Nonparametric estimation when income is reported in bands and at points", Working Paper 94, ERSA, available at http://www.econrsa.org/papers/w_papers/wp94.pdf.

---- (2014) "Analysis of employment, real wage, and productivity trends in South Africa since 1994", ILO Conditions of Work and Employment Series, Working paper No. 45, March 2014 available at http://www.ilo.org/public/libdoc/ilo/2014/114B09_23_engl.pdf

----- and Farah Pirouz, (2013) "The measurement of earnings in the post-Apartheid period: An overview", DataFirst technical paper 23, September 2013, available at http://www.datafirst.uct.ac.za/images/docs/DataFirst-TP13_23.pdf

Yu, Derek (2011), "Some factors influencing the comparability and reliability of poverty and inequality estimates across household surveys, Paper presented to MASA conference, available at http://www.aceconferences.co.za/MASA%20FULL%20PAPERS/Submission%20-%20Derek%20Yu1.pdf.

The **Research Project on Employment, Income Distribution and Inclusive Growth (REDI3x3)** is a multi-year collaborative national research initiative. The project seeks to address South Africa's unemployment, inequality and poverty challenges.

It is aimed at deepening understanding of the dynamics of employment, incomes and economic growth trends, in particular by focusing on the interconnections between these three areas.

The project is designed to promote dialogue across disciplines and paradigms and to forge a stronger engagement between research and policy making. By generating an independent, rich and nuanced knowledge base and expert network, it intends to contribute to integrated and consistent policies and development strategies that will address these three critical problem areas effectively.

Collaboration with researchers at universities and research entities and fostering engagement between researchers and policymakers are key objectives of the initiative.

The project is based at SALDRU at the University of Cape Town and supported by the National Treasury.

Consult the website www.redi3x3.org for information on research grants and scholarships.

Tel: (021) 650-5715

www.REDI3x3.org

REDI3X3

Research Project on Employment, Income Distribution & Inclusive Growth

Towards an integrated response to unemployment, inequality and poverty