Introduction to the South African Revenue Service and National Treasury firm-level panel

Duncan Pieterse,[1] C. Friedrich Kreuser,[2] and Elizabeth Gavin[3]

May 2016

United Nations University World Institute for Development Economics Research

wider.unu.edu

Abstract: The South African Revenue Service and National Treasury Firm-Level Panel is an unbalanced panel data set created by merging several sources of administrative tax data received during 2015. The four data sources that constitute the panel are: (i) company income tax from registered firms who submit tax forms; (ii) employee data from employee income tax certificates submitted by employers; (iii) value-added tax data from registered firms; and (iv) customs records from traders. These data sets constitute a significant and unique source for the study of firm-level behaviour in post-apartheid South Africa. We review the key data sources used to construct the panel, highlight some important questions that arise as a result of panel construction, discuss the biases in the resulting data, compare key aggregates in the panel to other data sources, and provide a descriptive overview of the tax records.

[1] Director: Microeconomic Policy, National Treasury of South Africa, Pretoria, South Africa, corresponding author: Duncan.Pieterse@treasury.gov.za; [2] Researcher, Stellenbosch University, South Africa; [3] Executive: Policy Research, South African Revenue Service, Pretoria, South Africa.

1   Introduction

The South African Revenue Service (SARS) and National Treasury (NT) Firm-Level Panel (hereinafter the 'SARS-NT Panel'), is an unbalanced[1] panel data set created by merging several sources of South African administrative tax data received during 2015.[2] The four data sources that constitute the SARS-NT Panel are: (i) company income tax (CIT) data from CIT-registered firms who submit CIT forms; (ii) employee data from employee income tax certificates submitted by employers (i.e. IRP5 and IT3a); (iii) value-added tax (VAT) data from VAT registered firms; and (iv) customs records from traders. In this paper we review the key data sources used to construct the SARS-NT Panel, highlight some of the important questions that arise as a result of panel construction, discuss the biases in the data, compare key aggregates in the SARS-NT Panel to other data sources, and provide a descriptive overview of the tax records.

The forms that tax-paying persons and entities submit to SARS contain various fields, which are used to construct the variables in the SARS-NT Panel. The CIT records contain firm characteristics, such as the sector in which a firm operates, financial information from their income statements and balance sheets, and tax information. The employee records from individual IRP5 and IT3a forms contain all employee related incomes, deductions, taxes, and payments made by the firm (such as skills development levy and unemployment insurance fund payments). The VAT records contain information about input and output VAT paid and payable by the firm as well as VAT specific information, such as VAT paid on capital expenditure. The trader and customs records contain information about the products imported and exported by traders, including the product codes, value and volume of goods, and origin and destination information. Together, these data sets constitute a significant and unique data source for the study of firm-level behaviour in South Africa.

Using administrative tax records for cutting-edge empirical research is becoming widespread internationally (see Card et al. 2010). First, administrative data offer much larger sample sizes than traditional survey data sources, and can be harnessed to generate more compelling research designs. Second, administrative records have an inherent longitudinal structure that enables researchers to follow firms and individuals over time and conduct credible public policy evaluation. Third, administrative data provide better information than is typically available for survey sources, which suffer from high and rising rates of non-response, attrition, and under-reporting. As a result, using administrative tax data to conduct policy research has become widespread in such countries as Finland, Denmark, Sweden, the Netherlands, France, Germany, and New Zealand.

By combining different sources of administrative tax data, the SARS-NT Panel covers various aspects of firm behaviour. In the construction of the panel, all variables that could be used to identify firms were removed to create an anonymized data set where the confidentially of firms and individuals is ensured.

---

[1] In an unbalanced panel the number of time periods is not the same for all firms. For example, this occurs when firms drop out of the panel or new firms join the panel.

[2] This could lay the foundation for a broader South African Administrative Database, which includes other sources of available tax data (such as information from the EMP501 and EMP201 forms and personal income tax) as well as other administrative data. See Barnes et al. (2007) for a comprehensive review of other sources of administrative data for South Africa.

The remainder of this paper is organized as follows: Section 2 explores the key elements of the SARS-NT Panel including which tax records are included, how a firm is a defined, as well as how the SARS-NT Panel is constructed; Section 3 contains a discussion of data characteristics including biases in the data; Section 4 compares key variables in the data to other data sources such as the Quarterly Employment and Financial Statistics released by Statistics South Africa; and Section 5 briefly draws out some key descriptive statistics to convey the richness of the data.

2 Key elements of the SARS-NT Panel

2.1 The tax records that constitute the SARS-NT Panel

CIT data form the backbone of the SARS-NT Panel. CIT forms are completed by all businesses resident[3] for tax purposes in South Africa and include information in respect of taxable income.[4] CIT records are drawn from the IT14 and ITR14 forms. The ITR14 form replaced the IT14 form on 4 May 2013. In its current form, the SARS-NT Panel is available for the tax years from 2008–14. Since all tax returns submitted by a tax-registered entity must be completed electronically or at a SARS branch within 12 months of its financial year end, data for the 2014 tax year are still somewhat incomplete. The CIT data are augmented with variables from the following databases: (i) customs records; (ii) VAT vendors; and (iii) employee income tax certificates.

Customs records

The customs records contain information on the import and export activities of trading entities. Entities that appear in the trader database include the following: local importers and exporters; foreign importers and exporters; rebate users;[5] 'drawback' manufacturers;[6] participants in the Automotive Production and Development Programme (APDP);[7] and firms who operate in an Industrial Development Zone (IDZ).[8]

Customs records are available at the transaction level for entities identified with a customs reference number. Each transaction entry includes information on the Rand amount imported or exported, the destination or origin, the quantity and unit of the product, as well as detailed product codes using the Harmonised System (HS) of the World Customs Organisation. Customs

---

[3] A business is resident in South Africa if it is incorporated or effectively managed in South Africa. Non-resident companies which operate through a branch or which have a permanent establishment within the Republic are subject to tax on all income from a source within South Africa.

[4] Not-for-profit organizations that receive preferential tax treatment are an exception provided they meet the requirements set out in the Income Tax Act. Once approved by SARS, the organization is registered as a public benefit organization (PBO), allocated a unique PBO reference number, and annually submits a separate IT12EI income tax return.

[5] Individuals and entities who are allowed to import goods without paying duties under certain conditions, e.g. for processing, manufacturing, and packaging with the intention of exporting them.

[6] All local persons who import goods into South Africa that are used in the manufacture of goods which are subsequently exported should register if they wish to apply for a drawback (refund) for exporting goods on which duties were already paid.

[7] APDP is a production incentive programme for the motor industry whereby components are imported under rebate item 317.03 for the manufacture of light motor vehicles and 317.06 for the manufacture of automotive components. Anyone who imports components for use in terms of such provisions should register for the APDP.

[8] Available at: http://www.sars.gov.za/ClientSegments/Customs-Excise/Processing/Pre-assessment/Registration/Pages/default.aspx.

reference numbers are linked to a CIT reference number using a link table provided by SARS. These firms are then matched to the CIT database as detailed in Section 2.3. While customs records are available to researchers at the transaction level, the SARS-NT Panel includes aggregated information for each customs reference number level for the tax year in question. The SARS-NT Panel includes the total Rand value of imports and exports, number of import origins and export destination, and number of unique products imported or exported.

Value-added tax data

VAT is an indirect tax on the consumption of goods and services in the economy and is charged at each stage of the production and distribution process. VAT is presently levied at the standard rate of 14 per cent on the supply of most goods and services and on the importation of goods. Some goods and services are subject to VAT at zero rate or are exempt from VAT. Although any person that carries on a business may register for VAT, VAT registration is only mandatory if a company's taxable supplies are in excess of R1 million in any consecutive twelve-month period. These businesses become vendors that act as the agent for government in collecting VAT. VAT-registered entities submit these forms on a monthly, bi-monthly, quarterly, semi-annual, or annual basis. The information contained in these forms is aggregated to an annual level after which the VAT reference number of an entity is linked to a CIT reference number to be matched to the CIT data. The SARS-NT Panel contains the aggregated information on all of the VAT fields including input and output VAT as well as more specific information such as VAT paid on capital expenditure.

Employee tax data

The employee database contains information from individual level employee tax certificates (IRP5 and IT3(a)) submitted by pay-as-you-earn (PAYE) registered entities. All employers must register with SARS within 21 business days after becoming an employer, unless none of the employees are liable for normal tax. An employer must issue an employee with an IRP5 or IT3(a) certificate where remuneration is paid or has become payable. Where no employee tax was deducted from remuneration and the employee receives more than R2,000 per month, an IT3a form is provided to an employee. If an employee earns less than R2,000 in a given tax year and no employee tax was deducted, the employee is not issued with an IRP5 or IT3a form. IRP5 certificates of all employees in a company must be submitted within 60 days of the end of the tax year. The IRP5 and IT3a forms issued by employers are reconciliation forms that include details of the total amount paid by that employer to the employee from different sources, as well as the total amount of tax paid, skills development levy payments, unemployment insurance fund (UIF) payments, as well as the periods worked in the year of assessment. Data from these forms are used to generate employment estimates and employee characteristics. Each IRP5 or IT3(a) submitting entity is identified through a PAYE reference number which is linked to a CIT reference number to be merged with the CIT database. The aggregation of employment data is described in Section 2.5.

2.2 What is a firm?

The richness of the tax records used to construct the SARS-NT Panel allows for several different ways to think about a firm: a firm could be a CIT-registered entity, a PAYE-registered entity (i.e. an employer), or a grouping of either. In the SARS-NT Panel a firm is defined as a CIT-registered entity that has completed an IT14 and/or ITR14 form. In other words, each unique CIT reference number is associated with a firm. Given this definition, Table 1 illustrates how one should think about employees, employers, and firms in the SARS-NT Panel.

Table 1: Defining a firm: Merging different tax records

Firm	Employer	Employee
CIT reference number 1	PAYE reference number 1	Individual IRP5 1
		Individual IRP5 2
		Individual IRP5 3
	PAYE reference number 2	Individual IRP5 3
		Individual IRP5 4
CIT reference number 2	PAYE reference number 3	Individual IRP5 1
		Individual IRP5 5
		Individual IRP5 6
No CIT reference number	PAYE reference number 4	Individual IRP5 7
		Individual IRP5 8
		Individual IRP5 9

Source: Authors' figure based on tax records.

First, a single CIT-registered entity (e.g. CIT reference number 1) can be associated with multiple PAYE reference numbers if, for example, that entity has multiple branches. As shown in the table, where several PAYE numbers can be linked to the same firm, the same tax reference number is allocated to these different PAYE reference numbers. In one case more than 40 PAYE reference numbers were linked to a single CIT reference number in one year. This ensures that multiple branches and employees of the same firm are allocated to that firm. As shown in Table 1, although a single CIT-registered entity can have multiple PAYE reference numbers, the converse is not possible.

Second, the same worker (e.g. individual IRP5 3) can appear in two PAYE-registered entities for the same firm—in this case two records are retained for this worker. Therefore, the employment numbers linked to a firm count the number of jobs, not individuals. There is also the case of IRP5 revisions where multiple IRP5 forms are available for some workers within the same PAYE-registered entity; this is discussed in Section 2.5. Furthermore, if individual IRP5 forms associated with reference numbers 1, 5, and 6 contain no employment income information, then the firm associated with CIT reference number 2 will have no workers (*IRP5_empl*=0), but listed as having three forms (*IRP5_forms*=3). For a PAYE reference number to be matched to a CIT reference number, it must have at least one IRP5 form associated with it.

Third, many workers can be employed at the same firm so multiple individual IRP5 numbers can be associated with one CIT reference number. Furthermore, more than one IRP5 certificate can be issued to an individual if that person has more than one job (either simultaneously or during consecutive stints in the same tax year). As shown in Table 1, individual IRP5 1 can be linked to both CIT reference numbers 1 and 2 (e.g. a consultant who works in both academia and the private sector). Similar to the case of individual IRP5 3, in this case individual IRP5 1 is counted as two workers, even though this is in fact one individual.

Fourth, in cases where a tax reference number does not have a PAYE number in the link table, a firm in the CIT data cannot be linked to a PAYE reference number in the IRP5. The lack of a corresponding PAYE reference number means that no IRP5 forms can be linked to a tax reference number. In this case that particular firm will have missing IRP5 variables (i.e. worker characteristics). There are many firms that could fit the description of having no workers; however a question remains about whether these are genuine firms. For example, an individual could register a company for their rental properties to benefit from lower company tax rates (compared to their marginal personal income tax rate). In this case, the associated CIT-registered entity may not have any employees if the owner does not report income received from the firm through an IRP5. Another example would be bank nominee companies, which are CIT-registered entities that hold significant assets on behalf of insurance firms, pension funds, and investment companies. These entities are unlikely to be registered for PAYE and anecdotal evidence suggests that these companies may even classify themselves as dormant.[9]

Finally, there are some IRP5 forms (7–9) that cannot be linked to a firm. These would include employees of government departments. These individuals are dropped from the SARS-NT Panel since no firm-level information is available for these observations. However, the worker data for these individuals are still available in the IRP5 panel and researchers are therefore still able to explore their characteristics. IRP5 forms that cannot be linked to a firm highlights the fact that defining a firm as a CIT-registered entity is merely one way of thinking about a firm. A researcher who is primarily interested in the employee information in the tax records may want to define a firm as an employer (i.e. a PAYE-registered entity). However, as shown in Table 1, if one wishes to use CIT information in this case one would have to think about how to treat cases such as CIT reference number 1, where one set of CIT data is available for two PAYE-registered entities.

