Ebrahim, Amina; Axelson, Chris

Working Paper

The creation of an individual panel using administrative tax microdata in South Africa

WIDER Working Paper, No. 2019/27

Provided in Cooperation with:
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: Ebrahim, Amina; Axelson, Chris (2019) : The creation of an individual panel using administrative tax microdata in South Africa, WIDER Working Paper, No. 2019/27, ISBN 978-92-9256-661-6, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki, https://doi.org/10.35188/UNU-WIDER/2019/661-6

This Version is available at:
https://hdl.handle.net/10419/211257

UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2019/27

The creation of an individual panel using administrative tax microdata in South Africa

Amina Ebrahim[1] and Chris Axelson[2]

March 2019

United Nations University World Institute for Development Economics Research
wider.unu.edu

Abstract: The availability of anonymized individual tax return data can contribute to a deeper understanding of the drivers behind the high levels of inequality and unemployment in South Africa. In the recent past, researchers have examined either payroll or personal income tax data. This paper explains the methodology behind the creation of an anonymized individual panel combining these data to produce a more complete overview of formal employment in South Africa. This paper provides a descriptive overview of the characteristics of the panel, notes some of the pitfalls in the data, and offers some ideas for research to highlight some of the potential.

Keywords: tax data, administrative data, individual panel
JEL classification: C81

Acknowledgements: We acknowledge excellent research assistance from Charl van Schoor. We also acknowledge assistance from Allan Davids with the geocoding of the IRP5 data.

[1]UNU-WIDER, Pretoria, South Africa and PhD Candidate, University of Cape Town, South Africa, corresponding author: amina@wider.unu.edu; [2]Economic Tax Analysis Unit, South African National Treasury, Pretoria, South Africa.

This study has been prepared within the UNU-WIDER project on 'Southern Africa – Towards Inclusive Economic Development (SA-TIED)'.

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9256-661-6

Typescript prepared by Lesley Ellen.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

South Africa has one of the highest levels of inequality in the world, due in part to structurally high levels of unemployment since the advent of democracy alongside wide disparities in wage income. Research on the trends in inequality and poverty, such as that by Leibbrandt et al. (2010) and Sulla and Zikhali (2018), illustrate that although poverty levels have decreased, inequality has generally increased since 1994. This paper describes the creation of a new individual panel using administration tax microdata (henceforth tax data), which can be used to provide greater insight into inequality and employment in South Africa.

The use of tax data for empirical economic research is gaining popularity internationally, as shown by Card et al. (2010). Tax data is both large in terms of the number of observations (hosting the full population of taxpayers) and multi-dimensional, fitting part of the definition of 'big data'. Connelly et al. (2016) suggest that tax data is a form of big data and can be seen as the 'big data' in the field of economics. Tax data is collected for the purpose of determining the tax liability of registered taxpayers, and not specifically for research. The data is typically unstructured, and in the South African case exists with little metadata. Even with similar challenges, research using tax data to evaluate public policy has become widespread in Europe and the USA (Card et al. 2010).

Administrative tax records are attractive to researchers as they are believed to host more accurate records, such as income information for individuals. At the same time, tax data often lacks any demographic information about individuals. Survey data on the other hand covers a wider range of demographic information but suffers from non-response and insufficient coverage of individuals, for example at the top end of the income distribution. In an ideal world we would be able to match the tax and survey data and take advantage of the strengths of both types of data. This is currently not possible as the tax data is anonymized and the legal framework around matching data is underdeveloped.

The regular collection of administrative records mean that administrative data are longitudinal in nature. Data that follows firms and individuals over time enables more reliable policy evaluation as it provides data before, during, and after policy implementation, enabling long-term follow up. From the moment an individual is a registered taxpayer they can be seen in the tax data until the day they leave formal employment. This allows for mobility and survival analysis in the labour market. The panel advantage is, however, not exclusive to tax data; survey data can also be longitudinal but sometimes suffers from attrition.

In this paper we present an anonymized individual panel created from the combination of payroll and personal income tax records. We use these two datasets to build a comprehensive picture of the income distribution for the formally employed in South Africa.

The individual panel seeks to expand upon the available administrative data in the matched employer–employee panel, the Company Income Tax-RP5 (CIT-IRP5) panel, as described in Pieterse et.al. (2018). By including information from the ITR12 (personal income) returns that are submitted by individuals to South African Revenue Service (SARS) (mainly through e-filing), we include additional income information such as self-employment income, rental income, and retirement annuity contributions. This new panel can provide further insights into the income distribution, how it has changed over time, as well as the structure and dynamics of formal employment and the impact of tax policy changes.

2 Background

The National Treasury, in conjunction with SARS and UNU-WIDER, has made anonymized tax data available for research at the National Treasury Secure Data Facility in Pretoria. This data includes personal, enterprise, and trade related tax information.

In the past, researchers have used either the IRP5/IT3(a) (payroll) or ITR12 (assessed personal) tax data. Wittenberg (2017) compares earnings information in the ITR12 data to those in the Quarterly Labour Force Survey (QLFS) data. He correctly suggests that using only the ITR12 data excludes lower income earners who are under the threshold for filing personal income taxes.

Hundenborn et.al. (2018) combine a 20 per cent sample of the ITR12 and the National Income Dynamics Survey data (NIDS) to examine the effects of top incomes on inequality. Tax data often lacks demographic information whereas surveys usually cover a range of socioeconomic factors. The authors admit that the two datasets show significant differences in the distribution of total taxable income. This is especially true, and graphically represented in their paper, at the top and bottom of the distribution.

Bassier and Woolard (2018) use both the aggregate personal income statistics from the ITR12 returns (available in Tax Statistics (SARS 2017)) and the Post-Apartheid Labour Market Series (PALMS) data (Kerr et al. 2017). Again, the authors combine these datasets as the tax data provides better information at the top end of the income distribution while the survey data does a better job of capturing those at the bottom end of the distribution.

Our motivation is clear: using only the ITR12 data systematically ignores workers not required to file individual taxes. By combining the ITR12 and IRP5/IT3(a)[1] (henceforth, IRP5) tax records we believe we have a more complete version of the income distribution of formal sector workers in South Africa, alongside detailed income information from retirees who receive income from a retirement fund and individuals who are self-employed and only submit ITR12 returns.

In the next section we describe the two data sources in more detail and provide information on how we choose to link them.

3 Data sources

3.1 IRP5 certificates

The IRP5 certificate, or Employee Tax Certificate, is submitted to SARS by employers who are registered for pay-as-you-earn (PAYE) in respect of each employee who received remuneration from the employer. Since the 2010/11 tax year it has been compulsory for employers to submit an IRP5 certificate for each employee, regardless of the level of remuneration. The IRP5 certificates are used to pre-populate information in the tax return of each individual (in the ITR12). Employers are required to submit an IRP5 for each employee three months after the end of the tax year (by the end of May each year). In practice the terms 'employer' and 'employee' cover a wider range of

[1] An IT3a is the same as an IRP5 except that it shows employee earnings from which tax has not been deducted.

institutions and people, such that IRP5 certificates are also submitted for payments to individuals for consulting fees or from retirement funds.

The IRP5 certificate contains all the relevant information needed to determine the tax liability for the individual, if that was the only income received. The tax liability is withheld by the employer and paid over to SARS on behalf of the individual (the PAYE tax amount). The information in the certificate would include items such as income, deductions, allowances, fringe benefits, medical scheme contributions, and age, which are used to determine the PAYE tax amount. The PAYE tax amount that is withheld by the employer is a provisional estimate of the tax liability for the individual.

3.2 ITR12 returns

The ITR12 return, or Personal Income Tax Return, is completed by the individual (or tax practitioner) and includes all the information related to calculating the final tax liability for that year of assessment. Income and deductions from employment from the IRP5 certificate are generally pre-populated in the return, while the individual will manually include additional income from self-employment, investment income, or other sources, and any further deductions. The PAYE tax amount that is determined in the IRP5 certificate is provisional, while the tax liability determined in the ITR12 is final. This may lead to refunds on assessment (or additional tax payments on assessment) if the withheld PAYE tax amount is greater (or lower) than the final calculated tax liability.

Not all taxpayers are required to complete an ITR12. In general, taxpayers are not required to fill in an ITR12 return if they only have employment income from one source, do not have investment income above the exempt thresholds, do not utilize additional deductions, and have an income below the compulsory submission threshold. The compulsory submission threshold was introduced at R120,000 for the 2007/08 tax year, was increased to R250,000 from the 2012/13 tax year, and further increased to R350,000 from the 2014/15 tax year onwards. The personal income section of the annual Tax Statistics publication by SARS and the National Treasury illustrate statistics generated from the ITR12 returns.

