A Guide to the Employers and the Self-employed Series

Andrew Kerr, DataFirst, February 2015

Introduction

The Surveys of Employers and Self-Employed (SESE) are a set of surveys that have been undertaken by Statistics South Africa in September 2001, September 2005, September 2009 and September 2013. A stacked cross-sectional dataset has been created by DataFirst from these four surveys and released publicly via the DataFirst web portal[1]. We have chosen to call this stacked cross sectional data set the Employers and Self-employed Series (ESS) 2001-2013. ESS can be downloaded and worked on immediately- it only requires registering on datafirst.uct.ac.za. This guide to ESS explains more about the surveys that underlie the series, discusses how this series was created and what variables are currently included.

More about the surveys

The Survey of Employers and Self-Employed are surveys of non-VAT registered businesses conducted by Statistics South Africa (Stats SA). The primary aim for Stats SA in running these surveys has been to obtain estimates of the size of value added in the informal sector for the compilation of the national accounts. Since no business register exists for businesses that are not registered, these businesses must be identified in another way. Stats SA identified non-VAT registered businesses through the Labour Force Surveys (in 2001 and 2005) and the Quarterly Labour Force Survey (in 2009 and 2013). Individuals who were surveyed in the (Q)LFSs and found to be running unregistered businesses were then targeted for a follow up survey- the SESE. This method corresponds to the 1 and 2 parts of a 1-2-3 survey (see ILO, 2013) The SESEs asked many detailed questions about the business run by the owner and enabled Stats SA to obtain value added in each unregistered business and thus an estimate for the whole unregistered economy.

Stats SA has made the SESE data publicly available, providing researchers with four cross sectional datasets of informal businesses that have relatively detailed data about each business. No other similar data exists for South Africa and thus it is a valuable resource for researchers wanting to better understand the formal economy. The data can also be linked to the Labour Force survey data collected as part 1 of this 1-2 survey and this further enriches the range of topics that can be studied using this data.

It should be noted that individuals could have more than one business. Each line in the data represents one business and so the same owner could be repeated several times on several lines of data. If researchers want to analyse the number of business owners rather than the number of businesses then they should remember how the data is set up and adjust their analysis accordingly. There will not be much difference, however, since most individuals report only owning one business.

Creating ESS from the SESEs

The dataset that has been created by DataFirst (ESS) is a stacked cross sectional data set containing data from both the SESEs and the (Q)LFSs. By this we mean that each of the four cross-sectional

[1] Funding for this project has been provided by the REDI project, funded by the National Treasury in South Africa.

SESEs undertaken by Stats SA have been cleaned, harmonised where possible, each SESE (except 2005) has had extra data from the (Q)LFSs added to it and then all these data have been put together in one data set. The data has been released to enable researchers to more easily access and use the wealth of data contained in the SESEs and the linked (Q)LFSs. This process is made transparent by releasing the files used to clean, harmonise and link the different datasets together with the actual data set.

Each of the four SESEs had its component parts merged together (if required). The variables were then renamed and relabelled for consistency across the four waves. Finally data from the relevant (Q)LFS was merged in using the PALMS dataset (see http://www.datafirst.uct.ac.za/dataportal/index.php/catalog/434 for more information on PALMS- which includes the three (Q)LFSs that were connected to three of the four SESEs). Each of the four datasets prepared in this way were then appended together. Data cleaning was also undertaken, along with further labelling.

The data was renamed and labelled using a set of do files and csv files in a method borrowed from Professor David Lam at the University of Michigan, who used this method to harmonise data from the South African Labour Force Surveys. These do files and csv files are available on the DataFirst portal, along with the micro data and this guide to the ESS. Doing the renaming and relabeling using this method is helpful because it makes clear which variables were asked in which wave and makes the do files used more structured, which may help new users of the data to quickly understand how the cleaning and harmonising was done. This method also lowers the time cost of including future SESEs into ESS.

Concerns about data quality

The 2009 SESE report released by Stats SA discusses the comparability of the 2001, 2005 and 2009 SESEs. It notes that whilst in 2009 and 2005 the QLFS and LFS were undertaken and then the SESE was administered several weeks later, in 2001 the SESE was undertaken immediately after the LFS interview if the owner or someone else who could answer was available. It then notes that "Because of these changes in the methodology, comparisons should be interpreted with caution." In discussions with Stats SA it was revealed that the key difference between 2001 and the other SESEs is that in 2001 enumerators were paid per SESE interview completed. There were thus very strong incentives to report those surveyed as self-employed in unregistered businesses.

