WIDER BACKGROUND NOTE | 2/2021

Building tax data for research

The South African experience

Amina Ebrahim[1] and Aalia Cassim[2]

June 2021

wider.unu.edu

Abstract: Countries need data and evidence to create, amend, and evaluate policy. South Africa has been at the vanguard of data collection in sub-Saharan Africa with strong and long-standing institutions collecting data for research purposes. Making tax administrative microdata (henceforth, tax data) available for research purposes places South Africa at the forefront of big data research for development and puts the country in a novel position relative to other developing countries. However, making the tax data available was not without its challenges as it required alignment between government departments, resources to be made available in a constrained fiscal environment, and legislation to align.

Key words: data collection, tax data, administrative data, South Africa

JEL classification: C81, H2

Acknowledgements: We gratefully acknowledge the research assistance provided by Grace Bridgman.

Note: On 6 July 2021, some hyperlinks and references were corrected.

[1] UNU-WIDER; [2] National Treasury, Pretoria, South Africa, corresponding author: aalia.cassim@treasury.gov.za

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

Information and requests: publications@wider.unu.edu

https://doi.org/10.35188/UNU-WIDER/WBN/2021-2

United Nations University World Institute for Development Economics Research

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

The United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

Introduction: data for development

Globally, one of the key factors associated with increasing use of data to inform policy-making has been the increased availability of new administrative data sources. In 2014, the National Treasury of South Africa (NT) pioneered the development of new tax administrative datasets to be used for policy-relevant research. These data have enabled, and continue to enable, several new avenues of research. In particular, the impacts of tax policy, industrial policy, and labour regulation have been considered.

In this note we share background on the South Africa tax data experience. We start by providing the context and setting of how the tax data came to be. Then, we share the lessons learnt in making the tax data available for research and conclude with some of the possibilities for the future of the tax data in South Africa and elsewhere.

The vision

There has been a steady increase in the fraction of published academic papers in leading economics journals using administrative data (see Chetty 2012). As data protection legislation develops alongside data protection technology and computing capacity, so has the access to administrative data. Increased researcher use of administrative data has had a dramatic improvement on the quality of research evidence available to policy makers.

There are many developed countries that make tax data available for research, such as the United States, New Zealand, the United Kingdom (UK) and Scandinavian countries, and a growing number of developing countries such as Uganda, Senegal, Rwanda, Brazil, Pakistan, and a few others. The level of access varies between countries. Countries either make data available online, hand over de-identified data to researchers, provide only synthetic data, or, as is the case in the UK and South Africa, provide the data onsite in a data lab.

In South Africa, traditionally, public policy research has been based on datasets produced by the South African Reserve Bank, Statistics South Africa (Stats SA), and independent data providers, such as the Bureau for Economic Research at Stellenbosch University, Quantec, and DataFirst at the University of Cape Town. Prior to the introduction of the data lab at the NT, the only other secure data facility used for economic research was the secure data centre at DataFirst.

Collecting and sharing data is not a new phenomenon in South Africa. Seen in the context of nationally representative surveys such at the National Income Dynamic Study (NIDS) and the General Household Survey (GHS), it may have been just a matter of time before the tax data in South Africa was made available. This does not take away from the serious effort mounted to make new datasets available, but points to the maturity of research efforts in South Africa.

The main reason for collecting new data is to understand a particular context where data was either of poor quality, scarce, or did not previously exist. Before the tax data was made available, there was no, or little, data on firms in the South African economy available. Many research studies made use of sector-level data, losing important nuance about firms in the process. Stats SA conducts enterprise-level surveys which are drawn from a population of tax records on VAT registered firms. The Quarterly Employment Statistics (QES) surveys collect information regarding the number of employees and total wage bill in each firm—however, the data has not been made publicly available and suffers from non-response.

In terms of labour data, survey datasets such as the Quarterly Labour Force Survey (QLFS), the October Household Surveys (OHS) and NIDS, have provided deep insights into the labour market. However, they suffer from issues of missing data, attrition, and non-response. The tax data improves on some of the shortcomings of the survey data but is limited to a study of the formal sector. Some of the advantages of tax data are discussed below:

Data quality: higher quality information due to low levels of missing data or attrition. The digital collection of data at point of origin, and data input automation is becoming the norm which improves accuracy over self-reported data. The data can measure certain features objectively for example, where information is legally required. This can avoid social desirability or recall biases of survey data.

