

Using Tax and Household Survey Data to Measure Earnings and Earnings Inequality

Andrew Kerr

School of Economics and DataFirst, University of Cape Town

Conference in Honour of Martin Wittenberg

October 2023

This paper has been funded by UNU-WIDER's SA-TIED programme.

I thank the UNU-WIDER research assistants who helped with remote work on the tax microdata in 2021.

Introduction

- In this research I use data from two sets of South African household surveys and administrative tax microdata to describe trends in earnings and earnings inequality.
- I show that the publicly available Quarterly Labour Force Survey (QLFS) earnings data is unreliable and produces highly implausible results.
 - By comparing to non-public QLFS earnings data
- The tax microdata that I use allow for a novel comparison with survey data that extends far down the earnings distribution.
- But I also highlight that there are several challenges that make comparisons between survey and tax data difficult.
- Some of these challenges are South Africa specific, but there are important lessons for researchers working in developing countries more broadly.

Martin's Influence: is all over this paper

- I'm trying to do careful descriptive measurement work using household survey data.
- I identify data quality issues in the surveys.
- I use PALMS code created by Martin to harmonise earnings data in QLFS and adapt this code for the General Household Survey.
- I extend the analysis inequality which was undertaken in Wittenberg (2017a, 2017b).
- I use analysis Martin undertook in his 2017 work comparing tax filing data and survey data (Wittenberg, 2017c).

Income and Earnings Inequality

- The importance of measurement challenges in describing earnings inequality trends is well known and regularly discussed in rich countries.
- Acemoglu and Autor (2011, HLE) note that changes in imputation and response rates in the US Current Population Survey matter for measuring earnings inequality trends.
- The rise in the use of administrative tax data has allowed researchers to compare income inequality in survey and admin data and to improve inequality measurement (Jenkins, 2017).

Earnings Inequality in South Africa

- Earnings inequality measurement has mostly been undertaken with household survey data in South Africa.
- Wittenberg (2017a, 2017b SAJE) described wage inequality trends from 1993 -2011, finding that
 - Wage inequality was extremely high.
 - Wage inequality was roughly stable since early 2000s using Gini or variance of log earnings, but this hid a more complicated pattern:
 - The bottom of the distribution has caught up to the middle, but the top moved away from the middle.

How close are the tax and survey data distributions?

- Jenkins (2017) shows that in the UK survey and tax data match closely up to the 99th percentile of the overall income distribution.
- Bassier and Woolard (2021) and Hundenborn et al. (2019) argue that this is true for South Africa at least up to the 90th percentile of the adult income distribution.
- I show that this is not true when using labour income, and that these conclusions are partly derived from incorrect use of the survey and tax data.
- This matters for important policy issues like where to set the minimum wage, in addition to inequality.

South African Household Survey Data sources

- Most work on SA earnings inequality uses survey data that is now available in the Post-Apartheid Labour Market Series (PALMS) (Kerr, Lam and Wittenberg, 2019)
 - Publicly available harmonised household surveys from 1993-2019
- The General Household Surveys (GHS) have not been used at all for earnings inequality measurement but are publicly available from 2002-2022.
- GHS and QLFS /PALMS have very similar questions about earnings and employment and in many years use overlapping sets of clusters to draw the sample.
- Both sets of surveys ask about market income from employment – ie before any deductions or transfers.

Tax microdata – IRP5 certificates

- All companies with employees are required to register with the South African Revenue Service (SARS) for PAYE tax.
- Each company must issue an IRP5 tax certificate to every employee that earnings more than R2000 per year, around \$100.
- So these are not just individuals filing returns or paying tax- it is (in theory) all employees in tax registered companies,
 - about 60% of all employment and 70% of all wage employment.

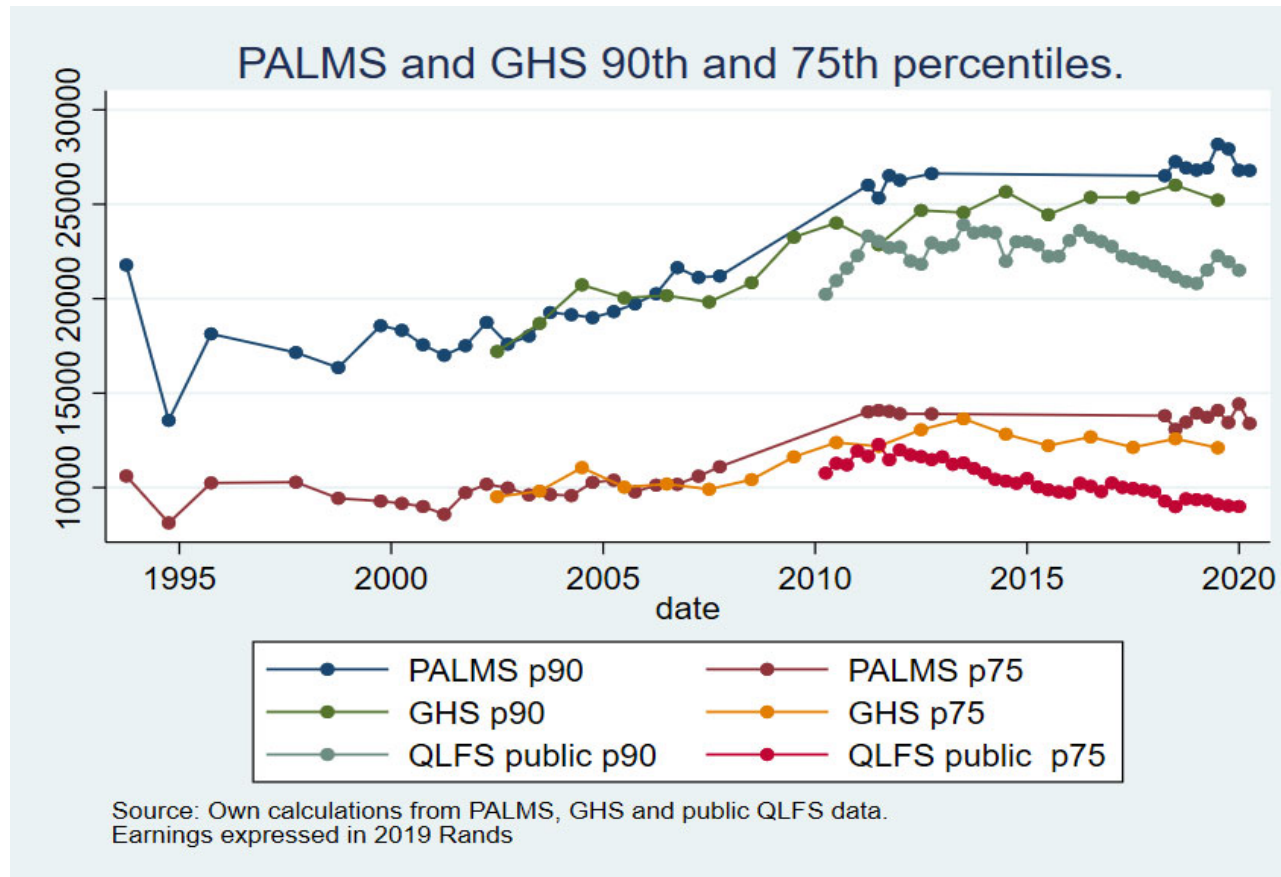
Unit and Item Non-Response and Imputation

