SALDRU

Southern Africa Labour and Development Research Unit

The top tail of South Africa's earnings distribution 1993-2014: Evidence from the Pareto distribution

Martin Wittenberg

Working Paper Series
Number 224, Version 1

About the Author(s) and Acknowledgments

Authors:
Martin Wittenberg: School of Economics, SALDRU and DataFirst, University of Cape Town, South Africa

Acknowledgments:
Funding from REDI3x3 for this project is gratefully acknowledged as are helpful comments from a REDI3x3 anonymous reviewer. Helpful feedback was also provided by participants at the Economic Society of South Africa biennial conference in Grahamstown, August 2017. Much earlier versions of this paper were presented at seminars at the University of Michigan and at a Data Quality workshop hosted by DataFirst in Cape Town.

This research report was first published in December 2017 as Working Paper 46 of the Research Project on Employment, Income Distribution and Inclusive Growth (REDI3x3), funded by the South African National Treasury and based at SALDRU, University of Cape Town: www.redi3x3.org

Recommended citation
Wittenberg, M. (2018). The top tail of South Africa's earnings distribution 1993-2014: Evidence from the Pareto distribution. Cape Town: SALDRU, UCT. (SALDRU Working Paper Number 224).

ISBN: 978-1-928281-85-6

Working Papers can be downloaded in Adobe Acrobat format from www.opensaldru.uct.ac.za. A limited amount of printed copies are available from the Senior Administrative Officer: SALDRU, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 1808, Fax: (021) 650 5697, Email: tania.hendricks@uct.ac.za

The top tail of South Africa's earnings distribution 1993-2014: Evidence from the Pareto distribution

Martin Wittenberg

Saldru Working Paper 224
University of Cape Town
March 2018

Abstract

We estimate the parameters of a Pareto distribution for South African earnings as measured through the October Household Surveys, Labour Force Surveys and Quarterly Labour Force Surveys, as assembled in the Post-Apartheid Labour Market Series (PALMS). We develop an outlier detection algorithm consistent with this distribution and then adjust the Gini coefficient for inequality in the top tail, using the robust estimation technique of Cowell and Flachaire. That procedure suggests that wage inequality is a bit higher than conventionally estimated. We also show that the top tail of the South African earnings distribution is 'thick tailed' and explore what that means. Our analyses show big shifts in the distribution in some of the surveys in ways that suggest measurement changes rather than changes in the underlying distribution.

JEL Codes: C13, C14, J31
Keywords: Pareto distribution; earnings; inequality

The top tail of South Africa's earnings distribution, 1993-2014: Evidence from the Pareto distribution*

Martin Wittenberg
School of Economics, SALDRU and DataFirst
University of Cape Town
South Africa

Abstract

We estimate the parameters of a Pareto distribution for South African earnings as measured through the October Household Surveys, Labour Force Surveys and Quarterly Labour Force Surveys, as assembled in the Post-Apartheid Labour Market Series (PALMS). We develop an outlier detection algorithm consistent with this distribution and then adjust the Gini coefficient for inequality in the top tail, using the robust estimation technique of Cowell and Flachaire. That procedure suggests that wage inequality is a bit higher than conventionally estimated. We also show that the top tail of the South African earnings distribution is "thick tailed" and explore what that means. Our analyses show big shifts in the distribution in some of the surveys in ways that suggest measurement changes rather than changes in the underlying distribution.

JEL Codes: C13, C14, J31
Keywords: Pareto distribution; earnings; inequality

1 Introduction

South Africa has long had the reputation of high levels of inequality. Many analysts (Bhorat, Van Der Westhuizen and Jacobs 2009, Leibbrandt, Woolard, Finn and Argent 2010, Leibbrandt, Finn and Woolard 2012, van der Berg 2011) concur that inequality in South Africa has not decreased in the post-apartheid era. Much of this literature has focused on the relationship between inequality and poverty. An issue that has received less academic attention is the fate of the relatively affluent. In popular imagery, however, the idea that "the rich get richer" (Blandy 2009) is fuelled both by newspaper reports of conspicuous consumption by the emerging black elite, as well as by the continued privileged position of South Africa's white population. Indeed Wittenberg (2017c) has shown that the top tail of the earnings distribution seems to have pulled away from the median over this period.

One of the difficulties in analysing this issue is that South Africa's household surveys are less suited to this purpose than for the analysis of poverty. Firstly, there are fewer rich individuals in

*Funding from REDI3x3 for this project is gratefully acknowledged as are helpful comments from a REDI3x3 anonymous reviewer. Helpful feedback was also provided by participants at the Economic Society of South Africa biennial conference in Grahamstown, August 2017. Much earlier versions of this paper were presented at seminars at the University of Michigan and at a Data Quality workshop hosted by DataFirst in Cape Town.

the society and so the chance of them being sampled is small. Furthermore those that are sampled are more likely to refuse to participate in surveys. The surveys attempt to compensate for this by "weighting" up those respondents that they do find. To the extent to which nonparticipants differ systematically from respondents, the resulting analysis may underestimate inequality. Secondly, even if they assent to be interviewed they are more reticent to divulge their earnings than people with lower incomes (Wittenberg 2017b). This is reflected *inter alia* in the fact that "bracket responses" are more common at the top end of the income distribution. Frequently bracket responses are given imputed values. Unfortunately this makes the analysis of this part of the income distribution sensitive to the nature of the imputations. This is likely to be particularly problematic for the top income category, where there are no bounds within which to impute. Thirdly, information about earnings other than labour income is likely to be poor, so that overall trends in inequality are likely to be understated (Wittenberg 2017a).

An additional issue that was flagged by Burger and Yu (2007) and Wittenberg (2014b, 2014a) is the problem of extreme values, many of which seem to be outright data errors. Cowell and Flachaire (2007) have noted that the reliable estimation of inequality measures needs to confront the dual issue of data contamination in the top tail as well as the sparseness of data points. They note that the top tails of income distributions tend to be "heavy-tailed" and random samples tend to underestimate the true weight in the upper tail. Consequently they suggest that the estimation of inequality should combine the typical nonparametric estimates (i.e. measuring inequality purely on the empirical distribution function) with a parametric estimation of the contribution of the top tail. This they do by using the Pareto distribution.

Our contribution in this paper is to estimate the parameters of a Pareto distribution of the distribution of earnings as measured in the Post-Apartheid Labour Market Series (PALMS) dataset (Kerr, Lam and Wittenberg 2016) and to assess how much the robust estimation technique of Cowell and Flachaire alters our understanding of earnings inequality. Neither of these have been done properly on South African survey data. In the process we will also highlight, yet again, the importance of measurement issues for the understanding of the trends. The paper also makes a number of methodological contributions. For instance we show how the Pareto distribution can be used to flag outliers.

The estimated parameters of the Pareto distribution are interesting in their own right. Jones (2015a, 2015b) has argued that there is a close connection between the dynamics of growth and the Pareto distribution. In his models the Pareto parameter turns out to be a function of the birth and death rates of firms. Fifty years earlier Mandelbrot (1960) argued that heavy-tailed distributions of the Pareto type are "stable" in the same sense that Gaussian distributions are, i.e. they are the sum of shocks each distributed also as Pareto-type. These distributions are likely to describe the data well in contexts where the outcome is the result of a small number of "big" shocks. Arguably earnings distributions fit this description, since increases due to promotions and changes in jobs tend to have a larger impact than incremental wage adjustments within a job category. In the case of self-employment income, where windfall gains are possible, this is even more likely to be the case. One of the features of these distributions is that they exhibit much higher levels of inequality than "thin-tailed" distributions like the Gaussian. Indeed extreme outcomes occur sufficiently frequently that the variance of the "stable" Paretian distributions have infinite variance. We will show below what this looks like in the context of South Africa's earnings distribution.

We begin our discussion with a brief review of the literature and of the data that we will be using. We then present our estimation strategy. This starts with a non-parametric view of the top tail of the distribution and continues to discuss three methods of estimating the Pareto parameter.

We use a pseudo-maximum likelihood technique in the rest of the paper. We turn to discuss the problem of outlier detection in the context of estimating Pareto parameters and then present the Cowell-Flachaire (2007) procedure. The results and discussion round off the paper.

2 Literature Review

The evolution of disparities in earnings between different groups have been discussed in a number of papers (Hlekiso and Mahlo 2006, Woolard and Woolard 2006, Burger and Yu 2007). Earnings inequality measures for South Africa have also been presented in a number of papers (Leite, McKinley and Osorio 2006, Heap 2009, Tregenna 2011, Tregenna and Tsela 2012). Wittenberg (2017b, 2017c) suggests that earnings inequality has increased over the post-apartheid period, but notes that measurement issues affect the reliability of these estimates. Some of the key points are:

- Changes in the instrument and sampling procedures:

  The October Household Surveys undersampled small households (Kerr and Wittenberg 2015) and as a result probably underenumerated certain types of workers (e.g. domestic workers living in the backrooms of their employers' homes). This problem is compounded by the fact that the LFSs found many more informal sector workers, particularly in agriculture (Neyens and Wittenberg 2016). This means a sharp disjuncture between the OHS and LFS wage and employment series, particularly in relation to self-employed workers.

- Extreme values

  Burger and Yu (2007) commented on the fact that some datasets seemed to contain many more "millionaires" than others. Wittenberg (2017b) discusses several possible "outlier detection" routines and shows that removing the extreme values has an appreciable effect on the mean of real earnings. His preferred method is based on a Mincerian regression. Observations with extreme standardised residuals (more than an absolute value of 5) are marked as outliers.

- Bracket responses

  Respondents that were unwilling to disclose a Rand earnings amount were given the option of specifying a range instead. Wittenberg (2017b) provides evidence that individuals who responded in brackets were more likely to be high earners. He also shows that imputing midpoints or means of the ranges (as much of the other literature does) is likely to distort both the estimate of the mean and of inequality measures. Instead he suggests that the bracket information can be used to reweight the point responses. Alternatively he suggests a multiple stochastic imputation routine.

- Missing values

  There are a number of respondents in the pre-QLFS surveys who supply neither Rand nor bracket information. Again these seem to be predominantly high income individuals. Wittenberg (2017b) argues that a multiple imputation routine provides the best way of dealing with these cases.

None of these approaches deals well with the situation of data contamination of a "heavy-tailed" distribution, such as the Pareto. The criterion for judging extreme values is based on what looks "extreme" in the context of a normal distribution. However this will lead to an over-rejection of

observations that might well look less extreme if the true distribution is Pareto. Secondly the multiple imputation routine is a version of a "hot deck", i.e. it is a draw from the empirical distribution function. This means that if there are too few high values to begin with (or if these have been over-zealously removed in the outlier detection routine) then they will not be created in the imputation. In this context the Cowell and Flachaire (2007) procedure looks more promising, since it takes the possibility of a heavy tail seriously.

