UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2020/99

Industry classification in the South African tax microdata

Joshua Budlender[1] and Amina Ebrahim[2]

August 2020

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

Abstract: This paper documents the industry classification variables in the anonymized tax microdata available for research at the National Treasury Secure Data Facility in Pretoria. It discusses how the variables in the data are related to the raw records captured in various tax forms and outlines the various industry classification systems. We discuss and present a recoding by which idiosyncratic industrial classifications are transformed into one comparable system. For each of the industry variables, we examine its internal consistency (across years and other industry variables), external validity (by comparison with other data sources), and completeness (for important subsets of the data). On this basis, we suggest a set of 'best' industry variables for researcher use based on the underlying raw variables, while noting potential issues with the major options.

Key words: administrative microdata, industry classification, ISIC 4, SIC 5, SIC 7

JEL classification: C8, Y10

Acknowledgements: We would like to thank Carol Newman (UNU-WIDER), John Rand (UNU-WIDER), Michael Kilumelume (UNU-WIDER), C. Friedrich Kreuser, Murray Leibbrandt (UNU-WIDER), Joseph Lukhwareni (Statistics South Africa), Catherine MacLeod (National Treasury), M. Nicholas Makgopa (SARS), Andrew Nell, Hayley Reynolds (National Treasury), Alexius Sithole (SARS), and participants of the SA-TIED Metadata work-in-progress meeting for assistance, comments, and suggestions at various stages of this project. Any errors of course remain our own. Joshua Budlender acknowledges support from UNU-WIDER for this project.

[1] University of Massachusetts, Amherst, corresponding author: jbudlender@umass.edu; [2] UNU-WIDER

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

ISSN 1798-7237 ISBN 978-92-9256-856-6

https://doi.org/10.35188/UNU-WIDER/2020/856-6

Typescript prepared by Joseph Laredo.

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.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

The SARS-NT panel contains four usable sources of industrial classification information. However, these sources vary in their firm coverage, comprise different industry classification systems, and often assign conflicting industries to any given firm. Researchers frequently do not know the origins of the different classification systems and must make relatively arbitrary decisions about which industry variable to use for their analysis. As part of the larger effort to document and systemize the SARS-NT panel, this paper reviews the different industry classification variables and attempts to create a set of 'best' variables for researcher use.

Our preferred industry variable is the 5-digit SIC 7 ITR14 Main Industry Code, with backwards and forwards imputation to assign this industry code to firms missing industry information prior to 2013, when the current Main Industry Code was introduced on the ITR14 form. This option is not without problems, the most serious being insufficient industry switching over time and the creation of a sample selection issue when it comes to pre-2013 firms (only firms that survive into 2013 will have industry information). We thus present an additional alternative approach, using predominantly CIT Profit Codes converted into SIC 5 codes, which is more internally consistent but is not as good a match with the external sources as the ITR14 Main Industry Code.

This effort has proved to be a larger task than anticipated. Initially, it was hoped that we could work with the industry variables in the existing panel, which a previous team of researchers had made substantial efforts to clean. However, these previously created variables lacked documentation of the procedures used in their creation, and it was difficult to back out the underlying records. We therefore preferred to start afresh and create industry classification variables from the raw data extraction ourselves, and then review and adjust these variables. During this process we kept in mind that the panel would be updated in the future using new extractions of tax records from SARS, and we attempted to design a system for creating 'best' industry variables that would be sustainable over time.

Throughout this paper we make a distinction between industry classification *variables*, which are industry information attached to each firm in the dataset, such as the ITR14 Main Industry Code or the VAT Activity Code, and industry classification *systems*, which are the coding schemes used to associate a particular industry with a given (alpha)numeric code, such as Statistics South Africa's Standard Industrial Classification revision 5 system (SIC 5) or the United Nations' International Standard Industrial Classification revision 4 system (ISIC 4).

In Section 2, we briefly give the history of the SARS-NT panel and the panel creation process. In Section 3 we review the main industry classification systems that were available or considered for the data. In Section 4 we discuss the different industry classification variables available in each dataset contributing to the SARS-NT panel. Starting with the Company Income Tax (CIT) data, we move on to the Value Added Tax (VAT), employment (IRP5), and Customs datasets. In Section 5 we explain how we convert the industry variables into a common industry system (the Statistics South Africa SIC 5 system), and in Section 6 we examine the internal and external consistency of each industry variable. We perform this last task by comparing aggregate industry-specific time series information according to each variable with equivalent statistics from the SARS Tax Statistics, and Statistics South Africa's Quarterly Labour Force Survey (QLFS), Quarterly Employment Statistics (QES), and Quarterly Financial Statistics (QFS). On the basis of the analysis, in Section 7 we discuss options for 'best' industry variables and present our preferred approaches, testing again for their internal and external consistency. Section 8 concludes.

2 The SARS-NT Panel

The SARS-NT panel is made up of a few distinct sources of South African tax data. Specifically, there are CIT records captured from IT14 and ITR14 forms, VAT data from VAT-registered firms, employment data from IRP5 and IT3a forms, and Customs data from traders. What we will call the 'old panel' was created from these datasets by UNU-WIDER, National Treasury, and South African Revenue Service (SARS) researchers in 2015. The authoritative reference to these data is Pieterse et al. (2018). The researchers involved in the creation of the old panel undertook a mammoth task. Without any comparable South African researcher-facing administrative data to guide them, they managed masses of (frequently inconsistently coded) raw data across a complex web of different forms, cleaned and created variables, developed systems for reconciling the different data sources, and produced a novel and uniquely valuable dataset and made it available for researcher use. The unprecedented nature of the task necessarily means that there was significant 'learning by doing', and the simultaneous demand for use of the dataset by researchers while it was being constructed meant that frequent revisions and adjustments to the data were unavoidable. With the difficult dataset creation process taking priority over documentation and systematization, the end result was a uniquely valuable South African dataset, but one that is inconsistently documented and difficult to reverse-engineer for researchers' intent on uncovering data-handling decisions. While Pieterse et al. (2018) explain most of the key decisions made in the panel creation, the origins of large parts of the panel nonetheless remain opaque, especially as many of the people involved in the initial dataset creation have since moved on to different roles.