- Household **unit** response rates were around 87-93% between 2002-2019 in both GHS and (Q)LFS.
- In the QLFS and GHS item non-response to earnings questions increased from around 7-9% in early 2000s to 29% in GHS 2019 and QLFS 2020 Q1 (almost unaffected by covid...).
- Because of the large item non-response rate imputation is required.
- Following Martin's work for PALMS, I undertake a form of hotdeck imputation in GHS and non-public QLFS
 - but use only one imputation rather than doing multiple imputation.

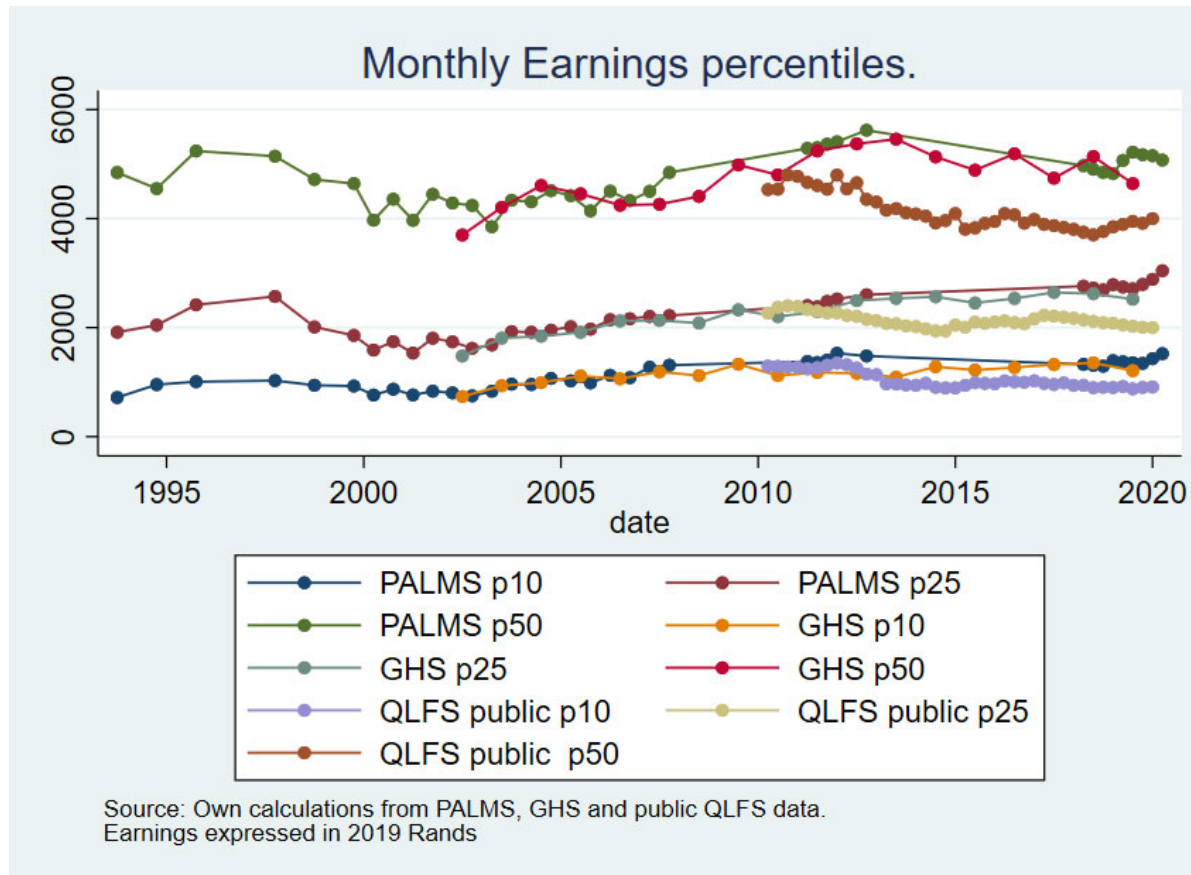
Notation

- In the figures that follow I'm calling the 3 sources of survey data GHS, PALMS and QLFS public.
- PALMS includes PSLSD, OHSs, LFSs (all public), and non-public QLFS for 2011-2012 and Q1 2018- Q1 2020.
 - So it does NOT include the QLFS earnings data currently in public PALMSv3.3
- What I call "QLFS public" has the public QLFS earnings data, which is imputed untransparently by Stats SA.
- GHS earnings is publicly available and whilst earnings is imputed for non-responders the unimputed data is there.
 - I use this and impute missing earnings in a very similar way to the way I impute QLFS non-public and LFS/OHS/PSLSD.
- All earnings are expressed in real values- December 2019