3.3 Linking the IRP5 and the ITR12

Linked administrative data are characterized by large sample sizes, with detailed data on groups less likely to be included in survey data and continuing information over a period of time (Harron et al. 2017). The advantages of a linked individual panel are well understood. Limitations to the tax data include missing information and data quality. These can be compounded when combining two datasets and we take care to describe how we create the linkages, where we fall short, or any bias is created.

The aim of creating the individual panel, alongside an explanatory technical paper, is to allow for researchers to access a version of the data that has been restructured and cleaned and fully explained, which would make it easier for researchers to focus on specific research questions rather than on data manipulation and queries on the content of the panel. By creating one panel for research, papers that use the data should be predominantly comparable as there is less space for different assumptions to be made by researchers at the data preparation stage. The process seeks to retain as much information from the original data as possible, while taking some steps to simplify what is available and restrict the size of the panel to make it manageable for analysis. The panel would be updated regularly, preferably on an annual basis, but each iteration will have a version number so that any research paper can be linked back to a particular panel to enhance transparency and reproducibility. The version of the panel on which this explanatory paper is based

includes IRP5 certificates from the tax year 2010/11 to tax year 2017/18 and ITR12 returns from the tax year 2010/11 to tax year 2016/17 (due to different filing deadlines), which represents data from over 142 million IRP5 certificates and 39 million ITR12 returns.

Information contained in the IRP5 certificates is used to pre-populate the ITR12 returns (if an ITR12 is filed), leading to a substantial amount of overlap between the two data sources. Table 1 indicates the total number of unique individuals in each tax year, and how many individuals have either only an IRP5 certificate, or only an ITR12 return, or both. Each year over 18 million ITR12 returns and IRP5 certificates are submitted to SARS. The IRP5 certificates comprise the majority, with a decreasing share of ITR12 returns over successive years, most likely due to the higher compulsory submission thresholds that have been introduced. The last four columns indicate the number of individuals who have both an ITR12 return and IRP5 certificate, or who have only one of each, as well as the number of unique individuals in each tax year.

Table 1: Number of ITR12 returns, IRP5 certificates and unique individuals

Tax year	ITR12 returns	IRP5 certificates	Total ITR12 and IRP5	Taxpayers with ITR12 only	Taxpayers with IRP5 only	Taxpayers with ITR12 and IRP5	Total unique taxpayers
2011	6,084,907	11,927,725	18,012,632	971,381	6,814,199	5,113,526	12,899,106
2012	6,359,048	12,471,832	18,830,880	935,185	7,047,969	5,423,863	13,407,017
2013	6,103,488	12,720,083	18,823,571	843,445	7,460,040	5,260,043	13,563,528
2014	5,806,204	12,997,547	18,803,751	754,605	7,945,948	5,051,599	13,752,152
2015	5,370,717	13,339,622	18,710,339	649,626	8,618,531	4,721,091	13,989,248
2016	4,800,344	13,240,102	18,040,446	542,648	8,982,406	4,257,696	13,782,750
2017	4,791,897	13,517,959	18,309,856	488,042	9,214,104	4,303,855	14,006,001
2018	0	13,015,472	13,015,472	0	13,015,472	0	13,015,472

Note: In this table, columns 2–4 include the cleaned numbers of observations from the IRP5 and ITR12 datasets. Columns 5–8 are the numbers of individuals in various categories. Individuals differ from observation as the data is at the job level and therefore allows for individuals with multiple (consecutive or simultaneous) jobs in a single tax year.

Source: Authors' own estimates based on ITR12 and IRP5 data.

Although there are over 4 million individuals included in both datasets each year, combining the data sources allows for a more complete view of income by including the group of individuals who are only located in each data source. Linking the different variables from each data source also creates a richer and more diverse source of information. For example, one could now investigate whether increases in income outside of employment would lead to individuals dropping out of formal sector employment.

The data is provided by SARS in the form of a SQL database that is around 1.8 TB in size. It would be difficult to work with data of this magnitude with in-memory statistical applications, such as Stata and R, due to limits of available RAM in most computers (the computer used to create the panel had 60 GB of RAM). As such, the process creates four separate datasets that try to follow 'tidy' data principles, as described in Wickham (2014), which reduces the size of the data (without losing information) and makes the data easier to filter and analyse.

In the next section we provide some descriptive information on the linked datasets that make up the individual panel.

4 Description of linked panels

Four linked panels are created: an 'ID panel' which contains the original anonymized identification variables for each IRP5 certificate and ITR12 return; an 'employment panel' where each row represents either a formal period of employment, a lump sum payment, or a payment from a retirement fund; a 'source of income panel' where each row represents the amount of one type of income per person per tax year; and an 'income panel' where each row represents the aggregated level of taxable income and tax liability, amongst other types of income, per person per year. Selected columns of data from the first individual for the 2013 tax year in each panel are represented in Tables 2, 3, 4, and 5, where the income and identification information has been randomly changed to preserve confidentiality. A full description of the variables in each panel is contained in the Appendix.

The 'ID panel' in Table 2 has a row for each submission that was made to SARS, and columns contain the identifying variables in each of those submissions. Table 2 shows that individual 'abcdek' (a derived identifier which is explained in more detail in Section 5.1) had three IRP5 certificates submitted to SARS in the 2013 tax year (rows with a value in the 'IRP5_ID' column), and the individual also filed an ITR12 return (row with 'NA' in the 'IRP5_ID' column).

Each row in the 'employment panel' in Table 3 represents an IRP5 certificate that was submitted to SARS and is uniquely identified through the 'IRP5_ID' variable. The 'employment panel' contains all the information that was provided in each certificate. The table indicates that the individual received non-retirement funding income ('NRFI') from 1 March 2012 to 30 June 2013 (122 days, as shown in 'Period worked'), on which they paid SDL (skills development levy) and UIF (unemployment insurance fund) payroll taxes, but no PAYE tax. The second certificate appears to show a one-off retirement payment as the employment period is one day and no payroll taxes were applied. The third certificate shows further non-retirement funding income from 1 June 2012 until the end of the tax year (9 months). In this case PAYE tax and UIF was withheld, but there is no SDL amount, which might imply the employer was exempt (if the total remuneration of the employer was less than R500,000).

The types of income by source code contained in each IRP5 certificate, as well as from the ITR12 returns, are shown in the 'source of income panel' in Table 4. This individual had 'normal taxable income' (3601) and an 'annual payment' or bonus (3605) from the first and third certificates in Table 3. The third certificate also included a 'general fringe benefit' (3801). The ITR12 return ('Assessed' in the 'IRP5_ID' column) has aggregated the information per source code from the IRP5 certificates. The second certificate indicates there was a pre-retirement withdrawal from a retirement fund after ceasing employment (3920), which corresponds with the employment end date on the first certificate. The first certificate also included a deduction for a contribution to a pension fund (4001). The panel also includes derived variables to indicate the type of income, whether the value is final, and whether the income is used in the taxable income calculation.

The 'income panel' in Table 5 presents the final income position of the individual after including all the information from the IRP5 certificates and ITR12 returns. For individuals who have filed an ITR12 return, the final taxable income amount and tax liability are taken from the ITR12 return, but this excludes any retirement lump sum payments. For individuals who did not have an ITR12 return for that year, the taxable income and tax liability amount is the aggregated amount of each value across all the IRP5 certificates in respect of that specific individual. Additional columns indicate the derived values for gross income, exempt income, income after exemptions, deductions, taxable income (income after exemptions and deductions), lump sums, and the tax on lump sums.

Table 2: IDs panel

Tax_year	ID_d	IRP5_ID	Date_of_birth	ID_number	Tax_reference_number	Taxpayer_category	PassportNo
2013	abcdek	255673340	01/01/1970	AXCJKZTACTCBK	CCCCCCCCCC	INDIVIDUAL	
2013	abcdek	244705789	01/01/1970	AXCJKZTACTCBK	CGBAXKTJCQ		
2013	abcdek	256183279	01/01/1970		CGBAXKTJCQ	INDIVIDUAL	
2013	abcdek	NA	01/01/1970	AXCJKZTACTCBK	CGBAXKTJCQ	INDIVIDUAL	

Source: Authors' own examples from the individual panel.