The Pareto distribution has been used informally in the analysis of South Africa's income distribution. For instance the practice of imputing incomes in South Africa's top income bracket at twice the value of its lower bound is based on a Pareto coefficient estimate of 2 obtained by Charles Simkins (Simkins, personal communication). The Pareto distribution has been used formally in a paper by Fedderke, Manga and Pirouz (2004) which attempts to critique estimates of inequality and poverty on the basis of South African household surveys. Unfortunately that analysis is bedevilled by several faults. Firstly, the authors attempt to calculate per capita incomes on data which are not really suited to that analysis. In particular the OHSs that are not linked to income and expenditure surveys provide information on labour market earnings, but not on other types of income. Secondly, in so far as the labour earnings are utilised, the analysis does not seem to deal at all with the issues of incomes reported in brackets. Indeed it seems clear from the authors' discussion of the 1996 October Household Survey (where income was **only** reported in brackets) that the authors merely imputed incomes at the midpoint of each bracket. What they did in the top category is unclear. It is evident that this imputation strategy will heavily affect the parameter estimates. Thirdly they used rather generous definitions of "tails", i.e. the threshold above which they estimated the parameter was probably too low, as our analysis below will suggest. Fourthly their estimation strategy is not well described but seems to be the regression approach discussed in section 4.2.2 below, which is not the most efficient method available.

More recently Alvaredo and Atkinson (2010) have investigated the top tail of South Africa's wealth distribution using tax data and estimated various Pareto coefficients in the process. These coefficients were estimated from the income shares of groups (the top 1% and top 0.1%) and not off microdata.

3 The Data

We make use of the Post-Apartheid Labour Market Series (PALMS), version 3.1 (Kerr et al. 2016). This dataset combines the labour market information from the Project for Statistics on Living Standards and Development (1993), the October Household Surveys from 1994 to 1999, the biannual Labour Force Surveys from February 2000 to September 2007, and the Quarterly Labour Force Surveys from 2008 through to the fourth quarter of 2015. Earnings figures were not released with the QLFSs in 2008 and 2009, nor were the 2015 ones available at the time of doing this research. The earnings information in 1996 was collected exclusively in brackets and consequently we excluded that survey. We therefore have usable information from 42 separate surveys. Given the fact that we will restrict our analyses to individuals earning more than R6000 per month (in real June 2000 values) we end up with around 75 000 individually reported incomes.

The PALMS dataset provides several useful tools for analysing earnings across time. Firstly it has attempted to harmonise definitions. Secondly, it provides a set of harmonised sampling weights to ensure that shifts in the dataset are not due to simple shifts in the demographic models that underpin the weights (Branson and Wittenberg 2014). Thirdly it calculates a set of "bracketweights" which can be used to reweight the reported Rand incomes to account for individuals who responded

in brackets (Wittenberg 2014b). Fourthly it marks extreme values using the standardised residuals from a Mincerian regression as diagnostics (Wittenberg 2014b). In this paper we do not use the multiple imputations also released with PALMS, since it isn't clear that the methodology is consistent with the attempt to measure the Pareto coefficient.

For the purposes of this paper we excluded all self-employed agricultural workers, since they are measured inconsistently across time. The number of this type of worker increases by over a million between Octtober 1999 and February 2000. Furthermore at the end of the LFS period, this type of employment almost vanishes again in the survey data (Neyens and Wittenberg 2016).

4 Methods

4.1 Properties of the Pareto distribution

The Pareto distribution is defined by the cumulative distribution function

$$F(x) = 1 - \left(\frac{x_0}{x}\right)^{\alpha} \tag{1}$$

where x_0 is the cut-off defining the "tail" of the distribution and α is the Pareto parameter. This can be rewritten as the simple power law

$$\log(1-p) = \alpha \log x_0 - \alpha \log x \tag{2}$$

where p is the cdf evaluated at x.

The pdf of the Pareto is

$$f(x) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}}, \text{ where } x \geq x_0 \tag{3}$$

and it follows that

$$E(x|x \geq x_0) = \frac{\alpha}{\alpha - 1} x_0$$

Note that this is defined only if $\alpha > 1$. It is easy to verify that this relationship holds for any cut-points higher than x_0, i.e.

$$E(x|x \geq x') = \frac{\alpha}{\alpha - 1} x', \text{ for any } x' \geq x_0 \tag{4}$$

It is worth noting that the variance of the Pareto distribution is defined only if $\alpha > 2$. It is in this sense that it is potentially "fat-tailed" - extreme outcomes happen with a sufficiently high probability that the variance can be undefined.

The quantile function is given by

$$Q(F; q) = \frac{x_0}{(1-q)^{1/\alpha}}$$

Substituting $q = .99$ into this we see that the 99^{th} percentile is at a value of $\frac{x_0}{(0.01)^{1/\alpha}}$. Using equation 4 it follows that the expected earnings of those above the 99^{th} percentile will be $\frac{\alpha}{\alpha-1}\frac{x_0}{(0.01)^{1/\alpha}}$. Multiplying this by the number of people in this group, (which for the top 1% is $0.01 * N$) we can work out the total income of the top 1%., i.e. it will be $\frac{\alpha}{\alpha-1}\frac{x_0}{(0.01)^{1/\alpha}}0.01N$. Similarly the total

income of the top 0.1% will be given by $\frac{\alpha}{\alpha-1}\frac{x_0}{(0.001)^{1/\alpha}}0.001N$. It follows that the ratio of these two, which is equal to the ratio of their shares, will be given by

$$\frac{S_1}{S_{0.1}} = 10^{1-\frac{1}{\alpha}} \tag{5}$$

where S_1 is the share of the top 1% and $S_{0.1}$ the share of the top 0.1%.

4.2 Estimation strategies

The equations provided above provide at least four different *parametric* approaches to estimating α, but they also suggest a simple nonparametric sense-check of the data.

4.2.1 Nonparametric approach

The power law formulation in equation 2 is useful as a starting point, since it provides a simple visual check on whether the parametric approach is sensible or not. We graph $\log(1-p)$ against $\log x$. If the relationship is approximately linear, then a Pareto distribution is a reasonable summary of the shape of the tail distribution.

4.2.2 Regression

The power law equation can also be used to estimate α, i.e. we regress $\log(1-p)$ on $\log(x)$. This approach is likely to be less efficient than the pseudo-maximum likelihood version that we will adopt.

4.2.3 Method of moments

The conditional moment equation (4) can be used to define a method of moments estimator, i.e.

$$\widehat{\alpha}_{MoM} = \frac{\overline{x}}{\overline{x} - x_0}$$

where $\overline{x}$ is the sample mean in the top tail. Monte Carlo simulation studies (available from the author) suggest that this procedure is likely to be considerably less efficient than maximum likelihood.

4.2.4 Share of the top 0.1% within the top 1%

The ratio of the shares given by equation 5 is used by Alvaredo and Atkinson (2010) to estimate the Pareto coefficient on tax data, i.e.

$$\widehat{\alpha}_{share} = \frac{1}{1 - \log_{10}(S_1/S_{0.1})}$$

This requires knowing how much of the total income goes to the top 1% and the top 0.1% respectively. Since S_1 and $S_{0.1}$ are derived from the conditional means above $Q(F; 0.99)$ and $Q(F; 0.999)$ respectively this is an estimator that depends on the ratio of two sample moments far up the distribution and is likely to be noisy on survey data.

4.2.5 Maximum likelihood

We can use the pdf given in equation 3 to derive the maximum likelihood estimator

$$\widehat{\alpha}_{ML} = \frac{1}{\frac{1}{n}\sum \ln x_i - \ln x_0}$$

This is related to Hill's estimator of the rate of decrease of the distribution function in the tail (Hill 1975). His estimator is

$$\widehat{\alpha} = \frac{1}{\frac{1}{r}\sum_{i=1}^{r} \ln y^{(i)} - \ln y^{(r)}}$$

where $y^{(1)}, y^{(2)}, ..., y^{(r)}$ are the r largest values ranked from largest downwards. The main difference between our approaches is that we effectively take a fixed cut-off x_0, whereas the Hill estimator allows it to vary with the data. We show below some sensitivity analyses in which we vary x_0, which is akin to varying r, which is normally taken to be some fixed proportion of the sample (Cowell and Flachaire, for example, use $n/10$).

The reason why we prefer a fixed cut-point (corresponding to a fixed level of real earnings) is that it allows us to more effectively compare what happens to the tail over time, since the "tail" has a fixed definition. Furthermore, as we show below, this allows us to check for the presence of outliers among the largest values. The Hill procedure, by contrast, needs to assume that everyone of the r largest values is measured correctly.

An additional complication arises in the estimation of this model, given that we are dealing with complex survey data and higher nonresponse among rich South Africans. The underrepresentation of white South Africans in the national surveys is likely to be particularly problematic when dealing with the top incomes. There is little option but to use the sample design weights adjusted for nonresponse. This problem is exacerbated by the fact that we do a second level reweighting to account for individuals who gave bracket responses. This means that the observations in our data are not independently and identically distributed. Consequently our estimation procedure is a pseudo-maximum likelihood one, i.e. we assume that the population moment condition

$$E\left[\frac{\partial \ln L}{\partial \alpha}\right] = 0$$

can be consistently estimated by the weighted sample moment condition

$$\sum_i w_i \frac{\partial \ln L_i}{\partial \alpha} = 0$$

where w_i are the sample weights adjusted for bracket response.

We implement the estimation procedure using Stata's maximum likelihood routine, which allows us not only to weight the data but to calculate standard errors robust to clustering. These standard errors are markedly bigger than they would have been under the assumption of independent sampling.

4.3 Dealing with Outliers

A key question for the empirical analysis is how to flag outliers without relying on criteria derived from the normal distribution for judging which observations are extreme. It is useful to rehearse

some of the standard approaches used in the outlier detection literature (see (Billor, Hadi and Velleman 2000) for a review). One simple approach, adopted, for instance by Cowell and Flachaire (2007), is to successively delete each observation and see the impact this has on the parameter estimates. Observations that have a disproportionate impact can be flagged as problematic. The problem is that this doesn't deal with the potential that data contamination may involve a *cluster* of problematic observations. Indeed the empirical work in Wittenberg (2014b) suggests that this is often the case. More generally the problem is that the presence of outliers will contaminate any statistics calculated to detect those outliers. Consequently a standard approach is to begin with a small subset of observations assumed safe from contamination and then to add in observations that are deemed to also be "safe" given the empirical information in the safe set. The BACON algorithm (Billor et al. 2000), for instance, judges observations to be safe based on their Mahalanobis distance, i.e. $\sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}})}$, from the safe set, where $\overline{\mathbf{x}}$ and $\mathbf{S}$ are the mean and covariance matrix calculated on the safe set and $\mathbf{x}_i$ is the vector under consideration. In the case of a univariate distribution this measure is just $\frac{|x_i - \overline{x}|}{s}$, i.e. it is akin to a t-statistic evaluating the probability of observing an observation of size x_i (or more extreme) given that the true mean is $\overline{x}$. Indeed these statistics are compared to a t or normal distribution to assess the probability of the observation coming from the same distribution.

This test statistic will not work in the case of a Pareto distribution for two reasons. Firstly in many cases (including the South African one as we will show) the Pareto parameter is in the range where the variance is not defined, so that asymptotically the t-statistic does not exist. Secondly given the focus of the Pareto distribution on the upper tail, the "safe sample" will be asymmetrically defined and not picked around the median of the distribution. This means that $\overline{x}$ from the safe sample will not be a reasonable or consistent estimate of the population mean.