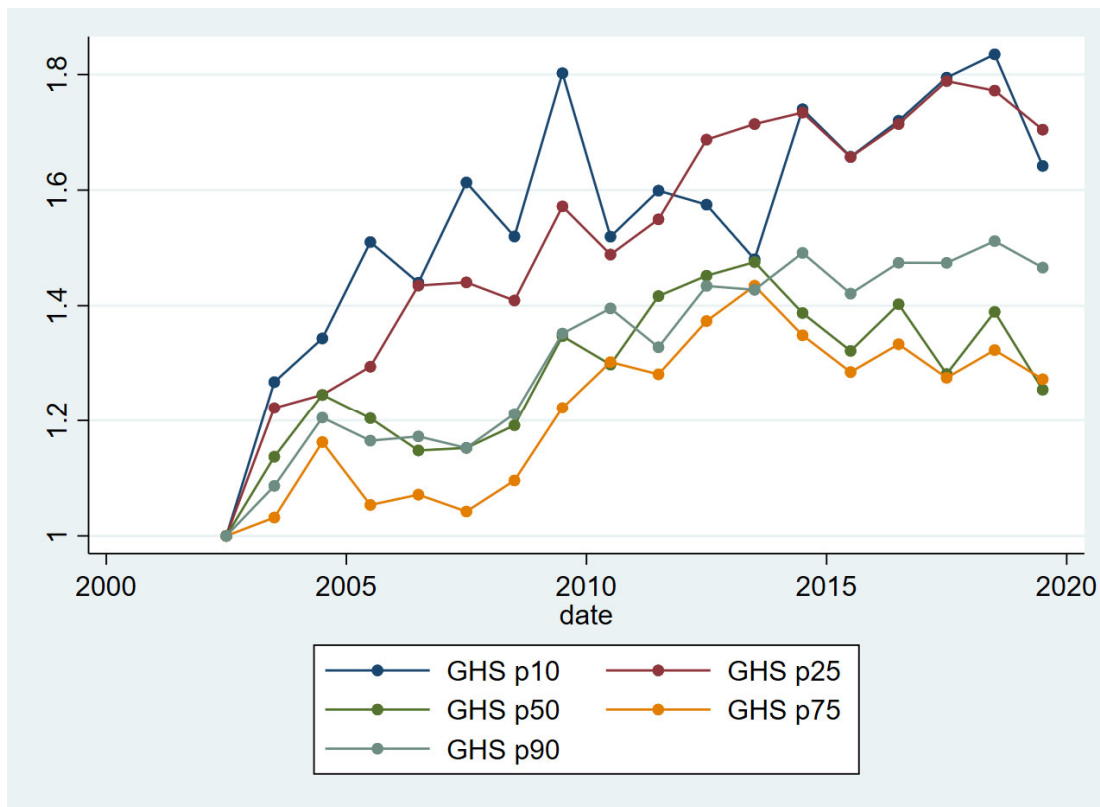
Earnings trends across the 3 data sources



Earnings trends across the 3 data sources



GHS relative changes



Public QLFS earnings data Problems- Q1 2020

Table 4: Q1 2020 Public and Non-Public Earnings Data Comparisons

Non-public Response Type	Correct Amt Imputed	Public data			Row Total
		Incorrect Amt Imputed	Refusal	Amt Imputed	
Bracket Only	10	61	28	0	100
Don't Know	0	0	36	64	100
Refusal	0	0	40	60	100

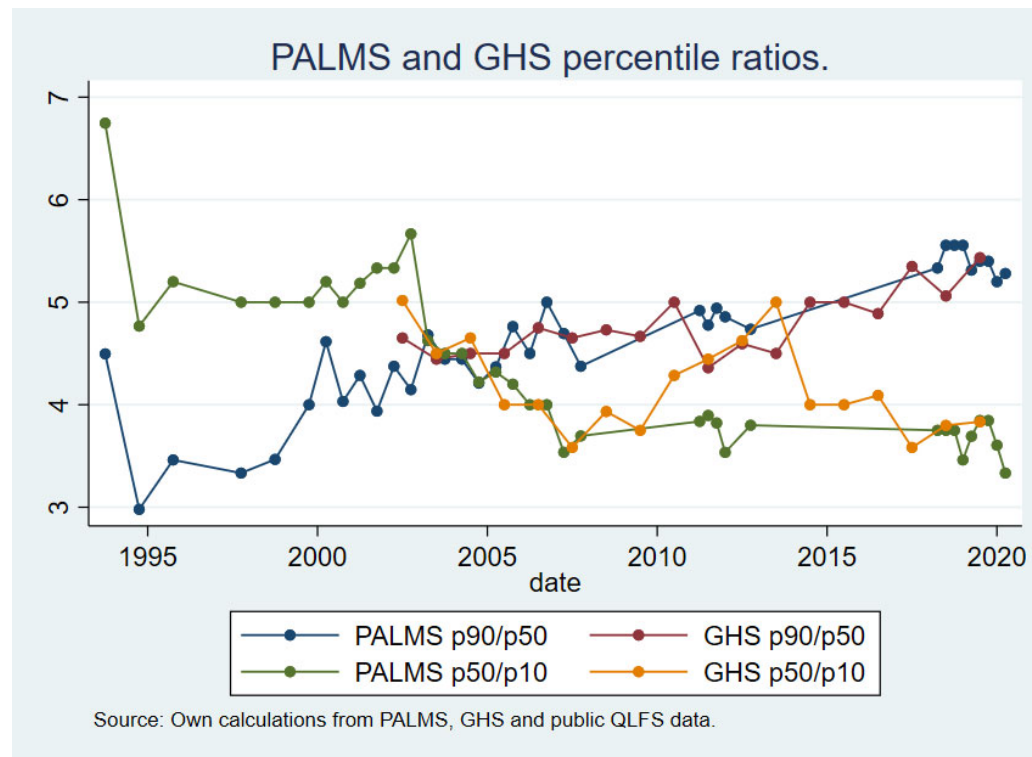
Note: Own calculations from public and non-public QLFS Q1 2020 earnings data.

