DataFirst

DataFirst Technical Papers

An analysis of the data quality of the Surveys of Employers and Self-Employed

by

Andrew Kerr

Technical Paper Series
Number 31

About the Author(s) and Acknowledgments

Andrew Kerr - Senior Researcher, DataFirst, University of CapeTown

This work was funded by REDI through a grant from the South African National Treasury. I thank Martin Wittenberg, Frederick Fourie, Ndivhuwo Gangazhe and Francois Roubaud for helpful comments.

Recommended citation

Kerr, A. (2015). An analysis of the data quality of the Surveys of Employers and Self-Employed. A DataFirst Technical Paper 31. Cape Town: DataFirst, University of Cape Town

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

DataFirst

An analysis of the data quality of the Surveys of Employers and Self-Employed

Andrew Kerr

DataFirst Technical Paper 31
University of Cape Town
June 2015[1]

In this note I highlight and explore the large differences between the different numbers of non-VAT registered employers estimated from the Survey of Employers and Self-Employed (SESE) survey, undertaken by Statistics South Africa in 2001, 2005, 2009 and 2013. I provide a number of data quality checks on the surveys. I conclude that whilst 2005-2013 are more comparable to each other the 2001 survey actually looks closer to the truth than the others, despite being a outlier compared to the later three SESEs.

The four SESEs were each linked to the LFS or QLFS undertaken at the same time. The LFS or QLFS enumerated 30 000 households, and any individual who reported themselves to be self-employed and operating a non-VAT registered business should have been interviewed again and the SESE survey administered to this owner. The SESE was a very detailed survey about the business which the individual reported he or she owned. The SESE is used by Statistics South Africa to obtain estimates of the size of value added in the informal economy and GDP figures are adjusted based on these estimates.

Thus every four years since 2001 a detailed picture of non-VAT registered businesses has been obtained from the SESE, whilst in the intervening periods only basic data is collected about these businesses in the (Q)LFS. However we can use the OHS, LFS and QLFS to construct a time series of the estimates of the total number of individuals reporting that they own a non-VAT registered business between 1995 and the present. This is shown in the solid line figure 1. The 4 dots represent the estimates from SESE of the total number of owners of non-VAT registered businesses. Second and third businesses have been excluded from this calculation so it should be comparable to the (Q)LFS number, because the SESE sample is all those who are identified as running non-VAT registered businesses in the (Q)LFS. The numbers from SESE and the (Q)LFS are quite different in both 2001 and 2009 and this is discussed further below.

[1] This work was funded by REDI through a grant from the South African National Treasury. I thank Martin Wittenberg, Frederick Fourie, Ndivhuwo Gangazhe and Francois Roubaud for helpful comments.

There are reasons to doubt the validity of the trends from SESE and the (Q)LFS of a huge decline in informal business ownership until 2009 and then a jump up in 2013. For the first SESE enumerators were paid for each SESE interview they undertook (personal communication with Statistics South Africa, 2012). This created incentives for the enumerators to find as many self-employed as possible. In the subsequent three SESEs the incentives worked the other way- enumerators were not paid for the extra work required to interview non-VAT registered business owners. Thus enumerators may have had incentives to underreport the number of non-VAT registered business owners.

The spike in 2001 has been noted by Devey et al (2006), although the trend was not as dramatic in the figure they showed because the authors were using the LFS and included individuals reporting themselves to be employees in non-VAT registered business when enumeration of these employees was not affected by the SESE-related extra payments.

An initial expectation may thus be that while 2001 was an overestimate of the true number of non-VAT registered business owners 2005-2013 may be an underestimate. Alternatively one might think that some respondents may prefer not to reveal self-employment or not consider the marginal work they do as a business but that in 2001 enumerators pushed these type of respondents harder than they did in later years to reveal their businesses and thus 2001 may be closer to the true value in that year than was the case in later years. The point of this data quality note is to explore whether this conclusion is correct and to help those wishing to use the data to do the best they can with what are obviously very different estimates across the four surveys.

Comparing reported employment in SESE with the reports of employees in the LFS

I undertake a number of data quality checks on the data in this research note.

As a first check I explore the differences between SESE 2001 and the 2001 LFS estimates of the number of self-employed individuals with non-VAT registered businesses as well as differences between the 2009 SESE and the 2009 QLFS. In 2001 the SESE number is around 10% higher than the LFS number whilst in 2009 it is around 10% lower than the QLFS number. The reason for these differences seems to be related to capturing of second jobs that are self-employment in non-VAT registered businesses in 2001 but not in other years. This again suggests that the incentives to SESE enumerators had unintended consequences. Around 10% of those identified as having a non-VAT registered business were actually listed as working as a wage worker in their main job in 2001. This is not true for the later SESEs.