Table 3: Employment panel (select columns only)

ID_d	Tax_year	IRP5_ID	RFI	NRFI	Total_tax_liability	Periods_worked	PAYE	SDL	UIF	Employment_start	Employment_end
abcdek	2013	256183279	0	16700	500	122	0	200	300	01/03/2012	30/06/2012
abcdek	2013	244705789	0	9200	550	1	0	0	0	30/06/2012	30/06/2012
abcdek	2013	255673340	0	65200	5000	9	3700	0	1300	01/06/2012	28/02/2013

Source: Authors' own examples from the individual panel.

Table 4: Source of income panel

Tax_year	ID_d	IRP5_ID	Source_code	Amount	Final_d	Category_d	Taxable_d
2013	abcdek	255673340	3601	62000	0	Normal_income	1
2013	abcdek	256183279	3601	14300	0	Normal_income	1
2013	abcdek	Assessed	3601	76300	1	Normal_income	1
2013	abcdek	255673340	3605	3200	0	Normal_income	1
2013	abcdek	256183279	3605	2200	0	Normal_income	1
2013	abcdek	Assessed	3605	5500	1	Normal_income	1
2013	abcdek	256183279	3801	200	0	Fringe_benefit	1
2013	abcdek	Assessed	3801	200	1	Fringe_benefit	1
2013	abcdek	244705789	3920	9200	0	Lump_sum_retirement	1
2013	abcdek	Assessed	3920	9200	1	Lump_sum_retirement	1
2013	abcdek	255673340	4001	4750	0	Deduction	NA
2013	abcdek	Assessed	4001	4750	1	Deduction	NA
2013	abcdek	Assessed	4102	3700	1	Withheld_tax_income	NA
2013	abcdek	Assessed	4115	550	1	Withheld_tax_retirement	NA

Source: Authors' own examples from the individual panel.

Table 5: Income panel (transposed)

ID_d	abcdek
Tax_year	2013
ITR12_taxable_income	77250
ITR12_tax_liability	2465
IRP5_PAYE_d	3700
IRP5_lump_sum_tax_d	550
Gross_income_d	91200
Exempt_income_d	0
Income_d	91200
Deductions_d	4750
Lump_sum_retirement_d	9200
Taxable_income_d	77250
Tax_liability_d	2465

Source: Authors' own examples from the individual panel.

5 Data structure and challenges

The data from the IRP5 certificates and ITR12 returns are unaudited, presenting some challenges when conducting any analysis. Only a handful of researchers have had access to the tax data and even fewer researchers have used the IRP5 or ITR12 datasets. We discuss below some of the existing issues in the data we have noticed and noted by other researchers. Some of these data errors are important for researchers to note as they will affect any econometric analysis. While we check for structural anomalies in the panel, we recognize that issues may still persist. The intention is to rectify problems identified by researchers in future versions of the panel where we may not have noticed any issue previously. We also acknowledge with each new year of tax data, changes in tax policy and reporting requirements will affect how the panel is adapted. In light of this, a new release of the panel will include documentation on changes made.

5.1 Deriving a unique identifier

The initial challenge in creating a panel from the IRP5 certificates and ITR12 returns is that they need to be linked according to each specific individual (i.e. across IRP5 certificates, within a single year between the IRP5 certificates and ITR12 returns, and over multiple years within, and between, each data source). This process would be simple if there was a single unique identifier between each data set and across years, but the existence of identification variables is not complete in each source for each individual, as described in Kerr (2018) and Ebrahim et al. (2017) for the IRP5 certificates. For example, identification variables include an individual's anonymized ID number, date of birth, and tax reference number. However, some IRP5 certificates only include the ID number from some employers and the tax reference number from other employers, while there may also only be tax reference numbers in the IRP5 certificates but ID numbers and tax reference numbers in the ITR12 return (as shown in Table 2). Linking the data by all three main identifiers would not identify a single person in these circumstances (as not all three identifiers would match). This method would lead to an overestimate of the number of individuals who are employed, the number of people with income and the level of churn in employment and would also lead to an underestimate of average income.

To avoid this potential bias, and reduce the number of missing cases, a derived unique identifier is created using a 'deterministic linkage' approach, which uses a step-by-step process to try and identify individuals across tax records. Harron et al. (2017) state how a deterministic approach may lead to missed matches, although false match rates are usually low (which may still create a bias, although to a lesser extent).

We combine the identification variables (tax year, date of birth, ID number, and tax reference number) from each source and each year together into one data set. The data is initially filtered to exclude rows without an ID number then ordered by tax year, date of birth, ID number, and tax reference number. A simple rule is applied to replace a missing tax reference number with the tax reference number in the row above if the date of birth and ID number of that row matches the row above. The process is then repeated for the date of birth and the tax reference number to fill out missing ID numbers, and for cases with only passport numbers. The expanded set of information on ID numbers and tax reference numbers is then re-joined with the original IRP5 certificates and ITR12 returns and can now be merged across all three main identification variables with far fewer missed matches.

By following this approach, the number of separate individuals across all years decreases from 33.8 million to 22.7 million. There are, however, further complications relating to each individual identifier as there are around 400,000 individuals who have the same date of birth and ID number but have different tax reference numbers. The above rule would create multiple individuals in these instances and randomly assign the lowest ranked tax reference number to all the remaining ID numbers, which is not ideal. Instead, for these cases it is likely that they do represent the same people (as their ID number and birthday are the same) and so the same unique identifier is attributed to the different tax reference numbers.

The derived identifier for the individual panel is *ID_d* and it is recommended that this variable is used by researchers as the main identifier for individuals in the dataset.

5.2 Geocoding

The IRP5 certificates contain both residential and workplace address fields. The postal codes for residential and workplace addresses are extracted and separately geocoded. Street name, street number and suburb are also contained in the data but due to the incompleteness and the large number of spelling errors in these fields the postal code field is used to conduct the geocoding.

We link the postal codes to other geographical aggregations including: province, municipality, district municipality, and census main place. This is done to facilitate spatial research and so that researchers can aggregate the IRP5 data to a geographical level, which promotes spatial research while still ensuring the anonymity of individuals and firms.

In the absence of the postal code shape files, the Google Maps (or OpenStreetMap) API is used to calculate the geographic midpoint of each postal code and find the geographical structure in which the postal code is located. This approach makes a few notable assumptions:

- The postal codes reported by Google Maps is correct;
- The midpoint of the postal code is an accurate reflection of the area of the postal code; and
- The postal code does not straddle multiple geographical structures, and the geographical structure in which the midpoint of the postal code is located is also the geographical structure which has the largest overlap with said postal code.

A list of all unique postal codes which exist in the data is created, and each of the postal codes is passed to the Google Maps API, which then returns the GPS coordinates of the midpoint of the postal code. There are approximately 3,400 unique postal codes recorded from the IRP5 data from a total of 10,000 postal codes in South Africa.

Using the longitude and latitude coordinates, the location is mapped to the geographical structure in which the coordinate falls. This is done using shapefiles of various geographical structures: province, local municipality, district municipality, and census main place from the South African Census 2011 (Statistics South Africa 2011).

This process generates a conjunction table including postal codes, provinces, district municipality, local municipality, and census main place. The postal codes in the IRP5 are then linked to the conjunction table, producing the results outlined in Table 6. The postal codes are then removed to ensure the anonymity is maintained in the data.

Below, in Table 6, we report the percentage of residential and workplace observations we are able to geocode.

Table 6: Percentage of personal and workplace geocoded, 2011–18

Tax Year	Residential	Workplace
2011	21%	1.3%
2012	34%	5.2%
2013	40%	73%
2014	32%	90%
2015	16%	91%
2016	16%	90%
2017	16%	88%
2018	15%	88%

Source: Authors' own estimates using the individual panel.

We think the residential location information is inconsistent, and very low across years in the data as it is not a required field for employers to inform SARS of the residential location of their employees. The workplace location information improves dramatically in 2013 and has been well populated in the later years. While the geocoding information is useful for workplace analysis, we do not recommend residential location analysis is conducted with this data due to the lack of information.

5.3 Nature of person

The ITR12 returns that are included in the panel specifically exclude returns from trusts and companies. The IRP5 certificates include clubs, partnerships, retirement funds, associations, and other types of entities that are required to submit IRP5 certificates. The panel assumes that these certificates are made on behalf of individuals who are either employed or paid by these institutions.

5.4 Start and end of employment

As indicated by the Business Requirements Specification document issued by SARS, it is mandatory for each IRP5 certificate to include the start and end of the employment period (SARS 2018). This information can then be used to determine the job duration of the individual and informs us on the hiring and separations of employees. There are, however, some errors in the start and end dates of employment. For example, there are cases where the employment start is after the employment end date. In this case we cannot verify the start and end dates of employment. Kerr (2018) describes this and other measurement errors in the start and end dates in more detail. He finds a drop in the number of individuals employed in the last two weeks of the tax year and sees the continuation of that drop into the next tax year.