Correct Amt Imputed means that the amount imputed was within the bracket the respondent gave.

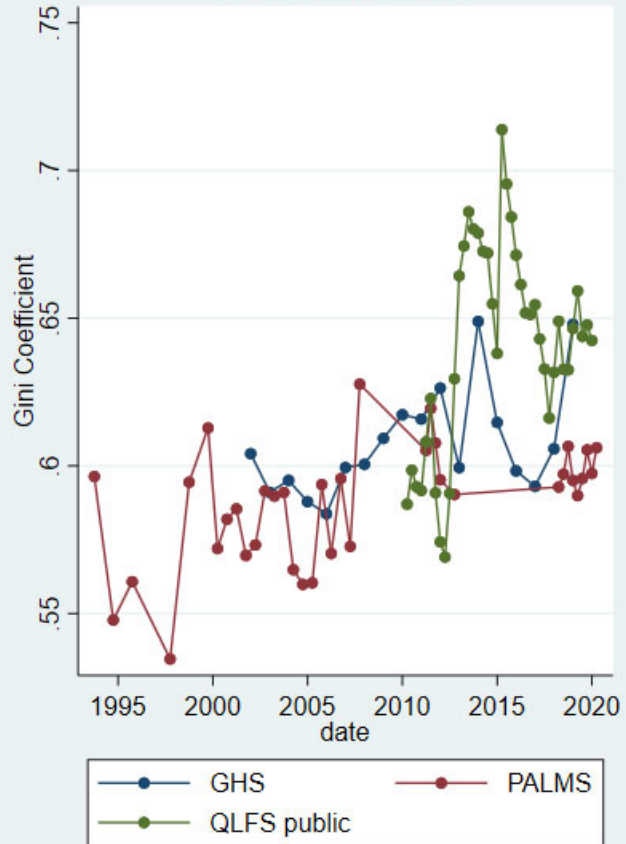
Incorrect Amt Imputed means that the amount imputed was not within the bracket the respondent gave.

Amt Imputed means that the amount imputed was from a refusal or don't know.

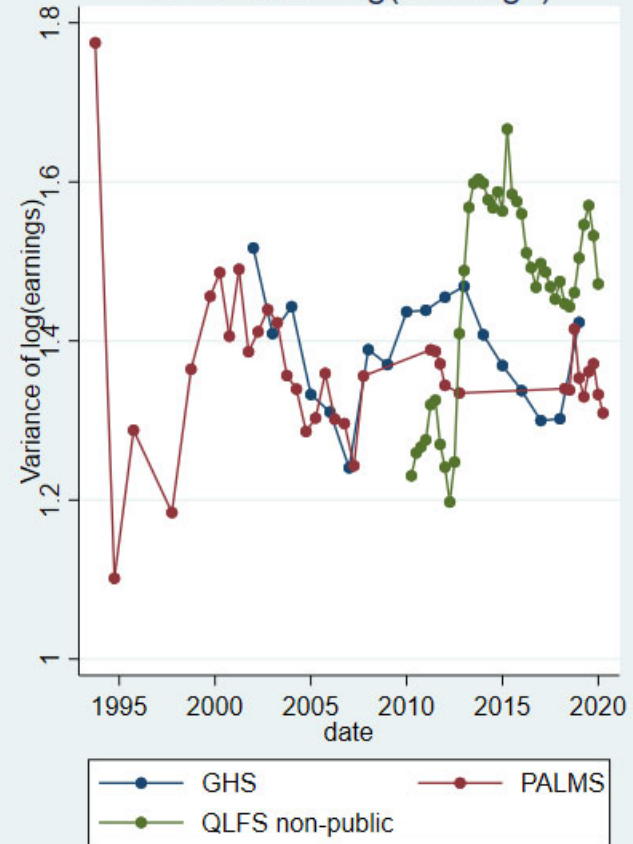
Trends in Earnings inequality



Gini coefficients



Variance of log(earnings)