This is borne out by the data- where the estimate of the total number of individuals running an unregistered business decreased from 2.2 million in 2001 to 1.7 million in 2005 and to 1.1 million in 2009. A halving of the number of informal self-employed businesses in 8 years is clearly unlikely to be what actually transpired. A separate piece of data quality research is being undertaken to document the differences between the different years and this will be released via the ESS page on DataFirst when it is complete.

Unit Non-Response Rates

Of all those individuals who were captured in the (Q)LFS and were eligible to be interviewed as part of the SESE, not all were interviewed for the SESE because of unit non-response. This could have been due to interview fatigue, a reluctance to give information to Stats SA enumerators etc. For those who did respond to SESE the difference in average SESE weights and the same weights in the

(Q)LFS gives us the non-response rate to SESE, for those who were interviewed in the (Q)LFS. Using the differences in average weights as the measure of non-response in 2001 the non-response rate of those who should have been in SESE and who were captured in the LFS was 95% whereas in 2009 and 2013 the response rate of those who should have been in SESE and who were captured in the QLFS was 81% and 82% respectively.

Item non-response (Missing Data)

Data that was missing due to the question being not applicable or not answered was coded in different ways across the different SESEs. In general 2001 and 2009 contain missing codes, for example 8888 or 99999, whilst 2005 simply lists these as missing in Stata (=.). Individual users of the data should be aware of these differences and the fact that the missing codes have generally been included and make their own decisions about what to do with missing data.

Data Versioning

Stats SA occasionally releases new versions of their data and it is thus important to keep track of which versions of the underlying SESE data were used in creating ESS.

For the creation of ESS version 1 for 2001we have used “SESE 2001 Operation_v1.2”, “SESE 2001 Location_v1.2” and “SESE 2001 Costs_v1.2”. These can be downloaded from the DataFirst website and were all last modified on the 23rd August 2013.

For 2005 we used the files "SESE 2005 Capital_v1.1.dta", "SESE 2005 Cost_v1.1.dta", "SESE 2005 Expenditure_v1.1.dta", "SESE 2005 Operations_v1.1.dta" "SESE 2005 Transport_v1.1.dta" and “SESE 2005 Business_v1.1.dta” These were all last modified on the 17th July 2012. There was a problem with the business number variable in the 2005 data that meant merging the separate files was not a simple undertaking. See the do file “sesemerge2005.do” for more details on how this issue was solved.

For 2009 we used the file “SESE 2009_v1.1.dta”, last modified on 21 May 2012.

For 2013 we used the file “SESE_2013_F1.dta”, downloaded from the Stats SA website on 15 October 2014.

Merging SESE and the (Q)LFSs

To create ESS we also merged in data from the relevant (Q)LFS. We used PALMS v2.1 to obtain this information and then merged it to each of the three SESEs. This merging process was simple in 2001, 2009 and 2013 where the household ids were identical in the relevant LFS or QLFS and the SESE. Unfortunately the ids used in 2005 SESE do not match easily onto the 2005 LFS data. Thus at the moment ESS contains no data from the LFS in 2005. We hope to obtain these data from Statistics South Africa relatively soon.

When releasing ESSv1.3 PALMS only had data going up to March 2012. We thus created a new version of PALMS and added in much of the required data. However this version of PALMS does not yet have income data and thus ESS 2013 does not have income data from PALMS at present. The income data from SESE is obviously included in ESS. The income data from PALMS for 2013 will be made available when a new version of PALMS is released publicly.

A description of the Stata do files and csv files used to create ESS

As noted above the data can be worked on straight away once it has been downloaded. For transparency, however, the code used to create these data are also available to download for those who wish to check how certain variables were coded or cleaned or for those wishing to add in, for example, more data from PALMS.

The first important do file is "SESEcreatedofiles.do" This takes the information from the csv files "SESE master codebook.csv" and "SESE Master value labels.csv" and creates a set of do files that do most of the renaming and value labelling to make the different SESEs consistent.