The reason for the 2009 SESE estimate of the total number of owners of non-VAT registered businesses being lower than that estimated from QLFS seems to be an incorrect weighting up for non-response in the SESE. I find that around 18% of those identified as running non-VAT registered businesses in the QLFS were not interviewed for SESE. Stats SA reports a non-response rate of 6.1% in 2009. It is not clear what explains this discrepancy. The usual method to correct for this type of unit non-response is to increase the weights of those who were interviewed for SESE. However it seems that the weights of these individuals were only increased by about 5%, leading to a smaller

estimate of non-VAT registered business owners in SESE than in the 2009 QLFS undertaken at the same time.

A second important data quality check is to compare the estimates of the number of employees reported by owners of unregistered businesses in SESE with the number of individuals reporting themselves to be employees in unregistered businesses in the LFS conducted at the same time. These are two different estimates of the same total and therefore they should be roughly similar. Unfortunately the question to employees about whether they were working in a non-VAT registered business was dropped in the QLFS so we cannot compare the totals from SESE 2009 and 2013. Where we can these are shown in figure 2. The estimates from SESE are much lower than those from the LFS- only roughly a half of the total employees who say they work in a non-VAT registered business are captured in SESE in 2005. The fraction is higher in 2001. This is one piece of evidence that suggests that whilst the 2001 SESE captured many more businesses than later SESEs it may actually have been closer to the true value than later surveys.

It is possible that those who work as employees do not know whether the business they work for is VAT registered and that the estimate from the questions to employees may be subject to measurement error. A priori, however, it is not clear that the error should lead to an overestimate of the number of employees who do work in a non-VAT registered business. This is therefore unlikely to explain the large discrepancy between the estimates from SESE and the LFS.

We can investigate this discrepancy further by comparing the actual distribution of non-VAT registered firm sizes reported in SESE 2001 and 2005 with the implied non-VAT registered firm sizes as reported by employees- who were asked what size firm they work for. These are shown in figure 3 below for 2001[2]. SESE seems to have under captured larger informal firms if the reports of employees in these larger, unregistered, firms are to be believed. If a similar pattern held in subsequent SESEs then SESE 2009 in particular looks to have massively underestimated employment in non-VAT registered firms, given the much smaller total number of firms and employees working in those firms.

However, it is not possible to check whether firms are reporting fewer employees than they actually have or whether large firms are just not being enumerated. Sampling theory provides one explanation for why it could be the latter. With a skewed firm size distribution large firms will not be captured in most samples without stratification and thus most samples will miss larger firms. Stratification by firm size is one way other firm surveys get around this issue (Statistics South Africa's firm surveys are stratified on size). However because of the lack of a list of informal firms there is no way to stratify except on the size reported by owners in the (Q)LFS. Owners of large firms may also be more likely to refuse to respond to SESE enumerators because they fear that the information obtained may be used to make them pay taxes.

The checks undertaken thus far have suggested that the SESEs underestimate employees in informal firms but we cannot know if this is due to firms reporting fewer employees than they have or because large firms are simply not being captured at all. The checks also do not speak to whether

[2] To calculate a firm size distribution from firm size reports from *employees* who report in intervals I divided the person weights by the average of the firm size reported in each category. The open category is 50+ employees and I divided the person weights of employees reporting in this category by 100.

the number of single person firms is an underestimate because, by definition, these firms do not have employees reporting that they work in these firms.

However, there are substantive differences in the estimated number of businesses with and without employees across the four SESEs, as shown in Table 1. It is clear that the number of no employees firms is more volatile across the surveys, although there was a very large jump in the number of employing firms between 2009 and 2013, such that this number is higher even than the 2001 estimate, despite the number of no employees firms being substantially lower than the 2001 estimate. The ratio of employers to employees also varies over the surveys- from roughly 3.1 in 2001 to 1.5 in 2013.

Marginality Checks

The above analysis suggested all SESEs under-captured employment and likely the actual numbers of firms. A possible explanation for the larger number of firms estimated from SESE 2001 compared to the other years is that enumerators found more marginal firms in this year. An analysis of the fraction of firms with and without employees is a check on whether the SESE 2001 total was much higher due to more marginal firms being enumerated. In SESE 2001 85% of the firms had no employees, whilst this fraction was similar in both 2005 and 2009- 83% and 81%. 2013 looks a bit more different- only 76% of the firms had no employees.

There are other ways to explore whether marginal businesses were better captured in 2001. More marginal businesses are likely to occupy the owners for a smaller number of hours each week. We can use the hours worked in the last week reported by the owner in 2001, 2009 and 2013 in the LFS or QLFS as a check on whether 2001 businesses were more marginal. The median number of hours worked was 48 in 2001, 46 in 2009 and 45 in 2013, whilst the 10^{th} percentile was 11 hours in 2001, 10 in 2009 and 10 in 2013. This suggests 2001 businesses were not more marginal based on the number of hours the owners reported working in the previous week. Figure 4 shows kernel densities for the number of hours worked in 2001, 2009 and 2013, further suggesting that businesses enumerated in 2001 were not more marginal than in later surveys.