2.3 Constructing the SARS-NT Panel

The SARS-NT Panel is created in the following steps: first, the raw data from both the IT14 and ITR14 forms are extracted from a static SQL database as provided by SARS. The first step is to combine the data from the IT14 and ITR14 forms and clean the data where necessary as illustrated in Section 2.3. There is an important difference between the IT14 and ITR14 forms, and this has an implication for how the SARS-NT Panel is constructed: the IT14 form contains the same set of fields for all firms irrespective of their size, whereas the ITR14 form includes different questions for firms of different sizes, with a greater degree of disaggregation for larger firms. Therefore, while a consistent set of data is available for all companies who completed the IT14, in the ITR14, medium and large companies complete and submit a comprehensive extended return, while micro- and small firms complete a shorter, simplified return.[10]

In Table 2 we show an example of the data for fixed property, plant, and equipment. The ITR14 requires micro- and small firms to submit data on the closing values of property, plant, and

---

[9] There are currently about 100 bank nominee companies approved by the Financial Services Board (FSB).

[10] A micro-business is classified as a company with a gross income (sales/plus other income) not exceeding R1 million and total assets (current and non-current) not exceeding R5 million, and which is not classified as a body corporate/share block company. A small business is classified as a company with a gross income not exceeding R14 million and total assets not exceeding R10 million, which is not classified as a body corporate/share block company or micro-business. A small business is not the same as a small business corporation as defined in section 12E of the Income Tax Act. If a company is not classified as a body corporate/share block company, micro-business, or small business, it will be classified as a medium to large business (i.e. gross income exceeding R14 million and/or total assets exceeding R10 million).

equipment in one field (without a separate field for other fixed assets), while medium and large firms are required to submit information on fixed property, plant and equipment, and other fixed assets in three separate fields. In contrast, the IT14 asks all firms to disaggregate the information into three separate property, plant and equipment, and other fixed assets fields.

Table 2: Example of raw ITR data

Firm ID	Property, plant and equipment (micro)	Property, plant and equipment (small)	Property (medium – large)	Plant and equipment (medium – large)	Other fixed assets (medium – large)
1	x				
2		y			
3			a	b	c

Source: Authors' figure based on CIT cleaning code.

In Table 3, we show how these fields are harmonized in the cleaning process. We combine only the property, plant and equipment fields from the ITR14 and IT14 for all firms to generate a single new property, plant and equipment field. The other fixed assets field is not included as it is not clear from the form that micro- and small firms will include the information. As small and micro-firms do not have reported information on other fixed assets, these fields are assigned missing values.

Table 3: Example of harmonized ITR data

Firm ID	Property, plant and equipment	Other fixed assets
1	x	.
2	y	.
3	a + b	c

Source: Authors' figure based on CIT cleaning code.

Therefore, where firms of different sizes submit the same information in different parts of the form, the fields are harmonized so that only one variable contains the matchable data. Unique source codes from ITR14 forms to IT14 forms (and vice versa) are matched to ensure that variables captured by the old IT14 are easily and directly related to the updated ITR14 form. Where no similar information exists the non-matched field is kept in its raw state.

Second, the IT14 and ITR14 variables are renamed and given the relevant prefixes (to indicate from which form the variables are drawn) and only the last revisions are retained for each firm. In other words, in all cases, earlier submissions of financial statements as part of the same tax

return are discarded in favour of the most recent submission in a particular year.[11] This removes many of the duplicates in the CIT panel due to multiple submissions by the same firm in a particular year. Furthermore, some cleaning is required because some firms may have submitted both IT14 and ITR14 forms within the same year. To address this, each IT14 and ITR14 variable is collapsed where the variable suffixes are the same and share the same numeric format. In all cases the ITR14 variable (where it exists) takes precedence over an IT14 equivalent as the ITR14 will have been submitted at a later date. The collapsed data are given a new prefix that corresponds to the variable category (for example, the capital variables from the balance sheet have the new prefix 'k_'). The data are then checked to ensure that a specific CIT-registered entity does not appear more than once in a particular year in the CIT panel. The variables remaining with IT14 and ITR14 prefixes are not matched and it is up to the researcher to decide how to include these variables and where to aggregate them.

Third, the finalized CIT panel is merged with the employer data from the IRP5 forms as well as the VAT and customs records (Section 2.4 contains a detailed discussion of how the IRP5 panel is created and merged). A link table provided by SARS, containing all tax reference numbers and the corresponding PAYE, customs, and VAT reference numbers, is used to implement the merge. In all cases, VAT and customs tax records are dropped if they have no tax reference numbers in the link table (e.g. VAT records for individuals and partnerships). In the final step, an unchanged labour broker data set, which contains a PAYE reference number and labour broker indicator, is merged into the SARS-NT Panel. It is important to note that all unique CIT-registered entities, or firms, are retained in the SARS-NT Panel made available to researchers.

2.4 Variable naming conventions

The variable naming conventions used in the SARS-NT Panel are aimed at simplifying its use and include prefixes and suffixes that identify variable type, original data set (where necessary), and other descriptors. For example, all variables in the IT14 and ITR14 forms that do not have an equivalent in the other form (for example, due to changes in the forms over time) are retained with the appropriate prefix (IT14 and ITR14), as well as the categorization prefix (for example ITR14_k_). The main prefixes are shown in Table 4.

---

[11] Firms are allowed to submit revisions of IT14 and ITR14 forms if they feel they have been inaccurate or if the firm made unreported changes.

Table 4: Summary of main prefixes

Prefix	Description
c_	Characteristic
k_	Capital items
e_	Equity items
t_deb	Tax items – debit
t_cred	Tax items – credit
y_	Income
x_	Expenses
irp5_	IRP5
l_	Labour
v_	VAT
cust_	Customs

Source: Authors' table based on cleaning code.

The CIT forms include tax computation fields that reflect adjustments to the accounting figures—these are given the prefix *t_deb* and *t_cred*, respectively. The *t_deb* entries are debit adjustments which decrease net profit or increase net loss. This category includes non-taxable amounts credited to the income statement, allowable adjustments, and special allowances not claimed in the income statement. The *t_cred* entries are credit adjustments that increase the net profit of the firm or decrease its net loss. This category includes non-deductible amounts debited to the income statement, allowances or deductions that were reversed, as well as adjustments to be added back into the income statement. It should be noted that for certain variables, for example depreciation or interest paid, the SARS-NT Panel will contain both the accounting variable from the income statement and the tax variable.

2.5 Timing and merging of data

The SARS year of assessment for individuals only—not companies—commonly referred to as a 'tax year' runs from 1 March to the end of February each year. Each firm submits a tax return for a tax year determined by their financial year end. In the CIT data each firm has a financial year end variable that indicates the month of the year that corresponds to the firm's financial year end. Most firms have financial year end data: in the IT14 we have the financial year end month and in the ITR14 we have the financial year end day and month.[12]

---

[12] In cases where a firm does not have a financial year end, it is assumed that the firm's financial year end coincides with that of the tax year (i.e. the end of February). In Table 5 we provide justification for this. The original financial year end data for the firm are available in the SARS-NT Panel.

Figure 1: SARS tax year
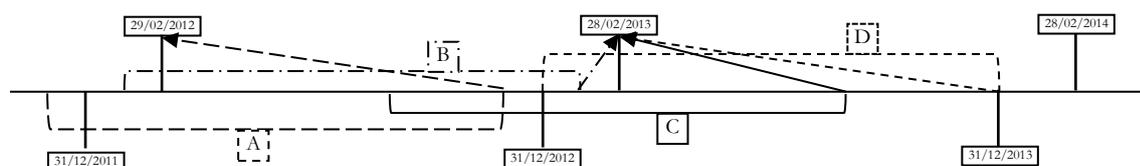
Source: Authors' figure based on SARS-NT (2015).

All firms with a financial year that ends in 2013 fall into the 2013 tax year (which runs from 1 March 2012 to 28 February 2013). In Figure 1, this would include both Firm B that has a financial year from 1 January 2013 to 31 December 2013, as well as Firm D that has a financial year from 1 February 2012 to 31 January 2013. As a result, both Firms B and D fall within the same tax year although they do not share directly comparable financial years. Therefore, while the tax year variable is useful for comparing data to those reported by SARS in their annual publication, the tax year variable is less useful when trying to align the reporting period of firms based on their financial year end.

In contrast to the CIT data, the IRP5 data report on employee characteristics for the year from 1 March to the end of February the following year, regardless of a firm's financial year end. As a result, a decision rule has to be employed to match (as closely as possible) a firm's financial year end to their IRP5 tax year. We take the month from the financial year end variable in the CIT data to create a new financial year variable that more closely aligns the IRP5 data to the actual financial year of the firm.

The new *finyear* variable is created as follows: in all cases where the financial year end variable is equal to or greater than 8, the firm's tax year (as classified by SARS) is pushed to the following tax year; and in cases where the financial year end variable is less than 8 (i.e. August), the firm's tax year remains as is. As shown in Figure 2, this means that Firms A and B will now both be moved forward to end in the financial year ending 28/02/2013. Firms C and D will also be moved forward to the year ending 28/02/2014, but as can be seen their period of overlap is considerably smaller than the overlap for A and B.

Figure 2: Generating the firm's financial year
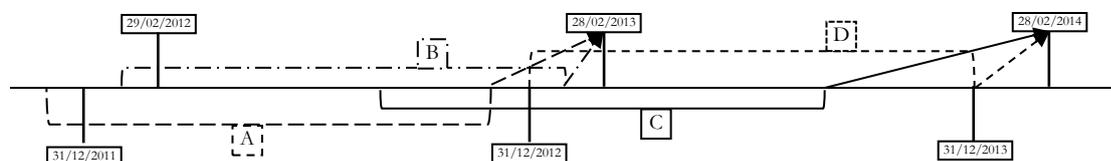
Source: Authors' figure based on SARS-NT Panel cleaning code.

This ensures that the IRP5 data corresponds to at least 6 months of the firm's financial year. This financial year variable (*finyear* in the data) is the most appropriate time variable to use in regression and other analysis as it is more comparable across firms within a given year. Table 5 shows that around 85 per cent of firms have their financial year end at the end of February. Due to the strict domination of firm financial years ending in February in each year, we assume a firm's financial year ends in February where it does not provide the details. In Table 5, the three largest other category of month ends are in June, March, and December. The proportions (and

9

numbers) of CIT returns by month end for the latter years in the panel are an indication of the lack of completeness of these years.

Table 5: Firm*'s* financial year end

Firm's Financial Year End	2008	2009	2010	2011	2012	2013	2014	Total
January	1 501	1 538	1 770	1 934	2 145	2 341	2 227	13 456
	(0.217%)	(0.223%)	(0.245%)	(0.27%)	(0.304%)	(0.351%)	(0.425%)	(0.285%)
February	590 955	598 529	619 473	616 717	601 239	566 966	453 924	4 047 803
	(85.615%)	(86.883%)	(85.895%)	(85.942%)	(85.213%)	(85.055%)	(86.539%)	(85.856%)
March	18 132	16 131	20 040	20 332	21 163	20 664	16 357	132 819
	(2.627%)	(2.342%)	(2.779%)	(2.833%)	(2.999%)	(3.1%)	(3.118%)	(2.817%)
April	5 387	5 312	5 641	5 683	6 106	6 320	5 351	39 800
	(0.78%)	(0.771%)	(0.782%)	(0.792%)	(0.865%)	(0.948%)	(1.02%)	(0.844%)
May	3 266	3 179	3 337	3 307	3 959	4 082	3 333	24 463
	(0.473%)	(0.461%)	(0.463%)	(0.461%)	(0.561%)	(0.612%)	(0.635%)	(0.519%)
June	28 848	27 190	28 408	27 794	27 199	25 523	20 472	185 434
	(4.179%)	(3.947%)	(3.939%)	(3.873%)	(3.855%)	(3.829%)	(3.903%)	(3.933%)
July	3 077	3 057	3 114	3 179	3 814	3 869	3 097	23 207
	(0.446%)	(0.444%)	(0.432%)	(0.443%)	(0.541%)	(0.58%)	(0.59%)	(0.492%)
August	7 947	6 514	8 096	7 924	8 357	7 307	4 800	50 945
	(1.151%)	(0.946%)	(1.123%)	(1.104%)	(1.184%)	(1.096%)	(0.915%)	(1.081%)
September	6 524	6 048	6 417	6 210	6 347	6 266	4 013	41 825
	(0.945%)	(0.878%)	(0.89%)	(0.865%)	(0.9%)	(0.94%)	(0.765%)	(0.887%)
October	2 399	2 299	2 459	2 451	2 687	2 818	1 893	17 006
	(0.348%)	(0.334%)	(0.341%)	(0.342%)	(0.381%)	(0.423%)	(0.361%)	(0.361%)
November	1 660	1 704	1 811	1 943	2 229	2 448	1 675	13 470
	(0.24%)	(0.247%)	(0.251%)	(0.271%)	(0.316%)	(0.367%)	(0.319%)	(0.286%)
December	20 553	17 393	20 633	20 120	20 326	17 982	7 389	124 396
	(2.978%)	(2.525%)	(2.861%)	(2.804%)	(2.881%)	(2.698%)	(1.409%)	(2.639%)

Source: Authors' figure based on SARS-NT Panel.