Our procedure starts from the assumption that the k smallest observations $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$ just above the cut-off x_0 constitute the safe sample. We pick $k = 100$, which is small in the context of the typical sample size in the top tail. We then obtain an initial estimate of the Pareto parameter $\widehat{\alpha}_1$ on the safe sample. We then calculate the probability of observing the observations $x_{(k+1)}, x_{(k+2)}, \ldots, x_{(n)}$ (or ones more extreme) on the assumption that the distribution is truly Pareto with parameter $\widehat{\alpha}_1$. The probability of observing $x_{(j)}$ or values higher than it will be given by

$$P\left(X \geq x_{(j)}\right) = \left(\frac{x_0}{x_{(j)}}\right)^{\widehat{\alpha}_1} \tag{6}$$

This probability can be compared to the empirical distribution function. Assume that w_j is the weight of observation $x_{(j)}$ and define the empirical cumulative weight $W_j = \sum_{i=1}^{j} w_{(i)}$. Then the empirical estimate of $P\left(X \geq x_{(j)}\right)$ assuming that $x_{(j)}$ is the *last* properly measured observation is

$$\widehat{p}\left(X \geq x_{(j)}\right) = \frac{w_j}{W_j} \tag{7}$$

We can compare the theoretical probability $P\left(X \geq x_{(j)}\right)$ to the empirical one $\widehat{p}\left(X \geq x_{(j)}\right)$. If the ratio is too small we would reject the idea that $x_{(j)}$ forms part of the same distribution as the safe sample. Our criterion is

$$\text{accept } x_{(j)} \text{ into the safe sample if, and only if, } \frac{P\left(X \geq x_{(j)}\right)}{\widehat{p}\left(X \geq x_{(j)}\right)} \geq \tau \tag{8}$$

The constant τ determines how easily the procedure accepts extreme values. Values of τ close to one will be less forgiving than values closer to zero. For our empirical analysis we used $\tau = \frac{1}{2}$. After the first iteration of the procedure a number of additional points will be flagged as part of the "safe sample". This new safe sample is then used to re-estimate the Pareto parameter to yield $\widehat{\alpha}_2$. The probability of observing points $x_{(j)}$ outside the safe sample are then recalculated according to formula 6 (with $\widehat{\alpha}_2$ rather than $\widehat{\alpha}_1$) and again compared to the empirical probabilities (equation 7) which may lead to yet further additions to the safe sample. The procedure terminates if either the entire sample is marked safe or the observations outside the safe sample have a much lower probability of occurring than their weight in the sample suggests. The final split into safe sample and outliers is internally consistent, in the sense that the Pareto parameter estimation is not contaminated by the outliers and the outliers look anomalous in light of the Pareto coefficient.

It should be noted that this outlier detection procedure will only capture anomalous observations outside the range of the "safe" values, i.e. if there is a data capture error (e.g. shifting the decimal point two places) which does not, however, move the observation far out into the top tail, it will not be caught by this procedure. This, of course, is equally true of most other univariate outlier detection algorithms. Furthermore unlike the regression procedure it does not take into account the values of any covariates. Lastly this procedure treats errors in the earnings distribution asymmetrically: implausibly large values will be marked as dubious and excluded from the analysis, while implausibly small ones will escape such scrutiny. Since we are less concerned about the bottom of the distribution this does not concern us here, but it would raise questions in a context where we want to investigate characteristics of the distribution as a whole.

4.4 Smoothing the estimation of the Pareto parameter

The estimation procedure outlined in the previous section looks at extreme values only in the context of one particular survey. Nevertheless in PALMS we have over fifty surveys, with earnings information in 42 of them. What may look anomalous in one survey may look less so when compared to adjoining periods. For this reason we also pool surveys within 8 quarters of a particular period and run the Pareto estimation/outlier detection algorithm outlined in the previous section on that pooled sample.

4.5 The Cowell-Flachaire procedure for robust estimation of means and inequality

Cowell and Flachaire (2007) argue that the potential of data contamination together with the fact that surveys are likely to underestimate the true importance of the top tail necessitate the use of hybrid estimation techniques. In particular they suggest that the distribution should be split into two: the top $(100p_{tail})\%$ and the bottom. Within the bottom part of the income distribution one would use the standard nonparametric estimation techniques, i.e. calculate the mean and inequality measures using the empirical distribution function. In the top part, however, one uses parametric estimates. More concretely, the population mean would be estimated as

$$\widehat{\mu} = (1 - p_{tail})\,\widehat{\mu}_0 + p_{tail}\widehat{\mu}_{tail} \tag{9}$$

The mean in the bulk of the distribution $\widehat{\mu}_0$ would be calculated in the usual way, but the mean of the upper tail $\widehat{\mu}_{tail}$ would be estimated as$\frac{\widehat{\alpha}}{\widehat{\alpha}-1}x^*$ (see equation 4) where x^* is the lower bound of

the upper tail as defined by the fraction p_{tail}. Effectively this discards the top $p_{tail}n$ observations and replaces them with the parametric estimate.

Cowell and Flachaire make the point that p_{tail} should be selected much smaller than the number of observations on which the Pareto parameter is estimated. Furthermore it should be selected so that $p \to 0$ as $n \to \infty$ to ensure consistency. In their analysis they pick $p_{tail} = 0.04 * n^{-\frac{1}{2}}$ which means that effectively $0.04 * n^{\frac{1}{2}}$ observations are not used in the "nonparametric" part of the estimation. In the case of PALMS this is a handful of observations per survey.

Cowell and Flachaire provide formulae for the Generalised Entropy or Atkinson inequality measures using the same general approach.

It can be shown that the Gini coefficient will be given by

$$Gini = (1 - p_{tail})(1 - s_{tail})G_0 + s_{tail} - p_{tail} + p_{tail}s_{tail}G_{tail}$$

where s_{tail} is the share of total income accruing to the top $(100p)\%$, G_0 is the Gini coefficient estimated nonparametrically on the bottom part of the distribution and $G_{tail} = \frac{1}{2\alpha-1}$ which would again be estimated using the Pareto coefficient. The total income accruing to the top tail is $\frac{\alpha}{\alpha-1}x^*p_{tail}N$ where N is the total population size. The total accruing to the bottom would be estimated in the standard way as $(1 - p_{tail})\widehat{\mu}_0 N$. The share s_{tail} can therefore be estimated as the ratio of $\frac{\alpha}{\alpha-1}x^*p_{tail}$ to $\widehat{\mu}$, which is estimated as in equation 9.

5 Results

5.1 Nonparametric Analysis

Our first look at the data is provided by Figure 1. The graph represents information from surveys two years apart. Several features are apparent in this diagram. Firstly many of these trajectories resemble straight lines for the bulk of the distribution, but change tack in the last few observations. Secondly we see that some surveys have a markedly different trajectory from the others. The October 1999 Household Survey is particularly noteworthy in this regard, but the third quarter of QLFS 2012 and the the third quarter of 2014 also look anomalous. Despite these problems linearity does not seem far-fetched and so we turn to parametric estimates.

5.2 Parametric estimates: regression, method of moments and maximum likelihood

Figure 2 presents three different approaches to the estimation of the Pareto coefficient discussed above. It is evident that the regression and method of moment estimates are more volatile survey-on-survey than the pseudo-maximum likelihood ones. To interpret these results it should be remembered that lower values correspond to much higher levels of inequality. Distributions with Pareto coefficients below 2 are so thick-tailed that they do not have a variance, and many surveys end up giving point estimates in this range, regardless of estimation method. It is also noticeable that there seem to be major changes in the size of the coefficient over implausible short time horizons. In particular the big reversal between October 1999 and February 2000 is astonishing.

The results of these preliminary analyses confirm our *a priori* assumptions that the pseudo maximum-likelihood approach will be the most reliable available. There are also ample indications that outliers and measurement issues will be important.