Another example of an incorrect start date is where the start date precedes the start of the tax year. The start date could be listed at 1 February 2012 in the 2013 tax year, which actually only begins

on 1 March 2012. One way in which researchers can overcome this is by manually moving the start date to 1 March 2012 in the 2013 tax year.

If duration of employment is of interest and not the start or end dates specifically, the periods worked information can be used. The IRP5 data includes the periods worked as well as the number of periods in the year of assessment. These are also mandatory fields according to the Business Requirements Specification document (SARS 2018). The number of periods in the year of assessment indicates how the employer divided the year and related to the intervals at which the employee is paid. This could be, for example, weekly, fortnightly, monthly, or daily. The periods worked then indicates the number of periods for which the employee worked. This, too, can be used as a measure of employment duration for the year. Pieterse et. al. (2018) have a good discussion on how they use this information to calculate weights for employment in the CIT-IRP5 panel.

We make no changes to these variables; researchers can very easily calculate weighting or correct the start and end dates of employment as they require for their research.

5.5 Date of birth and calculated age

There is a small percentage of observations with missing data in the date of birth variable. Often the population of interest in research is for those of working age, but, in the data, we see approximately 0.18 per cent of the data includes those below 15 years of age in the IRP5 certificates. We regard this as a data entry error in the date of birth variable as it is illegal for children under the age of 15 to be employed. Approximately 7 per cent of the individuals per tax year are over the age of 65 and there is a very small percentage of observations with ages over 99 years or below 10 years. We think that this is most likely due to data entry errors but, since these errors cannot be corrected, we recommend that these observations are dropped for research purposes.

5.6 Source codes

The Business Requirements Specification document indicates that SARS restricts the number of income source codes reported in the IRP5 to 20 and the number of employer and employee deduction codes to 12. To this end SARS has identified source codes and sub source codes which need to be aggregated to main codes when they exceed 20. In cases where the source codes exceed these limits, the employer then combines amounts into the main codes and SARS would not receive information on the sub source codes. There is no indicator or flag for when this has taken place in the data.

5.7 Gender

The gender variable in the IRP5 data is derived at SARS before the anonymization process. The 7th digit of the South African Identity Number (ID number) indicates the gender of the person. This automatically means that observations with no ID number are missing in the gender variable. There are two main reasons why we think some observations do not have ID numbers. First, the data might include foreign workers in South Africa who do not have ID numbers, but the data includes their passport number. Second, the IRP5 data includes non-natural entities, and therefore they should be removed for any labour market analysis. Researchers should note that any work using the gender variable or done at the gender level will be biased toward South African workers.

Ebrahim and Lilenstein (forthcoming 2019) include a more detailed discussion of the gender variable and how it can be used for research on the labour market.

5.8 Revisions to submissions

The IRP5 data includes multiple submissions per person per year in some cases, as certificates are re-submitted to SARS with updated information. SARS includes a revision number for each submission. The panel only includes IRP5 certificates with the higher revision number, which is taken to be the final submission.

5.9 Changes to forms and tax policy amendments

The tax data has not been collected with the primary aim of research, but rather for revenue collection purposes. Revenue agencies try to reduce the administrative burden where possible to minimize compliance costs, which may result in changes in the type of information submitted from year to year, leading to periods of insufficient data for researchers. Similarly, tax policies change, resulting in changes in taxable income definitions and adjustments to the type of information that is required to be submitted. As a result, researchers need to take extra care when making inferences from changes in the panel over time.

The changes to source codes for the IRP5 certificates are shown in the Business Requirements Specification document (SARS 2018). The document shows numerous changes over the years, some of which may have a material impact on any analysis. For example, as highlighted in Kerr (2018), the source codes for retirement fund income (such as 3603 and 3610) were amalgamated into 3601 (normal income) for the 2009/10 to 2011/12 tax years. Any analysis of changes in employment income would need to correct for the additional income from retirement funds in 2010/11 and 2011/12 in the panel.

The previous example shows the impact of changes in reporting; however, adjustments also need to be made due to tax policy changes. For example, a significant change to the taxation of retirement funds was enacted from 1 March 2016. Although there was no change to the source code for pension fund contributions (which remained as 4001), the amount reported under this line item changed from only indicating employee contributions to pension funds to both employer and employee contributions to pension funds. Provident fund contributions (4003) also became deductible for the first time. These changes have substantially altered the amounts, and interpretation, of the same source codes and any analysis of these contributions over time would need to make the requisite adjustments.

Ideally, further work needs to be done to adequately catalogue all the changes to the forms and the changes to tax policies to make it easier for researchers to understand how these adjustments affect any analysis when using the panel. At the minimum, researchers using the individual panel need to note this information and make a decision on the extent to which this affects their work before using the data for their analysis.

5.10 Outliers and errors

The panel has been checked from an overall structural perspective, but no changes have been made to alter particular data points which may be erroneous, due either to errors in capturing the data or errors from taxpayers when submitting returns. All the data in the panel (other than the derived variables) is directly from SARS. A lot of subjective decision making would be required to make adjustments to what one thinks is an error, which in our view could lead to further mistakes and complications. At present, researchers will need to use their own judgement to make adjustments for outliers and potential errors. Although not included in this panel, a future version could include a flag to indicate whether a particular entry warrants further investigation.

6 Reliability of the individual panel

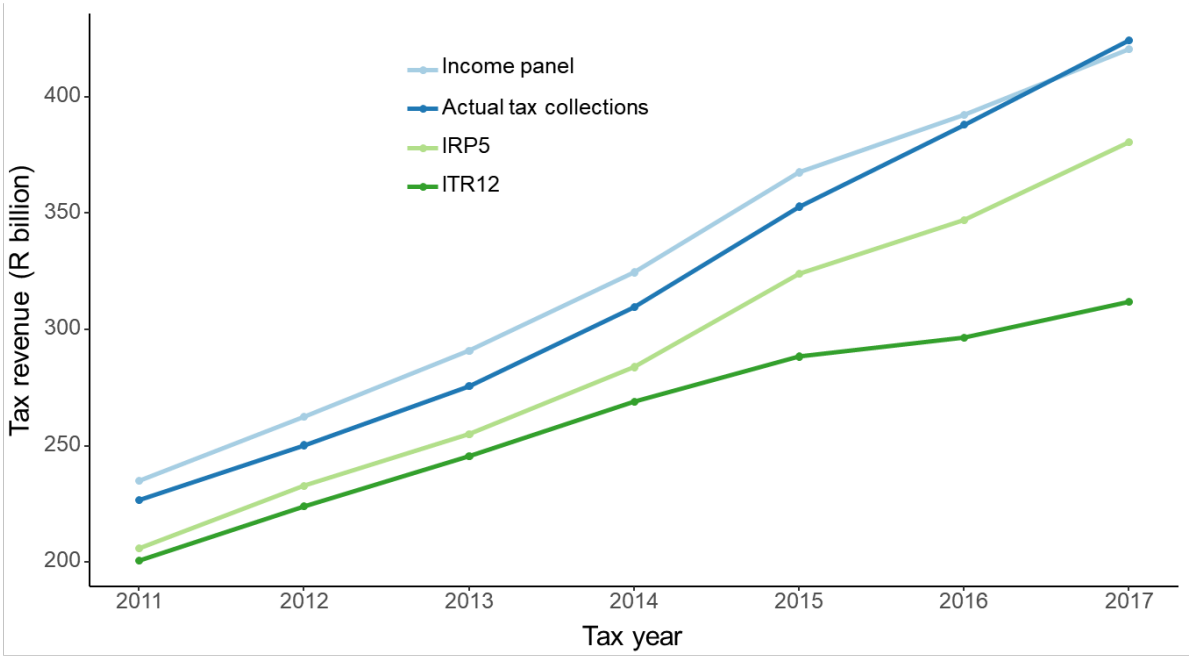
Survey data often have a weighting variable to gross up each measure to reflect the population, but the individual panel reflects raw microdata with no corrections to match to an overall control total. Confidence in using the individual panel for analysis (especially from a macro perspective) will be heavily dependent on whether the panel reasonably captures total incomes and total tax revenues.