The approach set out above to link data sets with different reporting periods is not ideal where we have actual date data. For example, most VAT returns pertain to either a single month or two

and the customs transaction data have exact dates, so aligning the time periods of these tax records to the CIT data should be less complex. Even in the case of the IRP data, we have periods employed to and from and could arguably use this to more accurately match the IRP5 records to the CIT data. However, creating the *finyear* variable remains useful for analysis where we think that increased granularity is associated with increased measurement error or where the date data are incomplete dependent on form year. The latter is the case for the IRP5 data where the periods employed to and from data are only well populated after 2009. In Section 2.5 we discuss the construction of the labour data by shifting the financial year and by using the employment dates. In the next iteration of the data we will also include this information for the customs and VAT data.

2.5 Labour data

While the CIT data contain information about the wages paid by the firm, there is no information about the number of workers or of the nature of their employment (e.g. when they started working). A firm-level labour indicator for the SARS-NT Panel is derived from the associated IRP5 forms (i.e. IRP5 and IT3a forms) submitted by PAYE-registered firms. In addition to the employment measures, labour data also include an employment tax incentive (ETI) indicator which indicates whether the individual receives the ETI. At present, no wage or income measures are used from the IRP5 forms. Twelve different employment measures are constructed. These indicators can be grouped into two sets: (i) the total periods set; and (ii) the dates set. For the total periods set, we create six employment measures by matching the IRP5 data to the CIT data using the timing and merging approach discussed in Section 2.4. For the dates set, we create another six employment measures where an employee is assigned to a firm's financial year based on when they started and finished a particular job spell. The trade-offs associated with the two approaches are discussed briefly at the end of this section.

The total periods set

The first set of six employment measures makes use of number of periods worked and total number of periods in year of assessment data in the IRP5 forms. Three unweighted variables are created: (i) the sum of all unique IRP5/IT3(a) forms per unique tax reference number entity (Equation 1); (ii) the sum of all unique IRP5/IT3(a) forms with positive and non-missing taxable income recorded as 3601 per unique tax reference number entity (Equation 2); and (iii) the sum of all unique IRP5/IT3(a) forms with positive and non-missing employment income per unique tax reference number entity (Equation 3).[13] A unique individual employed by a firm is identified by a unique South African identification (ID) number to PAYE reference number match. After the ID–PAYE link is made, the data are cleaned by CIT reference number. Therefore, an individual employed at two different branches of the same firm will be counted as two employees. As indicated in Section 2.2, the employment numbers for the firm count the number of jobs by PAYE-registered employer, not the number of individuals. Where a South African ID number is not available the passport number is used and where the passport number is not available a certificate number is used. Where an ID–PAYE reference number match appears more than once in a given year, the number for the highest period worked is used unless the total weight of the worker is smaller than or equal to one, as discussed below. Where no period worked number exists, the highest 3601 income figure is used; otherwise the first that appears in the data is used. In other words, only one record is selected from multiple certificates reflecting the same combination of individual ID and firm in a year if there is no 'periods worked'

---

[13] The 3601 code refers to amounts paid to an employee for services rendered, overtime, and pension and paid on a regular basis in tax years 2010–13.

information. It should be noted that for the tax years 2010, 2011, and 2012 income associated with pensions and retirement annuities was also included in the 3601 income code along with salaries and wages paid to an employee. During these years the count of employees using the 3601 code will thus yield an over-count of employment. However, this is likely to occur largely in the finance sector. This measure is also included as an indicator along with other income sources for the employment income indicator.[14]

$$irp5\_forms_{f,t} = \sum_i Form_{i,f,t} \tag{1}$$

$$irp5\_empl3601_{f,t} = \sum_i Form_{i,f,t} \times (3601\ Income_{i,f,t} > 0) \tag{2}$$

$$irp5\_empl_{f,t} = \sum_i Form_{i,ft} \times (Employment\ Income_{i,f,t} > 0) \tag{3}$$

The second set of three measures uses employee weightings to augment the unweighted measures. The weighted employment measures intended to account for cases where individuals were not employed for the full duration of a particular year, which is the case for a significant proportion of the tax certificates. An employee's weighting is a measure of the employment spell in a given year and is calculated using Equation (4).

$$weight_{i,f,t} = \frac{Total\ Periods\ Worked_{i,f,t}}{Total\ Periods\ in\ Year\ of\ Employment_{i,f,t}} \tag{4}$$

Calculating these weightings is complicated by two factors: first, employers report the total periods in year of employment differently, with some using days, weeks, months, or years. Furthermore, since the different ways in which the number of periods worked (i.e. days, weeks, or months) are not stated categorically, the creators of the Panel used their discretion to determine whether the period of work is reported in days, weeks, or months. As a result, the manner in which the period of work is reported may produce differences in the weightings that are unrelated to actual differences in periods worked. For example, three individuals who all work 11 months of the year could have different weightings depending on the reporting period used (334 days divided by 365, 11 months divided by 12, or 48 weeks divided by 52).

Second, multiple IRP5 forms are available for some workers at the same PAYE-registered entity, due either to revisions of the employee's IRP5 form, or an employee having multiple IRP5s filed because they are hired for different durations of work within the same period. For example, the

---

[14] These other income source codes (in addition to 3601) used to determine labour cost are: 3605, 3606, 3607, 3615, 3616, 3701, 3703, 3707, 3717, 3718, 3801, 3802, 3808, 3809, 3810, 3813, 3814, 3815, 1816, 3820, 3821. For a list of definitions, see the SARS website, available at: http://www.sars.gov.za/TaxTypes/PIT/Tax-Season/Pages/Find-a-Source-Code.aspx.

same individual can show up in the data with four IRP5 forms where each form indicates this individual worked three months of the year. When an individual has more than one IRP5 form, a sum of the number of periods worked is taken; when this sum is greater than one, the highest number of the different periods worked is used as it is believed that this must have been a revised IRP5 that was filed, and when this sum is less than one, the sum is used as the period worked (it is therefore assumed that these are consecutive periods worked and not revisions).[15] The weighted employment information is constructed using the following equations, with (5) corresponding to (1), (6) corresponding to (2), and (7) corresponding to (3).

$$irp5_{forms_{f,t}} = \sum_i Form_{i,f,t} \times weight_{i,f,t} \tag{5}$$

$$irp5\_empl3601_{f,t} = \sum_i Form_{i,f,t} \times (3601\ Income_{i,f,t} > 0) \times weight_{i,f,t} \tag{6}$$

$$irp5\_empl_{f,t} = \sum_i Form_{i,ft} \times (Employment\ Income_{i,f,t} > 0) \times weight_{i,f,t} \tag{7}$$

The 'total periods set' is matched to the CIT panel through the CIT reference number by moving the data to the corresponding year, as described in Section 2.4.

The dates set

The second set of employment measures uses the periods employed to and from to weight workers and assign them to a firm in its financial year. This is to ensure that the number of employees created for each firm better corresponds to the actual reporting period for the income statement and balance sheet information in the CIT data of the firm. Where a person reports employment start (end) dates that fall before (after) the corresponding tax year, the start (end) of that person's period is brought forward (backwards) to match the corresponding tax year for the CIT return. In other words, any form starting at 15 February 2007 for the tax year ending on 28 February 2009, will be altered to report that the person started on 1 March 2008. Then the firm's financial year end is merged into the IRP5 data at an annual level. To clarify this process, we provide Figure 3 for the case where a person started employment on 01/04/2012 in a firm with a financial year starting on 01/01 and ending on 31/12 in a given year. This employee is reported in the IRP5 as having started employment on 01/04/2012 and ended employment on 28/02/2013 for the tax year 2013 and was reported to start on 01/03/2013 and end on 28/02/2014 for the tax year 2014. The person's weighted employment date variable for the tax year 2012 will thus be 275/366. For the 2013 tax year, the person's weighted employment date variable will be the sum of periods D and C, i.e. 365/365.

---

[15] There is currently no indicator of revision in the data and one therefore needs to make the assumption that when periods worked sum to less than one that these are consecutive periods worked. However, it is theoretically possible that these are merely revisions of the IRP5 form submitted for the same individuals. An obvious way to improve the data is to obtain a revisions indicator from SARS and update the IRP5 panel accordingly.

Figure 3: Employment date matching
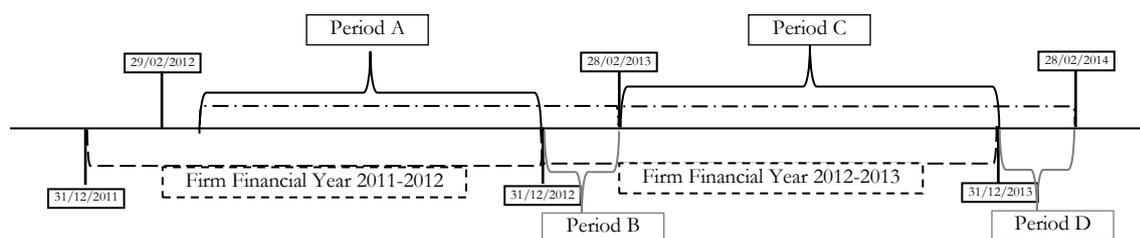
Source: Authors' figure based on SARS-NT Panel cleaning code.

Three data sets are constructed for each IRP5 year, one containing only the data on persons employed within a tax reference number entity year, one containing information on persons' employment data before the start of the tax reference number entity year and one containing information persons' employment data after the end of the tax reference number entity year. A new data set is then constructed by appending the employment data of workers after the previous year, the employment data on persons within the current year and the employment data before the next year into the same data set. From here the earliest employment start date and latest employment end date is assigned to each unique employment spell. The unweighted employment variables are then constructed as in equations (1)–(3) above, the main difference now being that only persons with available date data are included and that the matching procedure places them in what is arguably a more accurate tax reference entity year. The weights are now constructed as:

$$date\_weight_{i,f,t} = \frac{Period\ Employed\ To_{i,f,t} - Period\ Employed\ From_{i,f,t}}{Days\ in\ Year_t} \qquad (8)$$

Comparing the aggregates in the total period and dates sets

In Tables 6 and 7 we report the aggregates of the unweighted and weighted employment measures, respectively. In neither of these tables is the sample restricted to firms with available data; the implications of additional restrictions will be seen in Section 4. In Table 6, we see that the dates set totals to around 10 per cent of the period set in 2009. This result is due to missing date data in 2009. The difference in the aggregates of the unweighted dates to period sets narrows as we restrict the employment measure. In Table 7 we see this result as well. The weighted employment measures are between 64 and 75 per cent of the unweighted measures in general.

14

Table 6: Aggregates of unweighted employment measures

Unweighted	Set	2009	2010	2011	2012	2013	2014
Forms	Period Set	8 254 732	10 362 781	11 623 808	11 730 337	11 740 680	10 252 700
	Dates Set	826 709	8 845 588	9 381 639	9 668 522	10 079 389	8 869 464
	% Dates to Period	(10.01%)	(85.36%)	(80.71%)	(82.42%)	(85.85%)	(86.51%)
Empl	Period Set	7 328 676	9 877 843	11 051 842	11 177 416	10 845 941	9 409 574
	Dates Set	636 211	8 640 199	9 254 054	9 550 786	9 952 394	8 739 590
	% Dates to Period	(8.68%)	(87.47%)	(83.73%)	(85.45%)	(91.76%)	(92.88%)
Empl3601	Period Set	6 902 205	9 418 081	10 545 941	10 632 115	10 293 971	8 917 674
	Dates Set	583 149	8 350 605	8 943 754	9 217 350	9 617 825	8 445 015
	% Dates to Period	(8.45%)	(88.67%)	(84.81%)	(86.69%)	(93.43%)	(94.7%)

Source: Authors' calculations based on SARS-NT Panel.