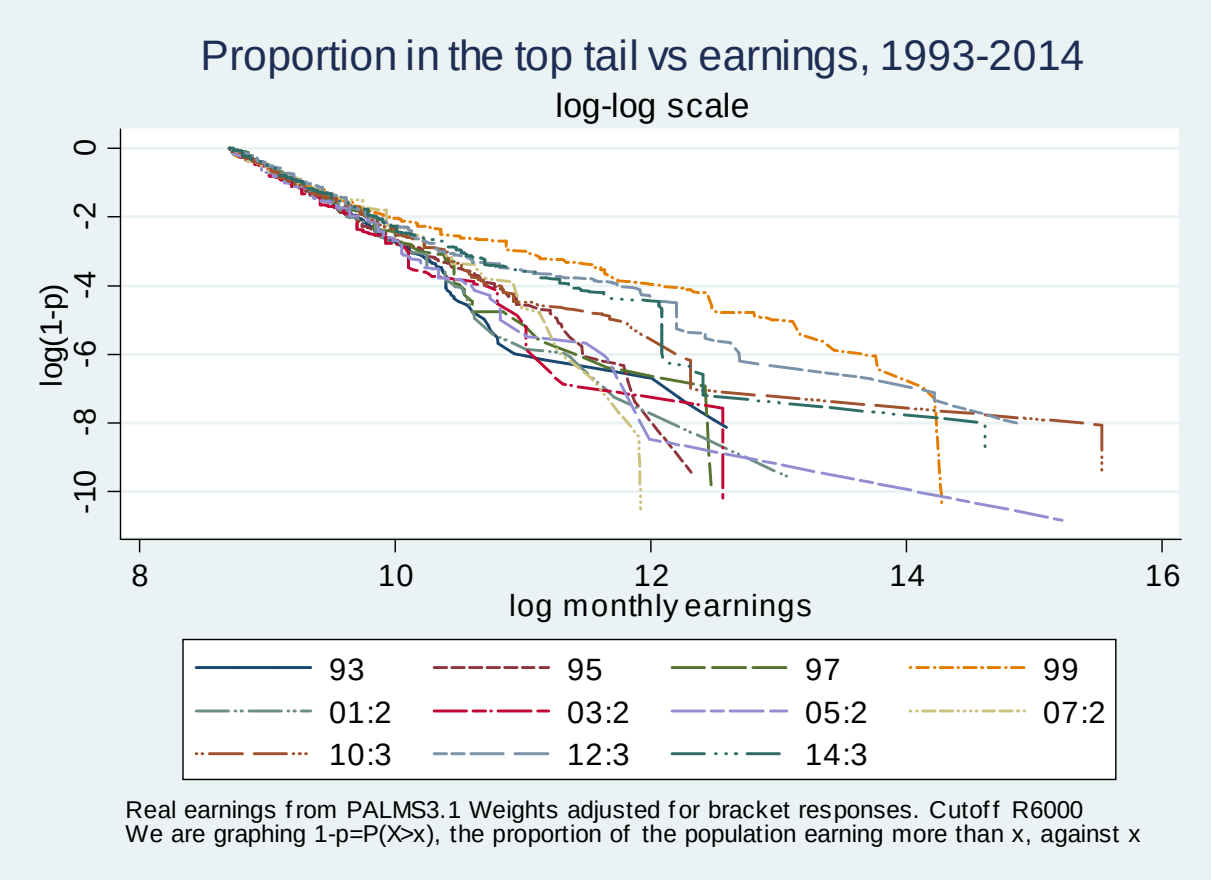

Figure 1: Top tail of the earnings distribution

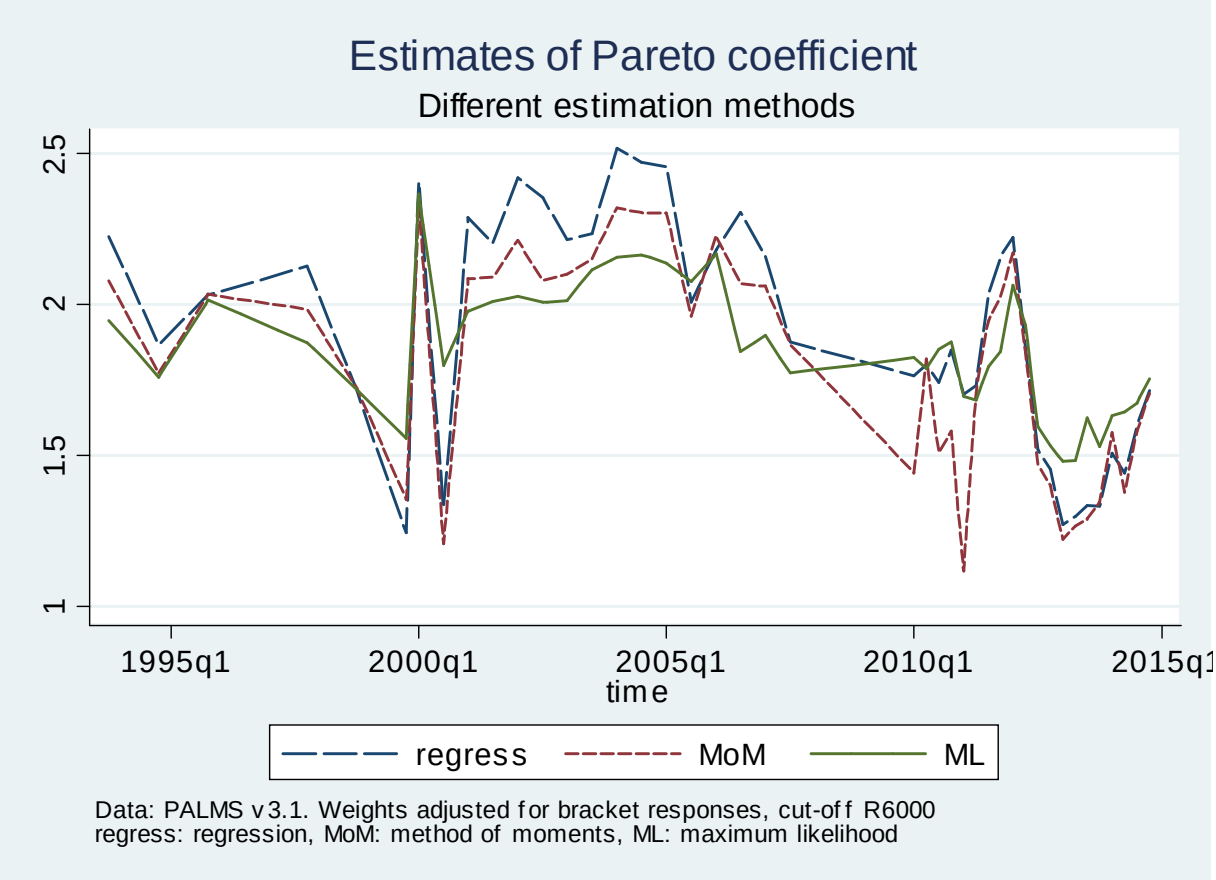

Figure 2: Comparing different estimation methods of the Pareto coefficient α

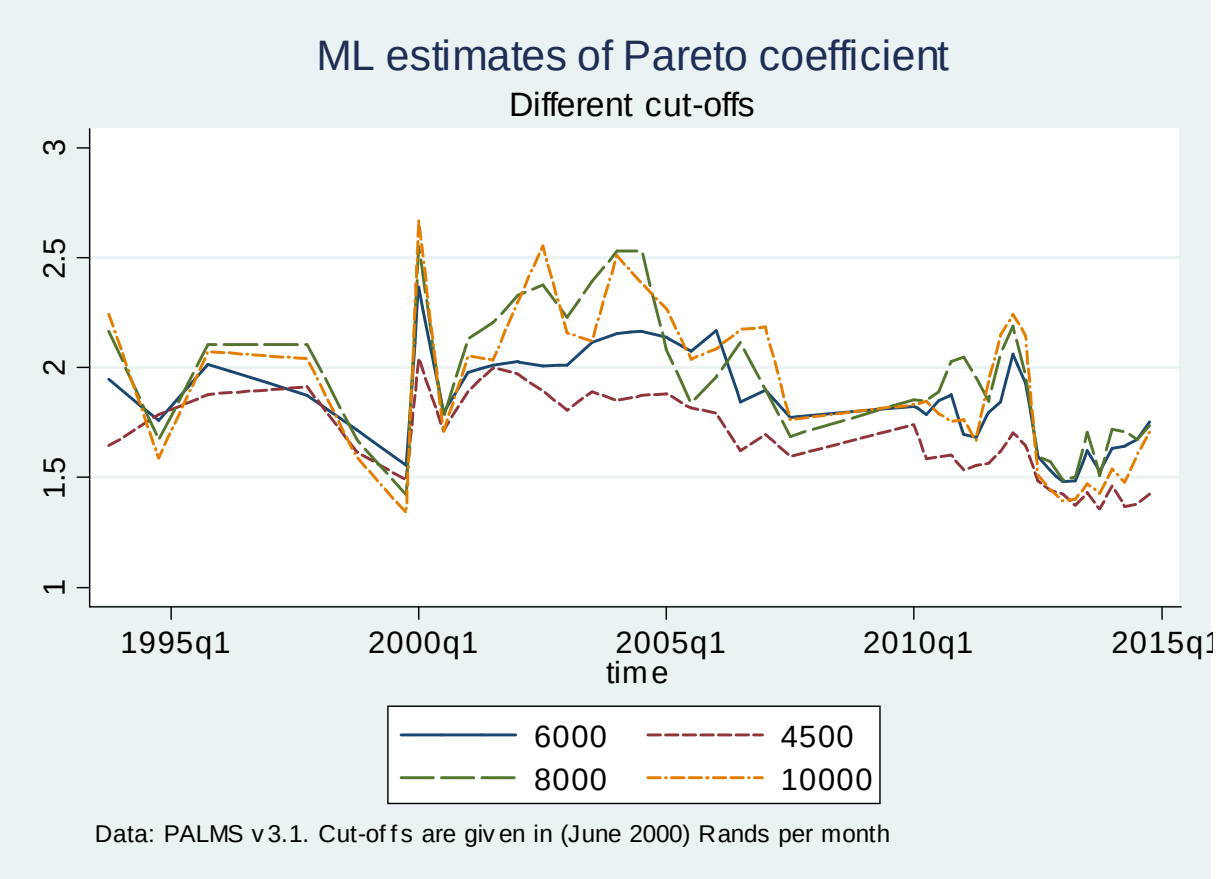

Figure 3: Different cut-offs for the top tail

5.3 The importance of the cut–offs

It is important to pick a value of the cut-off above which the Pareto coefficient remains relatively stable. In Figure 3 we present some evidence on how sensitive the results are to different choices of the boundary between the "top tail" and the rest. It appears that with the lower cut-off of R4500 per month the estimated Pareto coefficient is noticeably lower in virtually all time periods. The cut-off of R6000 provides lower Pareto estimates in the earl 2000s when compared to higher cut-offs, but this is not a consistent pattern. Given that there is a trade-off between the size of the sample on which the coefficient is estimated and the stability of the parameter, we thought that there was little evidence in favour of going to yet a higher value.

5.4 Outlier detection

We noted in section 2 that extreme values in the wage data have the potential to influence the estimated mean and indeed inequality measurs (Burger and Yu 2007, Wittenberg 2017b). The noise evidenced in Figure 1 also suggest that the data change from one survey to the next in ways which may be related to the presence of outliers. We adopted three approaches to outlier detection. Firstly we used the regression approach implemented in PALMS and discussed in Wittenberg (2017b) The procedure used was to estimate a Mincerian wage regression and to flag observations as outliers if their standardised residual was bigger than five This approach assumes the approximate normality of the residuals. It resulted in the removal of 345 observations (across all waves) from the analysis. Secondly we flagged observations as outliers based on the iterative procedure described in Section 4.3. This procedure led to the removal of only 61 observations. Interestingly enough four of these were not flagged by the regression routine, because key control variables were missing. The third approach was the smoothed approach utilising adjoining datasets, as described in Section 4.4. That approach was even more conservative, flagging only twenty-six observations as outliers.

The impact of the difference between the regression and the Pareto approach can be seen in Figure 4. Both panels should be compared to the "raw" distribution shown in Figure 1. It is evident that the most egregious "zig-zags" in the distribution have been eliminated. Nevertheless it is also clear that the regression approach has cleaned out high values more aggressively than the Pareto approach developed in this paper. Indeed there are no tail values left above a log value of 13 (around R440 000 per month). It is worth noting, since this will be of some importance in the discussion later, that the October 1999 and 2012 quarter 3 trajectories on the extreme right of both diagrams have been aggressively pruned by the regression approach, whereas they are largely intact on the Pareto approach. Indeed that is hardly surprising given that these approximate straight lines in this log-log space reasonably well.

5.5 The impact on the estimation of α

At the end of the iterative outlier detection algorithm we obtain an estimate of the Pareto parameter α which is not contaminated by outliers, but as Figure 5 shows the estimates are for the most part hardly affected. This is undoubtedly due to the fact that not that many observations were flagged as outliers through this routine. It is apparent that the sharp change in estimates (in particular that between October 1999 and February 2000) will lead to corresponding sharp changes in the inequality estimates obtained via the Cowell-Flachaire estimation procedure.