Time span: longitudinal tracking over long periods. Routine collection of tax data means the data is more representative and may solve an Achilles' heel of many potential surveys and experiments: attrition.

Depth: samples sizes are large, which means that the data captures the breadth of formal economic activity across several useful characteristics as the forms are long and detailed. The size of administrative datasets can make it possible to run experiments that have more treatment arms without loss of statistical power and to detect effects that are small or heterogeneous between groups.

Costs: low cost of collecting data. Tax information is provided to the revenue authority as part of a tax declaration.

Novel data: These data allow for a better assessment of several firm and individual characteristics that previously researchers were unable to study robustly, for example wealth inequality (Chatterjee et al. 2020), firm-level implementation of the minimum wage (Piek et al. 2020), and job and firm characteristics of labour brokers (Cassim and Casale 2018).

Tax data improves on survey data in many aspects but suffers its own issues—ultimately the datasets are complementary.

How did it happen?

The development of the tax data in South Africa has taken places through a partnership between the National Treasury, the South African Revenue Service (SARS), and UNU-WIDER. The initial partnership, the Regional Growth and Development in Southern Africa project, ran between 2014–17. This is where access to the tax data was initiated. The scope of the project, and the data that was made available, expanded in 2018 under the Southern Africa – Towards Inclusive Economic Development (SA-TIED) programme, which will run until June 2021 and is an expanded partnership that now includes also the International Food Policy Research Institute; The Delegation of the European Union to South Africa; The Department of Planning, Monitoring, and Evaluation; The Department of Trade, Industry, and Competition; and Trade and Industrial Policy Strategies.

2014: Tax data access starts small, as a pilot. Three computers made up the data lab and a team of young data scientists and researchers developed the tax data for research for the first time.

2015: News spreads of the possibilities that the tax data presents and with that a rapid increase in demand for access to the tax data. Two additional computers are purchased, and data access is expanded.

2016: Datasets are developed hand-in-hand with policy makers and academic researchers. The first set of research papers are published.

2017: The demand for data access far exceeds the computers in the data lab. The National Treasury funds the National Treasury Secure Data Facility (NT-SDF) on the basis that increased security is required given the nature of the data available and better computing power is required to improve processing capacity to work with large data.

2018: Full time staff are hired to further develop the data, build, and manage the secure data facility.

2019: NT-SDF opens to the research community. There are large increases in data usage, new datasets are released, and more staff join to improve the quality of the available data.

2020: Expansion of secure data facility supported by UNU-WIDER. Research assistants provide remote assistance due to the pandemic and data consultants begin to support the team which prepares tax data extractions at SARS.

What has the development made possible in terms of research?

It is now possible to do types of research not possible seven years ago. These developments have enabled many new avenues of research, such as how behavioural factors affect optimal policy design, how to credibly evaluate long-run effects of landmark social programmes and tax incentives, or how to make changes to these programmes informed by a better understanding of the levers of impact. In the process, they have dramatically improved the quality and breadth of evidence used to inform policy.

Examples of policy-relevant research findings

Labour markets

- South Africa's key job creation policy for youth, the Employment Tax Incentive, has had limited impacts on the employment of youth (Ebrahim et al. 2017; Ebrahim and Pirttilä 2019; Bhorat et al. 2020)
- The amendments to South Africa's overarching labour legislation to protect temporary workers improved working conditions for those that remained employed; however, this was potentially at the expense of the loss of employment for a large number of workers (Cassim 2020).
- The large increase in agricultural minimum wages in South Africa resulted in decreased employment by 14 percentage points following the minimum wage increase (Piek et al. 2020).

Tax policy

- The graduated corporate tax rate was found to not necessarily be effective in encouraging economic activity due to evidence of tax avoidance or tax planning (Boonzaaier et al. 2019).
- The take up of research and development incentives remains among larger and older firms despite high returns. Improving governance and accessibility of the incentive is required to have a wider impact on firms (Steenkamp et al. 2018).

Industrial policy

- Chinese competition has negatively impacted on South African firms and those that invest in capital equipment, innovation, and training are more resilient to competition (Torreggiani and Andreoni 2019).