Table 7: Aggregates of weighted employment measures

Weighted	Set	2009	2010	2011	2012	2013	2014
Forms	Period Set	5 662 848	7 571 684	8 510 044	8 509 723	8 562 579	7 479 914
	Dates Set	533 222	6 449 147	6 941 520	7 031 719	6 786 838	5 871 761
	% Dates to Period	(9.42%)	(85.17%)	(81.57%)	(82.63%)	(79.26%)	(78.5%)
Empl	Period Set	4 970 807	7 133 033	7 980 241	8 004 830	7 741 076	6 691 479
	Dates Set	468 786	6 371 686	6 867 859	6 962 412	6 716 598	5 808 192
	% Dates to Period	(9.43%)	(89.33%)	(86.06%)	(86.98%)	(86.77%)	(86.8%)
Empl3601	Period Set	4 660 259	6 808 911	7 619 487	7 609 874	7 350 016	6 354 152
	Dates Set	423 409	6 179 644	6 667 307	6 768 114	6 523 418	5 633 386
	% Dates to Period	(9.09%)	(90.76%)	(87.5%)	(88.94%)	(88.75%)	(88.66%)

Source: Authors' calculations based on SARS-NT Panel.

The above section compares the employment aggregates, but does not give an indication of the within firm relationship between the dates and periods sets. In this section we show the results of regressions on the employment indicators developed for total periods and dates sets for all firms from 2010–14. For earlier years only the total period indicators are reliably available, but the completeness of fields on the IRP5 forms since 2010 allows the relationship between these indicators for later years to be investigated. In Table 8 we show the regression results of a function as specified in Equation (9) where type is either *forms, empl, empl3601, forms_weight, empl_weight,* or *empl3601_weight.* For example, we conduct a regression of the labour variable generated for *forms* in the periods set on the labour variable generated for *forms* in the dates set and relate it to the firm's tax year. This gives us an idea of the nature of the relationship between the two variables and whether it changes over time.

$$irp5\_TYPE = \sum_{y=2010}^{2014} \beta_y \ taxyear \ y + \sum_{y=2010}^{2014} \beta_{dates,y} \ IRP5\_TYPE\_date \ \times taxyear \ y \qquad (9)$$

Table 8 shows that the hypothesis that the coefficient of the date variable is statistically different from unity can be rejected in all cases. While this result, along with the positive coefficient of the year dummies, can be interpreted as the period set being larger than the dates set in general (as will be seen below), it should be noted that the coefficients of the date variables are still slightly above unity.

In terms of intertemporal differences, we cannot reject the hypothesis that $\beta_{dates,2012} = \beta_{dates,2013}$ in all cases. Further, we cannot reject the hypothesis that $\beta_{dates,2010} = \beta_{dates,2014}$ for unweighted *empl3601*, weighted *empl* and weighted *empl3601*. These results, along with decline in $\beta_{dates,y}$ as y increases indicate that the measures become more consistent over time, with the exception of 2014.

Table 8: Regressions on periods set employment value

Variables	TYPE					
	Forms	empl	empl3601	forms weight	empl weight	empl3601 weight
$\beta_{dates,2010}$	1.089***	1.088***	1.082***	1.054***	1.053***	1.045***
	(0.00155)	(0.00154)	(0.00152)	(0.00170)	(0.00170)	(0.00164)
$\beta_{dates,2011}$	1.065***	1.064***	1.059***	1.034***	1.033***	1.026***
	(0.00128)	(0.00128)	(0.00124)	(0.00169)	(0.00168)	(0.00166)
$\beta_{dates,2012}$	1.060***	1.059***	1.055***	1.028***	1.028***	1.022***
	(0.00123)	(0.00123)	(0.00120)	(0.00154)	(0.00154)	(0.00148)
$\beta_{dates,2013}$	1.058***	1.058***	1.053***	1.026***	1.026***	1.021***
	(0.00119)	(0.00119)	(0.00116)	(0.00150)	(0.00151)	(0.00147)
$\beta_{dates,2014}$	1.041***	1.041***	1.037***	1.050***	1.050***	1.046***
	(0.000932)	(0.000936)	(0.000903)	(0.00196)	(0.00197)	(0.00192)
β_{2010}	0.509***	0.490***	0.0993***	0.691***	0.664***	0.299***
	(0.0215)	(0.0210)	(0.0201)	(0.0179)	(0.0174)	(0.0165)
β_{2011}	0.460***	0.449***	0.0591***	0.723***	0.699***	0.323***
	(0.0186)	(0.0183)	(0.0173)	(0.0188)	(0.0183)	(0.0176)
β_{2012}	0.441***	0.433***	0.0360**	0.765***	0.745***	0.361***
	(0.0177)	(0.0175)	(0.0167)	(0.0178)	(0.0174)	(0.0164)
β_{2013}	0.456***	0.445***	0.0645***	1.233***	1.203***	0.815***
	(0.0180)	(0.0177)	(0.0168)	(0.0200)	(0.0196)	(0.0188)
β_{2014}	0.534***	0.533***	0.154***	0.908***	0.894***	0.517***
	(0.0141)	(0.0140)	(0.0131)	(0.0210)	(0.0207)	(0.0196)
R-Squared	0.955	0.956	0.957	0.931	0.932	0.934
Adjusted R-Squared	0.955	0.956	0.957	0.931	0.932	0.934
Observations	1137376	1137376	1137376	1137376	1137376	1137376
Clusters	314067	314067	314067	314067	314067	314067
P-Value for H_0:						
$\beta_{dates,2010} = 1$	0	0	0	0	0	0
$\beta_{dates,2011} = 1$	0	0	0	0	0	0
$\beta_{dates,2012} = 1$	0	0	0	0	0	0
$\beta_{dates,2013} = 1$	0	0	0	0	0	0
$\beta_{dates,2014} = 1$	0	0	0	0	0	0

$\beta_{dates,2010} = \beta_{dates,2011}$	0	0	0	0	0	0
$\beta_{dates,2010} = \beta_{dates,2012}$	0	0	0	0	0	0
$\beta_{dates,2010} = \beta_{dates,2013}$	0	0	0	0	0	0
$\beta_{dates,2010} = \beta_{dates,2014}$	0	0	0.14	0	0.1	0.91
$\beta_{dates,2011} = \beta_{dates,2012}$	0	0	0	0	0	0.02
$\beta_{dates,2011} = \beta_{dates,2013}$	0	0	0	0	0	0
$\beta_{dates,2011} = \beta_{dates,2014}$	0	0	0	0	0	0
$\beta_{dates,2012} = \beta_{dates,2013}$	0.2	0.1	0.31	0.2	0.21	0.37
$\beta_{dates,2012} = \beta_{dates,2014}$	0	0	0	0	0	0
$\beta_{dates,2013} = \beta_{dates,2014}$	0	0	0	0	0	0

Source: Authors' calculations based on SARS-NT Panel.

Where a researcher is interested in employment data from before 2010, the researcher will be forced to use the periods set because the periods employed to and from data are only well populated after 2009. Where the researcher only requires data after 2010, we leave the choice of appropriate variable to the researcher. In the dates set, the date indicators may be estimated with less precision but are distributed around the correct mean, while the periods set may be measured with more accuracy but dispersed around the wrong mean temporally. It should further be noted, that measurement error is likely to increase when weighted data are used.

2.6 Using the deflators

Four deflators from the South African Reserve Bank are merged into the SARS-NT Panel. The following quarterly deflators are currently available in the SARS-NT Panel at the one-digit (major division) level: (i) gross value added; (ii) gross capital formation; (iii) consumer price index (CPI); and (iv) producer price index (PPI). The base year for the deflators is 2012 and the deflators are merged into the SARS-NT Panel using the tax year of the firm. These deflators do not provide full coverage of all sectors. For example, the gross value added deflator is available for all sectors, while the gross capital formation deflator is not available for agriculture. The CPI and PPI are available for very few sectors at the two-digit level. Unfortunately not all sectors, even at the one-digit (major division) level have disaggregated deflators available for them. Therefore, in addition to the four deflators mentioned above, economy-wide CPI and PPI deflators are also provided. These deflators are merged in for ease of use and researchers are expected to review the deflators and tailor them to suit their specific needs.

3 Data characteristics

3.1 Overview of the data

Table 7 shows the total number of entities available per year, as well as how many of these are matched through the merging process. First, we show the total number of entities per tax year. This number includes CIT-registered entities (or firms), as well as other entities (such as PAYE-registered entities that we are unable to link to a CIT reference number—see Table 1). Second,

we show that 68–79 per cent of these entities are found in the CIT data in each tax year. These are the CIT-registered entities that we consider to be firms in the context of the SARS-NT Panel.

Only 21–23 per cent of the firms in the CIT data have a match to the IRP5 data. In other words, only 21–23 per cent of the firms in the SARS-NT Panel can be matched to a PAYE-registered entity with worker information. More than 50 per cent of firms with matching IRP5 data also have matching VAT data, while around 41 per cent of firms with matching IRP5 data have no matching VAT or customs data. Only around 2 per cent of all firms have matching IRP5, VAT, and customs data. Around 14 per cent of firms in the CIT panel have matching VAT and/or customs data but not matching CIT data, with the majority of these firms having VAT data. Note that the relative incompleteness of CIT returns for 2014 is evident in the linkage rates with customs records, VAT returns, and PAYE information.[16]

The Tax Statistics Bulletin 2015 (SARS-NT 2015) provides an indication of the number of firms active in terms of providing CIT, VAT, and PAYE returns in any year. More CIT returns are made than VAT or PAYE returns, in part due to tax administration requirements. Thus while the number of firms that are matched through the merger process seems low, as shown in Section 4, aggregate variables such as sales and employment compare well to other sources of data (suggesting that the absence of firms not retained is unlikely to meaningfully affect analysis). Researchers should use their discretion to determine how best to further manipulate the SARS-NT Panel to get to a more accurate sense of the number of firms that are useful for firm-level economic analysis.

The second part of Table 9 shows the PAYE, VAT, and customs entities not in CIT—these are entities that have a CIT reference number, but their reference numbers do not match an entity captured in the CIT in a given year. As discussed in Section 3.3, understanding how many PAYE, VAT, and customs entities cannot be linked to CIT is important to gain some insight to the selection issues in the SARS-NT Panel that may be a potential source of bias.

---

[16] As much of the data was extracted prior to the end of 2015, the set of returns for 2014 would not be expected to be complete.

Table 9: Total entities in the SARS-NT Panel per tax year: Merging other data sets

	2008	2009	2010	2011	2012	2013	2014	Total
Total entities in SARS-NT Panel	921 538	937 302	945 417	924 338	900 877	871 798	762 820	6 268 705
Firms in CIT	690 249	688 894	721 199	717 594	705 571	666 586	524 531	4 714 624
Percentage of Firms in SARS-NT Panel	(74.9%)	(73.5%)	(76.28%)	(77.63%)	(78.32%)	(76.46%)	(68.76%)	(75.21%)
CIT with no IRP5, VAT, or customs	422 188	415 572	450 699	458 445	459 266	434 538	343 160	2 983 868
Percentage of Firms in CIT	(61.16%)	(60.32%)	(62.49%)	(63.89%)	(65.09%)	(65.19%)	(65.42%)	(63.29%)
CIT Firms with IRP5	150 800	152 342	155 868	158 279	153 331	150 595	120 242	1 041 457
Percentage of Firms in CIT	(21.85%)	(22.11%)	(21.61%)	(22.06%)	(21.73%)	(22.59%)	(22.92%)	(22.09%)
CIT with IRP5 match and no VAT or customs	68 414	60 416	63 256	65 315	63 582	62 475	50 148	433 606
CIT with IRP5 and VAT but no Customs	78 978	69 350	68 052	67 424	63 846	62 251	50 609	460 510
CIT with IRP5 and Customs but no VAT	1 362	9 505	10 438	10 883	11 039	11 048	8 348	62 623
CIT with IRP5, VAT, and Customs	2 046	13 071	14 122	14 657	14 864	14 821	11 137	84 718
CIT without IRP5, but with VAT and/or customs	117 261	120 980	114 632	100 870	92 974	81 453	61 129	689 299
Percentage of Firms in CIT	(16.99%)	(17.56%)	(15.89%)	(14.06%)	(13.18%)	(12.22%)	(11.65%)	(14.62%)
CIT with VAT match and no IRP5 or customs	116 317	110 485	103 264	90 154	82 169	72 012	54 397	628 798
CIT with Customs and no IRP5 or VAT	472	5 395	6 001	5 849	6 072	5 380	3 785	32 954

CIT with VAT and customs but no IRP5	472	5 100	5 367	4 867	4 733	4 061	2 947	27 547
PAYE, VAT and customs entities not in CIT	231 289	248 408	224 218	206 744	195 306	205 212	238 289	1 554 081
Percentage of Firms in SARS-NT Panel	(25.1%)	(26.5%)	(23.72%)	(22.37%)	(21.68%)	(23.54%)	(31.24%)	(24.79%)
Entities with unmatched IRP5	61 889	65 859	63 374	68 938	73 312	83 883	108 292	525 547
Percentage of Firms not in CIT	(26.76%)	(26.51%)	(28.26%)	(33.34%)	(37.54%)	(40.88%)	(45.45%)	(33.82%)
IRP5 with no CIT, VAT, or customs	34 321	35 652	35 052	39 452	42 452	48 572	60 589	296 090
IRP5 matched with VAT, but no CIT or customs	27 568	27 031	25 329	26 038	26 719	30 017	38 871	201 573
IRP5 matched with customs but no CIT or VAT	0	1 483	1 465	1 634	1 996	2 538	4 114	13 230
IRP5 matched with customs and VAT, but no CIT	0	1 693	1 528	1 814	2 145	2 756	4 718	14 654
Entities with no IRP5	169 400	182 549	160 844	137 806	121 994	121 329	129 997	1 028 534
Percentage of Firms not in CIT	(73.24%)	(73.49%)	(71.74%)	(66.66%)	(62.46%)	(59.12%)	(54.55%)	(66.18%)
VAT with no CIT, IRP5, or customs	169 344	169 303	147 021	123 835	105 818	103 635	110 392	933 963
Customs with no CIT, IRP5, or VAT	43	10 058	10 818	11 157	13 265	14 612	15 899	75 852
VAT matched with customs, but no CIT or IRP5	13	3 188	3 005	2 814	2 911	3 082	3 706	18 719

Source: Authors' calculations based on SARS-NT Panel. Note that the decline in number of VAT matching and non-matching firms over time is likely due to an incorrect extraction as these numbers do not match 'SARS' tax statistics *(*SARS-NT 2015).