Other checks on the marginality of 2001 businesses are also possible. Figures 5 and 6 compare reported sales and profits across the four surveys for those that reported positive sales, deflated to 2011 rands. Although real sales were lower in 2001 than in other years there is little evidence from the figures that the left hand tail is much fatter than in the later surveys. However, the 2013 density suggests that 2013 businesses were much larger than in the other three surveys, which makes sense given the larger fraction of 2013 firms with employees, noted above. The number of owners reporting zero profits or sales are shown in rows 1 and 3 of Table 2. Clearly 2001 has many fewer firms reporting zero profits or sales compared to later surveys.

Table 3 reports a number of other firm characteristics associated with marginality- the percentage of firms undertaking investment, the percentage of firms with no accounts, the percentage of businesses operating in the owner's home and the percentage of firms with paid employees. 2001 businesses look more marginal on all of these characteristics, particularly for investment and operating inside the home. Thus there is some indication that 2001 businesses were more marginal, but when measured by profits, sales or hours worked the 2001 firms do not look any more marginal.

There is therefore only mixed evidence that the 2001 total is much higher because enumerators in 2001 pushed owners of marginal businesses to report these as work more than enumerators in later surveys.

Other Data Quality Checks

It is also possible to check some aspects specifically related to data quality. In theory every owner of a non-VAT registered business enumerated for the LFS should be interviewed for SESE. However not all were, due to refusals or possibly mistakes in enumeration. The seventh row of Table 2 shows non-response rates for SESE, conditional on an individual reporting owning a non-VAT registered business in the (Q)LFS[3]. This is the fraction of self-employed who report non-VAT registered businesses in the (Q)LFS but are then not re-interviewed for the SESE. These response rates are not published in the Stats SA SESE reports and they suggest that the 2001 response rates were substantially higher than in 2009 or 2013. It is not possible to link SESE to the LFS in 2005 so we cannot know the response rate in this year. The Stats SA SESE release for 2005 only reports the LFS response rate.

If one believed that in 2001 enumerators responded to the incentives to increase their own wages at the expense of data quality this might show up in firm accounting data that was poorly measured relative to later years. A first pass at this is to look at the fraction of firms reporting numbers for profits and sales that are the same, which could only be true if costs were zero. Row 5 of Table 2 shows the percentage of businesses reporting the same numbers for sales and profits. This percentage is much higher in 2001 than in other years. This is a clear indication of a data quality problem in 2001 especially, but also in other years. It is not clear how this occurred. It may be that either the respondent or the enumerator did not understand the concepts of profits, sales and costs.

Row 6 of Table 2 shows that in all years there was a sizeable chunk of firms that reported no costs. This was highest in 2001 and lowest in 2005 but in 2001 was nowhere near the fraction of firms reporting equal values for profits and sales. Table 2 also shows that there has been a change in the way SESE surveys reports sales and profits- these were shown to be missing in 2001 and 2005 but in 2009 and 2013 only zeros are to be found in the data. This suggests that some of the zeros in these surveys may actually be missing and we cannot tell which are truly zero and which are missing. A suggestion to Stats SA would be to go back to reporting whether the data was missing or zero.

Another check on the quality of the accounting data is to see how much of the variation in revenues or profits a simple regression can explain across each SESE. If the fraction explained is much lower in 2001 this would be evidence that 2001 revenues and profits were poorly measured. The last two rows of Table 2 show the R squared of regressions of profits and revenues on the same set of controls for each SESE. These controls were a set of dummies for the age, location and industry of the firm and the race of the owner. 2001 does look somewhat worse than the subsequent surveys but the differences are not large, particularly with regards to profits.

[3] This is not a true non-response rate since it is possible, as discussed above for 2001, that individuals may have been running a non-VAT registered business as a second job and thus been eligible for SESE but refused to respond. There is no way of identifying such individuals in the data and thus the non-response rate discussed is a lower bound on the true non-response rate.

As a further data quality check it is also possible to explore whether the large increases in 2001 occurred across the country or only in specific provinces. Table 4 shows the changes between the LFS before the 2001 SESE and the 2001 SESE, as well as the changes between SESE 2001 and the following LFS. The results suggest that the changes were much larger in Free State, KZN and North West Province, with very little change occurring in the Western Cape. These results suggest that enumerators in some provinces responded very differently to the SESE-related incentives. But it is still not clear evidence of cheating- especially given that SESE finds substantially less employment in non-VAT registered firms than the reports from employees in the LFS. It does suggest that supervision or enumeration quality varies across provinces.