5.6 Smoothed Pareto estimation

We noted above that pooling datasets around a particular time "window" (eight quarters) flags even fewer observations as outliers. Nevertheless this procedure has a much stronger impact on the estimation of the Pareto coefficient, as adjoining surveys with quite different tail characteristics (e.g. October 1999 and LFS 2000:1 will be pooled). The impact is shown in Figure 6. The key question is whether pooling the datasets is just smoothing over major measurement shifts that would otherwise be clearly visible. This is particularly evident in this case since the discontinuity between the OHS and LFS series has been removed. One thing which the smoothing does make clear is that with the exception of the LFSs the Pareto coefficient is typically below 2 and more likely in the region of 1.8. Indeed if we were to pool all the available datasets we would get an estimate of 1.79. The LFSs are different in a number of other ways from the other datasets. They find more marginal

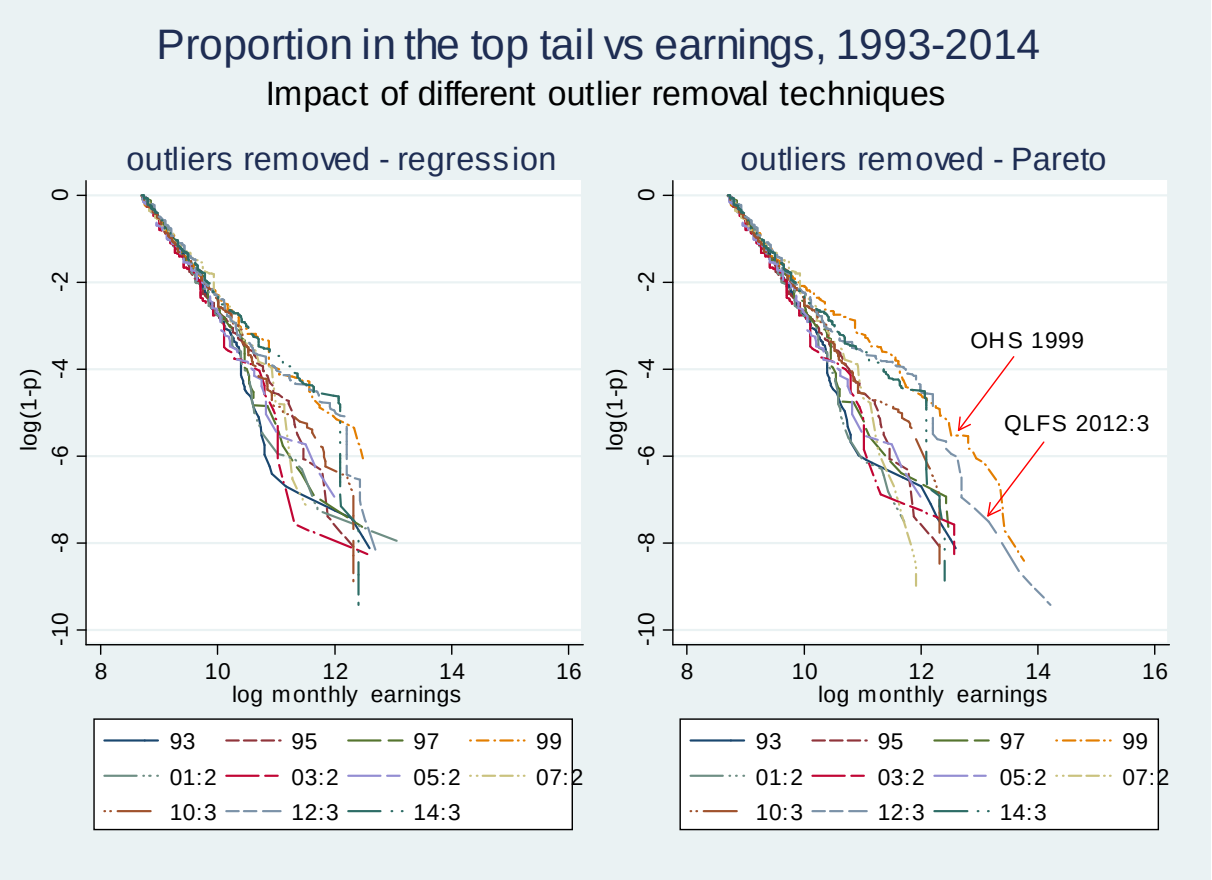

Figure 4: Impact of the outlier detection algorithm on the extent of the top tail

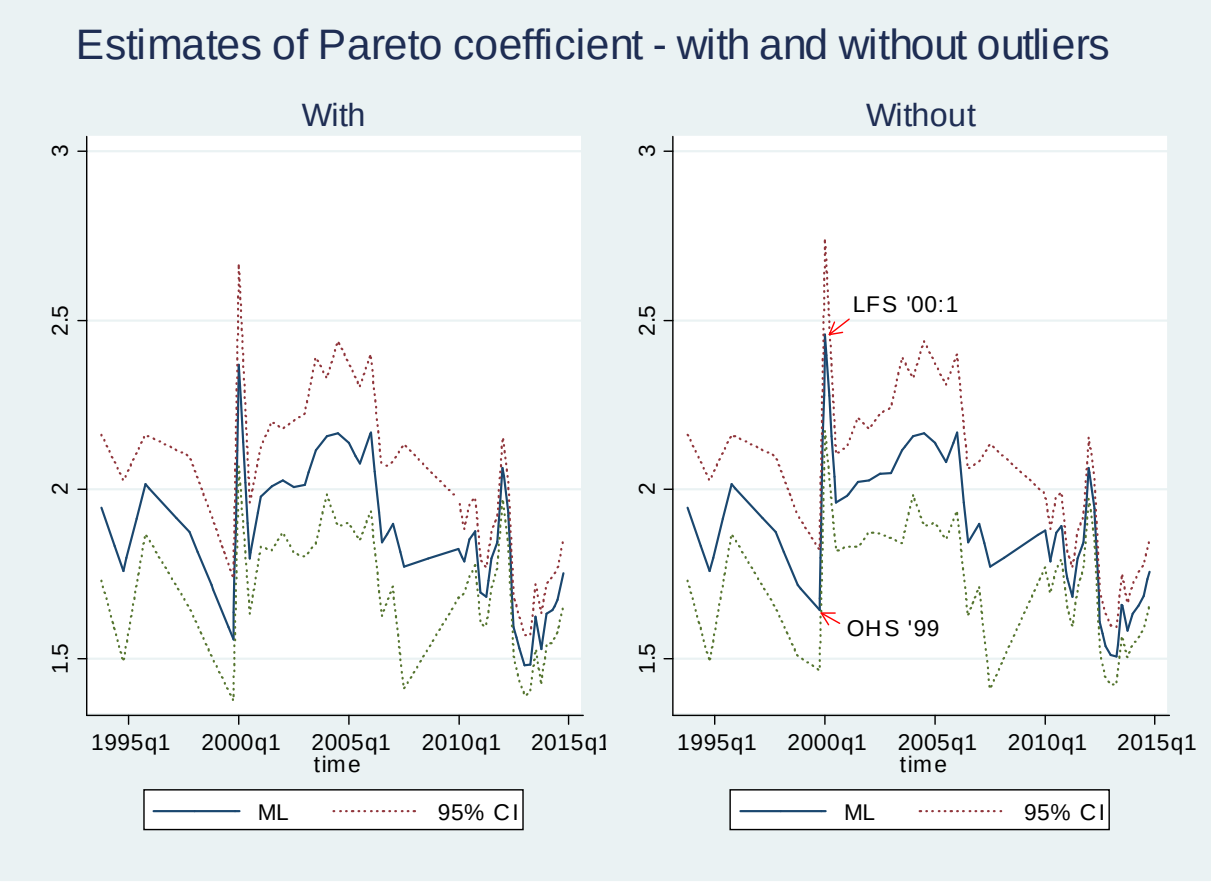

Figure 5: The impact of outlier removal on the estimation of the Pareto coefficient α

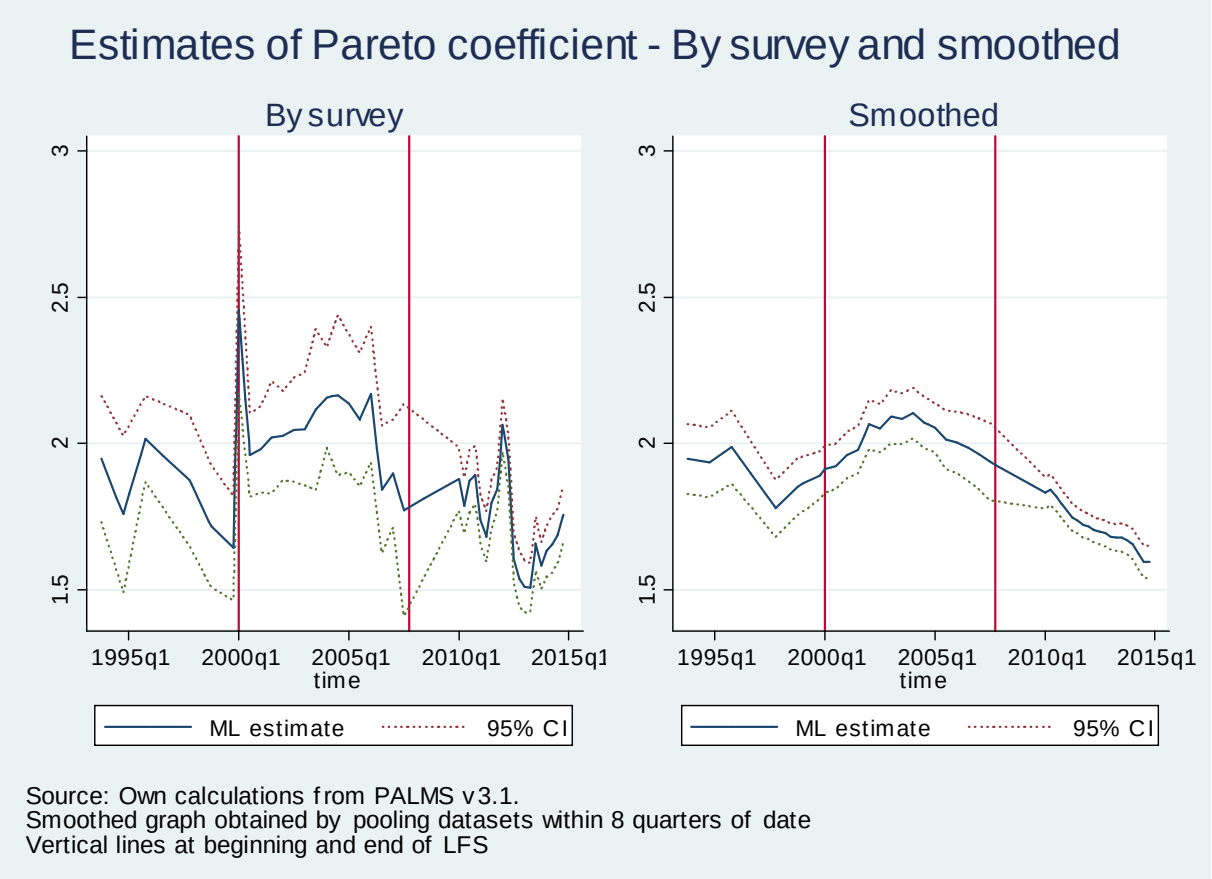

Figure 6: The impact of smoothing the Pareto estimates

forms of employment and, as Neyens and Wittenberg (2016) note, this is particularly pronounced in agriculture. Figure 6 suggests that this increase in the bulk of the lower tail is counterbalanced by a much thinner top tail. It should be noted that this is not a consequence of more low earning workers being recorded: the **conditional** distribution within the upper tail does not depend on what happens in the lower tail. One of the major ways in which the LFSs differ from both the OHSs and the QLFSs, is that they had the same question for self-employment earnings as for wages. It is plausible that this may have led to a poorer enumeration of top incomes.

More provocatively Figure 6 also suggests that the Pareto coefficient is coming down strongly in the recent past, suggesting that the top tail has "thickened". Given some of the measurement shifts that seem to have occurred in the more recent QLFSs (see for instance Kerr and Wittenberg 2017) one should be a bit cautious before jumping to this conclusion.