Table 10 illustrates the completeness of key variables in the SARS-NT Panel. When we show firms with non-missing sales only, we retain 63–71 per cent of the firms in the SARS-NT Panel. However, in 2008 we lose 97 per cent of the firms. This suggests that there is currently a problem with the 2008 data, which requires further investigation.

Table 10: Total firms per financial year and completeness of key variables

	2008	2009	2010	2011	2012	2013	2014	Total
Firms in the SARS-NT Panel	690249	688894	721199	717594	705571	666586	524531	4714624
with non-missing sales	22526	436019	462830	511640	501599	462973	361246	2758833
Percentage of Previous Category	(3.26%)	(63.29%)	(64.18%)	(71.3%)	(71.09%)	(69.45%)	(68.87%)	(58.52%)
and non-zero sales	7023	314794	323213	317635	308053	296018	240861	1807597
Percentage of Previous Category	(31.18%)	(72.2%)	(69.83%)	(62.08%)	(61.41%)	(63.94%)	(66.68%)	(65.52%)
Percentage of all Firms in CIT	(1.02%)	(45.7%)	(44.82%)	(44.26%)	(43.66%)	(44.41%)	(45.92%)	(38.34%)
and non-missing fixed capital stock	7023	273985	286089	303147	301782	295973	240840	1708839
Percentage of Previous Category	(100%)	(87.04%)	(88.51%)	(95.44%)	(97.96%)	(99.98%)	(99.99%)	(94.54%)
and non-zero fixed capital stock	5609	263596	268511	262145	253393	245251	199962	1498467
Percentage of Previous Category	(79.87%)	(96.21%)	(93.86%)	(86.47%)	(83.97%)	(82.86%)	(83.03%)	(87.69%)
Percentage of all Firms in CIT	(0.81%)	(38.26%)	(37.23%)	(36.53%)	(35.91%)	(36.79%)	(38.12%)	(31.78%)
and non-missing cost of sales	5609	195696	209833	252188	253361	245220	199911	1361818
Percentage of Previous Category	(100%)	(74.24%)	(78.15%)	(96.2%)	(99.99%)	(99.99%)	(99.97%)	(90.88%)
and non-zero cost of sales	3426	177335	181553	178176	172391	163002	132958	1008841
Percentage of Previous Category	(61.08%)	(90.62%)	(86.52%)	(70.65%)	(68.04%)	(66.47%)	(66.51%)	(74.08%)
Percentage of all Firms in CIT	(0.5%)	(25.74%)	(25.17%)	(24.83%)	(24.43%)	(24.45%)	(25.35%)	(21.4%)
and non-missing labour data	1321	101379	104142	105939	102675	100445	82053	597954
Percentage of Previous Category	(38.56%)	(57.17%)	(57.36%)	(59.46%)	(59.56%)	(61.62%)	(61.71%)	(59.27%)
Percentage of all Firms in CIT	(0.19%)	(14.72%)	(14.44%)	(14.76%)	(14.55%)	(15.07%)	(15.64%)	(12.68%)

Source: Authors' calculations based on SARS-NT Panel.

Firms with non-missing and non-zero sales constitute between 43 and 46 per cent of all firms in the SARS-NT Panel, indicating a large number of inactive firms. This is somewhat expected since the total firm numbers in the first row include three categories of CIT-registered entities that are less immediately useful for firm-level economic analysis: (i) dormant companies; (ii) share block companies; and (ii) body corporates.[17] SARS defines a dormant company as a firm that has not actively traded for the full year of assessment (i.e. if the company traded for part of the year of assessment, the company will not be regarded as a dormant company). Dormant

---

[17] A share block company is a company that operates a share block scheme, which is a scheme where a share in the scheme confers a right to the use of or interest in immovable property. A body corporate is a legal entity comprising all owners of units in a property scheme, such as an apartment building.

firms are easily identified in the IT14 and ITR14 forms, which contain questions to identify these firms. A body corporate/share block company indicator is available for the ITR14 data only, so to identify these firms from the IT14 forms, researchers have to use other characteristics of the data. The total firm numbers would also include other CIT-registered entities that are more difficult to identify and may complicate firm-level analysis. The bank nominee companies discussed in Section 2.2 is one example, as these CIT-registered entities report significant assets, but are unlikely to report other financial or tax information.

Between 87 and 100 per cent of firms with non-missing and non-zero sales data also have non-missing fixed capital stock data, with between 82 and 97 per cent of these firms also having non-zero fixed capital stock data.[18] Only 24 to 26 per cent of all firms in the SARS-NT Panel also have non-missing and non-zero cost of sales in addition to non-missing and non-zero sales and capital data. Between 57 and 62 per cent of all firms with non-missing and non-zero sales, cost of sales, and fixed capital data also have non-missing labour data. This figure implies that firms with available sales, capital, cost of sales, and labour data constitute between 14.5 and 15.7 per cent of all firms in the SARS-NT Panel.

3.2 Potential sources of bias caused by selection in the SARS-NT Panel

Although the SARS-NT Panel is closer to the actual population of firms and individuals in the formal economy than comparable sources of survey data, and therefore enables certain economic analysis that was not possible before, there are various sources of selection in the SARS-NT Panel which may lead to bias in estimated results.[19]

Only tax-registered firms and individuals are eligible to be in the SARS-NT Panel. Within that category, and with all the caveats listed above, only firms and individuals that actually submit a tax return will appear in the SARS-NT Panel. This implies that we will very likely not have a great deal of information on firms that may not be registered for tax such as informal firms, very young firms, or very small firms.

Since CIT returns must be completed within 12 months from the firm's financial year end, the 2014 tax year is still likely missing some submissions. Later years of the SARS-NT Panel will likely always under-report the true number of firms in the SARS-NT Panel. In Figure 4 we show the distribution of log real sales over the tax years in the sample. In the figure we see increased bunching around R1 million for both real and nominal sales over time. It is thus unclear how much of this difference is due to changes in tax policies or changes in firm behaviour.[20]

---

[18] Non-missing capital means that the firm reports at least one value for property, plant and equipment, or other fixed assets in their balance sheet.

[19] However, the few hundred thousand firms reported here are significantly more than the approximately 20,000 firms surveyed in the Quarterly Financial Statistics (QFS) and Quarterly Employment Statistics (QES) (see QES StatsSA 2015b; QFS StatsSA 2014c) and therefore provide a better basis for firm-level analysis. The coverage is further enhanced by the inclusion of non-VAT-registered firms, which are typically excluded from the QES and other surveys.

[20] $\ln(1\,000\,000) \approx 13.8$.

Figure 4: Distribution of log nominal and real sales per year
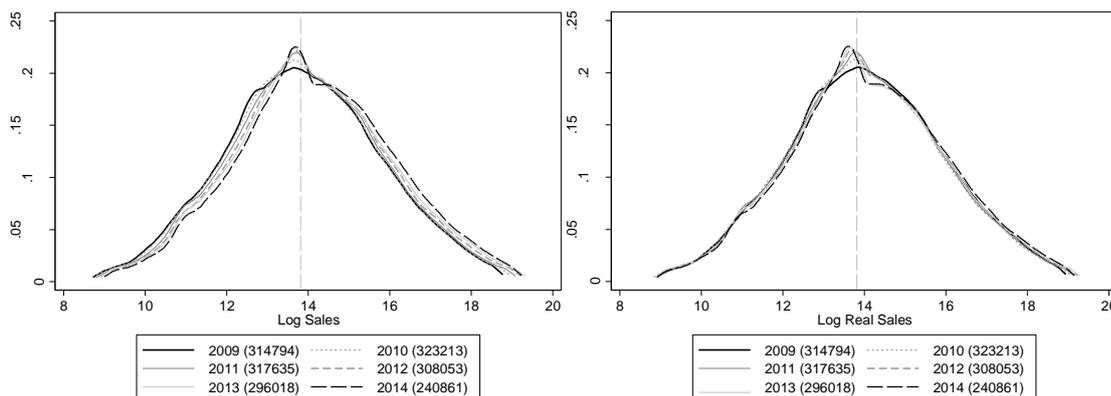
Sources: Authors' calculations based on SARS-NT Panel. We exclude the bottom 1% and top 1% of firms in each year. The reference line on the x-axis is at ln(1000000).

The VAT and customs records are available only for the VAT and trader subsamples, respectively, where there is a selection effect towards larger firms that may induce bias. In panel A of Figure 5 we show that firms registered for VAT are distributed around a higher mean than firms not registered for VAT, as expected given the non-zero threshold (R1 million) in turnover for compulsory registration for VAT. Furthermore, we can clearly see the bunching of firms not registered with sales around R1 million, although a substantial portion of firms earning more than this amount are also not reporting VAT. In panel B, we again observe firms with customs data being more normally dispersed around a dramatically higher mean than non-customs registered firms.

Figure 5: Distribution of sales for VAT and customs registered firms

A   B
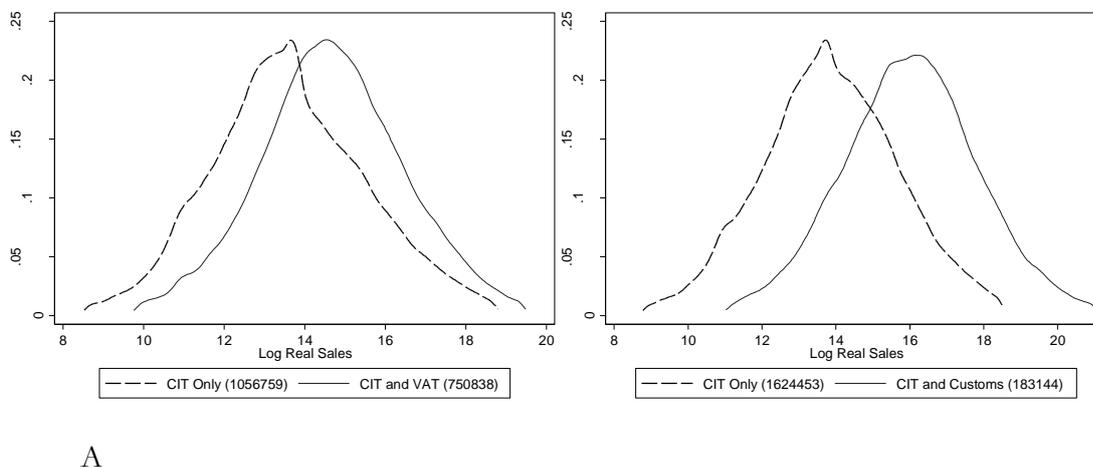
Source: Authors' calculations based on SARS-NT Panel.

In Figure 6, we show that firms with matching IRP5 data are distributed around a higher mean than firms without IRP5 data. In this context, researchers must be careful when thinking of potential sources of bias in the SARS-NT Panel caused by selection.

23

Figure 6: Distribution of sales for IRP5 matching firms
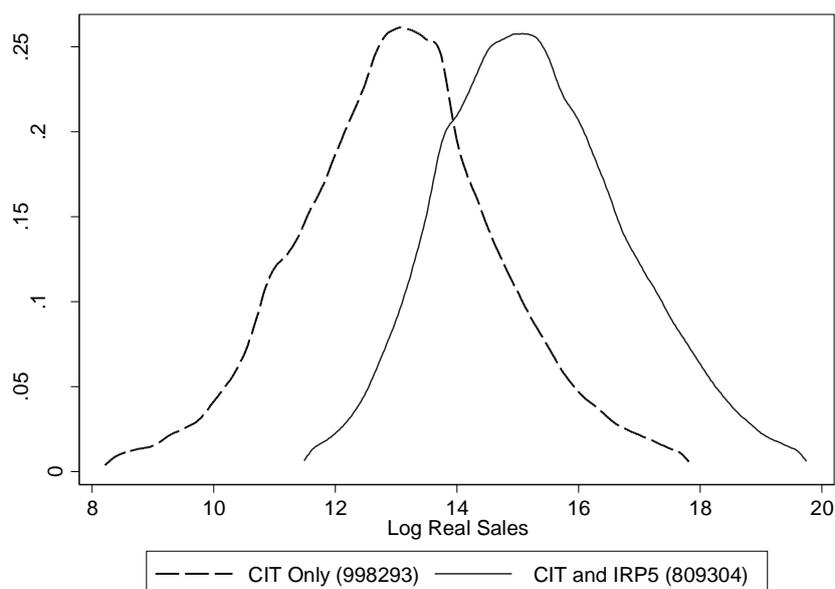
Source: Authors' calculations based on SARS-NT Panel.

