Guidelines to authors for compiling their data appendix and citing the tax data

October 2021

In 2019 the National Treasury Secure Data Facility (NT-SDF) started recommending that researchers include a data appendix in their research paper outlining their use of the administrative tax microdata available in the NT-SDF. The data appendix is like any other appendix, it provides more detailed information on a specific aspect of the research, at the end of your research paper. The data appendix provides the necessary information for replication of research in the NT-SDF and highlights any specific data cleaning or manipulation decisions. The data appendix will assist in the consistent use of the tax data across research projects so that future researchers can learn from previous experiences solving challenging data issues or improving on methods to create new useful variables.

Recommended items to include in the data appendix:

- Place of access: NT-SDF or SARS
- Name of the dataset used: SARS-NT panel, IRP5, Individual Panel, etc.
- Version of the dataset used: version 1, version 3.5, etc.
- Name and version of software used as well as any additional packages: Stata, R, python, etc.
- Period the data was accessed: date of first access to the last date of access.
- List of variables used: Include the list of variables from these datasets as well as those created or derived.
- How the data was cleaned: Describe important aspects of data cleaning, for example, dropped all dormant firms, keep observations with earnings income, data imputation, etc.
- Disclosure statement: Indicate that access was provided under a non-disclosure agreement, that your output was checked so that no firm or individual would be compromised, and your results do not represent any official statistics (NT or SARS) in a similar way that the views expressed in your research are not necessarily the views of the National Treasury or SARS.

Budlender and Ebrahim (2021) provide a good example of a data appendix that meets these requirements.

We encourage researchers to include similar information for any additional datasets used.

How to cite the data and data guides:

To encourage good research practice, it is recommended that researchers cite the use of the tax data, and any other data, in the research paper: (i) reference the data in your data section and as the source in tables and figures, (ii) cite the data in your reference list.

An important part of using the data is understanding how to cite it correctly. There are various ways to cite datasets. Below is a suggested format for the tax data:

Name of producer, date of distribution. Dataset name and years of coverage [dataset]. Version number. Place of production: Producer [producer], date of production. Place of distribution: Distributor [distributor], date of distribution.

Below we have examples of how to cite the data and the papers that describe the data:

a) Datasets

In text: National Treasury and UNU-WIDER (2019).

Reference list:

National Treasury and UNU-WIDER (2019). 'CIT-IRP5 Firm-Level Panel 2008–2017 [dataset]. Version 3.4'. Pretoria: South African Revenue Service [producer of the original data], 2018. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2019.

A full list of data references (for the most recent versions) of all datasets is included at the end of this document.

b) Working papers related to the data

Pieterse, D., F. C. Kreuser, and E. Gavin (2016). 'Introduction to the South African Revenue Service and National Treasury Firm-Level Panel'. WIDER Working Paper 2016/42. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2016/085-0

Ebrahim, A., and C. Axelson (2019). 'The Creation of an Individual Panel Using Administrative Tax Microdata in South Africa'. WIDER Working Paper 2019/27. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2019/661-6

Kerr, A. (2020) 'Earnings in the South African Revenue Service IRP5 Data'. WIDER Working Paper 2020/62. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/819-1

Budlender, J., and Ebrahim A. (2020). 'Industry Classification in the South African Tax Microdata'. WIDER Working Paper 2020/99. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/856-6

Kilumelume, M., H. Reynolds, and A. Ebrahim (2021). 'Identifying Foreign Firms and South African Multinational Enterprises'. WIDER Technical Note 2021/1. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-1

Ebrahim, A., C. F. Kreuser, and M. Kilumelume (forthcoming 2021). 'The Guide to the CIT-IRP5 Panel Version 4.0'. WIDER Working paper 2021. Helsinki: UNU-WIDER.

Ebrahim, A., C. Axelson, D. Brink, and G. Bridgman (forthcoming 2021). 'A Guide to the Individual Panel Version 4.0'. WIDER Technical Note 2021. Helsinki: UNU-WIDER.

Stata autogenerate code/Template

The program library on the server includes a Data Appendix template that is available to researchers to generate Open XML, Markdown, Latex, or PDF file that can be added to your research paper. The template can be customized.

More info

If researchers have created new variables that will be useful to other researchers, the NT-SDF will consider incorporating the new variables in the next version of the panel and make these available to all researchers going forward. New datasets created from the tax data can also be contributed in the same way and made available to other researchers.

Reporting issues in the data remains to be recorded on the Problem Sheet on the NT-SDF server, and the data team will resolve issues and report back through that mechanism. Please also send an email to sa.datateam@wider.unu.edu with the subject "Data error reporting" so that the team can work on these issues even when not physically present in the lab. If researchers have unearthed an issue in the data that relies on a subjective decision and have figured out how to resolve it, the

team encourages researchers to describe the issue in the Problem Sheet and the solution for future researchers to follow.

For researchers contracted through UNU-WIDER, please speak to the UNU-WIDER project assistant and relevant data lab staff regarding ethics approval for research conducted using the tax data. do-files required as per the UNU-WIDER contract will be placed in the program library folder on the NT-SDF server. The do-file will be available to researchers using the data lab. We encourage researchers to add a README file to explain the use of the do file shared.

Lastly, please ask the NT-SDF team for assistance with citing the datasets, particularly older datasets not listed below.

Reference List for datasets available at NT-SDF

National Treasury and UNU-WIDER (2019). 'Individual Panel 2011–2018 [dataset]. Version 2019_1'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2019.

National Treasury and UNU-WIDER (2021). 'CIT-IRP5 Firm-Level Panel 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2021.

National Treasury and UNU-WIDER (2020). 'CIT Firm-Level Panel 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'IRP5 Worker-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'IRP5 Firm-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'Customs Transaction-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'Customs Firm-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'VAT Firm-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'VAT Transaction-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

Budlender, J., and A. Ebrahim (2019). 'Industry Variables Supplemental Data [dataset]. Version 1.0'. Pretoria: National Treasury and UNU-WIDER [distributor of the dataset], 2019.

Brink, D., and M. Kilumelume (2021). 'Deflator Variables Supplemental Data [dataset]. Version 1.0'. Pretoria: National Treasury and UNU-WIDER [distributor of the dataset], 2021.