Since the initial panel creation in 2015, the team managing the panel has received regular new raw data—so-called 'extractions'—from SARS. The 'new panel', which incorporates both the old and new SARS data, broadly follows the structures and programs used for the old panel, but the lack of clear documentation and programming workflow from the old panel has hindered efforts to ensure consistency. In some cases, such as where researchers have uncovered systemic problems or where lack of documentation creates too much uncertainty about what particular variables mean, the new panel does not attempt to recreate old panel classifications but instead starts from a blank slate. The industry classification is one such area, with inconsistency in the old panel industry classifications having been noted by Budlender (2019), and the origins and coding of some old panel industry variables being unclear. This affects not just any research that explicitly uses industry classification, but also any research that uses producer price index (PPI) adjustments, as these indices are industry-specific. The purpose of this paper is to carefully develop a protocol and best practice for industry classification, which will be documented for researchers and should fit seamlessly into the new systematized and documented workflow currently being developed for panel updating.

3 Industry classifications systems

For readers requiring further background, we briefly review the usual structure of industry classification systems in Appendix A. Readers requiring additional information are referred to Statistics South Africa (2012).

3.1 Specific industry classification systems

Many different industry classification systems exist. The International Standard Industrial Classification (ISIC) system, through its various revisions, published by the United Nations Statistics Division is the leading international standard, but national statistical agencies frequently

publish additional systems that are optimized for local conditions. This is the case in South Africa, where Statistics South Africa has published various editions of the local Standard Industrial Classification (SIC). Additionally, any entity may have its own particular industry classification system, as is the case with SARS, which implements 4-digit codes for the purposes of industry classification.

Industry classification schemes must be updated over time to reflect the changing industrial make-up of actual economies, with some industries disappearing over time and other new industries emerging. It is these changes that prompt the various revisions of the ISIC codes and editions of the Statistics South Africa SIC codes. Currently the most up-to-date ISIC code is the ISIC 4th Revision (ISIC 4), published in 2008, while the most recent Statistics South Africa SIC code is the SIC 7th edition (SIC 7), published in 2012.[1] These classifications are very similar, with Statistics South Africa adapting ISIC 4 to create SIC 7.[2] The immediately preceding ISIC revisions were revisions 3 (ISIC 3, published 1989) and 3.1 (ISIC 3.1, published 2002), while the preceding Stats SA SIC system was edition 5 (SIC 5), published in 1993 and an adaptation of ISIC 3.

While previous versions of the SARS-NT dataset included some industry variables that use the ISIC 4 system, after correspondence from other coding systems, we do not use this system ourselves for the reasons discussed in Section 5. We therefore limit our descriptions of industry classification systems below to SIC 7 and SIC 5, as well as what we call the SARS Profit Code and Activity Code systems (see below). As mentioned, ISIC 4 is extremely similar to SIC 7, so our discussion of SIC 7 covers much of what would be said about ISIC 4 in any case.

SIC 7 system

The SIC 7 coding scheme categorizes firms into a hierarchical five-level structure of mutually exclusive categories described in Table 1. The most aggregated level consists of 21 'Sections', each of which is assigned an alphabetical code. For ease of reference and consistency with previous systems we still refer to this as the '1-digit' level. The next level of disaggregation is the 88 'Divisions', which are 2-digit numeric codes, and which nest the subsequent numeric 3-digit 'Groups', 4-digit 'Classes', and 5-digit 'Subclasses' in the usual fashion.[3]

The SIC 7 system is well documented in Statistics South Africa's (2012) official SIC 7 manual, but this document is not amenable to machine reading, a significant problem for the purposes of creating industry variable labels or concordance tables. After correspondence with Statistics South Africa we received a machine-readable Excel file that contains usable SIC 7 industry labels up to

[1] Researchers frequently misconstrue the '4' in 'ISIC 4' or the '7' in 'SIC 7' as indicating the number of digits in the industrial classification. This is incorrect: these numbers indicate revision/edition numbers.

[2] These codes are almost identical up to the 4th digit level, the only differences being related to Statistics South Africa's moving 'Retail sale of automotive fuel in specialized stores' from 'Retail trade, except for motor vehicle and motorcycles' (Division 47) to 'Wholesale and retail trade and repair of motor vehicles and motorcycles' (Division 45) and a few knock-on effects. Additionally, Statistics South Africa adds 5-digit categories in the SIC 7 classification (ISIC 4 only goes up to 4-digit). A document outlining the differences (from correspondence with Statistics South Africa) is available from the authors upon request and in the NT-SDF.

[3] For example, one can tell that a firm in SIC 7 5-digit code 12345 is in Division 12, Group 123, Class 1234, and Subclass 12345, but in order to determine its Section one needs to look at documentation that divides the 88 Divisions into the 21 Sections.

the 4-digit level. Labelled SIC 7 5-digit codes are then scraped from the SARS website and merged to create a complete labelled classification.[4]

Table 1: Industry categories in the SIC 7 system

Categories	Number of categories	Level
Section	21	'1-digit-level'
Divisions	88	2-digit level
Groups	238	3-digit level
Classes	419	4-digit level
xs	>500	5-digit level

Note: Statistics South Africa (2012) does not indicate how many distinct subclasses there are in the SIC 7 system. After cleaning, we found 521 distinct SIC 7 subclasses in the ITR14 data, out of 537 subclass descriptions from SARS (some of which are not appropriate for firm-level classification).