With respect to data collection, minor changes to both the IT14 and ITR14 forms are made on a continuous basis due to legislative and other changes. For example, on 6 December 2014, SARS updated the ITR14 return to allow firms to disclose the energy efficiency savings deduction in terms of Section 12L of the Income Tax Act. As a result, the available fields (i.e. variables) change over time and not all variables are available for the entire period. As indicated in Section 2.3, the IT14 form contained the same set of fields for all firms, irrespective of their size. Therefore, a consistent set of data is available for all companies who completed the IT14 but not for the ITR14 which includes questions on the first page that customize the contents of the ITR14 for completion. Finally, while all original data are retained in the SARS-NT Panel, the variables could be measured with error that in a way that is intentional (to minimize tax) or unintentional (based on a misunderstanding of the question or a data entry error).

4   Comparing the Panel to other sources of South African firm data

The SARS-NT Panel contains four potential variables that can be used to classify a firm into a sector: (i) profit code from the CIT data (a SARS-specific industry variable); (ii) industry code from the CIT data drawn from the ITR14 forms (SIC7 format); (iii) main income source code from the IRP5 data (SIC7 format); and (iv) HS6 code for all traders which we aggregate.[21] The HS code is a product, not sector, code, but can be mapped to sector codes using the appropriate concordance table. As expected, the IRP5 industry is available for fewer firms since only firms with matching IRP5 data, at some point, will have had an industry code. A separate technical appendix released with the SARS-NT Panel includes a detailed discussion of how the different sector variables compare to other sources of industry data (see Kreuser et al. 2016).

In this section we compare the sales and employment data by sector to Statistics South Africa's Quarterly Financial Statistics (QFS) (see Appendix I) and Quarterly Employment Statistics

---

[21] Industry codes from the CIT records are available for the ITR14 forms only, not for IT14.

(QES) (see Appendix II) to determine how the SARS-NT Panel compares to other data sources. We show the total sales by sector from the SARS-NT Panel using the industry codes from the IRP5 data set and compare these to total sales data from the QFS. Sales in the SARS-NT Panel refers to the sales amount reported in the income statement of the firm and sales in the QFS includes the value of sales, amounts received for work done, and amounts received for services rendered.

In Table 11 we provide a breakdown of the total nominal sales per industry in the panel against the values reported in the QFS. All values are nominal. Comparing the unrestricted aggregates for sales we observe that the agriculture, construction, wholesale and retail, and community and social service sectors have annual aggregates considerably lower than those reported in the QFS. The mining, manufacturing sector, electricity and finance sectors have higher aggregates in the panel than in the QFS. These differences are in part due to the assignment of different classifications of firms than those used by Statistics South Africa. For example, when reporting to SARS, retailers that trade in manufactured goods may classify their workers as being employed in the manufacturing sector, whereas the QFS allocates the firms where these workers are employed as being in the wholesale or retail trade sectors. Furthermore, we observe that between 14 and 22 per cent of all sales are in firms within unassigned industries, that is, firms which did not report the industry of their employees or had an equal number of employees in their largest industries. Once controlling for firms with attendant data, however, this figure reduces to between 1.5 and 2.5 per cent. These results indicate that a large portion of the unassigned firms are due to missing labour data.

Table 11: Total sales by sector, 2009–14 (R million)

Industry	Data	2009	2010	2011	2012	2013	2014
	No Restrictions	73 925	82 445	98 003	116 927	132 375	111 292
Agriculture	Has All Key Variables	63 127	71 978	91 696	111 245	128 773	96 482
	% of No restriction	(85.39%)	(87.3%)	(93.56%)	(95.14%)	(97.28%)	(86.69%)
	No Restrictions	115 544	349 399	338 687	475 346	485 586	196 065
	% of QFS		(110.41%)	(82.96%)	(102.85%)	(104.46%)	(36.81%)
	Has All Key Variables	103 542	281 667	316 187	461 654	475 501	189 646
Mining	% of No restriction	(89.61%)	(80.61%)	(93.36%)	(97.12%)	(97.92%)	(96.73%)
	% of QFS		(112.36%)	(129.11%)	(100.12%)	(97.76%)	(280.88%)
	QFS		316 470	408 239	462 196	464 842	532 674
	No Restrictions	1 691 229	1 822 079	2 008 004	2 304 378	2 482 885	1 384 159
	% of QFS		(120.8%)	(120.56%)	(121.49%)	(123.66%)	(60.75%)
	Has All Key Variables	1 590 103	1 596 412	1 951 777	2 247 531	2 445 201	1 174 451
Manufacturing	% of No restriction	(94.02%)	(87.61%)	(97.2%)	(97.53%)	(98.48%)	(84.85%)
	% of QFS		(94.48%)	(85.34%)	(84.39%)	(82.11%)	(194%)
	QFS		1 508 332	1 665 577	1 896 778	2 007 836	2 278 403

	No Restrictions	33 869	56 809	148 646	177 116	201 297	172 146
	% of QFS		(63.93%)	(131.35%)	(132.33%)	(135.39%)	(101.18%)
Electricity, Gas, and Water	Has All Key Variables	28 186	41 129	145 359	175 772	200 423	167 774
	% of No restriction	(83.22%)	(72.4%)	(97.79%)	(99.24%)	(99.57%)	(97.46%)
	% of QFS		(216.04%)	(77.85%)	(76.15%)	(74.18%)	(101.41%)
	QFS		88 856	113 166	133 842	148 675	170 142
	No Restrictions	149 847	150 555	152 469	167 928	185 462	153 905
	% of QFS		(59.45%)	(61.32%)	(66.67%)	(70.85%)	(50.17%)
Construction	Has All Key Variables	128 710	133 366	138 811	155 668	177 643	146 761
	% of No restriction	(85.89%)	(88.58%)	(91.04%)	(92.7%)	(95.78%)	(95.36%)
	% of QFS		(189.88%)	(179.13%)	(161.81%)	(147.36%)	(209.04%)
	QFS		253 242	248 652	251 888	261 776	306 785
	No Restrictions	828 203	1 151 355	1 121 369	1 292 088	1 367 016	1 074 757
	% of QFS		(71.43%)	(61.92%)	(61.52%)	(61.32%)	(43.28%)
Wholesale, Retail, Catering, and Accommodation	Has All Key Variables	757 964	1 085 962	1 081 191	1 239 057	1 346 173	1 014 009
	% of No restriction	(91.52%)	(94.32%)	(96.42%)	(95.9%)	(98.48%)	(94.35%)
	% of QFS		(148.44%)	(167.51%)	(169.5%)	(165.6%)	(244.87%)
	QFS		1 611 949	1 811 128	2 100 242	2 229 219	2 483 012
	No Restrictions	191 072	433 906	479 812	534 273	538 980	394 693
	% of QFS		(100.36%)	(103.28%)	(100.14%)	(92.59%)	(60.85%)
Transport, Storage, and Communication	Has All Key Variables	130 485	320 622	442 418	504 685	534 337	362 770
	% of No restriction	(68.29%)	(73.89%)	(92.21%)	(94.46%)	(99.14%)	(91.91%)
	% of QFS		(134.85%)	(105.01%)	(105.71%)	(108.94%)	(178.79%)
	QFS		432 360	464 587	533 515	582 122	648 586
	No Restrictions	504 498	768 342	768 099	880 637	908 585	630 293
	% of QFS		(160.98%)	(164.01%)	(166.33%)	(148.22%)	(83.42%)
Financing, Insurance Real Estate, and Business Services	Has All Key Variables	386 747	582 170	712 792	835 517	875 861	567 386
	% of No restriction	(76.66%)	(75.77%)	(92.8%)	(94.88%)	(96.4%)	(90.02%)
	% of QFS		(81.99%)	(65.7%)	(63.37%)	(69.99%)	(133.16%)
	QFS		477 303	468 311	529 456	612 986	755 544

Community and Social Services	No Restrictions	78 826	104 202	107 315	118 780	120 653	93 142
	% of QFS		(88.13%)	(72.72%)	(71.87%)	(68.73%)	(51.12%)
	Has All Key Variables	56 143	81 091	98 490	113 353	117 993	80 645
	% of No restriction	(71.22%)	(77.82%)	(91.78%)	(95.43%)	(97.8%)	(86.58%)
	% of QFS		(145.8%)	(149.85%)	(145.81%)	(148.77%)	(225.92%)
	QFS		118 232	147 583	165 277	175 540	182 193
Unassigned Industry	No Restrictions	745 620	1 459 954	1 211 168	1 133 644	1 052 646	871 156
	% of All No Restricted		(22.89%)	(18.83%)	(15.74%)	(14.08%)	(17.14%)
	Has All Key Variables	78 255	64 435	81 330	98 558	116 738	97 971
	% of No Restriction		(4.41%)	(1.56%)	(1.62%)	(1.82%)	(2.33%)
	% All Key Variables		(1.51%)	(1.61%)	(1.66%)	(1.82%)	(2.51%)
Economy	No Restrictions	4 412 633	6 379 046	6 433 572	7 201 117	7 475 485	5 081 608
	Has All Key Variables	3 323 262	4 258 832	5 060 051	5 943 040	6 418 643	3 897 895
	No Restrictions Assigned Industries	3 667 013	4 919 092	5 222 404	6 067 473	6 422 839	4 210 452
	Has All Key Variables Assigned Industries	3 245 007	4 194 397	4 978 721	5 844 482	6 301 905	3 799 924
	No Restrictions – QFS Industries	3 593 088	4 836 646	5 124 401	5 950 546	6 290 464	4 099 160
	% of QFS		(100.62%)	(96.19%)	(97.98%)	(97.03%)	(55.72%)
	Has All Key Variables – QFS Industries	3 137 494	4 029 867	4 532 003	5 048 131	6 173 132	3 703 442
	% of No restriction	(87.32%)	(83.32%)	(88.44%)	(84.83%)	(98.13%)	(90.35%)
	% of QFS		(83.84%)	(85.07%)	(83.12%)	(95.22%)	(50.34%)
	QFS		4 806 744	5 327 243	6 073 194	6 482 996	7 357 339

Source: Authors' calculations based on SARS-NT Panel and QFS. Both QFS and SARS-NT Panel figures for community and social services exclude government institutions.

Furthermore, the sudden decline in values for 2014 is further evidence that there are still outstanding forms in this year. We show that once restricting our sample to firms with all key variables we only lose around 16 per cent of total sales in the panel. In terms of total sales values we are capturing between 97 and 101 per cent of the values for the economy. When restricting our panel to firms with key variables, we have between 80 and 86 per cent of all sales in the QFS. It should be noted, however, that the financial services sector is not directly comparable since the QFS excludes firms in the financial intermediation, insurance, and business services, subsectors not classified elsewhere. Thus, our estimates are likely over-counting representation against the economy-wide aggregates.