6.1 Accuracy compared to actual tax collections

Figure 1 compares actual personal income tax collections with the amounts of tax collected as represented in the IRP5 data, the ITR12 data and the 'income panel'. The ITR12 returns alone represent the smallest amount of personal income tax collections across all the years in the panel. Although the ITR12 returns include taxpayers who do not have an IRP5 certificate (such as the self-employed), they exclude individuals who have paid tax through the IRP5 process but who have not filed a tax return and also exclude taxes paid on retirement fund withdrawals. In the latest year, the undercounting from using only the ITR12 return is close to 30 per cent. The tax liabilities from the IRP5 tax certificates are higher but remain lower than the actual personal income tax collections for each year, as these certificates exclude income that is not linked to an employer or retirement fund administrator. This discrepancy highlights the importance of combining both ITR12 and IRP5 information to get a more holistic view of total income and total tax collections. The main advantage of the individual panel is that it allows researchers to add individuals from the IRP5 data to the ITR12 data to get a more detailed view of the overall income distribution.

The total tax liability from the 'income panel' in Figure 1 represents the ITR12 tax liability for those individuals with an ITR12 return, and the IRP5 tax liability for those without an ITR12 return. In both cases, the taxes paid on retirement fund withdrawals in the IRP5 are added to the total tax liability amount. The tax liability from the 'income panel' is around 4 per cent higher than actual personal income tax collections for the 2010/11 to 2014/15 tax years, while the values are similar for 2015/16 and 2016/17. These two figures should not be identical, since actual revenue collections are shown on a cash-flow basis, while tax return liabilities are on an accrual basis. Individuals who made additional payments on assessment may then contribute to tax revenues in a later year for a return that is associated with a previous year (which may overstate revenues in the panel data, as in years 2010/11 to 2014/15). However, there may also be late ITR12 returns which would only be reflected in the panel at some future date and would understate total revenues (which may have occurred for the 2015/16 and 2016/17 tax years). Overall, it appears that the individual panel is a relatively good reflection of total taxable income according to aggregate figures.

Figure 1: Comparison of revenue outcome with results from tax data



Source: Author's own estimates using the individual panel, IRP5, ITR12, and National Treasury tax collections data.

6.2 Accuracy of derived variables

The 'income_panel' creates derived income variables to show gross income, deductions, taxable income, and others. These derived variables use the underlying source codes to recreate the calculation that is used to calculate the final taxable income upon which the tax tables are applied. The ITR12 returns provide the final taxable income amount as used in the assessment, but no final amount is provided for the IRP5 certificates. For the ITR12 returns, of the 39 million returns between 2011 and 2017, around 98.7 per cent of the derived taxable income amounts were correct.

After recalculating the taxable income amount, this value can be applied to the prevailing tax table in each year.[2] From 2012 onwards, the medical tax deductions available to individuals for excessive medical expenses and contributions to medical aids was changed to a medical tax credit (which works in the same way as a rebate and is taken off after the initial tax calculation). Data on the value of the medical tax credit is not available in the current panel for years 2013 to 2015 in the ITR12 return, resulting in a large increase in the number of incorrect calculations in these years. The other years shows a far smaller number of incorrect tax calculations, which could be due to assessed losses (which are also not included in the current panel) or some other element that is not included. The IRP5 calculations have a higher proportion that do not match, but that is expected to a large degree as payrolls do not know of other types of incomes and also assume each employee will work the full year when estimating the amount of tax to withhold on a monthly basis. As such, many IRP5 certificates would have an incorrect level of withholding tax. Table 7 below indicates the percentage of incorrect tax liability calculations from the ITR12 and the IRP5.

[2] The C++ function to run the tax calculation can be found here.

Table 7: Proportion of incorrect tax liability calculations per type of return

Tax year	ITR12	IRP5
2011	2.1%	18.4%
2012	1.9%	19.7%
2013	23.2%	17.3%
2014	11.2%	15.3%
2015	22.3%	17.7%
2016	4.6%	18.8%
2017	6.1%	19.1%
2018		25.6%

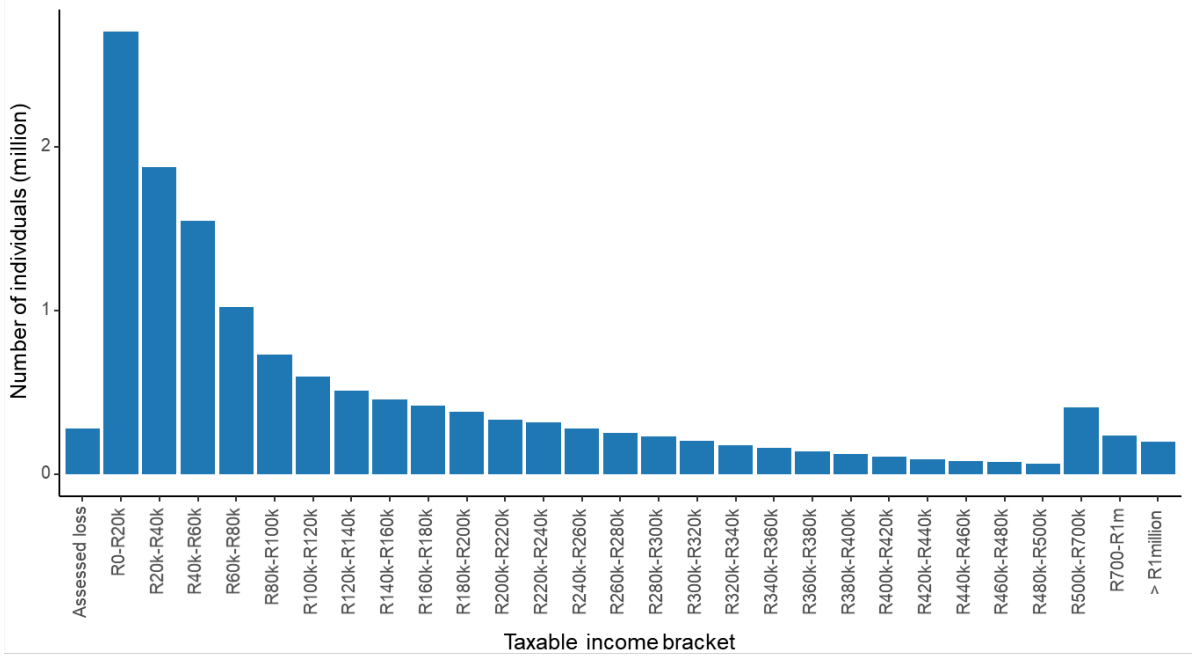
Source: Authors' own estimates of IRP5 and ITR12 data.

The tax calculation can be improved, especially in terms of adding medical tax credits, but for the vast majority of those on assessment the final tax amount aligns with the underlying calculation from data in the panel.

7 Implications for research

The individual panel creates a new data source to investigate the income distribution in South Africa. For example, Figure 2 shows the taxable income distribution in 2016/17, using the derived taxable income variable with most of the taxable income brackets with bins of R20,000. It is noticeable that over 2.7 million of the 14 million individuals in 2017 have a taxable income of less than R20,000. The top three highest brackets in Figure 2 have far wider bins but are still smaller in size compared to the lowest brackets, with close to 200,000 individuals with taxable incomes greater than R1 million.

Figure 2: Taxable income distribution for 2016/17 tax year

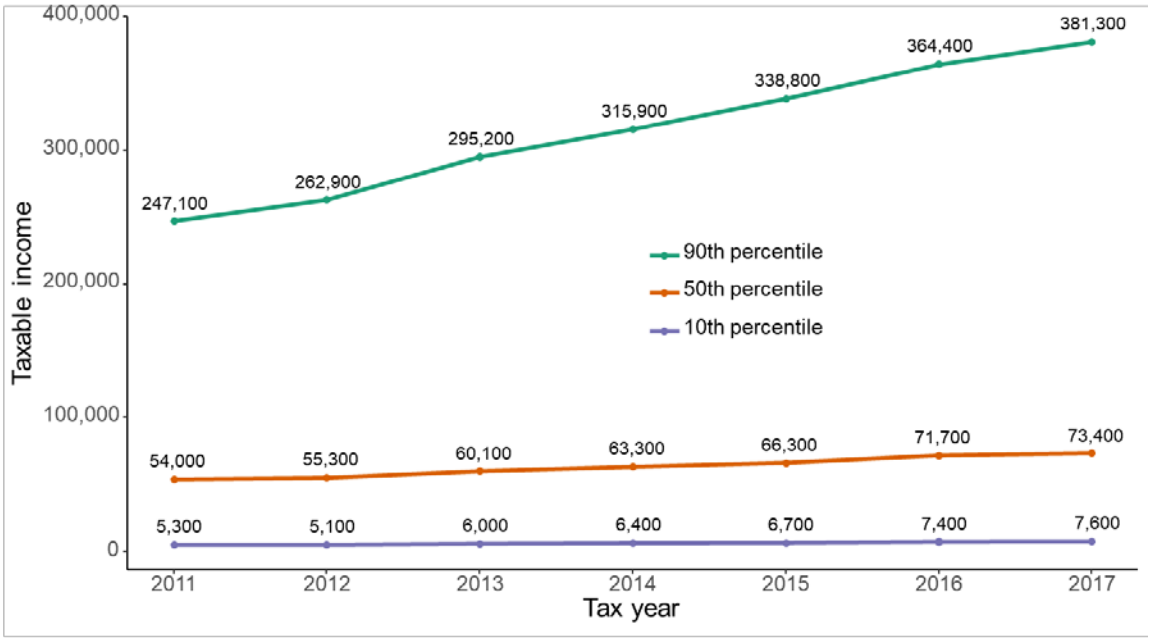


Source: Authors' own estimates using the individual panel.