Source: Statistics South Africa (2012).

SIC 5 system

Users of South African data will likely be most familiar with the SIC 5 classification system, as it (or some minor adaptation of it) is currently used for almost all South African economic datasets in which industries are classified, including the Quarterly Labour Force Surveys (QLFS), National Income Dynamics Study (NIDS), Quarterly Employment Statistics (QES), Quarterly Financial Statistics (QFS), Input-Output data, October Household Surveys (OHS) 1996–1998, and 2011 Census.[5]

The most aggregated level consists of 10 'Major Divisions', each associated with a single-digit code (including 0). The remaining levels, going down to a 5th-digit-level and indicated in Table 2, are all numerically nested in the usual fashion, so that given a 5-digit code one can easily read off its Major Division, Division, Major Group, Group, and Subgroup.

Table 2: Industry categories in the SIC 5 system

Categories	Number of categories	Level
Major Divisions	10	1-digit-level
Divisions	50	2-digit level
Major Groups	158	3-digit level
Groups	314	4-digit level
Subgroups	466	5-digit level

Source: authors' calculations using Statistics South Africa (1993).

[4] The complete classification is available from the authors upon request and in the NT-SDF. 5-digit SIC codes were scraped from https://www.sars.gov.za/TaxTypes/PAYE/ETI/Pages/SIC-Codes.aspx. Readers should note that an alternative source of labelled SIC 7 codes online, the Statistics South Africa 'SIC 7 Coder' available at https://apps.statssa.gov.za/Web_Sic7/Docs/sic%207%20coder.xls, is (as of 24 February 2020) plagued with significant data entry errors and should not be used without careful correction.

[5] As noted in Kerr and Wittenberg (2019), the 1999 OHS and the Labour Force Surveys (LFS) used a similar but slightly different industry categorization system. It seems that the same system is used for the General Household Surveys (GHS). It is not clear where this system comes from; the OHS 1999 and GHS 2002 documentation suggests that it is based on 'ISIC 1993', which may mean an ISIC 3 or SIC 5 classification, but the given codes match neither system.

The SIC 5 system is well documented in Statistics South Africa (1993), an interactive html manual. We scrape across the clear structure of the html file to create a usable spreadsheet with the full SIC 5 categories and descriptions.[6]

SARS Profit Codes

What we call the SARS Profit Code system is a 4-digit industry classification code apparently produced by SARS. CIT guides instruct users to choose the applicable 'profit code' from SARS 'main source of income' codes (as usual, eFilers choose from a drop-down list).[7] It is clear from examination of the CIT data that this classification refers to the 4-digit 'source codes' listed on the SARS website.[8] We refer to this as the SARS Profit Code system. However as discussed below, there is another very similar SARS 4-digit classification that is sometimes referred to as a 'source code'. What distinguishes the Profit Code system is a structure where coding is mostly distinct for profits and losses in a given industry, with even numbers indicating profits and adjacent odd numbers losses, as indicated in Table 3.[9] For the purposes of industrial classification we do not want this distinction, and so collapse the profit and loss distinctions into one category.

Table 3: Illustration of SARS Profit Code system

Code	Description	Profit/Loss
0310	Vegetable, Animal Oils & Fats	Profit
0311	Loss—Vegetable, Animal Oils & Fats	Loss
0312	Grain Mill Products	Profit
0313	Loss—Grain Mill Products	Loss

Source: authors' illustration.

The SARS Profit Code system allows aggregation of the 443 4-digit categories (sometimes called 'sub-activities') into 34 2-digit super-categories (sometimes called 'main activities'), using the first 2 digits of the 4-digit code.[10] There are thus only 2 levels of aggregation in this classification system. There is a 35th 2-digit category, which has 4-digit codes ranging from 3501 to 3534.[11] These seem to be aggregated categories such that if a firm knows it is in main activity 04 (Textiles), for example, but not which branch of textiles it should fit under, then it can give its Profit Code as 3504.

The 'profit codes' are not well documented in comparison with the Statistics South Africa classifications discussed above, and we were unable to find any discussion of their creation, history, or methodology. We can, however, create value labels for the full classification by scraping codes and descriptions from the SARS website for the 4-digit sub-activities, while manually transcribing 2-digit main activities.[12] These two-digit codes are available in Appendix B. With only two levels

[6] The complete classification is available from the authors upon request and in the NT-SDF.

[7] SARS has been using the terms 'profit code' and 'activity code' since the early 1990s, but the former are sometimes referred to as 'main industry codes' or 'source codes'—names that are also often used by SARS for very different coding systems—and the latter as 'trade classification codes' or 'main income source codes'—the second of these also being used for a very different coding system. The various appellations are discussed further below.

[8] Available at https://www.sars.gov.za/TaxTypes/PIT/Tax-Season/Pages/Find-a-Source-Code.aspx.

[9] The only exceptions to this structure are some codes in the range 3001–3025, which relate to various tax-exempt organizations, and codes greater than 3500, which are discussed below.

[10] This is 443 distinct 4-digit categories *after* collapsing the profit- and loss-specific codes into one code per category.

[11] Code 3535, which sometimes appears on source code lists, does not appear to be relevant for industry classification; it is meant to indicate 'Members of a CC/Directors of Company'. We drop it from the system.

[12] The website used for the 4-digit codes was https://www.sars.gov.za/TaxTypes/PIT/Tax-Season/Pages/Find-a-Source-Code.aspx (South African Revenue Service 2020), while 2-digit categories were manually transcribed after

of industry aggregation, very little documentation, and no use outside of the SARS CIT forms, this system has little appeal for the purposes of statistical analysis.