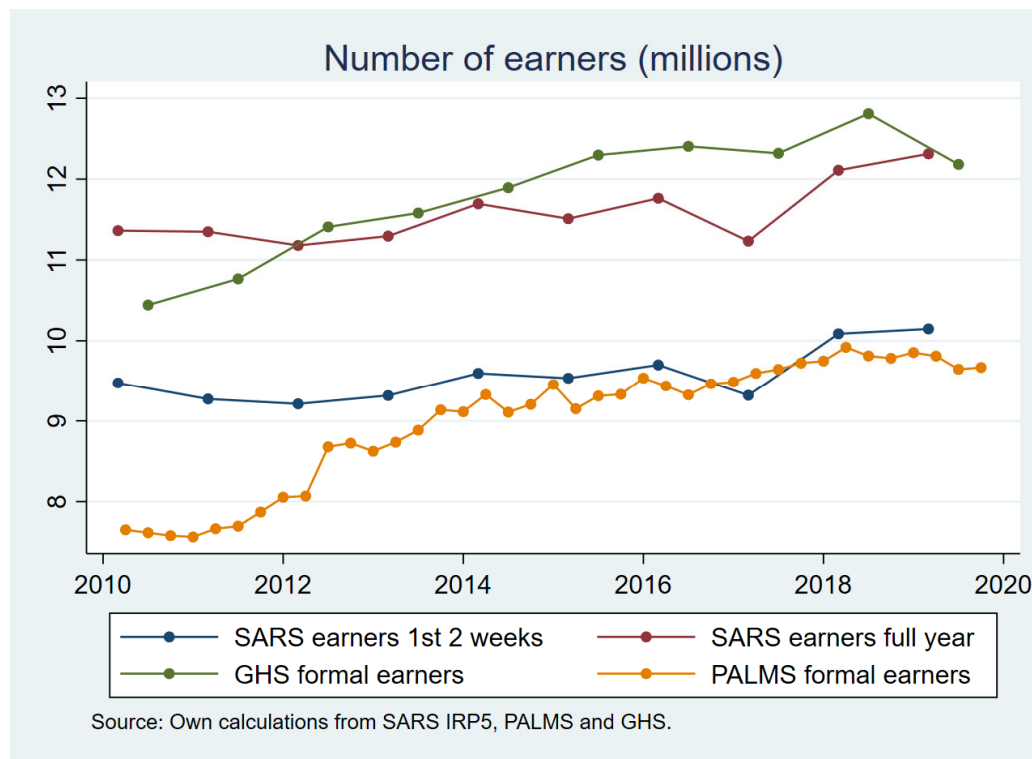


Comparing Survey and Tax Data

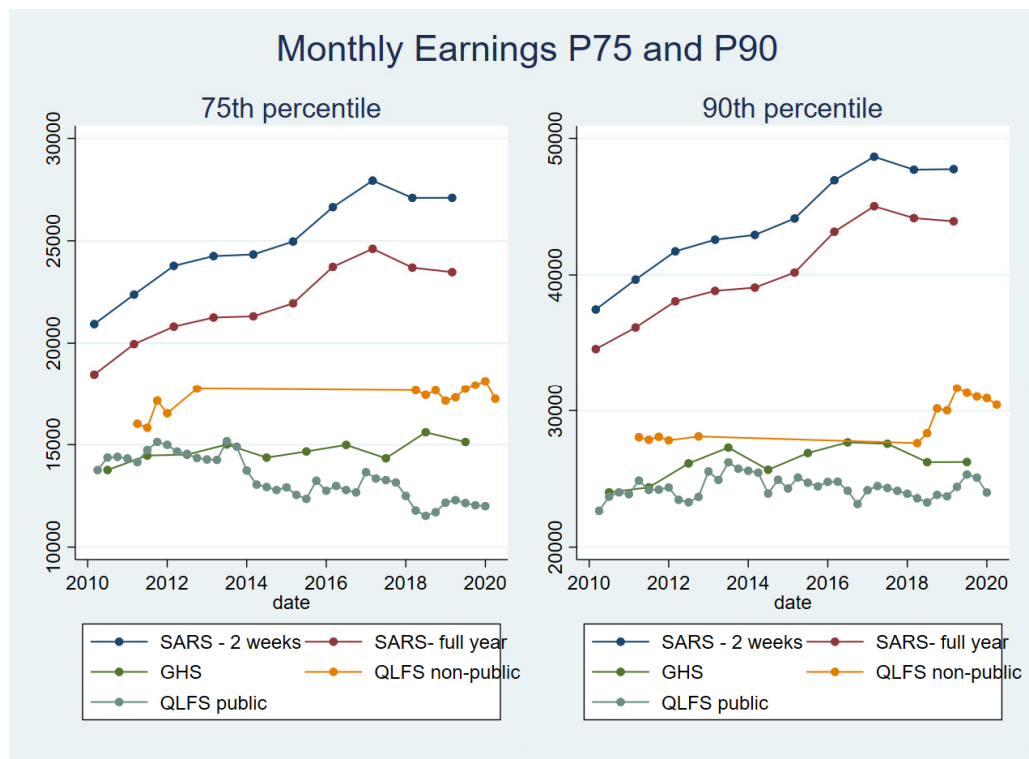
Comparisons between tax and household survey data

- We need the same hypothetical groups of employed in the QLFS, GHS and tax data to legitimately compare earnings across these data sources.
- We thus want to limit the surveys to **employees** in tax registered firms.
- Two issues:
 1. Tax data should be limited to those employed in a similar reference period as surveys- which ask about employment in the last week.
 - If we use all those appearing in the tax data at any point in the year, then we end up with a very different group of employees
 - They will include those employed for short periods who are more likely to be lower earners.
 2. The GHS questions about employers and contracts are limited, and result in a group that is too large.
 - QLFS has enough questions to get very close to tax data numbers.