Comparisons with other data sources

Another possible check on the SESE data quality is to compare the estimates from SESE with those from other sources of data. The obvious one is the National Income Dynamics Study. NIDS is a nationally representative panel survey conducted in 2008, 2010 and 2012. Thus we can compare estimates from SESE 2009 and 2013 with the 3 estimates from NIDS. Although the NIDS survey instrument is different from the QLFS they are similar enough to allow comparisons. Both ask about self-employment. NIDS has one question that asks whether each business is registered for either VAT or income tax. We use this question to identify unregistered businesses but because income tax and VAT are asked about in one question the resulting estimated totals will be slightly lower than if we were able to identify non-VAT registered firms only. NIDS also has a shorter proxy questionnaire for those individuals who cannot be interviewed – a household member is asked questions about the non-responding members. This proxy questionnaire does not ask the question about non-VAT registered businesses. In estimating the number of non-registered businesses in NIDS we thus assume that the same proportion of self-employed identified through the proxy questionnaire are non-registered as those self-employed we can identify as either registered or not through the adult questionnaire.

Table 5 shows the estimates of total self-employment from SESE and NIDS. NIDS estimates more self-employment in 2008 than SESE 2009. However the NIDS trend is downwards – 2010 estimates 300 000 less unregistered self-employed than 2008 and 2012 is nearly 300 000 fewer again. By contrast the SESE estimates a substantial increase between 2009 and 2013. It is not clear what to make of these differences. It is possible that the panel nature of NIDS means that respondents are fatigued and therefore report less self-employment in the later waves. It could also be that attrition means the sample in later waves is not representative of the population of self-employed individuals. Or it could be that SESE 2009 was an underestimate of the true total and that the upward trend is an artefact of this 2009 underestimate. Without other data it is not possible to say anything concrete about the differences in trends and levels between the two surveys.

Conclusions

This data quality note has highlighted a number of statistics from the four SESEs that are cause for concern. 2001 seems to not be comparable with later SESEs, due to the much higher number of firms enumerated in this survey. I noted that this was likely due to the monetary incentives given to survey enumerators in 2001. I also showed that the large decrease between 2001 is exaggerated in SESE compared to the (Q)LFS, due to the large number of wage workers identified as owning non-VAT registered businesses in 2001 and not later years, as well as a lack of upweighting for non-response in SESE 2009.

The comparison of employment in SESE firms with employment reported by employees in non-VAT registered businesses in the LFS suggests that all SESEs undercount employees working in non-VAT registered businesses. This may be because owners do not fully disclose the size of their enterprises, because owners of large firms refuse to be interviewed or because the largest firms are not captured in the four samples that have been drawn for SESE. Whatever the reason all four SESEs seem to under count larger firms, and thus are unlikely to adequately capture total employment, revenue, profit or value added in non-VAT registered firms.

I noted that it was possible that the incentives given to enumerators in 2001 made it likely more marginal firms were enumerated. There was mixed evidence for this hypothesis. There was a larger concentration of owners reporting low hours worked, profits and sales in later SESEs, the opposite to what was expected if the marginality hypothesis was correct. However 2001 did have the highest fraction of firms reporting no employees, as well as the lowest rates of reported investment and having accounts, and the highest fraction of firms located in the owner's home.

I also undertook a number of other checks on the quality of the data. 2001 looked of worse quality in some, but not all, respects. 2001 had the highest fraction of firms reporting equal values for profits and sales but had the highest response rate. I noted that in 2001 and 2005 one can separate zeros from missing data but this is not possible in 2009 and 2013. This is one way in which the quality of the later surveys is worse but this could easily be fixed by Stats SA. Using the simple regressions of profits and sales on a few basic firm characteristics however, the R squared values for 2001 were lower than in other years, possibly indicating lower data quality.