SARS Activity Codes

What we call the SARS Activity Code system appears to be a modification of the Profit Code classification discussed above. Like the 'profit code', it is a 4-digit code, where each 4-digit code represents a 'sub-activity', which can be aggregated into 34 2-digit 'main activities'.

The Activity Code classification system is almost exactly the same as the Profit Code system at the 2-digit 'main activity' level: the only difference is that the 26th category in the Activity Codes is 'Public administration', whereas in the Profit Codes it is 'Long-term insurers'. There are also slight differences in the description of the 30th main activity 2-digit category, for tax-exempt organizations, but it is not clear whether these differences in description imply different categories.

These differences at the 2-digit main activity level necessarily imply changes at the 4-digit sub-activity level, but these differences are quite straightforward to identify and reconcile. A more significant difference concerns the structure of the 4-digit sub-activity codes. Unlike the Profit Code system, the Activity Code system does not record profits and losses separately. When there is an Activity Code directly equivalent to a Profit Code, only the even-numbered Profit Code is used. However, categories are not always directly equivalent, as sub-activities in the Activity Code system are sometimes aggregated combinations of sub-activities in the Profit Code system. When this is the case, these 4-digit aggregated sub-activities in the Activity Codes are coded in multiples of 5. Therefore, all sub-activity codes in the Activity Codes are multiples of either 2 or 5.

Documentation of the Activity Codes is again poor. We have been unable to find any discussion of the codes, and we are not aware of any easily machine-readable official documentation linking codes and category descriptions. We did, however, find the *VAT/EMP 403 Vendors and Employers Trade Classification Guide* (SARS, no date a), which is a document providing industry codes and descriptions for the system. It is not easily machine readable for the purposes of creating a usable labelled system, and codes could be extracted only imperfectly. The document does, however, confirm the correctness of a spreadsheet included in the old panel documentation, and we use this spreadsheet to create a full, labelled, machine-readable classification of the Activity Code system.[13]

Like the SARS Profit Codes, the SARS Activity Codes have little appeal for the purposes of statistical analysis.

comparing categories at https://www.taxtim.com/za/tax-guides/definitions/business-code-table (TaxTim 2020) with the SARS Activity Codes described below.

[13] Among the documentation left by the creators of the old panel was a spreadsheet titled 'MAININCOMESOURCECODES'. We confirm that the system in the MAININCOMESOURCECODES matches the system described in the VAT/EMP 403 form mentioned above (SARS no date a), and on this basis use the machine-readable MAININCOMESOURCECODES to create our own labelled Activity Codes classification, which is available from the authors upon request and in the NT-SDF.

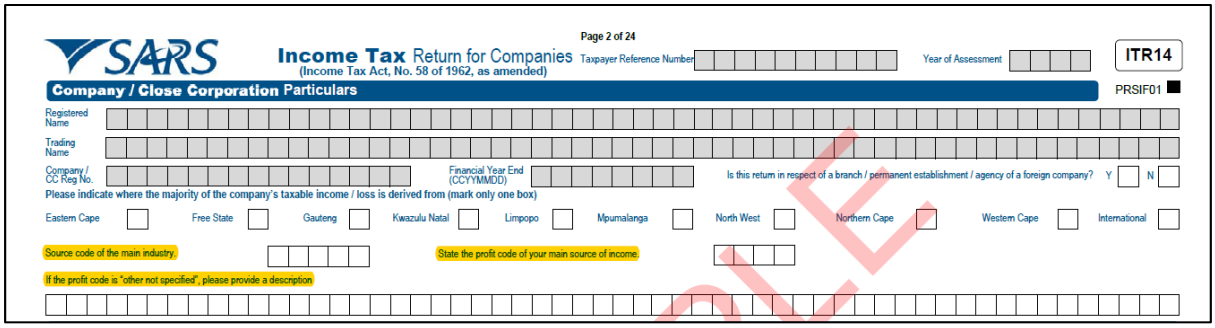
4 Industry variables

4.1 Company Income Tax (CIT) data—ITR14

The CIT data form the backbone of the SARS-NT panel. It is CIT entities that constitute 'firms' in the panel, other data being aggregated to the firm level when datasets are combined. There are two sources of CIT data in the panel: the IT14 and ITR14 forms. IT14 forms were discontinued and replaced by ITR14 forms in May 2013 (Pieterse et al. 2018). CIT records reflecting data from before the 2013 switchover are usually drawn from IT14 forms, but not exclusively, as ITR14 forms are frequently used for revisions. Records after the switchover date should all be recorded using ITR14 forms. The forms are similar and the original panel creators made significant efforts to harmonize variables across these forms, but some differences remain. Industry data is one area where there are differences. Ultimately the IT14 forms do not have a usable industry classification system except for Profit Codes, which persist in the ITR14 and are discussed below, and so discussion of the IT14 forms specifically is relegated to Appendix B.

The ITR14 form records two sources of firm industry information: a 5-digit code in response to the field 'Source code of the main industry', which we refer to as the ITR14 Main Industry Code; and a 4-digit code in response to the field 'State the profit code of the your main source of income', which we refer to as the ITR14 Profit Code. A snapshot of the industry fields on the ITR14 form is in Figure 1.

Figure 1: Income Tax Return for Companies (ITR14)



Note: the fields have been checked and are the same in the versions ITR14 v2013.0.13, ITR14 v2014.0.5, ITR14 v2015.00.28, ITR14 v2016.00.19, ITR14 v2017.00.24, and ITR14 v2018.00.05.

Source: ITR14 v2014.0.5 form, available from SARS.

Companies have access to a guide document, available on the SARS website, which explains how to complete the ITR14 form. A snapshot of this guide is provided by Figure 2.