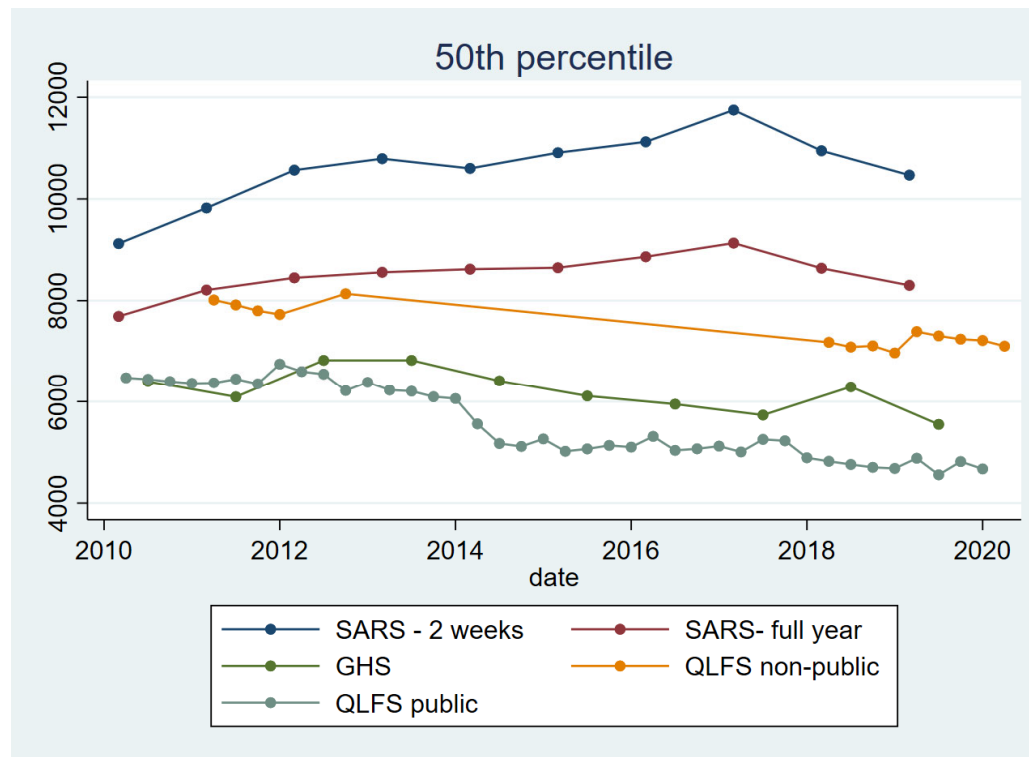
Number of “Formal” Employees in each data source



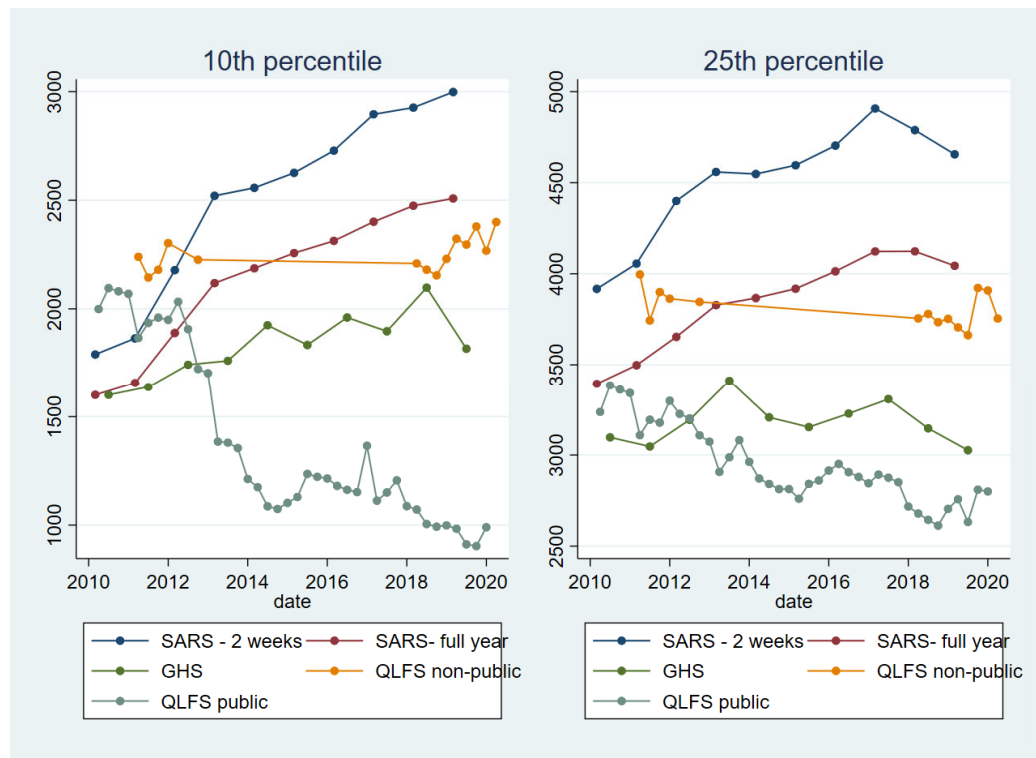
Formal sector distribution comparisons- p75 and p90 and p90