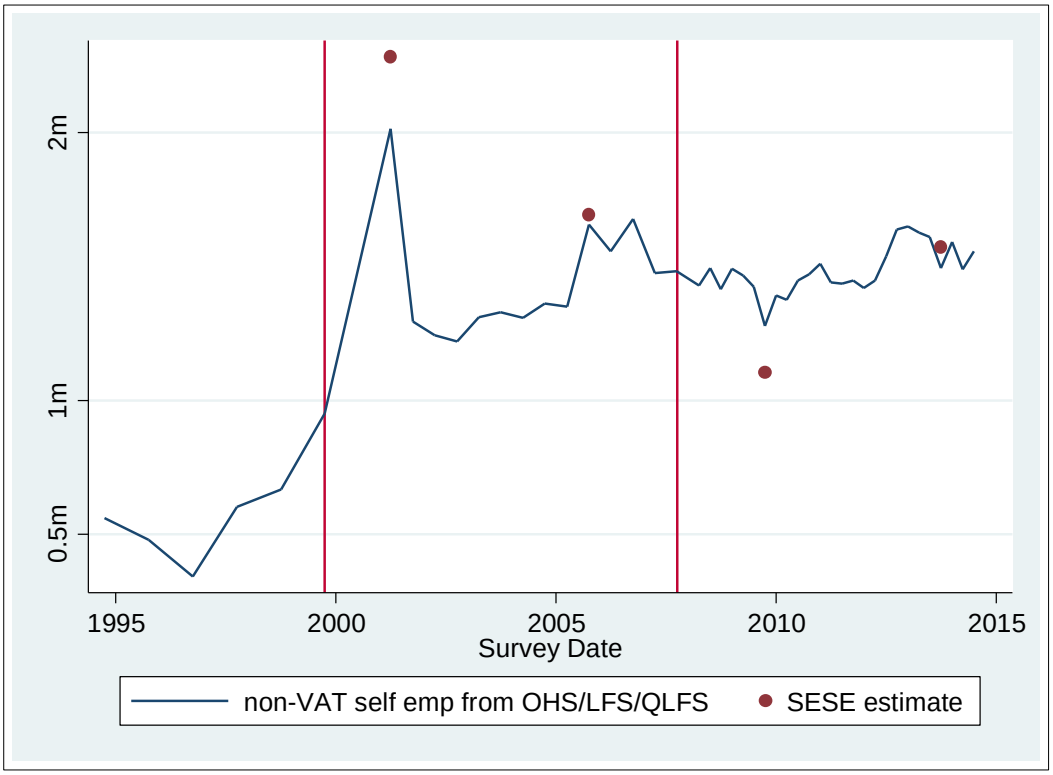
The four SESEs should definitely not be used for comparisons of total informal self-employment or employment in informal firms over time- as a result of the 2001 incentives but also the odd drop down in 2009 and then the spike in 2013, and the large gap between employment reported by employees in the LFS and employment reported by SESE owners, which is much smaller. Nevertheless there is valuable data in the SESE, which can be used to describe informal firms, and any contradictions between the years can be investigated and described.

References

Devey, R., Skinner, C. and Valodia, I. (2006), Definitions, data and the informal economy in South Africa: a critical analysis, in V. Padayachee, ed., `*The Development Decade?: Economic and Social Change in South Africa, 1994-2004*', HSRC Press, chapter 15, pp. 302-323.

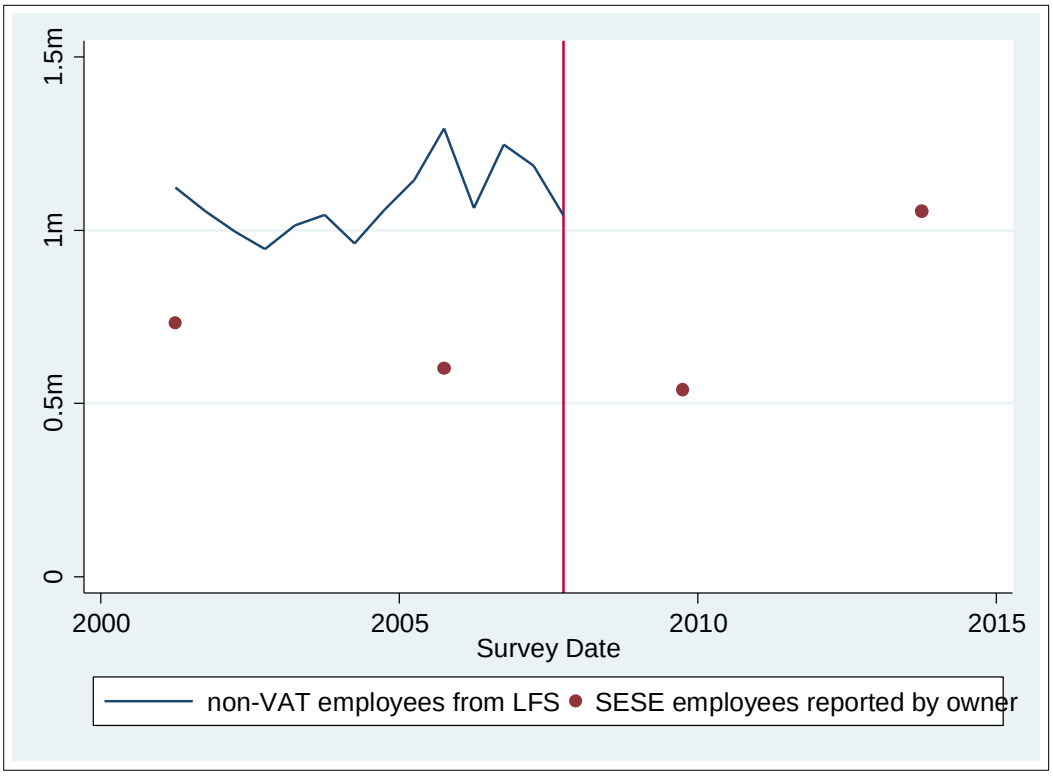
Figures

Figure 1: Trends in non-VAT registered self-employment from OHS/LFS/QLFS and SESE