Table 12: Total employment by sector, 2009–14

Industry	Data	2009	2010	2011	2012	2013	2014
Agriculture	No Restrictions	200 391	274 785	312 063	327 948	351 285	317 352
	Has All Key Variables	114 313	161 894	221 147	233 720	251 301	174 969
	% of No restriction	(57.04%)	(58.92%)	(70.87%)	(71.27%)	(71.54%)	(55.13%)
Mining	No Restrictions	344 687	500 557	566 959	591 440	557 842	346 610
	% of QES	(66.61%)	(102.26%)	(112.44%)	(113.9%)	(107.17%)	(69.05%)
	Has All Key Variables	186 884	405 324	431 036	522 572	478 448	197 901
	% of No restriction	(54.22%)	(80.97%)	(76.03%)	(88.36%)	(85.77%)	(57.1%)
	% of QES	(36.11%)	(82.8%)	(85.48%)	(100.64%)	(91.92%)	(39.42%)
	QES	517 500	489 500	504 250	519 250	520 500	502 000
Manufacturing	No Restrictions	1 432 349	1 925 470	2 049 257	2 105 915	2 176 483	1 752 840
	% of QES	(111.88%)	(160.59%)	(176.17%)	(182.69%)	(189.55%)	(151.2%)
	Has All Key Variables	940 640	1 528 518	1 691 440	1 738 020	1 770 285	1 125 972
	% of No restriction	(65.67%)	(79.38%)	(82.54%)	(82.53%)	(81.34%)	(64.24%)
	% of QES	(73.47%)	(127.48%)	(145.41%)	(150.77%)	(154.17%)	(97.13%)
	QES	1 280 250	1 199 000	1 163 250	1 152 750	1 148 250	1 159 250
Electricity, Gas, and Water	No Restrictions	94 547	123 802	127 897	163 255	137 089	126 226
	% of QES	(159.57%)	(218.15%)	(221.47%)	(273.23%)	(221.11%)	(213.94%)
	Has All Key Variables	37 180	54 670	97 791	135 132	118 538	95 292
	% of No restriction	(39.32%)	(44.16%)	(76.46%)	(82.77%)	(86.47%)	(75.49%)
	% of QES	(62.75%)	(96.33%)	(169.34%)	(226.16%)	(191.19%)	(161.51%)
	QES	59 250	56 750	57 750	59 750	62 000	59 000
Construction	No Restrictions	316 911	413 843	407 502	412 355	449 521	453 086
	% of QES	(67.68%)	(97.84%)	(99.94%)	(96.46%)	(104.18%)	(89.45%)
	Has All Key Variables	231 545	309 749	318 620	334 482	357 685	305 784
	% of No restriction	(73.06%)	(74.85%)	(78.19%)	(81.12%)	(79.57%)	(67.49%)
	% of QES	(49.45%)	(73.23%)	(78.14%)	(78.24%)	(82.89%)	(60.37%)
	QES	468 250	423 000	407 750	427 500	431 500	506 500

Wholesale, Retail, Catering, and Accommodation	No Restrictions	1 069 305	1 528 882	1 658 675	1 694 316	1 794 771	1 730 760
	% of QES	(62.32%)	(92.67%)	(100.25%)	(101.06%)	(105.93%)	(94.17%)
	Has All Key Variables	684 841	1 137 080	1 313 354	1 334 082	1 362 697	1 100 930
	% of No restriction	(64.05%)	(74.37%)	(79.18%)	(78.74%)	(75.93%)	(63.61%)
	% of QES	(39.91%)	(68.92%)	(79.38%)	(79.58%)	(80.43%)	(59.9%)
	QES	1 715 750	1 649 750	1 654 500	1 676 500	1 694 250	1 838 000
Transport, Storage and Communication	No Restrictions	288 716	380 143	440 607	432 424	492 897	436 840
	% of QES	(79.21%)	(106.56%)	(122.73%)	(118.47%)	(130.66%)	(97.18%)
	Has All Key Variables	111 071	265 954	364 874	385 473	443 160	349 433
	% of No restriction	(38.47%)	(69.96%)	(82.81%)	(89.14%)	(89.91%)	(79.99%)
	% of QES	(30.47%)	(74.55%)	(101.64%)	(105.61%)	(117.47%)	(77.74%)
	QES	364 500	356 750	359 000	365 000	377 250	449 500
Financing, Insurance Real Estate, and Business Services	No Restrictions	2 410 765	3 481 601	4 081 309	3 986 049	3 373 953	2 849 933
	% of QES	(126.75%)	(194.88%)	(227.94%)	(217.97%)	(183.54%)	(142.44%)
	Has All Key Variables	1 134 222	2 025 641	2 677 921	2 851 740	2 652 820	1 809 945
	% of No restriction	(47.05%)	(58.18%)	(65.61%)	(71.54%)	(78.63%)	(63.51%)
	% of QES	(59.63%)	(113.39%)	(149.56%)	(155.94%)	(144.31%)	(90.46%)
	QES	1 902 000	1 786 500	1 790 500	1 828 750	1 838 250	2 000 750
Community and Social Services	No Restrictions	500 535	766 279	838 086	842 532	855 891	837 526
	% of QES	(23.4%)	(34.87%)	(36.92%)	(36.23%)	(36.03%)	(33.91%)
	Has All Key Variables	163 162	258 349	334 743	350 844	370 612	242 996
	% of No restriction	(32.6%)	(33.71%)	(39.94%)	(41.64%)	(43.3%)	(29.01%)
	% of QES	(7.63%)	(11.76%)	(14.75%)	(15.09%)	(15.6%)	(9.84%)
	QES	2 138 750	2 197 250	2 269 750	2 325 500	2 375 500	2 469 500
Unassigned Industry	No Restrictions	670 470	482 481	569 487	621 182	656 209	558 401
	% of All No Restricted	(9.15%)	(4.88%)	(5.15%)	(5.56%)	(6.05%)	(5.93%)
	Has All Key Variables	196 551	223 556	347 479	424 238	444 946	301 327
	% of No Restriction	(29.32%)	(46.33%)	(61.02%)	(68.3%)	(67.81%)	(53.96%)
	% All Key Variables	(5.17%)	(3.51%)	(4.46%)	(5.1%)	(5.39%)	(5.28%)

	No Restrictions	7 328 676	9 877 843	11 051 842	11 177 416	10 845 941	9 409 574
	Has All Key Variables	3 800 409	6 370 735	7 798 405	8 310 303	8 250 492	5 704 549
	No Restrictions Assigned Industries	6 658 206	9 395 362	10 482 355	10 556 234	10 189 732	8 851 173
	Has All Key Variables Assigned Industries	3 603 858	6 147 179	7 450 926	7 886 065	7 805 546	5 403 222
Economy	No Restrictions – QES Industries	6 457 815	9 120 577	10 170 292	10 228 286	9 838 447	8 533 821
	% of QES	(76.46%)	(111.79%)	(123.93%)	(122.42%)	(116.47%)	(94.98%)
	Has All Key Variables – QES Industries	3 489 545	5 985 285	7 229 779	7 652 345	7 554 245	5 228 253
	% of No restriction	(54.04%)	(65.62%)	(71.09%)	(74.82%)	(76.78%)	(61.27%)
	% of QES	(41.31%)	(73.36%)	(88.1%)	(91.59%)	(89.43%)	(58.19%)
	QES	8 446 250	8 158 500	8 206 750	8 355 000	8 447 500	8 984 500
	No Restrictions – QES Industries	5 957 280	8 354 298	9 332 206	9 385 754	8 982 556	7 696 295
	% of QES	(94.45%)	(140.14%)	(157.19%)	(155.66%)	(147.93%)	(118.13%)
Economy Excluding Community	Has All Key Variables	3 326 383	5 726 936	6 895 036	7 301 501	7 183 633	4 985 257
	% of No restriction	(55.84%)	(68.55%)	(73.88%)	(77.79%)	(79.97%)	(64.77%)
	% of QES	(52.74%)	(96.07%)	(116.14%)	(121.1%)	(118.31%)	(76.52%)
	QES	6 307 500	5 961 250	5 937 000	6 029 500	6 072 000	6 515 000

Source: Authors' calculations based on SARS-NT Panel and QES. QES industries exclude agriculture and the QES figures for community and social services will include employment in government, whereas only employers with CIT data are included in the SARS-NT Panel.

Table 12 compares the total unweighted employment variable, as defined by Equation (3), per sector against Statistics South Africa's QES. The QES is the only representative employment survey at the firm level and therefore more appropriate than the Quarterly Labour Force Survey (QLFS) for comparison to the SARS-NT Panel.[22] We show the employment number both for the full set of firms as well as for a restricted set of firms with all key variables. We observe similar trends in over- and underestimation relative to the QES estimates, with the construction, wholesale and retail, and community and social services sectors being under-represented. We do, however, observe dramatic over-representation for manufacturing; electricity, gas, and water; transport and finance. Excluding the community and social services sector shows that the unrestricted QES data dramatically understate the total employment figures in the economy by

---

[22] Despite the QES being the most appropriate comparator to the SARS-NT Panel, there are other issues with the QES to keep in mind. QES employment numbers are significantly lower than formal sector employment in the QLFS (the QLFS numbers are a little higher than those reported in the SARS-NT Panel, which is expected given that some non-compliance is reflected in the tax statistics). The QES also omits private sector employment by non-VAT registered employers (QES StatsSA 2015b). While the contribution of these firms to value added may not be large, the relative contribution to employment is larger.

between 10 and 40 per cent. The restricted sample is capturing between 79 and 85 per cent of the QES sample, when including community and social services. When excluding this industry we see a higher count of workers in general. Furthermore, we display the number of workers in an unassigned industry, that is, workers in a firm which did not report their industry in the IRP5 forms, or firms with equal employees in different industries. These firms are around 5 per cent of the total employees in most cases, with between 29 and 69 per cent of these workers being in firms with available data. Since there is also an industry measure in the VAT sample, further work is required to determine how best to use the various industry measures in the SARS-NT Panel to create a comparable, consistent, and accurate sector classification measure for all firms in the SARS-NT Panel.

5   Initial descriptive statistics

In the remainder of the paper we provide some initial descriptive statistics of the firms in the SARS-NT Panel by looking at the relationship between sales and employment, value added, and the capital–labour ratio, as well as value added per worker and trading status. In most figures we demean (i.e. the mean of the variable for all firms for a specific year is subtracted from the individual firm values, in other words $x_{it} - E[x_{it}|t]$) to make the them comparable over time (see Equation (10) in Section 5.2). We focus on the period from 2009–14, given the concerns identified with the 2008 data in Table 8. Where applicable, we deflate all values using the economy-wide CPI values.

5.1   Sales and employment

In Figure 7 we show the relationship between sales and weighted employment from 2009−14.

Figure 7: Distribution of sales by weighted employment, 2009−14
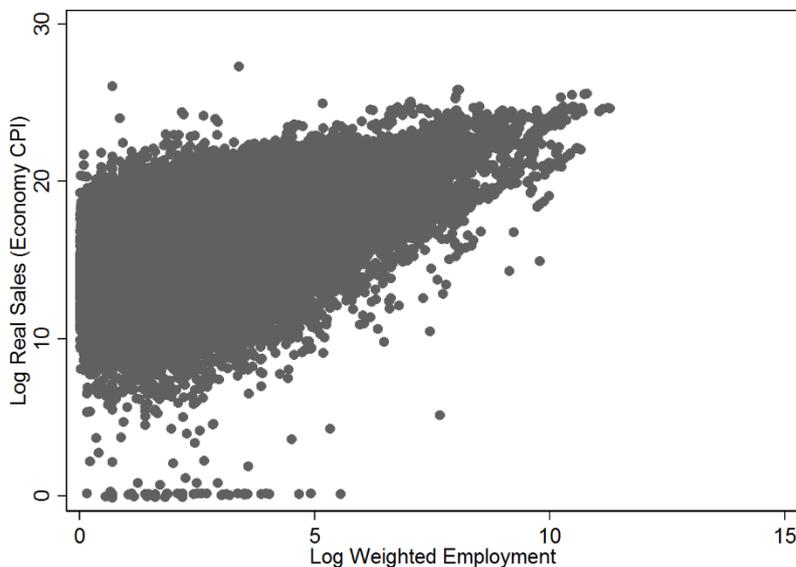
Source: Authors' calculations based on SARS-NT Panel. Only firms with more than 1 weighted employee are included in the figure.

Figure 7 shows a positive correlation between firm size and sales, and indicates that as employment increases, the variance in sales decreases. It is also clear that there are fewer larger

firms and the distribution tightens as employment increases, which may suggest a stronger relationship between sales and employment for larger firms.

5.2 Value added and the capital–labour ratio

Figure 8 provides insight into the relationship between the capital–labour ratio and firm value added. In this figure demeaned log value added (y) is constructed using Equation (10) and demeaned log capital ratio (x) is constructed using Equation (11).

$$y_{i,t} = Log\big(Real\ Sales_{i,t} - Real\ Cost\ of\ Sales_{i,t}\big) - E[Log\big(Real\ Sales_{i,t} - Real\ Cost\ of\ Sales_{i,t}\big)|t] \qquad (10)$$

$$x_{i,t} = Log\left(\frac{Real\ Fixed\ Capital\ Stock}{irp5\_empl_{i,t}}\right) - E[Log\left(\frac{Real\ Fixed\ Capital\ Stock}{irp5_{empl_{i,t}}}\right)|t] \qquad (11)$$

Figure 8: Value added per worker and the capital–labour ratio
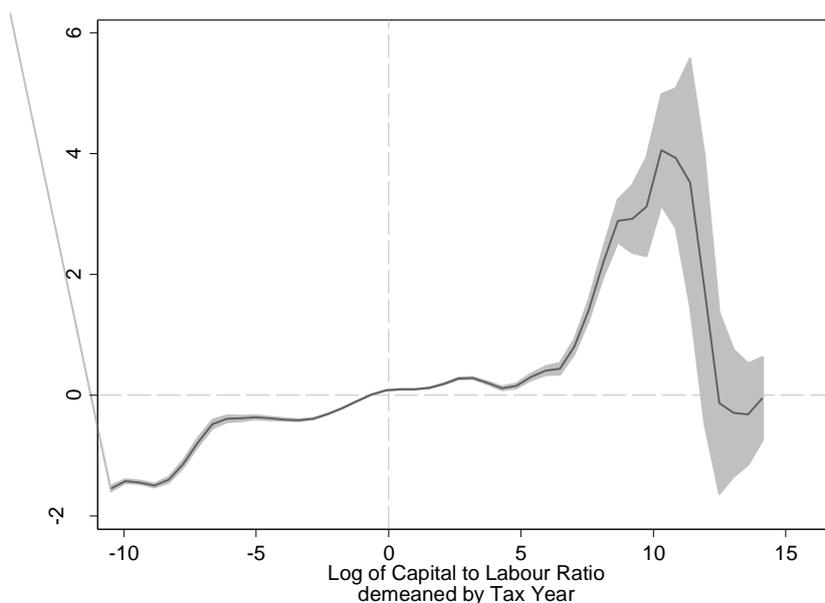
Source: Authors' calculations based on SARS-NT Panel. We only report results for firms with positive and non-missing log capital–labour ratios and firms with one or more weighted employee. Firms must also have positive value added. The mean of each tax year is calculated conditional on both the log capital–labour ratio and log value added information for firms being available. The grey area indicates a confidence level of 95%.