Figure 2: Guiding text to complete the ITR14 industry fields

- **Source code of the main industry**
  - A pop-up list with all the Standard Industry Codes (SIC) will be displayed on eFiling and when the agent captures the information in the SARS branch. For non-eFilers, the Company Representative/Public Officer completing the ITR14 return manually, can access the Standard Industry Codes (SIC) booklet on www.statssa.gov.za
- **State the profit code of your main source of income**
  - A pop-up list with all the main source of income codes will be displayed to select on eFiling. If the company is dormant this field will be pre-populated with code 9994 and locked. For non eFilers that are not dormant, the Company Representative/Public Officer completing the ITR14 return manually, can access the following main source codes by entering "Find a source code" on the SARS website www.sars.gov.za.
- **If the profit code is "other not specified", please provide a description**
  - If the profit code of your main source of income ends with '98', this is a mandatory free text field (maximum length 56 blocks).

Source: External Guide: How to Complete the Income Tax Return (ITR14) for Companies IT-GEN-G01 Revision 8 (SARS, 2013: 15).

ITR14 Main Industry Code

What we call the ITR14 Main Industry Code comes from a 5-digit field (SARS 2019). CIT entities must enter the 'Source code of the main industry' on the ITR14 form, as highlighted in Figure 1. This field accommodates a 5-digit SIC 7 code. We label the 5-digit numeric codes with the SIC 7 industry classification we have created (as described above) and use the classification to generate 1-digit, 2-digit, 3-digit, 4-digit, and 5-digit labelled SIC 7 variables.

Not all of the codes that come out the ITR14 extraction from SARS fit neatly into the SIC 7 codes, and there is a need for some cleaning.[14] A substantial proportion of firm observations have missing Main Industry Codes—at least 40 per cent per year in our cleaned data. For the vast majority of firms, this is because the industry variable is either system-missing or a 3400 code in the underlying SARS extraction.[15] Many of these are dormant or otherwise inactive firms. However, we also set the industry variable to 'missing' in clear cases of data entry error, indicated by the Main Industry Code having no equivalent category in the SIC 7 coding as available from Statistics South Africa. However, this is a very small proportion of total firms—less than 0.1 per cent.[16]

[14] A case in point is that in the new extractions, substantial numbers of firms are classified with code 3400, which is not a SIC 7 category. These 3400 codes were not present in old extractions and were the cause of some alarm, SARS being unable to explain their origin. Comparison with the old data shows that almost all firms with a 3400 Main Industry Code in the new data had a system-missing Main Industry Code in old extractions, and this seems to be simply a new placeholder for missing values. We therefore replaced all 3400 codes with missing values.

[15] See previous footnote.

[16] The extremely low rate of data entry error is likely because firms using e-filing must select their Main Industry Code from a drop-down list (SARS 2013).

ITR14 Profit Code

What we call the ITR14 Profit Code comes from a 4-digit field for the 'profit code of your main source of income', also indicated in Figure 1 (SARS 2019). It uses the SARS Profit Code system as outlined in Section 3.1. While the Profit Code certainly uses a different classification system than the Main Industry Code, it is unclear whether the two codes should always reflect the same underlying feature of a firm. While we have been told that there can on occasion be legitimate differences between a firm's 'main industry' and its 'main source of income', we have been unable to determine what prompts legitimate divergence in these respects—it likely depends on what exactly defines a 'main industry', which is unclear to us.

As was the case for the Main Industry Codes, some cleaning of the Profit Codes is necessary. A Profit Code of 9994 indicates a dormant company, so we replace it with a missing value for its industry classification. We also replace Profit Codes by 'system missing' when they are clearly data entry errors, such as any that are greater than 3535, or codes that do not match the Profit Code labels we merge in. For raw records in the 3501–3534 range, we set them as missing at the 4-digit level but use this information to assign a 2-digit-level code, in line with the discussion of these codes in Section 3.1.[17] A very small proportion of firms report Profit Codes ending in two trailing zeros—such as 0400—despite no codes of this nature being in the SARS classification. For these codes, we check to see if the firm Main Industry Code concords with the implied main activity from the first two digits of the Profit Code—in this case textiles (04). If it does, we keep the 2-digit code as indicating the main activity (textiles in this example) but set the 4-digit code to missing. In cases where we have an erroneous Profit Code with trailing zeroes, like 0400, and the Main Industry Code does *not* match the implied 2-digit main activity, we set both the main activity and sub-activity to missing. This affects a very small proportion of total firms.

4.3 Value Added Tax (VAT) data

VAT is an indirect tax levied on consumption, charged at each stage of the production and distribution process. While VAT is due for all consumption of goods and services (with the exception of some zero-rated items), becoming a VAT-registered entity is mandatory only if a company's taxable supplies exceed R1 million in any 12-month period (Pieterse et al. 2018). VAT-registered entities act as vendors that collect VAT on government's behalf and submit forms at least annually if not more frequently. As mentioned earlier, firms are defined as CIT entities in the SARS-NT panel, and VAT data are merged on this basis. Multiple VAT records often exist per CIT entity, but this does not cause any problems for our purposes, as the VAT industry variables are constant within CIT entities, and so the data are collapsed at the CIT entity level. VAT information, and thus industry information from the VAT data, is available only for the subset of CIT firms registered for VAT.

The VAT data in the SARS extractions include information on two distinct industry classifications (SARS 2019).[18] However, the VAT extraction from SARS is an extraction of the VAT 201 form data, which do not have any industry classification fields. It is therefore not clear where the industry variables in the extraction come from, but we suspect that they are provided by the firm in a previously completed SARS form—such as the VAT 101 registration form. This could then

[17] The 34 2-digit categories for the similarly structured Activity Codes are listed in Table C1 in Appendix C.