Figure 3 shows taxable incomes at the 90th, 50th, and 10th percentiles. Further research, following on from Wittenberg (2017) and Bassier and Woolard (2018), could be undertaken to assess how the income distribution from the panel, especially at the top end, compares to surveys in South

Africa such as the Income and Expenditure Survey, the QLFS, and the National Income Dynamics Survey. Previous research to compare these data sources has either used the aggregated Tax Statistics published by SARS and National Treasury, or a sample of micro data from one or two years, both of which have relied on the ITR12 returns alone. Notwithstanding some of the limitations of using tax data, more in-depth and accurate comparisons can be made by using the full population of tax records over multiple years, from both the ITR12 returns and the IRP5 certificates, that is available in this panel.

Figure 3: Taxable incomes at the 90th, 50th and 10th percentiles



Source: Authors' own estimates using the individual panel.

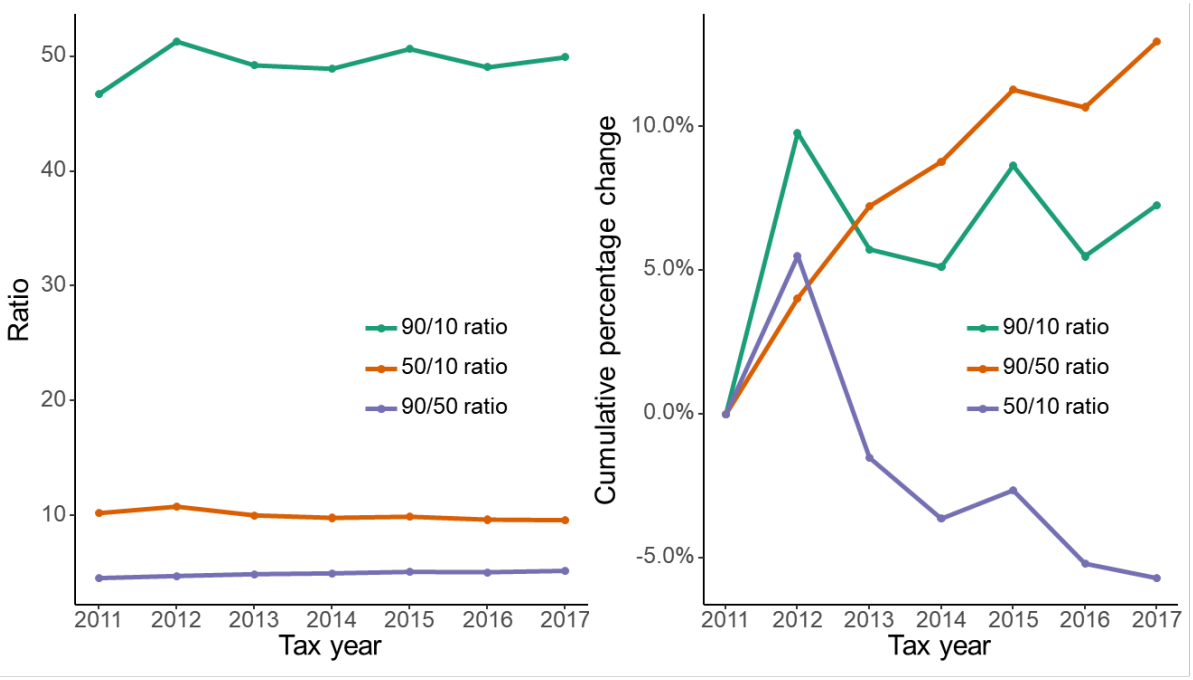
As part of the income distribution research, the longitudinal and detailed nature of the information in the panel will provide greater insights into the trends and drivers of the high levels of inequality in South Africa over the past few years, providing an extension to other work, such as that by Leibbrandt et al. (2010). As an example, Figure 4 shows the income ratios from the taxable income percentiles used in Figure 3. The 90/10 ratio in the first figure shows that individuals at the 90th percentile earn almost 50 times more than individuals at the 10th percentile, highlighting the extreme levels of inequality in South Africa. However, after a large increase in 2012, the 90/10 ratio has remained relatively flat in the following years up until the 2017. Decomposing this change with the 50th percentile in the second figure shows that although the 90/10 ratio has remained quite flat, this has masked large changes in the 90/50 and 50/10 ratios since 2012. The 90/50 ratio has increased by around 10 per cent since 2012, while the 50/10 ratio has decreased by a similar amount, suggesting that growth in the lowest income has kept pace with those at the 90th percentile, while those at the median have fallen behind.

These results, and other future research results, should be interpreted carefully as they represent only a subsection of the population (those who have filed returns with SARS). To fully capture the bottom end of the distribution we would need data on incomes in the informal sector which are currently only captured in the QLFS, but there would be scope to combine this data with survey data to provide one picture of the full income distribution, such as in Neri and Hecksher (2018) for Brazil. A further extension would be to combine this with national accounts data to create a set of distributional national accounts, following Piketty et al. (2017).

The panel also has extensive information on formal employment, which can feed into debates around levels of remuneration (such as the minimum wage negotiations in the recent past) and can allow for a greater variety of research avenues by linking other sources of income (and deductions) to those who are formally employed.

The tax data is well suited to providing evaluations of tax policy changes, such as the change to medical tax deductions, to medical tax credits, and the large changes to retirement fund deductions. The recent changes to personal income tax rates, including the introduction of a new top rate of 45 per cent, make this data well suited to estimating elasticities of taxable income, which could potentially be decomposed in a similar fashion to Piketty et al. (2014).

Figure 4: Income ratios



Source: Authors' own estimates using the individual panel.

It should be noted that the potential research from this data would still face some limitations, as not all investment income is available in these returns. Only investment income above the thresholds is usually included in the ITR12 returns. Also, dividend income is exempt as a withholding tax applies at the level of the company before the dividend is distributed to taxpayers. Taxpayers are generally required to fill in the exempt dividends they receive, but since it does not alter the tax calculation, many do not. SARS does, however, receive IT3(b) and IT3(s) certificates from financial services firms which details dividends, interest, and capital gains, but this data is not included in the current panel. An important extension would be an attempt to include information from these additional submissions in later versions of the panel.

We have highlighted some of the policy relevant research ideas that can be immediately undertaken using the individual panel. The scoping papers by Ebrahim et. al. (2019), Ebrahim and Lilenstein (forthcoming 2019), and Chatterjee (forthcoming 2019) are other sources of research ideas that could be implemented using the individual panel.

8 Conclusion

The availability of the up-to-date tax data allows for the real-time evaluation of policy, potentially allowing for policy changes if or where required. One major impediment to research in South Africa has been the access to administrative data. The Nordic countries lead with the model of consistent identifiers over various administrations and the infrastructure model to deal with large big data. While not publicly available, the ability for researcher to apply to access the tax data in a secure facility is an enormous step for a developing country.

Administrative data is not collected for the purpose of research, and the tax data we use to create the individual panel is more complex and messier than most datasets researchers will encounter. Despite our efforts to clean and produce a reliable dataset, time and effort will still be required by researchers to clean and enable that data to produce reliable estimates in their analysis.

This technical paper should be used alongside the individual panel, and researchers are encouraged to get in touch with the authors regarding issues so that they can be corrected in later versions of the panel.

References

Bassier, I., and I. Woolard (2018). 'Exclusive Growth: Rapidly Increasing Top Incomes Amidst Low National Growth in South Africa'. REDI3x3 Working Paper 47. Cape Town: SALDRU.