Note: the vertical lines show the last OHS (in October 1999) and then the last LFS (in September 2007)

Figure 2: Trends in total employees in non-VAT registered businesses from LFS and SESE



Note: the vertical line shows the last LFS (in September 2007)

Figure 3: Implied and Actual firm size distributions from SESE 2001 and LFS February 2001

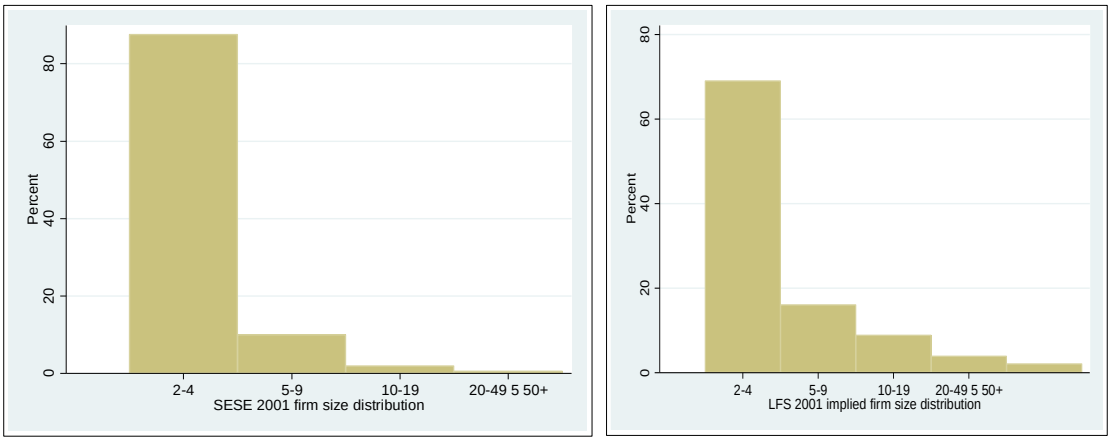
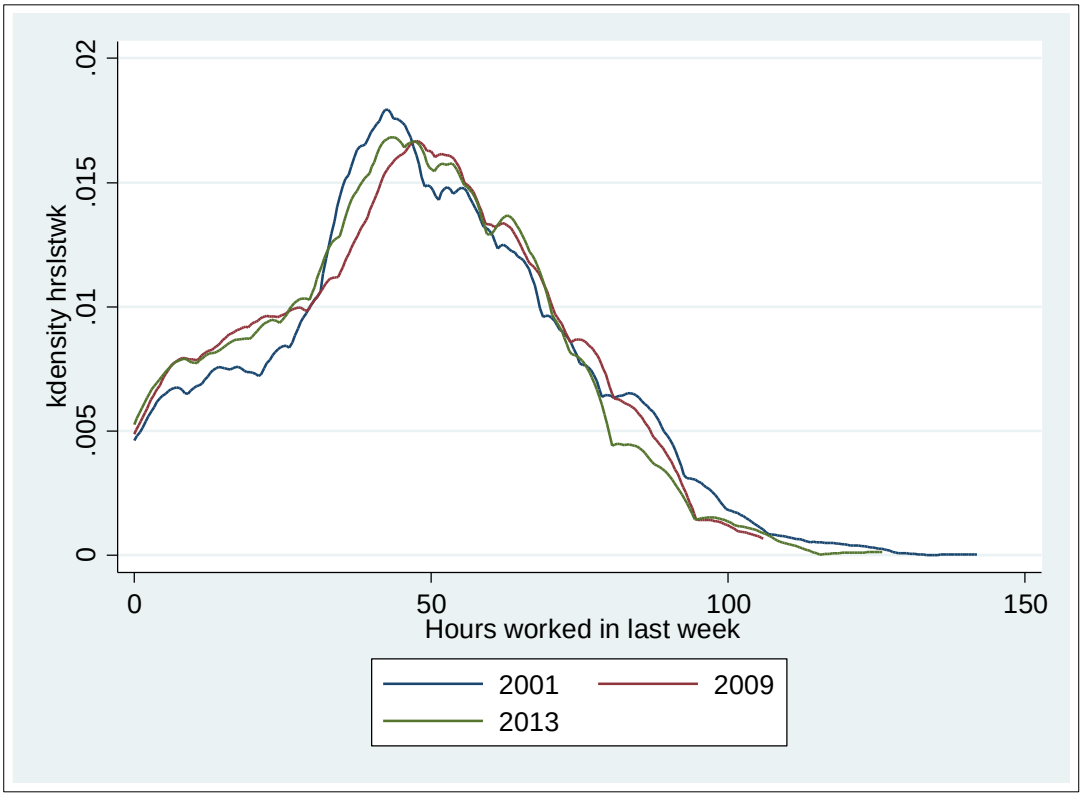


Figure 4: Hours worked in the last week



Note: 2005 is not included because it is not possible to link SESE 2005 with the corresponding LFS from that year.