Competition policy

- Average mark-ups across the economy appear to have risen between 2010–14. Larger firms, higher-intensity exporters, and firms with greater sales shares charge higher markups than comparator firms in South Africa, even after controlling for efficiency (Dauda et al. 2019).
- Studying the relationship between product market competition and labour market outcomes in South Africa, higher employment concentration in high-markup sectors is associated with higher unemployment and lower likelihood of transitions from unemployment to employment (Amodio et al. 2020).

Value beyond research

A spillover effect of the development of the data facility at the National Treasury has been the opportunity for young economists and data scientists to get involved in making tax data available for research purposes. The skills acquired traverse those used in big data analytics and academia which opens new opportunities for research to answer questions for which little information previously existed by using administrative data. Research assistants employed at the lab have gone on to work in national statistics departments and to pursue PhDs in economics and data science.

Several South African and international master's and PhD students have used the novel data in their degree training. This includes staff from the National Treasury and SARS. In addition, new training programmes have been made available to SARS, NT, and the team running the secure data facility. These include training in Stata, R, microdata curation for secure data centres and synthetic data preparation.

Lastly, the NT-SDF itself acts as a meeting point for local and international researchers and policy makers working on the data, who develop synergies between one another that contribute to generating new knowledge. More formally, workshops have been run to share knowledge and find solutions to critical questions that remain on the data.

Lessons

Research which has been done using the tax data has increasingly been drawn on by policy makers, for example, in examining the impact of the employment tax incentive (Ebrahim and Pirttilä 2019), the minimum wage (Piek et al. 2020), a better understanding of wealth inequality (Chatterjee et al. 2020), the returns to the R&D incentive (Steenkamp et al. 2018), and how small businesses structure their businesses in light of tax thresholds (Boonzaaier et al. 2019). However, the process in making the data available was not without its challenges and below we draw on some of the lessons learnt over the past few years.

First, building a data centre is costly as sufficient processing power is required and it is important to get the technical configurations correct. If done right, additional processing capacity can be easily added to the current system and the data centre can last for many years, making a significant contribution to both the academic and policy community.

Second, the South African experience has shown that it is possible to make sensitive tax data available for research without compromising the anonymity of firms or individuals. Many processes are in place at SARS and the NT-SDF to protect the data. These include non-disclosure agreements between the NT and the researcher, and checking of output before sharing with researchers. This bodes well for the possibility of using other administrative datasets in the future and initiating similar tax projects in other countries.

Third, development of tax data takes time, and a slow and steady approach is preferable. The NT-SDF made data available early on to researchers and—while a lot of time was taken to construct the CIT-IRP5 panel—there are still several issues that remain in the data. For example, certain variables are not well populated and there are inconsistencies in how firms are classified across the panel. Researchers working on the tax data then became key resources to assist with improving the quality of the panel. In hindsight, users of the data may have been better served if datasets were better prepared before granting access to a broad group of researchers. Early access to the data for some researchers can help with data development and was required for specific cases, such as analyses of key policy decisions on the employment tax incentive and national minimum wage where research was presented at the National Economic Development and Labour Council.[1] A more strategic pacing and sequencing in granting access to the data could have better served the user community.

Fourth, there has been high turnover of staff supporting work on the tax data, which means that institutional memory is often lost. Going forward, building technical capacity across many individuals within a number of institutions will be critical to the sustainability of the NT-SDF. Academic capacity in the data science and economics fields is required to analyse the tax data, but in addition to this, information and communications technology (ICT) capacity and efficient administrative skills are required to run and manage such a facility.

Lastly, making administrative data available requires 'buy-in' from government departments and agencies and—given that making data available is often outside of the core mandate of the entities—the process of securing it requires patience and perseverance. The value of making

[1] The National Economic Development and Labour Council is the forum through which government, business, labour and community organisations co-ordinate on issues related to the economy and development.

administrative data available for research is not always clear to all stakeholders. This means that government departments will need support to develop and implement such initiatives.

Future of the data

In the short term, it is expected that there will be further improvements to the quality of the data available. Automation in tax form completion will translate into fewer data errors and more reliable data.

Researchers can expect to see new and additional guides available, a typical shortcoming of administrative data. Stats SA recently agreed to host the metadata for the tax data, and it is expected that this collaboration will grow.

There is a clear demand for training and capacity building on the use of tax data. The next phase will be to develop training courses based on the tax data to further increase the usefulness of the data to researchers within government and to young academics.