The files "sesemerge2001.do", "sesemerge2005.do", "sesemerge2009.do" and "sesemerge2013.do" put each SESE together, run the do files that rename and relabel the data and then merge in the PALMS data.

The file "appendsese.do" then puts the different SESEs together, recodes some of the missing data and renames some variables to emphasise that they are from PALMS and not the actual SESE.

References

ILO, 2013. Measuring informality: A statistical manual on the informal sector and informal employment. Available at http://www.ilo.org/stat/Publications/WCMS_222979/lang--en/index.htm

Variable Description for ESS v1.3

Variables are listed below in the order that they appear in the dataset.

Missing codes are not considered valid when describing the valid ranges below- valid range refers to the values that researchers would use when using a particular variable.

uqnr

Household identifier. Unique across waves. It is the same as the household id supplied by Stats SA in each SESE.

personnr

Person number within the household interview in the LFS or QLFS. Valid range 1-18.

weight

The Stats SA person weight released with each SESE. This differs from the Stats SA person weight released for the relevant (Q)LFS because of non-response to the SESE of those who did respond to the (Q)LFS.

busnumber

The business number of a particular owner. Each line of data represents 1 business. The vast majority of owners reported only 1 business so in most cases each line of data also represents 1 owner, but this is not always true because of the possibility of multiple businesses. Valid range 1-3.

year

The year of the SESE. Valid range: 2001, 2005, 2009 or 2013.

gender

Gender of the individual. 1=male, 2=female.

popgroup

Population group of individual. 1=African/black, 2=Coloured, 3=Indian, 4=White, 5=Other, 9=Unspecified.

location

Firm's location. Valid range 1-99.

paylocation

Does the owner pay for the location for business purposes . 1= yes, 2=No, 8=Not applicable, 9=unspecified.

recordtype

Type of record keeping. Valid range: 0-4.

expseparate

Are business expenditures recorded separately from hh expenditures? Valid range 1-9

license

Does bus.have permits or licenses. 1=yes, 2=no.

incometax

Is bus. registered for income tax? 1= yes, 2= No, 9= unspecified.

uif

Is bus. registered for UIF? 1= yes, 2= No, 9= unspecified.

monthoperate

Number of months firm operated in last year. Valid range: 0-12.

whynooperate

Main reason the business did not operate in some months? Valid range: 1-99.

firmage

Age of the firm, categorical variable. Valid range: 1-6.

startreason

Main reason you started the business? Valid range: 1-99.

singleowner

Is the business owned by 1 owner? Valid range: 1= single owner, 2= multiple owners.

partnersinhh

Are business partners in the household? Valid range: 1= yes, 2=no.

numhhpartners

Number of partners in the household. Valid range: 0-21

partnersinothhh

Are there partners in other households? Valid range: 1= yes, 2=no.

numothpartners

Number of partners outside the household. Valid range: 0-39.

emppaid1yr

Number of paid employees 1 year ago. Valid range 0-45.

empunpaid1yr

Number of unpaid employees 1 year ago. Valid range 0-15.

emppaid

Number of paid employees in last 7 days. Valid range 0-82.

empunpaid

Number of unpaid employees in last 7 days. Valid range 0-61.

empunpft

No. of full time unpaid employees. Valid range 0-39.

emppft

No. of full time paid employees. Valid range 0-45.

empunpt

No. of part-time unpaid employees. Valid range 0-61.

empppt

No. of part-time paid employees. Valid range 0-80.

emppmale

No. of paid male workers. Valid range 0-30.

emppfemale

No. of paid female workers. Valid range 0-60.

empunpmale

No. of unpaid male workers. Valid range 0-8.

empunpfemale

No. of unpaid female workers. Valid range 0-80.

emppblack

No. of black paid employees. Valid range 0-82.

empunpblack

No. of black unpaid employees. Valid range 0-60.

emppcoloured

No. of coloured paid employees. Valid range 0-60.

empunpcoloured

No. of coloured unpaid employees. Valid range 0-8.

emppindian

No. of Indian paid employees. Valid range 0-60.

empunpindian

No. of Indian unpaid employees. Valid range 0-80

emppwhite

No. of white paid employees. Valid range 0-13

empunpwhite

No. of white unpaid employees. Valid range 0-50.