Figure 5: Log of sales in last month

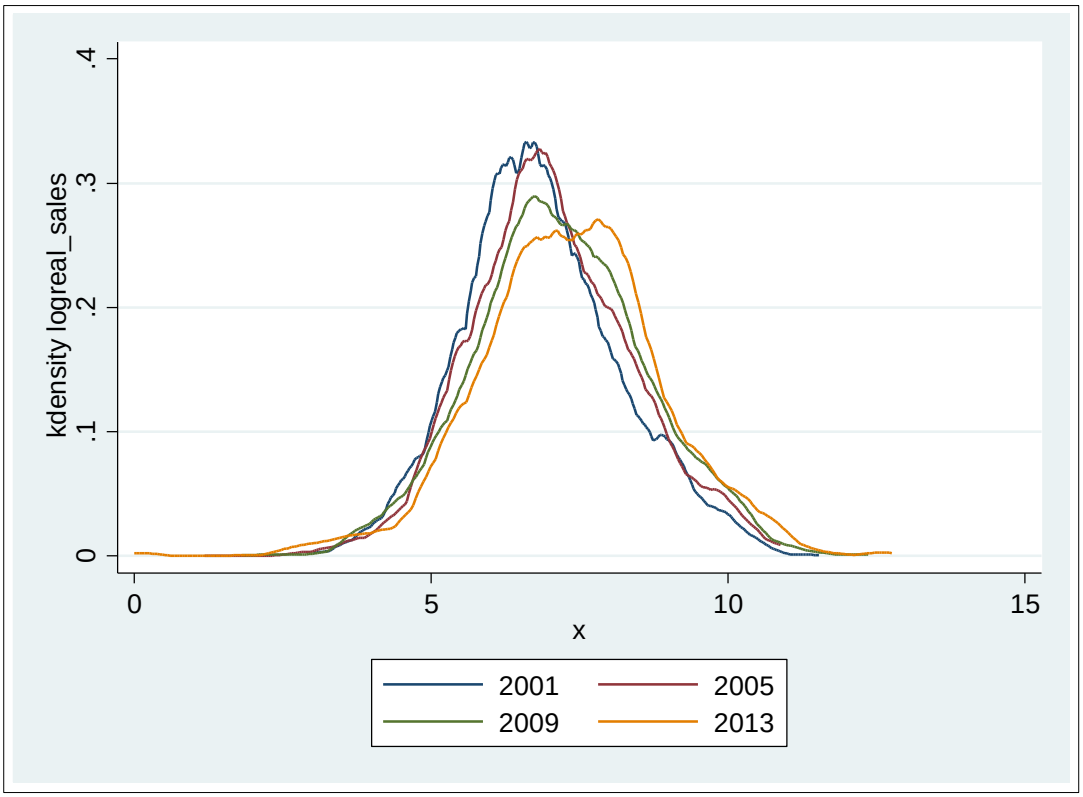
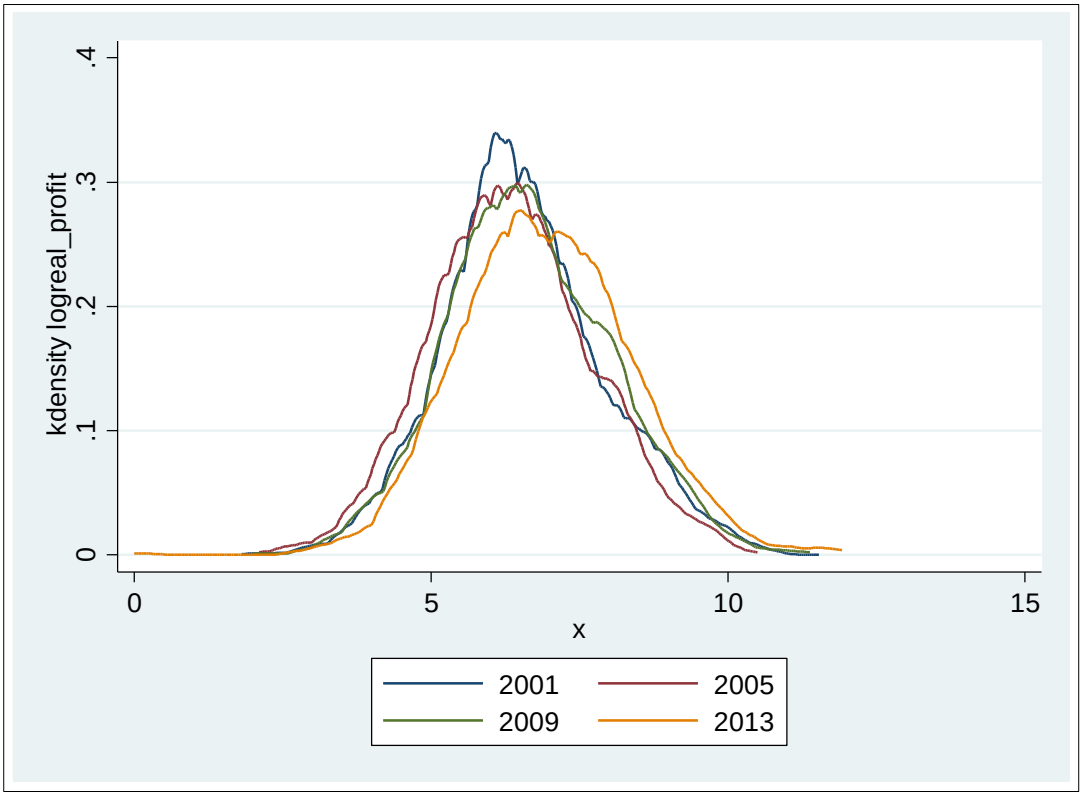