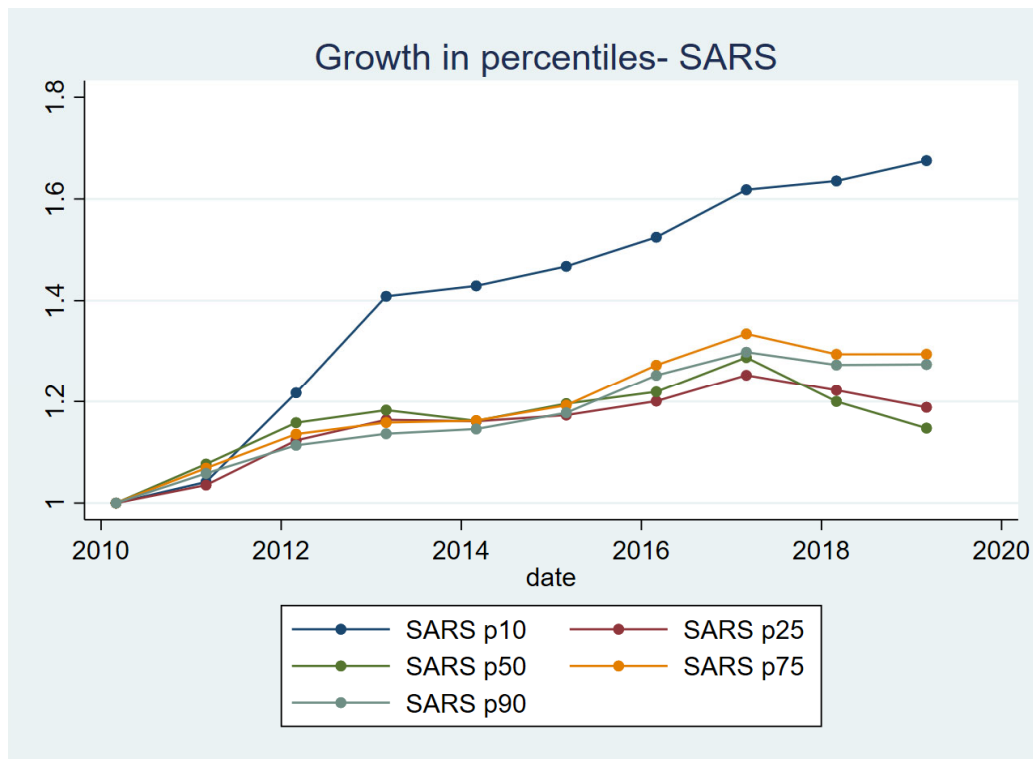
Formal sector distribution comparisons- p50