Card, D., R. Chetty, M.S. Feldstein, and E. Saez (2010). 'Expanding Access to Administrative Data for Research in the United States'. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas. Available at SSRN: http://dx.doi.org/10.2139/ssrn.1888586.

Chatterjee, A. (forthcoming 2019). 'Measuring Wealth Inequality in South Africa: An Agenda'. WIDER Working Paper. Helsinki: UNU-WIDER.

Connelly, R., C.J. Playford, V. Gayle, and C. Dibben (2016). 'The Role of Administrative Data in the Big Data Revolution in Social Science Research'. *Social Science Research*, 59: 1–12.

Ebrahim, A., R. Gcabo, L. Khumalo, and J. Pirttilä (2019). 'Tax Research in South Africa'. WIDER Working Paper 2019/9. Helsinki: UNU-WIDER.

Ebrahim, A., M. Leibbrandt, and V. Ranchhod (2017). 'The Effects of the Employment Tax Incentive on South African Employment'. WIDER Working Paper 2017/5. Helsinki: UNU-WIDER.

Ebrahim, A., and K. Lilenstein (forthcoming 2019). 'Gender and the South African Labour Market'. UNU-WIDER Working Paper. Helsinki: UNU-WIDER.

Harron, K., C. Dibben, J. Boyd, A. Hjern, M. Azimaee, M.L. Barreto, and H Goldstein (2017). 'Challenges in Administrative Data Linkage for Research'. *Big Data & Society*, 4(2): 1–12.

Hundenborn, J., I. Woolard, and J. Jellema (2018). 'The Effect of Top Incomes on Inequality in South Africa'. WIDER Working Paper 2018/90. Helsinki: UNU-WIDER.

Kerr, A. (2018). 'Job Flows, Worker Flows and Churning in South Africa'. *South African Journal of Economics*, 86: 141–66.

Kerr, A., D. Lam, and M. Wittenberg (2017). 'Post-apartheid Labour Market Series' [dataset]. version 3.2. Available at: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/434/accesspolicy (accessed 15 January 2019).

Leibbrandt, M., I. Woolard, A. Finn, and J. Argent (2010). 'Trends in South African Income Distribution and Poverty since the Fall of Apartheid'. *OECD Social, Employment and Migration Working Papers*, 101: 1–91.

Neri, M.C., and M. Hecksher (2018). 'Top Incomes' Impact on Inequality, Growth and Social Welfare.' WIDER Working Paper 2018/137. Helsinki: UNU-WIDER.

Pieterse, D., E. Gavin, and C.F. Kreuser (2018). 'Introduction to the South African Revenue Service and National Treasury Firm-Level Pane..' *South African Journal of Economics*, 86: 6–39.

Piketty, T., E. Saez, and S. Stantcheva (2014). 'Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities'. *American Economic Journal: Economic Policy,* 6(1): 230–71.

Piketty, T., E. Saez, and G. Zucman (2017). 'Distributional National Accounts: Methods and Estimates for the United States'. *The Quarterly Journal of Economics*, 133(2): 553–609.

SARS (South African Revenue Service and National Treasury) (2017). *2017 Tax Statistics*. ISBN: 978-0-621-45970-8. Pretoria: SARS and NT.

SARS (South African Revenue Service) (2018). *Business Requirement Specification: PAYE Employer Reconciliation* (August 2018 release). Pretoria: SARS.

Statistics South Africa (2011). *Census 2011 Geography Metadata*. Pretoria: Statistics South Africa.

Sulla, V., and P. Zikhali (2018). 'Overcoming Poverty and Inequality in South Africa: An Assessment of Drivers, Constraints and Opportunities'. Washington, DC: World Bank Group.

Wickham, H. (2014). 'Tidy Data'. *Journal of Statistical Software*, 59(10): 1–23.

Wittenberg, M. (2017). 'Measurement of Earnings: Comparing South African Tax and Survey Data'. SALDRU Working Paper 212. Cape Town: SALDRU.

Appendix

Data dictionary

The individual panel contains income and employment information from returns submitted to the South African Revenue Service (SARS) between the 2010/11 and 2017/18 tax years. The data is taken directly from the ITR12 tax returns that are submitted by individuals and from the IRP5 employee certificates that are submitted by employers. ITR12 information is available up to 2016/17, while IRP5 information is available up to 2017/18.[3]

The panel is a SQL Server database that contains four linked panels, which are described in the following tables. Variables with a name that ends with '_d' are derived variables, while all other variables reflect information taken directly from submissions to SARS. As far as possible, if no information is provided, the entry is recorded as NA—a 0 would represent information that is given to SARS, but with an amount of 0.

IDs_panel: Panel where each row represents identification variables for each submitted return. Information is used to create one derived identifier for each individual that is used in the other panels (to avoid cases where the same individual is seen as two separate people if identification indicators change from year to year). Consists of 179,441,076 rows and 8 columns.

Variable name	Example	Description
ID_d	abcdek	Derived variable: The value is given to individuals who have the same date of birth, but who may be missing either an ID number or a tax reference number or passport number in their returns for a particular year. For example, an individual with an ID number and tax reference number in 2017/18 but with only an ID number or a tax reference number in any of the previous years would be allocated the same derived identifier.
Tax_year	2017	Year in which the tax year ends. For example, 2017 refers to the 2016/17 tax year which runs from 1 March 2016 to 28 February 2017.
Date_of_birth	1970/01/01	Specific day, month, and year.
ID_number	AXCJKZTACTCBK	Anonymized national identity number as provided by Department of Home Affairs.
Tax_reference_number	GCBAXKTJCQ	Anonymized unique SARS tax reference number. Allocated on date of registration with SARS.
Taxpayer_category	INDIVIDUAL	Includes INDIVIDUAL and RETIREMENT FUND, where the latter indicates whether payment is made by retirement fund administrator.
IRP5_ID	120836934	Indicates the unique IRP5 ID number allocated by SARS when the return is submitted.
Passport_number	""	Anonymized passport number (if no ID number is supplied).

Source: Authors' own illustration of the IDs panel.

[3] The final filing date for IRP5 returns is the end of May in the following tax year (May 2018 for the 2017/18 tax year), while the final filing date for ITR12 returns is the following January (January 2019 for the 2017/18 tax year).

Source_code_panel: Panel where each row represents one type of income per person per year, based on the source code provided. Entries include amounts per source code from both ITR12 and IRP5 returns. Consists of 634,153,527 rows and 8 columns.

Variable name	Example	Description
ID_d	abcdek	Derived variable: Refer to IDs_panel.
Tax_year	2017	Year in which the tax year ends. For example, 2017 refers to the 2016/17 tax year which runs from 1 March 2016 to 28 February 2017.
IRP5_ID	440801595	For IRP5 returns, this variable indicates the unique IRP5_ID number allocated by SARS when the return is submitted. For ITR12 returns the value will be 'Assessed'.
Source_code	3601	Source code provided in the ITR12 or IRP5 return.
Amount	164390	Amount provided in relation to the source code in the return. For 'Business income' and 'Investment income', every odd source code represents a loss. To ease calculations, values in the 'Amount' variable were changed from positive to negative for loss source codes.
Final_d	0	Derived variable: ITR12 returns indicate the final assessment and override IRP5 returns. All ITR12 source codes will have a value of 1, while IRP5 source codes will have a value of 0 if the taxpayer filed an ITR12 return for that year and a value of 1 if no ITR12 return was filed that year. Note that useful information (such as provident fund contributions under source code 4003) may be shown in the IRP5 returns and have a value of 0 for this variable if an ITR12 return was filed but will not be included in the ITR12 information (as it is included in source code 4029). Rows with a value of 1 will only reflect information needed for the final tax calculation.
Category_d	Normal_income	Derived variable: Each source code is split into one of 11 categories, which are used to replicate the tax calculation. The categories are 'Normal income', 'Business income', 'Investment income', 'Fringe benefits', 'Allowances', 'Lump sum income', 'Lump sum retirement', 'Deductions'. Source codes for each category: 'Business income' - 102 to 3499 'Normal income' - 3601 to 3699 'Allowance' – 3701 to 3768 'Fringe benefit' – 3801 to 3866 'Lump_sum_retirement' – 3901 to 3957 'Deduction' – 4001 to 4051 'Withheld_tax_income' – 4101 to 4150 'Investment_income' 4201 to 4276, 4290 and 4292 'Activity income' 4278 to 4287 But specific source codes in each group are then recategorized: 'Withheld_tax_retirement' – 4115 'Contributions' (which are amounts not used in the final tax calculation) – 4005 and 4472 to 4475 For tax years 2017 and later, 'Contributions' - 4001 to 4004 and 4006, 4007 For tax years before 2017, 'Contributions' – 4003.
Taxable_d	1	Derived variable: A value of 1 indicates whether a source code amount would be taxable, with 0 otherwise. Non-taxable source codes are 3000 to 3021, 3602, 3604, 3609, 3612, 3652, 3659, 3662, 3696, 3703, 3705, 3714, 3755, 3764, 3814, 3865, 4203, 4203. Source codes in 'Deduction', 'Contributions', 'Withheld_tax_income', 'Withheld_tax_retirement' cannot be either and are given an NA.