5.7 Cowell-Flachaire robust estimation of the mean

How big a difference does the semi-parametric technique of Cowell and Flachaire make for the estimation of the mean? A first look at the impact is given in Figure 7. It is evident that the semi-parametric technique does not adequately deal with all types of data contamination. The spikes in October 1999 and September 2000 that exercised Burger and Yu (2007) have not been removed. Part of the problem may very well be that there are too many contaminating observations. We therefore need to combine the technique with some prior outlier detection algorithm. The parametric part of the procedure is therefore geared more at ensuring that the weight of the tail is not underestimated, rather than in removing problematic observations. The impact of the Cowell-Flachaire robust estimation together with different outlier removal routines is shown in Figure 8. All of the outlier removal routines get rid of the spike in September 2000, but the October 1999 one is removed only with the regression technique. Given what we showed earlier this is not surprising: the problem of OHS 1999 is not one or two outliers, but the entire position of the top tail seems different than in other years. This means several things: first there are many more extreme values to begin with (raising the mean); secondly these don't look anomalous in relation to each other, hence they are less likely to be removed; and thirdly the Pareto coefficient will be particularly low ensuring that the parametric component of the Cowell-Flachaire technique will add weight to the top part of the distribution even if some of the "outliers" are removed.

This raises several questions: Where did all these high earners come from? Why are they not present in the other surveys? Which of the surveys captures the top tail better? There are several potential hypotheses for these different trends:

- The "long tail" in OHS 1999 (and some of the other surveys) is an artefact of survey specific processing errors, e.g. the decimal point in the earnings figures was not captured properly so that a big enough group of observations had their incomes transposed upwards to create a problem (this is raised by Burger and Yu 2007, p.6). Such a lapse in quality control would be troubling, but wouldn't raise deeper issues for the analysis of the data. The simple remedy would be to remove much of the top tail, as the regression outlier detection method ends up doing.

- The "long tail" in OHS 1999 is due to changes in the questionnaire. Wittenberg and Pirouz (2013) note that self-employment income in 1999 was captured only as "total income" whereas in 1997 and 1998 there were separate questions about "gross turnover" and "expenses". From 2000 only gross income was asked, whether or not the person was self-employed or working for someone else. Wittenberg and Pirouz (2013) did not detect a major difference in the distributions between 1997–1998 and 1999, but they did not focus on the top. If the shift is due to a measurement change, then the more aggressive regression-based outlier detection routine would again be appropriate.

- The OHS 1999 was the first survey to be run with the 1998 master sample. As Kerr and Wittenberg (2015) document, this led to a better enumeration of small households. Could this have led to a better enumeration also of the top tail? That is less certain. What does seem clear is that the loss of a separate "self-employment" earnings question in the LFSs would have off-set any gains from better enumeration. So OHS 1999 is different from the earlier surveys in terms of sampling, but also different from later ones in respect of the questionnaire. Consequently there is at least a small chance that OHS 1999 might have better captured the

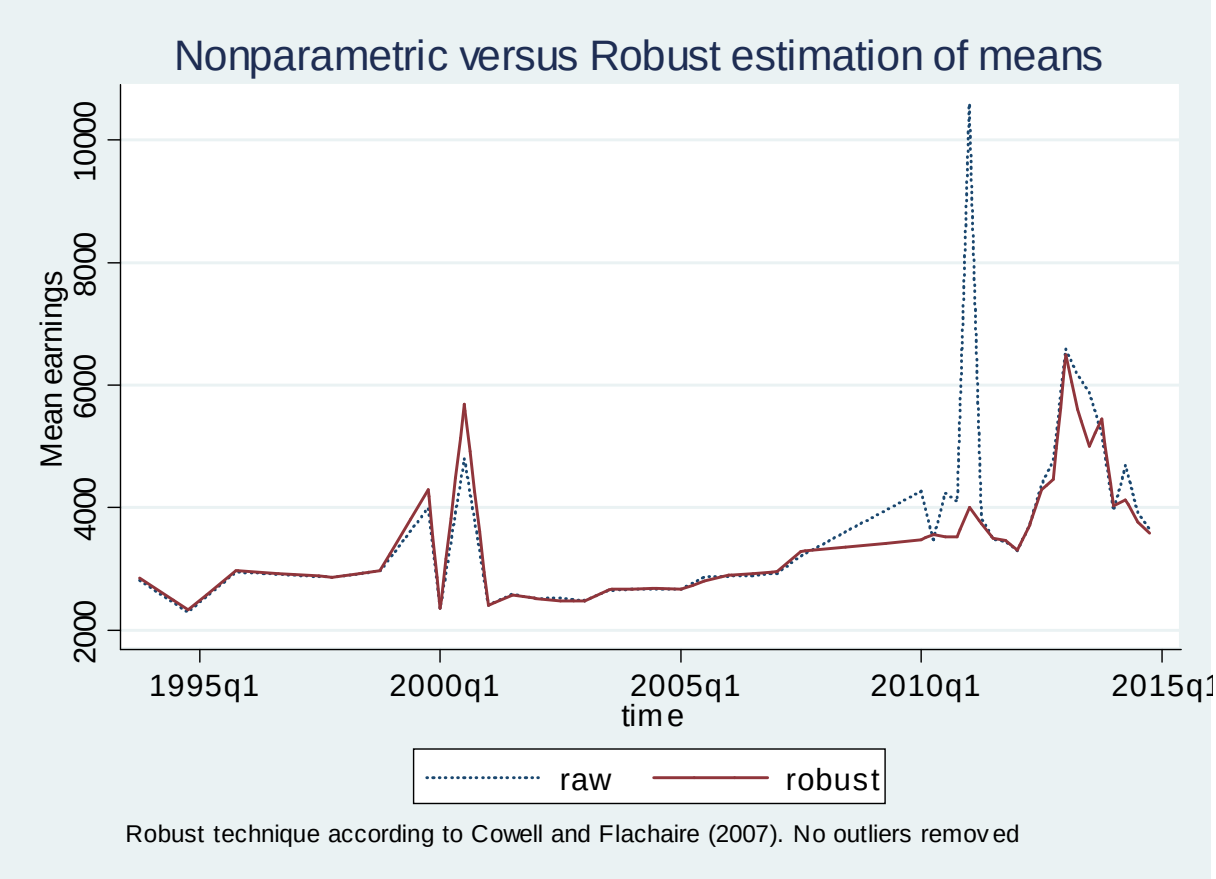

Figure 7: Average earnings estimated by the semi-parametric technique of Cowell and Flachaire (2007)

top tail than the other surveys around it. If this were true, it would be a mistake to remove the top tail.

The OHS 1999 is the most clear-cut case, but we saw in the right panel of Figure 4 that QLFS2012:3 and QLFS2014:3 also had much longer tails. In Figure 8 it is also evident that the spikes around 2014 and 2012 are not properly removed by the Pareto based outlier detection programme. Similar questions about whether these changes are due to processing (including imputations), sampling, field work or some other measurement changes arise.

5.8 Cowell-Flachaire robust estimation of the Gini

How does the robust technique fare in relation to the estimation of the Gini? Figure 9 shows that in virtually all cases the robust estimation technique increases the estimated coefficient slightly. In

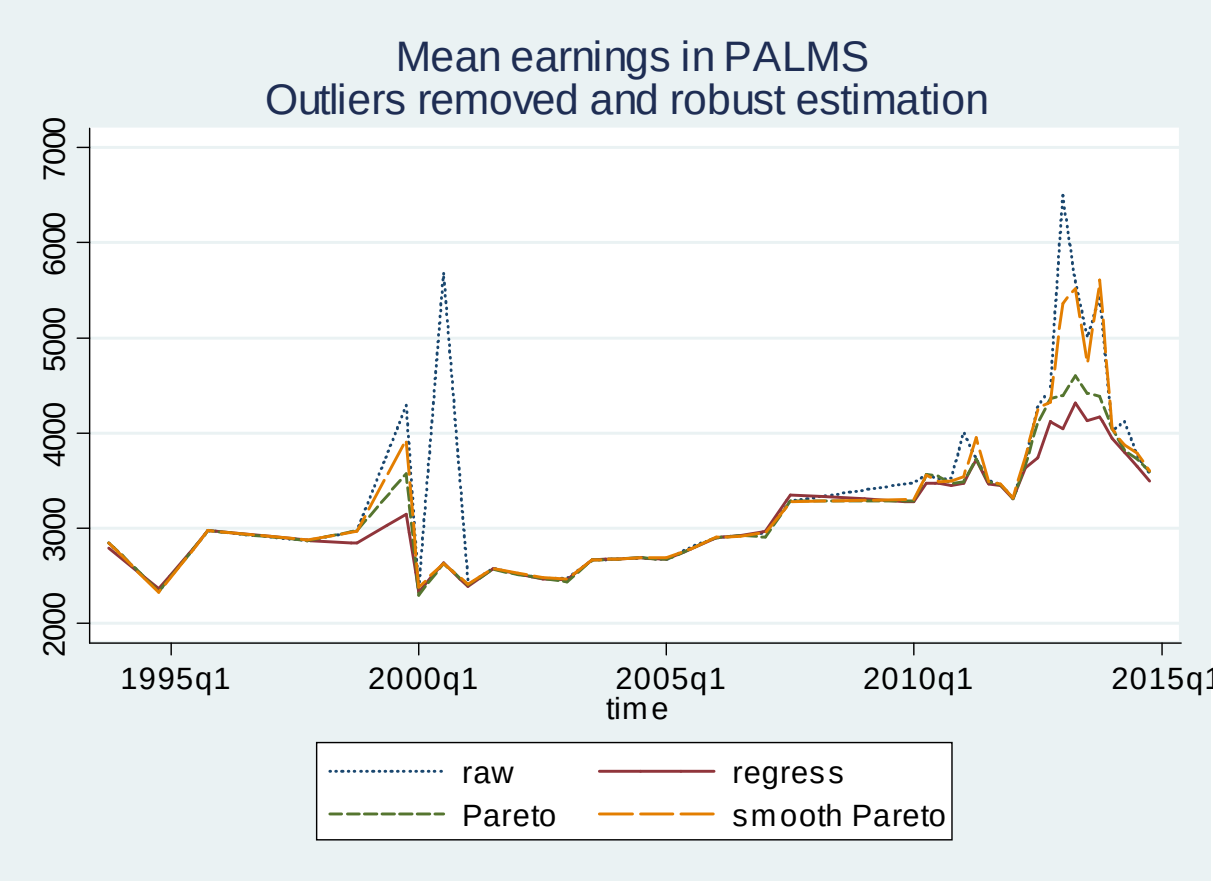

Figure 8: Combining outlier detection and the semi-parametric estimation technique of Cowell and Flachaire