Formal sector distribution comparisons- p10 and p25

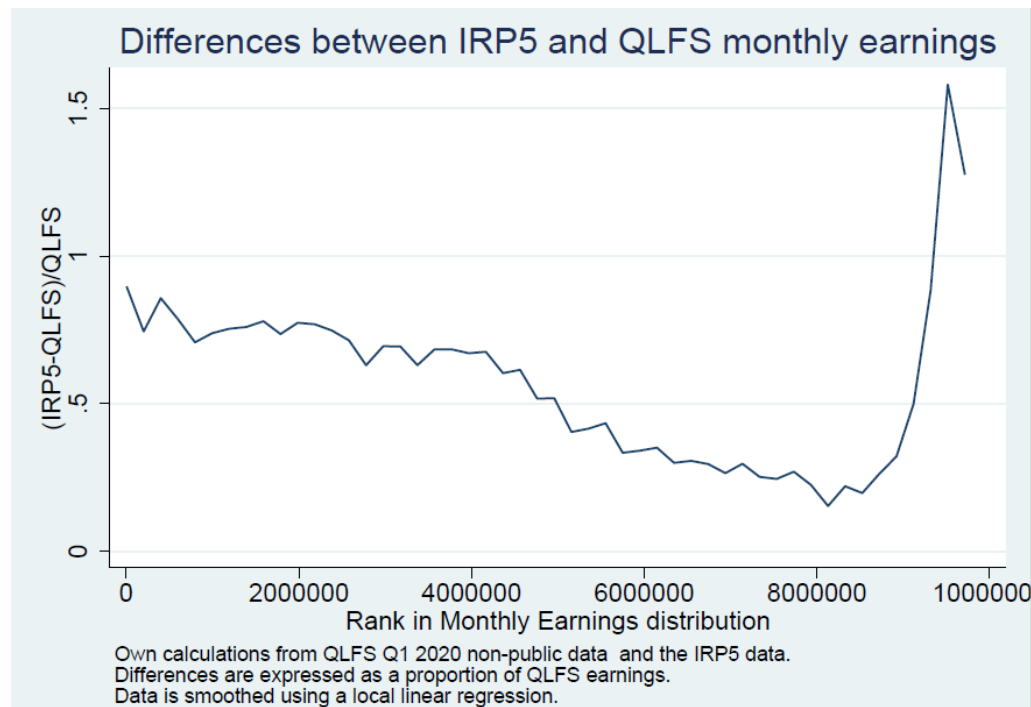


Growth in percentiles- SARS data



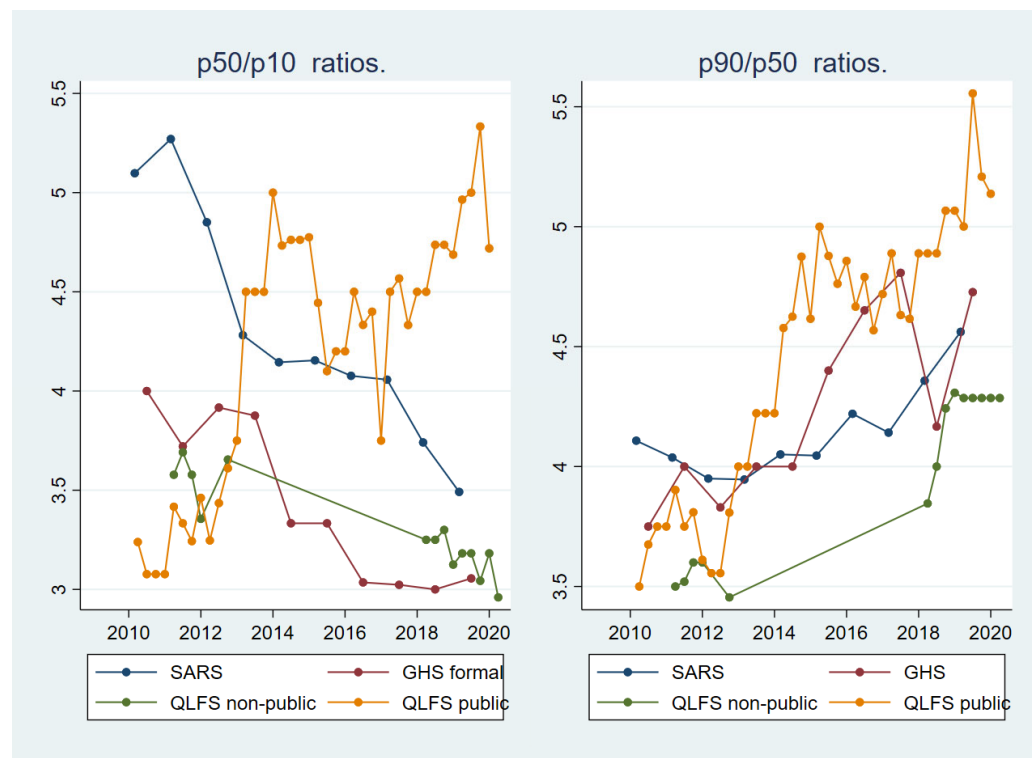
GDP per capita was constant over this period – around R75000 pa

QLFS– tax data comparison across the distribution.

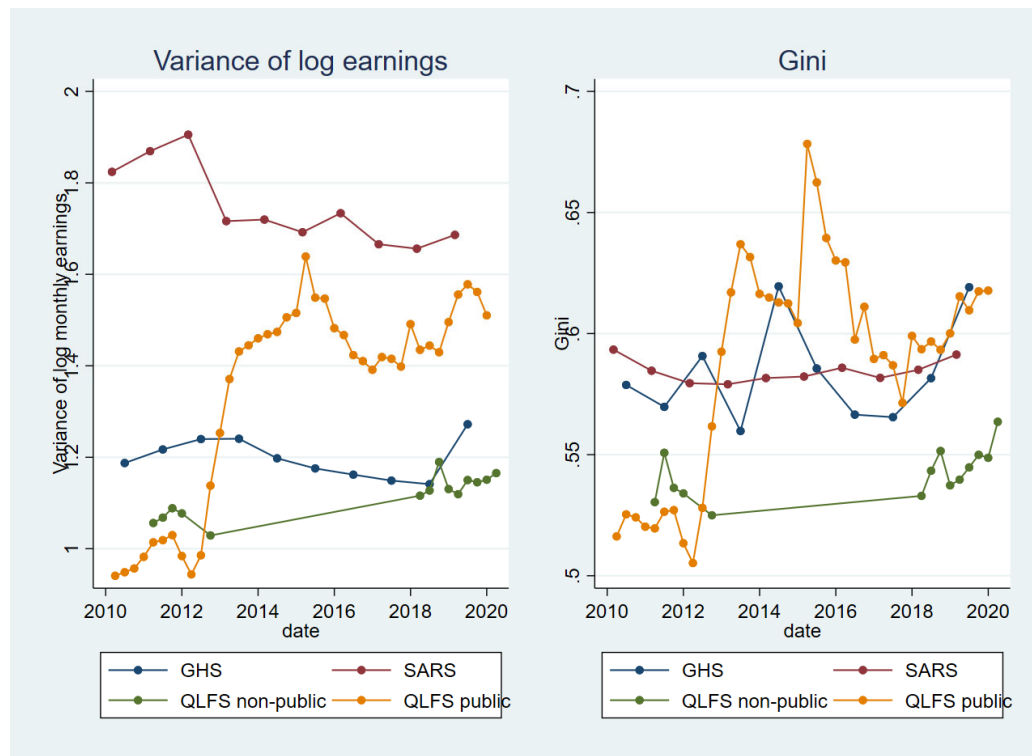


This method was undertaken by Wittenberg (2017c)- but for tax filers only in 2011.

Percentile ratios for formal employees



Gini and Variance of log earnings for formal employees



Conclusions

- The public QLFS earnings data is very poor quality.
 - We urgently need a public release of QLFS unimputed earnings data.
- Putting a harmonised set of GHSs into the public domain would be useful.
- The South African tax microdata allow an unusual comparison of tax and survey data across almost all the formal sector, which is a large % of overall employment in SA.
- Earnings is under-reported in household surveys far down the earnings distribution.
- Inequality in earnings is mostly higher in tax data but the trends are broadly similar (ignoring the public QLFS)-
 - Gini is stable or increased a little
 - Variance of log earnings is stable or decreased a little
 - P10 moved closer to P50, and p90 moved away from p50- a continuation of the longer run trend identified by Wittenberg (2017a, 2017b).

Wittenberg References

- Wittenberg, M. (2017a). 'Wages and Wage Inequality in South Africa 1994–2011: Part 1–Wage Measurement and Trends'. *South African Journal of Economics*, 85(2): 279–97.
- <https://doi.org/10.1111/saje.12148>
- Wittenberg, M. (2017b). 'Wages and Wage Inequality in South Africa 1994–2011: Part 2–Inequality Measurement and Trends'. *South African Journal of Economics*, 85(2): 298–318.
- <https://doi.org/10.1111/saje.12147>
- Wittenberg (2017c). 'Measurement of Earnings: Comparing South African Tax and Survey Data'. REDI3x3 Working Paper 41. Cape Town: SALDRU, University of Cape Town.