Source: Authors' own illustration of the Source Code panel.

Employment_panel: Panel where each row represents on employee certificate. Note this may include both certificates for the calculation of pay-as-you-earn (PAYE) in relation to employment, as well as PAYE on lump sum incomes (such as tax on early withdrawals from a retirement fund). Consists of 137,162,333 rows and 37 columns.

Variable name	Example	Description
ID_d	abcdek	Derived variable: Refer to IDs_panel.
Tax_year	2017	Year in which the tax year ends. For example, 2017 refers to the 2016/17 tax year which runs from 1 March 2016 to 28 February 2017.
Date_of_birth	1971/09/01	Specific day, month, and year.
Gender	M	
Taxpayer_category	INDIVIDUAL	As per IDs_panel.
IRP5_ID	440801595	The unique IRP5_ID number provided by SARS when the return is submitted.
PAYE_number	QBCKQKCACK	The PAYE reference number of the employer / administrator submitting the IRP5 certificate.
Certificate_number	70307270….	Unique number of certificate that is allocated by the employer / administrator.
Main_income_source_code	2605	Source code for main type of income from employer.
Non_taxable_income	0	
Retirement_funding_income	0	Amount used to determine limit for pension fund contributions before 2016/17.
Non_retirement_funding_income	0	Amount used to determine limit for retirement annuity fund contributions before 2016/17.
GrossRemunerationAmnt	202850	Amount used to determine limit for contributions to any retirement fund from 2016/17 onwards.
Total_tax_liability	22030	Includes all payroll taxes, SDL, UIF, SITE and PAYE.
Total_periods_worked	12.0	Could be in days, months, or years and varies from 1 to 365.
SITE	0	Withheld tax called Standard Income Tax on Employees. Phased out from 2012/13.
PAYE	16830	Pay-as-you-earn withholding tax.
PAYE_lump_sum	0	Withheld taxes on lump sum payments.
SDL	1680	Skills Development Levy contribution.
UIF	3520	Unemployment Insurance Fund contribution.
Employment_start	2016/03/01	
Employment_end	2017/02/28	
ETI	0	Employment tax incentive amount.
Medical_scheme	0	Medical tax credit claimed through IRP5.
Province_d	North West	Derived variable: From postal address postal code.
District_muni_d	Dr Kenneth Kaunda	Derived variable: From postal address postal code.
Local_muni_d	City of Matlosana	Derived variable: From postal address postal code.
Main_place_d	Kanana	Derived variable: From postal address postal code.
Bus_province_d	North West	Derived variable: From business address postal code.
Bus_district_muni_d	Dr Kenneth Kaunda	Derived variable: From business address postal code.
Bus_local_muni_d	City of Matlosana	Derived variable: From business address postal code.
Bus_main_place_d	Kanana	Derived variable: From business address postal code.
Tax_reference_number_CIT	QGBXJCJGAZ	Tax reference number of employer.
Multiple_firms_flag	0	
Number_of_PAYE_numbers	2	Number of different PAYE reference numbers for that one employer.

Source: Authors' own illustration of the Employment panel.

Income_panel: Panel where each row represents one individual per year. Provides a view of total income and total tax paid per individual. Consists of 108,648,162 rows and 19 columns.

Variable name	Example	Description
ID_d	'abcdek'	Derived variable: Refer to IDs_panel.
Tax_year	2017	Year in which the tax year ends. For example, 2017 refers to the 2016/17 tax year which runs from 1 March 2016 to 28 February 2017.
Date_of_birth	1970/01/01	Specific day, month, and year.
Age_d	46	Age at end of tax year.
ITR12_taxable_income	165440	Taxable income directly from ITR12.
ITR12_tax_liability	16275	Tax liability from ITR12.
ITR12_MTC	0	Medical tax credit from contributions to medical aids from ITR12.
ITR12_MTC_expenses	0	Medical tax credit due to medical expenditure from ITR12. Data in this version of the panel only includes medical tax credits from expenses for the years 2015/16 and 2016/17.
IRP5_PAYE_d	16850	Derived variable: Withheld taxes from IRP5 (excluding SDL and UIF and lump sums), aggregated by individual across certificates.
IRP5_lump_sum_tax_d	0	Derived variable: Withheld lump sum taxes in IRP5, aggregated by individual across certificates.
IRP5_MTC_d	0	Derived variable: Medical tax credit claimed in IRP5, aggregated by individual across certificates .
Gross_income_d	202850	Derived variable: Sum of amounts in 'Business income', 'Normal income', 'Allowance', 'Fringe benefit', 'Lump_sum_income', 'Lump_sum_retirement', 'Investment_income' and 'Activity_income' from Category_d in Source_code_panel, per individual per year where Final_d equals 1.
Exempt_income_d	0	Derived variable: Sum of amounts in Gross_income_d which are non-taxable, i.e. where Taxable_d is 0.
Income_d	202850	Derived_variable: Gross_income_d - Exempt_income_d.
Deductions_d	37410	Derived variable: Sum of 'Deductions' category where Final_d equals 1.
Lump_sum_retirement_d	0	Derived variable: Sum of 'Lump_sum_retirement' category where Final_d equals 1.
Taxable_income_d	165440	Derived variable: Income_d – Deductions_d – Lump_sum_retirement_d. Lump sums on retirement are taxed according to different lump sum tax tables.
MTC_d	0	Derived variable: Final medical tax credit amount. Takes the maximum of medical tax credits from contributions from either the ITR12 or IRP5 returns and adds medical tax credits from expenses.
Tax_liability_d	16275	Derived variable: If there was an ITR12 return, uses ITR12_tax_liability, otherwise uses IRP5_PAYE_d. Excludes lump sum taxes on retirement.

Source: Authors' own illustration of the Income panel.

It is envisaged that the panel is updated twice a year—once in March/April after the submission of the ITR12 returns at the end of January (e.g. next update will be version 2019_1), and again in July/August after the submission of the IRP5 returns at the end of May (2019_2).

Details of the source codes and their descriptions can be found in the guides from SARS. The IRP5 guide can be found here, while the ITR12 guide can be found here, and a lookup service for source codes can be found here. An additional table named 'Source_codes' is included in the SQL database that provides a longer description of each source code available.

Examples of using the panel in R:

Most research will require summaries or subsets of the information from the different panels. It is recommended that aggregations, subsets, and joins are done in the database before any data is transferred to RAM (loaded into R or STATA) as it is much quicker than loading the data into RAM first. Some examples are shown below— see here for some other use cases.

```
# Examples using individual level panel (in R)
library(DBI)
library(dplyr)

# Connect to SQL database
con <- dbConnect(odbc::odbc(), "Individual_panel_2018_2")

# Example to show total tax paid per year from Income_panel
tbl(con, "Income_panel") %>%
  group_by(Tax_year) %>%
  summarise(Revenue = sum(Tax_liability_d, na.rm = TRUE))

# Example to get individuals with provident fund contributions from 2016/17 (Sourc
e_code_panel),
# sum up the total provident fund contributions per individual,
# join with their taxable income, tax liability and age (Income_panel) and
# then sum amount of employment tax incentive claimed for them, if available (Empl
oyment_panel)
# and import final subset into R
Test_data <- tbl(con, "Source_code_panel") %>%
  filter(Tax_year == 2017, Source_code == 4003) %>%
  group_by(ID_d) %>%
  summarise(Amount == sum(4003)) %>%
  inner_join(tbl(con, "Income_panel") %>%
               filter(Tax_year == 2017) %>%
               select(ID_d, Taxable_income_d, Tax_liability_d, Age_d),
             by = "ID_d") %>%
  left_join(tbl(con, "Employment_panel") %>%
              filter(Tax_year == 2017, ETI > 0) %>%
              group_by(ID_d) %>%
              summarise(ETI = sum(ETI, na.rm = TRUE)),
            by = "ID_d") %>%
  collect()
```