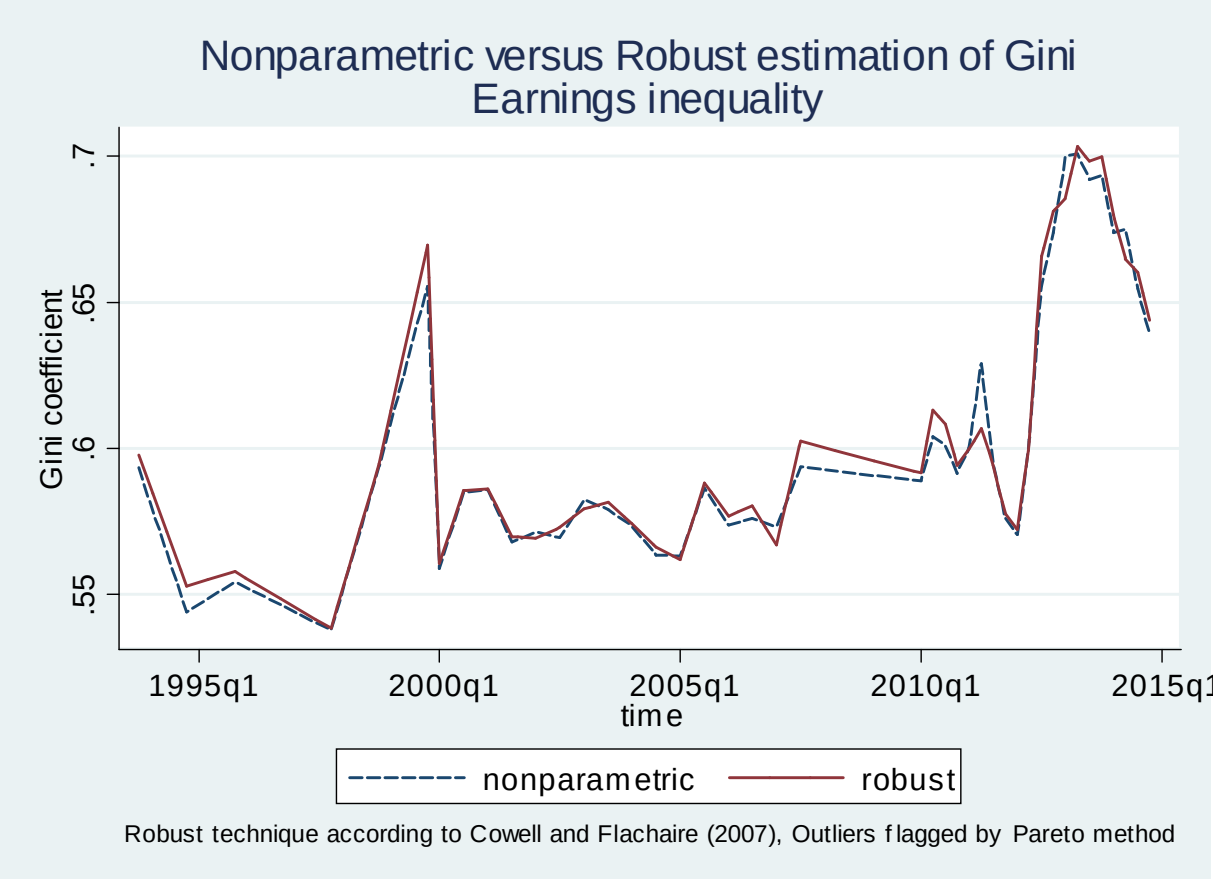

Figure 9: Semi-parametric estimation of Gini coefficient according to Cowell-Flachaire (2007) technique

a number of years the Gini is raised by a percentage point. This is noteworthy. But these are small effects when compared to the big differences that we see between different surveys. Again the huge increase in measured inequality in 1999 and the precipitous drop thereafter are striking.

The impact of the different outlier detection methods is shown in Figure 10 which mirrors the results in relation to mean estimation. It is clear that the spikes in the 1999 and 2014 Gini coefficients are not due to just one or two problematic data points. The more aggressive outlier removal routine embodied in the regression approach smooths over these bumps, but the underlying issue remains: we need to establish why the top end of the distribution seems so systematically different in these surveys.

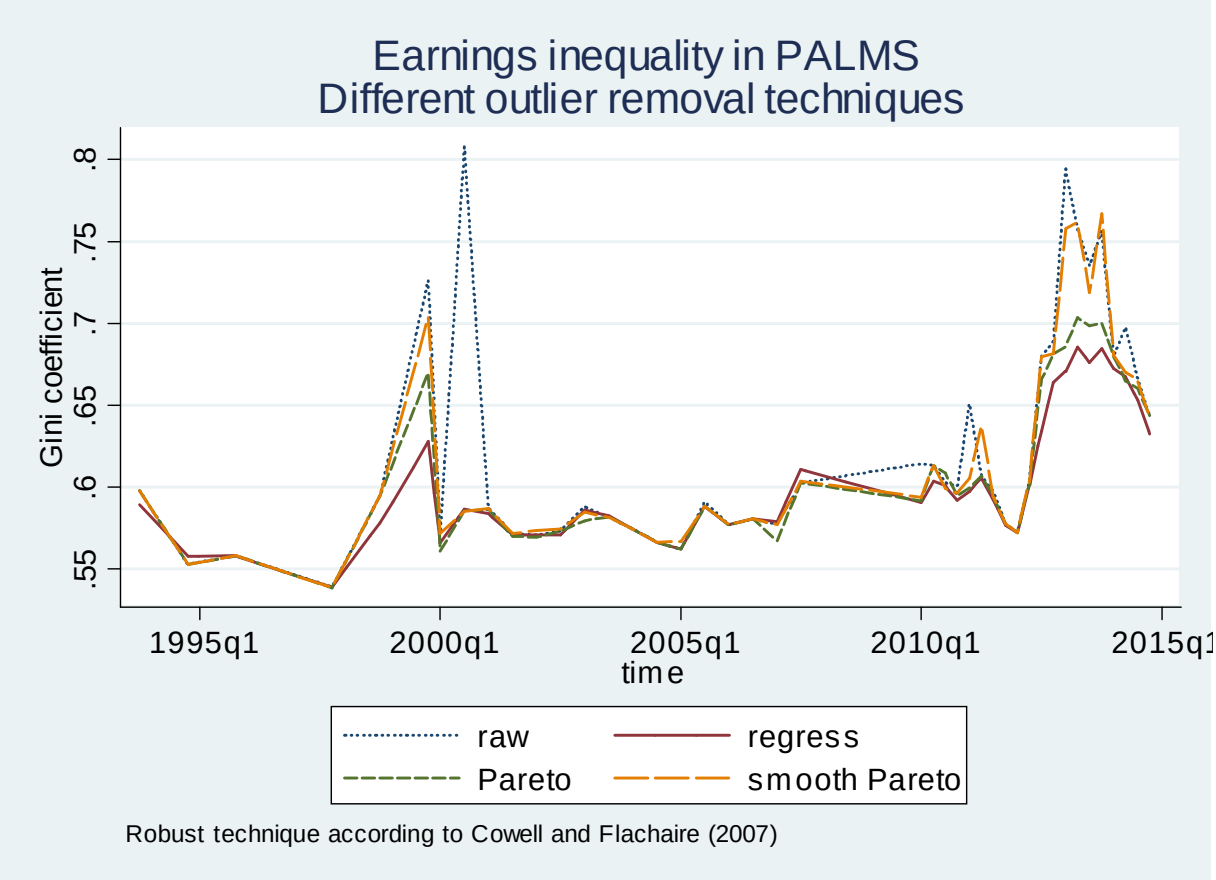

Figure 10: Semi-parametric estimation of the Gini coefficient combined with outlier removal

6 Discussion

Our discussion has cycled back to the key issue of measurement. It seems clear that in some of these surveys the "Data Generating Process" was somewhat different from those in other years. Previous discussions have already noted a number of discontinuities between October 1999 and February 2000. The location of the upper end of the earnings distribution can be added to those. Similarly there seem to be differences in some of the waves of the QLFS. Whereas in the OHSs and LFSs the bracket responses and missing data can be clearly identified, the QLFS data was released with full imputations. It is not clear how those imputations were done, but it seems likely that a "hot deck" (Andridge and Little 2010) was used This means that missing information is filled in from observations that are deemed to be "similar" in terms of the observable characteristics. This, however, creates the possibility that one or two problematic high income figures get multiplied, leading to a shift in the entire top tail.

Measurement and changes in measurement are first-order effects in trying to understand changes in the earnings distribution over time. Data contamination is clearly another issue. Both the regression and the Pareto routines pick up isolated cases of contamination. Arguably the Pareto approach is better if one wants to understand the nature of the top tail of the distribution. Nevertheless given the fact that the measurement shifts seem to involve entire clusters of problematic data, the more aggressive approach embedded in the regression approach seems to be more appropriate.

Despite these problems our estimates suggest that the tails of the income distribution are "heavy" and seem to have even become heavier over time. What does that mean? One way of thinking about this is in terms of the "mean excess function", defined as $E\left[X-u|X\geq u\right]$ (see Ghosh and Resnick 2010). The expression $X-u$ is the "excess" above the level u, given that X is bigger than u. In the context of income distributions we might think about u as being the richest income thus far observed, e.g. the income of Bill Gates. $E\left[X-u\right]$ is therefore the expected margin by which the next record breaker (i.e. the "next Bill Gates") will beat that income. In the case of a "thin-tailed" distribution like the Gaussian, the mean excess function converges to zero, as u increases. This means that the next record breaker will have an income that looks a lot like the current record holder, just a bit bigger. In the case of a "heavy-tailed" distribution, however, the mean excess function increases with u. That means that the current level of the record is no good guide to how large the next record could be. The log-normal distribution is "heavy-tailed" in that sense. So is the Pareto distribution[1]. Although the log-normal is "heavy-tailed" it is regular in other senses, i.e. it has finite moments. This means that the probability of encountering that next record breaker becomes vanishingly small. By contrast the Pareto distribution with $\alpha<2$ has no variance. That means that there are many more extreme values than there will ever be with a log-normal. Table 1 shows this based on data from the first quarter of 2011. The Pareto parameter for that period is 1.7 whereas the parameters of the log-normal (also estimated by pseudo-maximum likelihood) are $\mu=7.5$ and $\sigma=1.3$. The threshold x_i values given in Table 1 are (like all the figures in this paper) converted to June 2000 real values. The nominal figures would be around 80% higher. In the "Count" columns we calculate the **expected** number of individuals N_i in each earnings bracket according to the particular distribution (displayed to the nearest integer) while in the "Total income" columns we calculate the **expected** total income accruing to those individuals. The information in the first row is taken from the empirical 2011Q1 distribution, i.e. there were around 11.3 million people earning less than R 6000 (real) or around R10 000 (nominal) per month.

The total number of earners in the top tail (i.e. above R6000) was around 2.1 million. We see,

[1] In the case of the Pareto distribution the mean excess function is $\frac{1}{\alpha-1}u$ provided that the mean exists, i.e. $\alpha>1$.

Table 1: Expected distribution of top earners according to log-normal and Pareto distributions

	Log-normal, $\mu = 7.5$, $\sigma = 1.3$		Pareto, $\alpha = 1.7$		Pareto, $\alpha = 1.55$	
Threshold x_i	Count N_i	Total income T_i	Count N_i	Total income T_i	Count N_i	Total income T_i
0	11 300 000	20 200 000 000	11 300 000	20 200 000 000	11 300 000	20 200 000 000
6000	1 242 563	10 372 011 595	1 453 649	11 763 490 476	1 382 829	11 255 667 078
12000	582 132	9 563 365 538	447 413	7 241 277 790	472 250	7 687 847 172
24000	206 556	6 680 062 164	137 708	4 457 529 349	161 278	5 250 954 363
48000	55 499	3 534 621 800	42 384	2 743 931 177	55 078	3 586 507 523
96000	11 513	1 460 318 597	13 434	1 765 281 355	19 408	2 566 848 753
200000	1 543	397 338 130	3 746	1 010 464 554	6 030	1 636 017 440
400000	177	89 825 846	1 153	622 013 895	2 059	1 117 432 842
800000	15	15 371 546	355	382 894 466	703	763 229 123
1600000	1	1 990 650	109	235 699 191	240	521 300 854
3200000	0	195 034	34	145 089 871	82	356 058 976
6400000	0	14 452	10	89 313 292	28	243 195 448
12800000	0	845	5	143 014 583	15	524 031 337
Total	13 400 000	52 315 116 198	13 400 000	50 800 000 000	13 400 000	55 709 090 909

Notes:
Information in the first row based on empirical 2011Q1 distribution.
Other rows: $N_i = N \cdot \Pr(x_i \leq X < x_{i+1})$, $T_i = N_i \cdot E(X | x_i \leq X < x_{i+1})$

however, that where they are located in the distribution differs markedly between the log-normal and the Pareto distribution. In the case of the log-normal, we would not expect to see even one individual with monthly earnings above 3.2 million (real) or 5.8 million (norminal). The expected count is just 0.052. The "total income" is non-zero because it is the product of this very small expected count times a much bigger expected level of earnings (R 4 million real). Comparing these entries to the ones with the Pareto (1.7) distribution, we see that the tail extends much higher up. We would expect to see 34 earners between R3.2 million and R6.4 million, ten between R6.4 million and R12.8 million and five above R12.8 million (real), which would be R23 million per month (nominal). As a result we see that with the Pareto distribution much bigger shares accrue to the top of the distribution.