[18] Only the VAT Activity Code is useful for our purposes, so it is discussed below, while the VAT Micro Sector Code is discussed in Appendix D.

prepopulate the VAT 201 form, explaining why VAT industry codes form part of the VAT 201 extraction from SARS.

VAT Activity Code

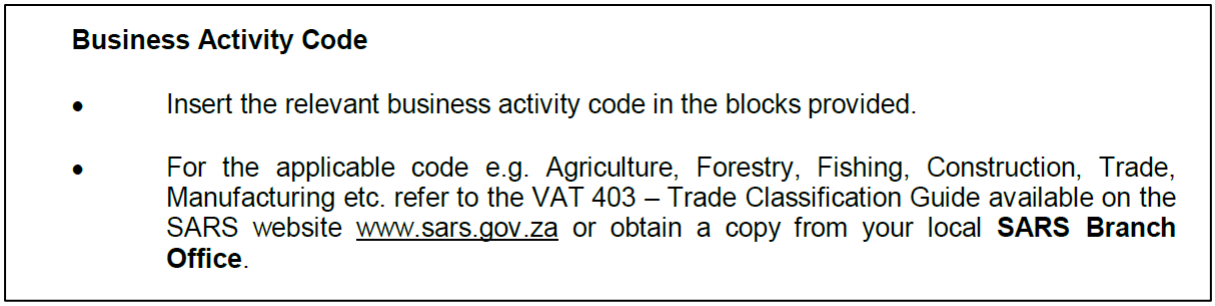
The SARS-NT data include a 4-digit 'industry classification variable', which we call the VAT Activity Code. We use this nomenclature because, on close inspection, it appears that the code derives from the VAT 101 form (used for VAT registration), and in particular from the 'business activity code' entry field (Figure 3). This code is to be selected from a list published in the *VAT/EMP 403 Vendors and Employers Trade Classification Guide* (SARS, no date a), per the instructions in the *Guide for Completion of VAT Registration Application Forms. VAT-REG-02-G01 Revision 10* (SARS no date a) (Figure 4).

Figure 1: Industry field on the VAT101 form

Source: VAT101 form.

The extremely low rate of firm industry switching in VAT Activity Codes discussed in Section 6 would seem to support the idea that this code comes from a registration process, rather than from a field filled in each year. The code is very well reported and requires very little cleaning.

Figure 2: Guide text to complete the VAT industry fields



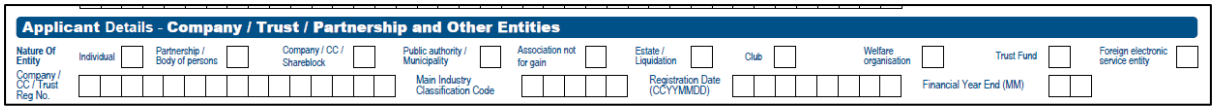
Business Activity Code

- Insert the relevant business activity code in the blocks provided.
- For the applicable code e.g. Agriculture, Forestry, Fishing, Construction, Trade, Manufacturing etc. refer to the VAT 403 – Trade Classification Guide available on the SARS website www.sars.gov.za or obtain a copy from your local **SARS Branch Office**.

Source: External Guide: Guide for Completion of VAT Registration Application Forms. VAT-REG-02-G01 Revision 10.

The VAT 101 form also asks for a 5-digit 'main industry classification code' (Figure 5), but we do not find anything analogous in the VAT extraction. Strangely, the SARS guide to the VAT 101 form (Figure 6) suggests again that the *VAT/EMP 403 Vendors and Employers Trade Classification Guide* (SARS, no date a) be used for the 'main industry classification code', but the VAT/EMP 403 guide refers only to a 4-digit code.

Figure 3: Industry field on the VAT101 form



Source: VAT101 form.

Figure 4: Guide text to entering the 'Main Industry Classification Code' on the VAT101 form

Main Industry Classification Code

- The main industry classification code refers to the main activity from which the applicant derives the majority of its business income (e.g. Agriculture, Forestry, Fishing, Construction, Trade, Manufacturing etc.) For detailed information refer to the VAT 403 – Trade Classification Guide available on the SARS website www.sars.gov.za or obtain a copy from your local **SARS Branch Office**.

Source: External Guide: Guide for Completion of VAT Registration Application Forms. VAT-REG-02-G01 Revision 10.

4.4 Employee tax data

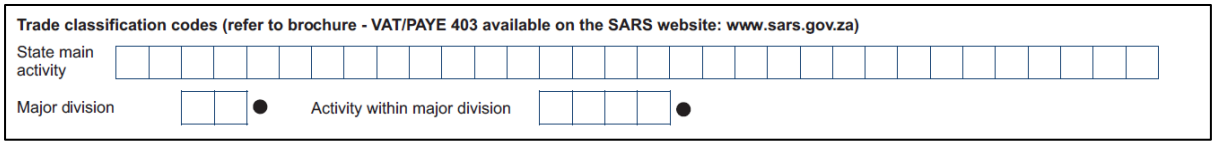
The employee tax data are structured according to job-level tax certificate records. These records are either IRP5 or IT3(a) certificates issued by employers via Pay-As-You-Earn (PAYE)-registered entities (Pieterse et al. 2018). Each PAYE-submitting entity is identified with a reference number, which is used to link the entity to a CIT firm. The process by which job-level records are merged into a firm-level database is not straightforward. There are frequently multiple records per individual, and it is not clear whether the records represent additional employment or just revisions of previous certificates. As detailed below, we draw from the methods used by the panel-updating team, which are quite similar to the method in Pieterse et al. (2018), to resolve these issues.