In the medium to long term there are plans to include new sources of administrative data that can further deepen the scope of research possibilities. For example, unemployment insurance fund information or education data could compliment the current data.

The large volume of such data also makes it much more amenable to cutting-edge analysis methods like machine learning, allowing for new classes of insight and inference such as artificial intelligence.

Tax information is often only required to be reported in yearly frequencies making it challenging to use for analysis of current or ongoing crises, such as the COVID-19 pandemic. Recent work has used the tax data to simulate the impact of COVID-19 on formal firms in South Africa (Lees et al. 2020). It is expected that the research from the tax data will be an important source of evidence about reactions to the pandemic, which be useful in pointing out policies that did (or did not) work in the case of future shocks.

References

Amodio, F., M. Di Maio, Y. Li, and P. Piraino (2020). 'Product Market Competition and the Labour Market: Evidence from South Africa'. WIDER Working Paper 2020/39. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/796-5

Bhorat, H., R. Hill, S. Khan, K. Lilenstein, and B. Stanwix (2020). 'The Employment Tax Incentive Scheme in South Africa: An Impact Assessment'. DPRU Working Paper 202007. Cape Town: Development Policy Research Unit (DPRU), University of Cape Town. Available at: http://www.dpru.uct.ac.za/sites/default/files/image_tool/images/36/Publications/Working_Papers/DPRU%20WP202007.pdf (accessed June 2021).

Boonzaaier, W., J. Harju, T. Matikka, and J. Pirttilä (2019). 'How Do Small Firms Respond to Tax Schedule Discontinuities?: Evidence from South African Tax Registers'. *International Tax and Public Finance,* Public economics and development action (special issue), 26(5): 1104–36. https://doi.org/10.1007/s10797-019-09550-z

Cassim, A. (2020). 'The Impact of Employment Protection on the Temporary Employment Services Sector: Evidence from South Africa Using Data From Tax Records'. WIDER Working Paper 2020/79. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/836-8

Cassim, A., and D. Casale (2018). 'How Large Is the Wage Penalty in the Labour Broker Sector?: Evidence for South Africa Using Administrative Data'. WIDER Working Paper 2018/48. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2018/490-2

Chatterjee, A., L. Czajka, and A. Gethin (2020). 'Estimating the Distribution of Household Wealth in South Africa'. WIDER Working Paper 2020/45. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/802-3

Chetty, R.. (2012). 'Time Trends in the Use of Administrative Data for Empirical Research'. NBER Summer Institute, July 2012. Available at: http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf (accessed June 2021).

Dauda, S., S. Nyman, and A. Cassim (2019). 'Product Market Competition, Productivity, and Jobs: The Case of South Africa'. Policy Research Working Paper 9084. Washington, DC: World Bank. http://hdl.handle.net/10986/33053

Ebrahim, A. M. Leibbrandt, and V. Ranchhod (2017). 'The Effects of the Employment Tax Incentive on South African Employment'. WIDER Working Paper 2017/05. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2017/229-8

Ebrahim, A. and J. Pirttilä (2019). 'Can A Wage Subsidy System Help Reduce 50 Per Cent Youth Unemployment?: Evidence From South Africa'. WIDER Working Paper 2019/28. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2019/662-3

Lees, A., G. Mascagni and M. Kilumelume (2020). 'Simulating the Impact of COVID-19 on Formal Firms in South Africa'. MTI Practice Note 9K. Washington, DC: World Bank. http://hdl.handle.net/10986/35052 (accessed June 2021).

Piek, M., D. von Fintel, and J. Kirsten (2020). 'Separating Employment Effects into Job Destruction and Job Creation: Evidence from a Large Minimum Wage Increase in the Agricultural Sector Using Administrative Tax Data'. WIDER Working Paper 2020/51. Helsinki: UNU-WIDER.https://doi.org/10.35188/UNU-WIDER/2020/808-5

Steenkamp, A., M. Schaffer, W. Flowerday, and J.Gabriel Goddard (2018). 'Innovation Activity in South Africa: Measuring the Returns to R&D'. WIDER Working Paper 2018/42. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2018/484-1

Torreggiani, S., and A. Andreoni (2019). 'Dancing with Dragons: Chinese Import Penetration and the Performances of Manufacturing Firms in South Africa'. WIDER Working Paper 2019/63. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2019/697-5