Figure 6: Log of profits in last month



Tables

Table 1: Number of firms with and without employees

	2001		2005		2009		2013	
	Number	CI	Number	CI	Number	CI	Number	CI
No employees	1947922	1860321, 2035523	1458936	1402310, 1515563	925740	868588.1, 982892.5	1215597	1142806, 1288387
Has Employees	335476	307787.3, 363166.3	289615	252778, 326453	211324	182624, 240023	385515	337444, 433586

Table 2: Data Quality indicators

Indicator	**2001**	**2005**	**2009**	**2013**
Zero sales	0.02%	4.56%	3.73%	4.66%
Missing sales	5.2%	1.6%	0	0
Zero profits	0.03%	5.97%	9.96%	7.79%
Missing profits	4.9%	0.8%	0	0
Profits=sales	38.1%	14.5%	20.3%	22.3%
Zero costs	16.5%	9.30%	15.0%	14.7%
Response rate relative to LFS or QLFS	95%	Cannot calculate	81%	82%
Profit regression R squared	19%	30%	22%	23%
Sales regression R squared	15%	27%	23%	19%

Table 3: Marginality Indicators

					P values		
Variable	**2001**	**2005**	**2009**	**2013**	**Test 2001=2005**	**Test 2001=2009**	**Test 2001=2013**
Investment	9.6%	18.1%	n/a	n/a	0	n/a	n/a
No accounts	81.6%	77.0%	78.7%	75.50%	0	0.047	0
Business located in the home	60.0%	49.6%	48.5%	47.3%	0	0	0
Paid employees	8.8%	13.2%	15.8%	19.4%	0	0	0

Table 4: Provincial changes for SESE 2001

	WC	EC	NC	FS	KZN	NW	Gau	MP	Lim
LFS 00:2	103535	158861	13001	60677	226489	80813	319078	117207	178748
LFS 01:1	112951	214539	16318	110303	510914	176121	494744	187514	272406
LFS 01:2	95869	191780	9986	64967	251172	93661	318861	114722	195500
% change between LFS 2 and LFS 3	9%	35%	26%	82%	126%	118%	55%	60%	52%
% change between LFS 3 and LFS 4	-15%	-11%	-39%	-41%	-51%	-47%	-36%	-39%	-28%

Table 5: SESE and NIDS estimates of non-registered self-employment

	NIDS			**SESE**	
	2008	2010	2012	2009	2013
Proportion of self-employed not registered	75%	73.90%	81.50%		
Total self-employed from proxy questionnaire in NIDS	138940	55191	92375		
Estimated no. of proxy respondents in non-registered self-employment (row1 *row 2)	104205	40786.15	75285.63		
Total non-VAT registered self-employed from adult questionnaire in NIDS	1298009	1102735	754918		
Total estimated non-VAT registered self-employment (row 3+row 4)	1402214	1143521	830203.6	1137064	1601112

About DataFirst

DataFirst is a data service dedicated to making South African and other African survey and administrative microdata available to researchers and policy analysts.

We promote high quality research by providing the essential research infrastructure for discovering and accessing data and by developing skills among prospective users, particularly in South Africa.

We undertake research on the quality and usability of national data and encourage data usage and data sharing.

DataFirst

www.datafirst.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 5708
info@data1st.org / support@data1st.org