suppliescost

Amt spent on supplies in last month. Valid range 0- 112000.

wages

Amt spent on wages in last month. Valid range 0- 70009.

inkind

Amt spent on inkind payments in last month. Valid range 0-8000.

trans

Amt spent on refunding employees transport costs in last month. Valid range 0-5500.

otherlabcost

Amt spent on other employee costs in last month. Valid range 0-1570.

rawmatamt

Amount spent on raw materials in the last month. Valid range: 0- 100000.

elec

Amount spent on electricity in the last month. Valid range: 0- 6200.

water

Amount spent on water in the last month. Valid range: 0- 6000.

fuel

Amount spent on fuel in the last month. Valid range: 0- 12000.

spares

Amount spent on spares in the last month. Valid range: 0- 9000.

rentprem

Amount spent on rent of premises in the last month. Valid range: 0- 6500.

rentequip

Amount spent on rent of equipment in the last month. Valid range: 0- 5400.

office

Amount spent on office supplies in the last month. Valid range: 0- 5500.

transport

Amount spent on transport of raw materials in last month. Valid range: 0- 9000.

repairs

Amount spent on repairs in last month. Valid range: 0- 40000.

busservices

Amount spent on business services in last month. Valid range: 0- 10000.

insmort

Amount spent on insurance and mortgage in last month. Valid range: 0- 9000.

protection

Amount spent on protection agencies in last month. Valid range: 0- 7000.

other

Amount spent on other in last month. Valid range: 0- 15000.

wagesown

Own salary (2009 and 2013 only). Valid range: 0- 150000.

inkindown

Own payment in kind (2009 and 2013 only). Valid range: 0- 70000.

transown

Own transport pay (2009 and 2013 only). Valid range: 0- 180000.

otherown

Own other renumeration (2009 and 2013 only). Valid range: 0- 8000.

sales

Value of gross sales in last month. Valid range 0- 350000.

otherinc

Amount of other income not from sales. Valid range 0- 300000.

profit

Net income/profit in last calendar month. Valid range: 0- 150000.

avgprofit

Average profit made in a month after deductions. Valid range: 0- 150000.

currentdebt

Business currently has debt? Valid range: 1= yes, 2=no, 9=unspecified.

debtsize

Size of current debt. Valid range: 0- 900000.

invmach

Did the business invest in machinery in the last year? 1=yes, 2=no.

invmachamt

Amount spent on machinery in the past year? Valid range: 0-30000

invtools

Did the business invest in tools in the past year? 1=yes, 2=no.

invtoolsamt

Amount spent on tools in the past year? Valid range: 0-75000

invvehicle

Did the business invest in vehicles, trailers, etc. for transport in the past year? 1=yes, 2=no.

invvehicleamt

Amount spent on vehicles, trailers, etc. for transport in the past year? Valid range: 0-9000000

invbuilding

Did the business invest in buildings and other structures in the past year? 1=yes, 2=no.

invbuildingamt

Amount spent on buildings and other structures in the past year? Valid range: 0-100000

invfurniture

Did the business invest in furniture in the past year? 1=yes, 2=no.

invfurnitureamt

Amount spent on furniture in the past year? Valid range: 0-19000

invother

Did the business invest in other capital items in the past year? 1=yes, 2=no.

invotheramt

Amount spent on other capital items in the past year? Valid range: 0-56000002E

startupk

Start-up capital required? Valid range 1= yes, 2=no.

ownstartupk

Own money used for Start-up capital? Valid range 1= yes, 2=no.

startupkamt

Amount needed for start-up capital. Valid range: 0- 800000.

startupkcoop

Borrowed from cooperative/stokvel for start-up capital? Only asked in 2001 and 2005. Valid range: 1= yes, 2=no.

startupkcoopamt

Amount borrowed from cooperative or stokvel for start-up capital. Valid range: 0-70000. Only asked in 2001 and 2005.

startupkother

Borrowed from other source for start-up capital? Valid range: 1= yes, 2=no. Only asked in 2001 and 2005.

startupkothamt

Amount borrowed from other source for start-up capital. Valid range: 0-75000. Only asked in 2001 and 2005.

startupkborrowany

Borrowed to start business? Valid range: 1= yes, 2=no. Only asked in 2009 and 2013, replaced the two types of borrowing questions in 2001 and 2005.

startupkborrowamt

Amount borrowed from any source for start-up capital. Valid range: 0- 750000. Only asked in 2009 and 2013, replaced the two types of borrowing questions in 2001 and 2005.

toilettype

Toilet facilities available. Valid range: 1-8. Not asked in 2009.

elecavail

Electricity available on site? Valid range 1= yes, 2=no. Not asked in 2009.

assistmostimp

What is the most important thing you need assistance with? Not asked in 2009. Valid range: 1-8.