IRP5 Activity Codes

The employee data come with 2- and 4-digit industry codes, which we refer to as IRP5 Activity Codes[19] (SARS 2019). There has been uncertainty about the origin of these codes for some time, going back to correspondence with SARS during the construction of the old panel. These codes seem to exactly match the SARS Activity Codes system discussed above, making some allowance for minimal data entry error and confusion with SARS employment codes.[20] This suggests that the IRP5 activity codes may come from EMP 101 or 102 forms, which must be submitted by employing entities in order to either register as employers or amend details of their registration (SARS no date c).

The EMP 101 and 102 forms require 2- and 4-digit 'main activity' codes (Figure 7), and explicitly refer taxpayers to the *VAT/PAYE 403* guide to complete these fields (Figure 8).

Figure 5: Industry field on the EMP101e form



Source: EMP101e Payroll Taxes Application for registration form.

[19] In the IRP5 data this variable is called 'main income source code'. This term is confusing, so we refer to it as Activity Code instead.

[20] We are referring to the income, allowance, benefit, and deduction codes that are used in a later section of the IRP5/IT3(a) form.

Figure 6: Guide text to entering industry information on the EMP101e form

TRADE CLASSIFICATION CODES

- Refer to brochure – VAT / PAYE 403 available on the SARS website, www.sars.gov.za, to see which activity and division codes are applicable to your business.
    - For foreign diplomatic or consular mission', the major division is "26" and the activity within the major division "2605".

STATE MAIN SECTOR AND ACTIVITY

- The main sector and activity from which the applicant derives the majority of its business income must be described.

Source: External Guide: Guide for completion of Employer Registration application EMP-REG-03-G01 Revision 5 (Page 13).

If these codes do in fact come from these EMP 101 or 102 forms, this implies that they are not worker-specific codes, but instead apply to the employing entity where the worker is based. However, in the data, these codes sometimes vary across individuals within any given CIT or PAYE entity. One explanation consistent with these being EMP 101 and 102 codes is that EMP 101 and 102 employment entities may not perfectly match CIT or PAYE entities. Evidence against these being EMP 101 and 102 codes entered during firm registration is that, unlike the analogous VAT Activity Codes, which we are fairly confident derive from the VAT 101 registration form, the IRP5 Activity Codes do noticeably vary over time per firm, as discussed in Section 6. This would seem to suggest they may come from some yearly-submitted form. In short, we do not know the origin of these codes, or whether they truly reflect worker-level characteristics, or some other kind of sub-CIT-entity firm-level feature.

Whatever their origin, the variability in these codes across individuals means that developing a firm-level industry variable requires some kind of aggregation. We first follow the panel protocols that drop records that are not jobs, or that are likely to be revisions rather than distinct employment spells. We then prioritize records that are associated with some kind of employment income (as opposed to dividends or pension income, for example), and take advantage of a rough per-year 'length of employment spell' variable, which can be backed out for each record. We assign firm-level IRP5 Activity Codes according to the following process, which is repeated separately for the 2-digit and 4-digit Activity Codes:

1. For a given year, determine the total number of employment days associated with each non-missing Activity Code per firm, by summing across records.
2. Assign to the firm the non-missing Activity Code that accounts for the largest number of employment days.
    a. If there is a tie such that no single code explains most days, assign the value -9 to the firm-level industry code.
3. Generate a firm-level variable that indicates the share of firm-level employment days accounted for by the assigned Activity Code.
    a. In the case of a -9 tie, this variable indicates the share of the split, for example 50 per cent in a two-way split.
4. For CIT entities that have no observations with employment income, only assign a firm-level Activity Code if all (non-employment) records within the CIT entity have the same Activity Code. If there are no employment income records and non-employment records have conflicting Activity Codes, assign the value of -9 to the firm-level Activity Code.

a. These non-employment CIT entities will have a missing value for the 'share of employment' variable created in Step 3.

Table 4 shows an illustration of this process.

Table 4: Illustration of method of creating firm-level IRP5 Activity Codes in the IRP5 data

Job	CIT ref.	No. of employment days	Activity Code	Employment days sum over activity code	Firm-level Activity Code	Firm-level Activity Code's share of employment days
1	AA	50	2	230		
2	AA	180	2			
3	AA	365	5		5	67%
4	AA	20	5	475		
5	AA	90	5			
6	BB	365	2	415		
7	BB	50	2		-9	50%
8	BB	60	17	415		
9	BB	355	17			

Source: authors' illustration.

The reason we assign -9 codes to some of the cases above is that this value indicates to researchers that there is industry information in the underlying IRP5 records for that firm, but that we could not assign an industry following our processes. Thus, if the researcher needs that industry information, they can delve into the IRP5 data themselves. This is in contrast to cases where Activity Code information is system missing in all IRP5 records and no underlying industry information is to be found.

We privilege records with employment information so that the industry assignment procedure has a clear and interpretable meaning. The duration data in the IRP5 records have an interpretable meaning (length of employment) only when that IRP5 record reflects employment income. These duration data are important, as the alternative of assigning a firm-level industry according to the modal record (which it seems has been used before) may be less reliable, as the results will be more sensitive to unrecorded certificate revisions. Additionally, even when there is no measurement error of this type, our method is more likely to accurately account for the bulk of a firm's activities because it generates a coherently 'weighted' modal industry, rather than assuming that all records are equivalently important. We nonetheless do not wish to discard industry information for firms without employment records, and hence we include them in our data as per Step 4. Researchers wishing to exclude these firms can easily do so by excluding firms with missing values for the 'employment share' variable created in Step 3.