In Table 1 we have also shown what a "thickening" of the tails might mean. In the last two columns we show the calculations for a Pareto distribution with parameter $\alpha = 1.55$, which is where the smoothed series in Figure 6 ends. We see that the number of super-rich increases – to the extent where total earnings in the economy would increase by about R5 trillion a month – but all of it accruing to the top.

The difference between the two Pareto distributions show why the Cowell-Flachaire (2007) procedure is potentially important for our estimates of the mean and of inequality measures. Empirically we saw that it made some difference to the estimated Gini coefficient. The impact might be larger on a different measure. Nonetheless it is clear that these impacts will be small relative to the huge shifts which seem to be due to measurement issues. If we can clean up the data properly it may turn out that these techniques may become more relevant.

Of course the fact that our earnings distribution is "heavy tailed" is of interest in its own right. It suggests that the labour market processes are capable of generating considerable inequality.

Extreme earnings are more common than the naive reliance on log-normal models would suggest. How unusual is South Africa's top tail? Feenberg and Poterba (1993, Table A-1, p.173) provide estimates of the Pareto parameter for the US economy for 1951-1990. For most of this period the parameter hovered above 2, but in the mid-1980s it came down sharply to finish in 1990 around 1.6. Against that backdrop a value of 1.8 for South Africa is not that remarkable. The difference, of course, is that this stretched upper tail exists in a context where earnings at the bottom are much lower so that overall inequality ends up being stratospheric.

7 Conclusion

Our analysis has once again thrown the issue of the measurement into sharp relief. There are no technical fixes, not even sophisticated ones, that can undo bad data collection/data entry or processing. But new techniques may help us identify where the problems are and that in turn may help us separate out mere measurement shifts from real changes in the underlying social processes. That in turn will help us describe and analyse the real long-run changes in the South African economy. In this paper we have shown that the "break" between the OHSs and the LFSs is even deeper than previous analyses have suggested. It appears that while the LFSs were much better at picking up marginal forms of work, they were less successful at capturing what happened at the top of the distribution. Arguably this was due to the loss of separate questions about self-employment earnings.

Another contribution of this paper is in developing a new diagnostic tool for outliers which does not blindly remove most extreme values. Instead it tries to identify the sort of extreme values that belong in the top tail and those that don't.

Finally, despite all the "noise" exposed in this paper, there are actually some fairly clear conclusions: the top tail of the earnings distribution is "heavy tailed" with a Pareto coefficient of around 1.8. There is no evidence that this tail is likely to thin out any time soon, in fact the evidence, for what it's worth, is that it is thickening. More substantively, a Pareto coefficient of this magnitude suggests that the distribution has a mean, but no variance. In essence the probability of observing extreme values does not die out sufficiently rapidly for the variance to remain bounded. It is the statistical reflection of the casual observation that there are quite a lot of very rich South Africans. The existence of this tail may very well give rise to the perception of the "rich getting richer" and that the "new South Africa" has failed the bulk of its population.

References

Alvaredo, F. and Atkinson, A. B.: 2010, Colonial rule, apartheid and natural resources: Top incomes in South Africa, 1903-2007, *Discussion Paper 8155*, Centre for Econonomic Policy Research.

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

Bhorat, H., Van Der Westhuizen, C. and Jacobs, T.: 2009, Income and non-income inequality in post-apartheid South Africa: What are the drivers and possible policy interventions?, *Working Paper 09/138*, DPRU, University of Cape Town. available at http://ssrn.com/abstract=1474271.

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

Blandy, F.: 2009, The rich get richer, Agence France Press, available at http://business.iafrica.com/features/476302.html.

Branson, N. and Wittenberg, M.: 2014, Reweighting South African national household survey data to create a consistent series over time: A cross-entropy estimation approach, *South African Journal of Economics* **82**(1), 19–38.

Burger, R. and Yu, D.: 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.

Cowell, F. A. and Flachaire, E.: 2007, Income distribution and inequality measurement: The problem of extreme values, *Journal of Econometrics* **141**, 1044–1072.

Fedderke, J., Manga, J. and Pirouz, F.: 2004, Challenging Cassandra: Household and per capita household income distribution in the October Household Surveys 1995-1999, Income and Expenditure Surveys 1995 & 2000, and the Labour Force Survey 2000, *Working Paper 13*, Economic Research Southern Africa. available at http://www.econrsa.org/system/files/publications/working_papers/wp13.pdf.

Feenberg, D. R. and Poterba, J. M.: 1993, Income inequality and the incomes of very high-income taxpayers: Evidence from tax returns, *Tax Policy and the Economy* **7**, 145–177.

Ghosh, S. and Resnick, S.: 2010, A discussion on mean excess plots, *Stochastic processes and their applications* **120**, 1492–1517.

Heap, A.: 2009, *Earnings inequality in South Africa: Decomposing changes between 1995 and 2006*, Master's thesis, Economics Department, University of Cape Town.

Hill, B. M.: 1975, A simple general approach to inference about the tail of a distribution, *The Annals of Statistics* **3**(5), 1163–1174.

Hlekiso, T. and Mahlo, N.: 2006, Wage trends and inequality in South Africa: A comparative analysis, *Labour Market Frontiers* **8**, 9–16. Published by S.A. Reserve Bank, available at https://www.resbank.co.za/Lists/News and Publications/Attachments/345/October 2006.pdf.

Jones, C. I.: 2015a, Pareto and Piketty: The macroeconomics of top income and wealth inequality, *Journal of Economic Perspectives* **29**(1), 29–46.

Jones, C. I.: 2015b, Simple models of Pareto income and wealth inequality, online appendix to Journal of Economic Perspectives article. Available at http://dx.doi.org/10.1257/jep.29.1.29.

Kerr, A., Lam, D. and Wittenberg, M.: 2016, Post-Apartheid Labour Market Series [dataset], DataFirst, University of Cape Town. Version 3.1.

Kerr, A. and Wittenberg, M.: 2015, Sampling methodology and field work changes in the October Household Surveys and Labour Force Surveys, *Development Southern Africa* **32**(5), 603–612.

Kerr, A. and Wittenberg, M.: 2017, Public sector wages and employment in South Africa, *Working Paper 42*, REDI3x3. Available at http://www.redi3x3.org.

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.

Leibbrandt, M., Woolard, I., Finn, A. and Argent, J.: 2010, Trends in South African income distribution and poverty since the fall of Apartheid, *Social, Employment and Migration Working Papers 101*, OECD. http://dx.doi.org/10.1787/5kmms0t7p1ms-en.

Leite, P. G., McKinley, T. and Osorio, R. G.: 2006, The post-apartheid evolution of earnings inequality in South Africa, 1995-2004, *Working Paper 32*, International Poverty Centre, UNDP.

Mandelbrot, B.: 1960, The Pareto-Lévy law and the distribution of income, *International Economic Review* **1**(2), 79–106.

Neyens, E. and Wittenberg, M.: 2016, Changes in self-employment in the agricultural sector, South Africa: 1994-2012, Working Paper, DataFirst University of Cape Town.

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

Tregenna, F. and Tsela, M.: 2012, Inequality in South Africa: The distribution of income, expenditure and earnings, *Development Southern Africa* **29**(1), 35–61.

van der Berg, S.: 2011, Current poverty and income distribution in the context of South African history, *Economic History of Developing Regions* **26**(1), 120–140.

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

Wittenberg, M.: 2014b, Wages and wage inequality in South Africa 1994-2011: The evidence from household survey data, *Working Paper 135*, Southern Africa Labour and Development Research Unit.

Wittenberg, M.: 2017a, Measurement of earnings: Comparing South African tax and survey data, *Working Paper 41*, REDI3x3. Available at http://www.redi3x3.org.

Wittenberg, M.: 2017b, Wages and wage inquality in South Africa 1994-2011: Part 1 – wage measurement and trends, *South African Journal of Economics* **85**(2), 279–297. http://dx.doi.org/10.1111/saje.12148.

Wittenberg, M.: 2017c, Wages and wage inquality in South Africa 1994-2011: Part 2 – inequality measurement and trends, *South African Journal of Economics* **85**(2), 298–318. http://dx.doi.org/10.1111/saje.12147.

Wittenberg, M. and Pirouz, F.: 2013, The measurement of earnings in the post-apartheid period: An overview, *Technical Paper 23*, DataFirst. available at http://www.datafirst.uct.ac.za/images/docs/DataFirst-TP13_23.pdf.

Woolard, I. and Woolard, C.: 2006, Earnings inequality in South Africa 1995–2003, *Employment, Growth and Development Initiative, Occasional Paper 1*, HSRC, Cape Town.

SALDRU

Southern Africa Labour and
Development Research Unit

The Southern Africa Labour and Development Research Unit (SALDRU) conducts research directed at improving the well-being of South Africa's poor. It was established in 1975. Over the next two decades the unit's research played a central role in documenting the human costs of apartheid. Key projects from this period included the Farm Labour Conference (1976), the Economics of Health Care Conference (1978), and the Second Carnegie Enquiry into Poverty and Development in South Africa (1983-86). At the urging of the African National Congress, from 1992-1994 SALDRU and the World Bank coordinated the Project for Statistics on Living Standards and Development (PSLSD). This project provide baseline data for the implementation of post-apartheid socio-economic policies through South Africa's first non-racial national sample survey.

In the post-apartheid period, SALDRU has continued to gather data and conduct research directed at informing and assessing anti-poverty policy. In line with its historical contribution, SALDRU's researchers continue to conduct research detailing changing patterns of well-being in South Africa and assessing the impact of government policy on the poor. Current research work falls into the following research themes: post-apartheid poverty; employment and migration dynamics; family support structures in an era of rapid social change; public works and public infrastructure programmes, financial strategies of the poor; common property resources and the poor. Key survey projects include the Langeberg Integrated Family Survey (1999), the Khayelitsha/Mitchell's Plain Survey (2000), the ongoing Cape Area Panel Study (2001-) and the Financial Diaries Project.

www.saldru.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5696
Fax: +27 (0) 21 650 5797
Web: www.saldru.uct.ac.za

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad
SPES BONA