Variables merged in from PALMS

Some of these are not asked in SESE so are extra data researchers might want to use, other variables are asked slightly differently or exactly the same in SESE and the (Q)LFS so researchers can use them

as a check on the answers given in both sources. 2005 PALMS data not currently merged in due to difficulties in merging SESE 2005 and LFS September 2005.

province

Province the owner’s household is located in. Valid range: 1-9.

ea

Enumeration Area the owner’s household is located in. Valid range: 1011001- 9.86e+07

stratum

Household Stratum in 2001. Valid range 1-18

stratum2

Household Stratum in 2009 and 2013. Valid range 101101 - 947405

urbrur

From PALMS: location type (2001 only). Valid range: 1=Urban, 2=rural.

urbrur2

From PALMS: location type (2009 and 2013 only). Valid range: 1=Urban formal, 2=Urban informal, 4= Tribal areas, 5=Rural formal.

metro

From PALMS: metro code (2009 and 2013 only). Valid range: 0-76.

ceweight2

Cross Entropy weight: ASSA2008 midyear estimates. Valid range: 1.154709- 5104.484

pweight

Stats SA person weight from (Q)LFS

inperson

Is the person responding in person? Valid range: 1= yes, 2=no.

marstat

Owner's Marital Status. Valid range: 1-4.

age

Owner's Age. Valid range: 15-105.

yrseduc

Owner's Years of education, derived variable. Valid range: 0-16.

educhigh1

Owner's Highest level of Education (2001 only). More detailed than yrseduc. Valid range: 0-22.

educhigh2

Owner's Highest level of Education (2009 and 2013 only). More detailed than yrseduc. Valid range: 0-26.

enrollment3

Owner's Enrollment in education (2001 only). Valid range: 1-3.

empstat1

Owner's Employment status, official. Valid range: 0-1.

empstat2

Owner's Employment status, expanded. Valid range: 0-1.

employer

Who does ... work for in main job (2001 only). Valid range1-5.

numworkers

Number of regular workers at job (2001 only). Categorical variable. Valid range: 1-7.

jobstartyear

Year started working in this job. Valid range: 1950-2013.

jobstartmonth

Month started working in this job. Valid range: 1-12.

jobindcode

1 digit industry code. Valid range: 1-10.

industry

3 digit industry code. Valid range: 10-990. 888 and 999 are Not applicable and missing codes respectively.

indus

1 digit industry code from SESE. Valid range 1-10.

jobocccode

1 digit occupation code. Valid range: 1-10.

occupation

4 digit occupation code. Valid range: 850-9333. 8888 and 9999 are Not applicable and missing codes respectively.

hrslstwk

Hours worked in total in last 7 days. Valid range: 0-142.

earnings

Consistent monthly earnings (rand amt) variable from PALMSv2. 2013 earnings data not currently included. Valid range: 2.5-80003.

outlier

Regression based outlier flag for earnings showing a studentised residual with value>5. Valid range: 0-1.

hhsize

Household size. (derived) from PALMS. Valid range: 1-26.

numemployed

Number of employed in hh excl business owner. (derived) from PALMS. Valid range: 0-10.

numwageemployed

Number of wage employed in hh excl business owner. Valid range: 0-4.

numselfemployed

Number of self-employed in hh excl business owner. Valid range: 0-6.

numkids

Number of <15 in the hh. Valid range: 0-13.

numold

Number of 65+ in the hh. Valid range: 0-4.

cpi

CPI from P0141 series by Stats SA, revised so that 2013 is the base year. To be used by researchers to make financial data from the four SESEs comparable. Valid range: .5004139 -1.