Figure 8 shows the steady increase in the value added by the firm as it becomes more capital intensive. A dramatic increase in the firm's value added is observed where the firm is 1,000 times more capital intensive than the average firm. This increase is not monotonic, however, as the relationship declines sharply where the firm becomes too capital intensive. The confidence level increases dramatically as the capital-to-labour ratio increases. This means that our confidence in these estimates reduces—something that is likely driven by a reduction in the number of firms.

5.3 Value added per worker and international trade

Figure 9 provides some insight into the customs data contained in the SARS-NT Panel. In panels A and B we show the relationship between the demeaned log of value added per worker and the number of different products exported and imported, respectively. The y-axis is calculated using Equation 12 to make the values comparable to firms not included in the customs data.

$$y_{i,t} = Log\left(\frac{Real\ Sales_{i,t} - Real\ Cost\ of\ Sales_{i,t}}{irp5\_empl_{i,t}}\right) - E\left[Log\left(\frac{Real\ Sales_{i,t} - Real\ Cost\ of\ Sales_{i,t}}{irp5_{empl_{i,t}}}\right)|t\right] \qquad (12)$$

In panel A we see that the value added per worker is increasing in the number of different products exported by the firm at the HS6 level. In panel B, we show that firms have larger increasing returns to value added per worker up to around 30 products imported, after which the value is escalating with increasing variance. The constants in these figures should be read in conjunction with the fact that a firm which exports no products imports at least one product, and a firm which imports no products is exporting at least one product. This implies that firms that import at least one product have a far greater premium than firms which export at least one product. In panel C, we show a fairly consistent increase in value added per worker as firms export to more countries, while panel D indicates a far more dramatic rise in value added per worker per import country.

Figure 9: Value added per worker and international trade

A   B

C   D
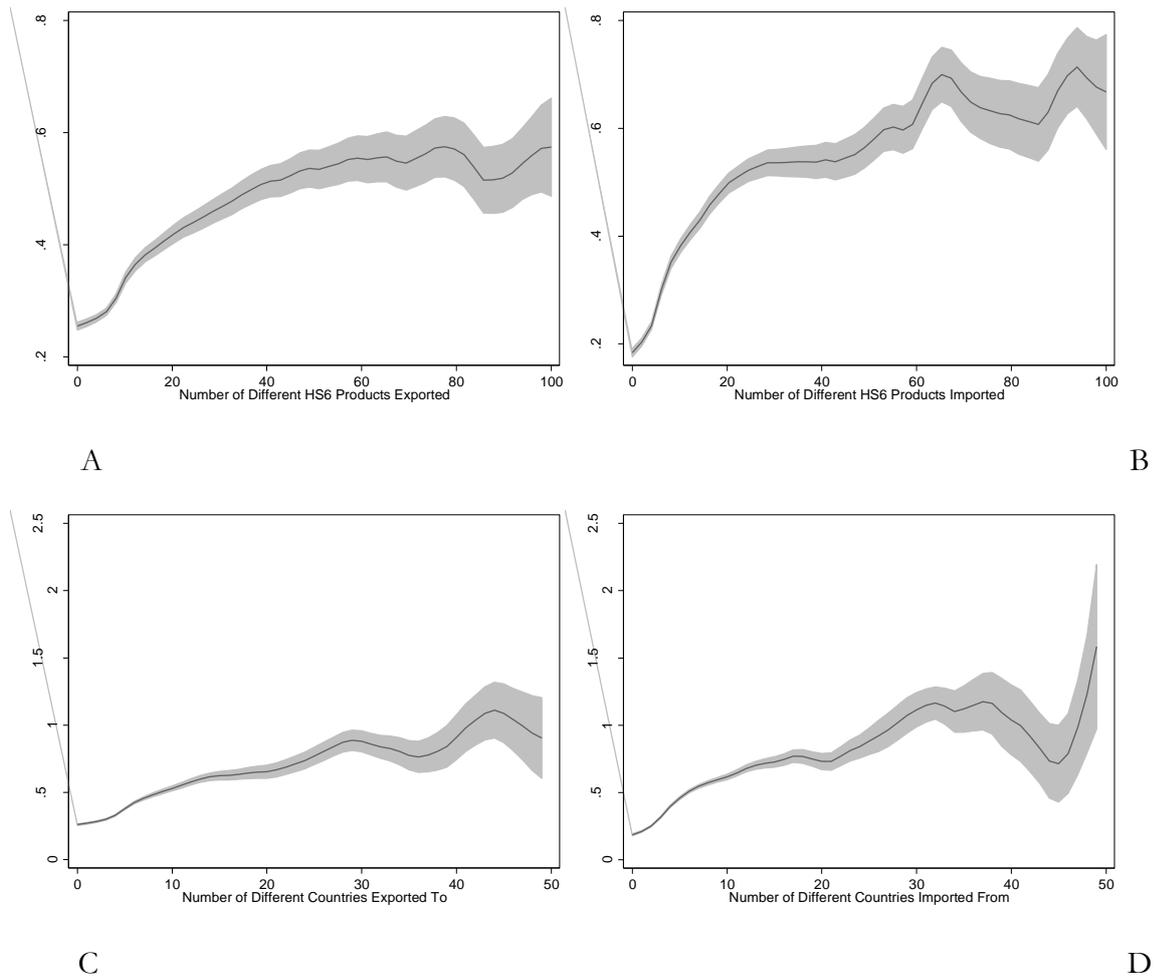
Source: Authors' *calculations based on SARS-NT Panel. All panels restrict the sample to firms with positive log value added per worker and firms with more than one weighted worker. The mean of value added per worker is calculated for all firms within a tax year for which the log value added per worker is positive and non-missing. Panels A and B display results for firms importing or exporting up to 100 different HS6 products. This figure constitutes around 98% of the trading sample. Panels C and D display results for firms importing from or exporting to up to 50 different countries only. This figure constitutes around 99% of trading firms. The grey area indicates a confidence level of 95%.*

References

Barnes, H., M. Noble, C. Dibben, C. Meth, G. Wright, L. and Cluver (2007). 'South Africa Microdata Scoping Study'. Working Paper 6. Oxford: Centre for the Analysis of South African Social Policy, University of Oxford.

Card, D., R. Chetty, M. Feldstein, and E. Saez (2010). 'Expanding Access to Administrative Data for Research in the United States'. Paper submitted to the National Science Foundation as part of its SBE 2020 planning activity. Available at: www.nsf.gov/sbe/sbe_2020/ (accessed 15 January 2016).

Kreuser, F., W. Flowerday, T. and Naughtin-Webb (2016). 'South African Revenue Service and National Treasury Firm-Level Panel: A Very Short Introduction to the Panel for Users'.

South African Revenue Service and National Treasury (SARS-NT) (2015). Tax Statistics, November 2015. Available at: http://www.treasury.gov.za/publications/tax%20statistics/2015/TStats%202015%20Inside%20WEB.pdf (accessed 15 January 2016).

Appendix I Quarterly Financial Statistics Data Sources (latest first)

StatsSA 2015a. P0044 – Quarterly Financial Statistics, December 2014. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=6092 [15/12/2015].

StatsSA 2014a. P0044 – Quarterly Financial Statistics, September 2014. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5972 [15/12/2015].

StatsSA 2014b. P0044 – Quarterly Financial Statistics, June 2014. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5911 [15/12/2015].

StatsSA 2014c. P0044 – Quarterly Financial Statistics, March 2014. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5844 [15/12/2015].

StatsSA 2014d. P0044 – Quarterly Financial Statistics, December 2013. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5790 [15/12/2015].

StatsSA 2013a. P0044 – Quarterly Financial Statistics, September 2013. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5714 [15/12/2015].

StatsSA 2013b. P0044 – Quarterly Financial Statistics, June 2013. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5633 [15/12/2015].

StatsSA 2013c. P0044 – Quarterly Financial Statistics, March 2013. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5558 [15/12/2015].

StatsSA 2013d. P0044 – Quarterly Financial Statistics, December 2012. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5491 [15/12/2015].

StatsSA 2012a. P0044 – Quarterly Financial Statistics, September 2012. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5414 [15/12/2015].

StatsSA 2012b. P0044 – Quarterly Financial Statistics, June 2012. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5342 [15/12/2015].

StatsSA 2012c. P0044 – Quarterly Financial Statistics, March 2012. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5193 [15/12/2015].

StatsSA 2012d. P0044 – Quarterly Financial Statistics, December 2011. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5491 [15/12/2015].

StatsSA 2011a. P0044 – Quarterly Financial Statistics, September 2011. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5132 [15/12/2015].

StatsSA 2011b. P0044 – Quarterly Financial Statistics, June 2011. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=5041 [15/12/2015].

StatsSA 2011c. P0044 – Quarterly Financial Statistics, March 2011. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=4951 [15/12/2015].

StatsSA 2011d. P0044 – Quarterly Financial Statistics, December 2010. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=4889 [15/12/2015].

StatsSA 2010a. P0044 – Quarterly Financial Statistics, September 2010. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=4830 [15/12/2015].

StatsSA 2010b. P0044 – Quarterly Financial Statistics, June 2010. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=4753 [15/12/2015].

StatsSA 2010c. P0044 – Quarterly Financial Statistics, March 2010. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0044&SCH=4685 [15/12/2015].

Appendix II Quarterly Employment Statistics Sources (latest first)

StatsSA 2015b. P0277 – Quarterly Employment Statistics (QES), December 2014. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=6268 [15/12/2015].

StatsSA 2014e. P0277 – Quarterly Employment Statistics (QES), September 2014. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5977 [15/12/2015].

StatsSA 2014f. P0277 – Quarterly Employment Statistics (QES), June 2014. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5915 [15/12/2015].

StatsSA 2014g. P0277 – Quarterly Employment Statistics (QES), March 2014. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5849 [15/12/2015].

StatsSA 2014h. P0277 – Quarterly Employment Statistics (QES), December 2013. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5770 [15/12/2015].

StatsSA 2013e. P0277 – Quarterly Employment Statistics (QES), September 2013. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5707 [15/12/2015].

StatsSA 2013f. P0277 – Quarterly Employment Statistics (QES), June 2013. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5617 [15/12/2015].

StatsSA 2013g. P0277 – Quarterly Employment Statistics (QES), March 2013. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5544 [15/12/2015].

StatsSA 2013h. P0277 – Quarterly Employment Statistics (QES), December 2012. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5483 [15/12/2015].

StatsSA 2012e. P0277 – Quarterly Employment Statistics (QES), September 2012. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5411 [15/12/2015].

StatsSA 2012f. P0277 – Quarterly Employment Statistics (QES), June 2012. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5327 [15/12/2015].

StatsSA 2012g. P0277 – Quarterly Employment Statistics (QES), March 2012. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5264 [15/12/2015].

StatsSA 2012h. P0277 – Quarterly Employment Statistics (QES), December 2011. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5189 [15/12/2015].

StatsSA 2011e. P0277 – Quarterly Employment Statistics (QES), September 2011. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5119 [15/12/2015].

StatsSA 2011f. P0277 – Quarterly Employment Statistics (QES), June 2011. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=5030 [15/12/2015]

StatsSA 2011g. P0277 – Quarterly Employment Statistics (QES), March 2011. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=4939 [15/12/2015].

StatsSA 2011h. P0277 – Quarterly Employment Statistics (QES), December 2010. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=4877 [15/12/2015].

StatsSA 2010d. P0277 – Quarterly Employment Statistics (QES), September 2010. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=4826 [15/12/2015].

StatsSA 2010e. P0277 – Quarterly Employment Statistics, 2nd Quarter 2010. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=4737 [15/12/2015]

StatsSA 2010f. P0277 – Quarterly Employment Statistics (QES), 1st Quarter 2010. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=4677 [15/12/2015].

StatsSA 2010g. P0277 – Quarterly Employment Statistics (QES), December 2009. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=4606 [15/12/2015].

StatsSA 2009a. P0277 – Quarterly Employment Statistics (QES), September 2009. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=4553 [15/12/2015].

StatsSA 2009b. P0277 – Quarterly Employment Statistics (QES), June 20109. [Online] Available: http://www.statssa.gov.za/?page_id=1854&PPN=P0277&SCH=4475 [15/12/2015].