The IRP5 Activity Codes do require some cleaning, which must be done prior to the firm-level assignment above. Principally this is about dropping the few records with codes that have no equivalent in the SARS Activity Code system, which is presumed to result from data entry issues or from confusion by firms between industry classification codes and other SARS codes (such as 'income source'). As is the case for the Profit Codes, records that have Activity Codes in the range 3501–3534 are assigned a corresponding 2-digit main activity but a system-missing 4-digit sub-activity. This is not the only reason main activity and sub-activity codes may differ in the IRP5 dataset we create. Because these codes are assigned to firms separately from the underlying IRP5 record-level data, there are some discrepancies even where both codes are non-missing. In some applications researchers will need sub-activities to be perfectly nested within main activities. In these cases, they should either exclude non-nested firms directly, or create a new 2-digit category directly from the 4-digit codes.

4.5 Customs data

A final possible source of industry information is Customs data, as mentioned in Pieterse et al. (2018). These transaction-level data contain detailed product codes in the form of 'Harmonized System' (HS) codes published by the World Customs Organization. Pieterse et al. (2018) suggest that these product codes can be mapped to sector codes using an appropriate concordance table, but we are unsure whether this is well advised. While the HS codes may be able to determine that a transaction involves a particular product, they cannot determine whether the activity associated with that firm is resale, manufacture, repair, or some other activity involving that product. There is therefore no immediate link to the type of industrial activity, which is what we are interested in here. After examining these data, and the available concordance tables, we therefore do not attempt to use them for industry classification.

5 Classification synthesis

Comparing the different industry variables with each other and with external data sources is a major part of this project. In order for this to be done, each variable needs to be converted to some standardized classification system. This in turn requires concordance tables. As far as we are aware, there are no official concordance tables linking the SARS Activity and Profit Codes to Statistics South Africa systems, while a SIC 5–SIC 7 linking table received in correspondence with Statistics South Africa requires significant manual adjustment to be useful for our purposes, as discussed below. The team that created the old panel also did not find official concordance tables, and manually matched some of the classification schemes. Given the weakness of official concordance tables, and the necessity that we do manual matching ourselves, the choice of a baseline classification system becomes especially important. We discuss below our choice of SIC 5 rather than SIC 7 or ISIC 4 as the base category for our comparisons in this paper.

The old panel team chose to use the ISIC 4 classification as a baseline, and created 1-digit-level ISIC 4 categories for all industry variables. There are many benefits to the ISIC 4 and SIC 7 systems over their predecessors, as one would expect from more modern systems. In particular, they include a more detailed treatment of services sector industries, especially ICT-related industries, as well as a more direct categorization of repairs industries. A disadvantage of the SIC 7 system is that, to the best of our knowledge, it is not currently used in any South African data except those produced by SARS, limiting comparability with existing data; the same disadvantage obviously applies to the ISIC 4 system, which, as far as we are aware, is also not used in any South African data. A further disadvantage of ISIC 4 is that it is not native to the tax data, so that all industry variables would require conversion, though the close relationship between ISIC 4 and SIC 7 mitigates this issue insofar as the SIC 7 ITR14 Main Industry Code is concerned.

The major benefit of the SIC 5 system is that, given its ubiquity (as discussed in Section 3), it facilitates easy comparison across South African datasets. The major disadvantage of the SIC 5 system is that, being a 1993 system, it is out of date. For example, there is minimal differentiation amongst services sector jobs, and that which does exist is often inappropriate in a world where ICT has transformed these jobs.

Another advantage of the SIC 5 system is that examination of the SARS Profit Code and Activity Code systems suggests that they are themselves modified versions of the SIC 5 system. Given that we create concordance tables ourselves, it makes sense to match the SARS codes to the closest standard classification system, in order to minimize error. There are cases where converting the SARS codes to the more detailed SIC 7 or ISIC 4 codes necessarily introduces error even at the

most aggregated level of classification, whereas when matching to the SIC 5 codes we can assign 3-digit-level SIC 5 codes in the vast majority of cases.

In general, this reflects the problem of matching a less granular system (such as the SARS codes) to a more granular system (SIC 7 or ISIC 4). For the creation of a base categorization for comparison, it is generally better to work in reverse, using a less granular system as the base. The disadvantage of this approach is of course precisely the fact that the base categorization is less detailed, and more detail is generally useful for researchers. However, given that the main purpose of our matched classification is to compare the industry variables with each other, as well as to external sources of South African data, we proceed with using SIC 5 as our base classification of choice.

We manually create concordance tables from the SARS Activity Codes and Profit Codes to SIC 5, with a particular effort to convert to at least 3-digit-level detail in the SIC 5 codes. A SIC 5–SIC 7 linking table we received from Statistics South Africa is not immediately amenable for our purposes of assigning a SIC 5 code for every SIC 7 code (or vice versa). The table contains multiple category descriptions per SIC 7 5-digit code, and category descriptions within a particular SIC 7 5-digit code may be linked to different SIC 5 5-digit codes. We therefore have to manually evaluate every SIC 7 5-digit code where this type of conflict occurs, and assign a SIC 5 code ourselves. In undertaking this process, we are guided mainly by the SIC 7 5-digit category descriptions available from SARS (see Section 3), as these are the category descriptions used by firms in the tax data. We again place a particular priority on assigning SIC 7 codes to at least a 3-digit SIC 5 category, if we cannot be more detailed.

We often cannot match a given SIC 7, Profit Code, or Activity Code to a particular level of the SIC 5 system. For example, while almost all categories of all variables can be assigned a 1-digit SIC 5 code, relatively few can be assigned a 5-digit SIC 5 code. In these cases, the SIC 5 version of that code will be missing at the 5-digit level, but existing at the 1-digit level. There are higher rates of non-matching when converting from the Profit and Activity Codes than from SIC 7, likely because we must create the concordance ourselves in these cases but can use the existing Statistics South Africa concordance as a base in the SIC 7 to SIC 5 conversion. Our concordance tables are available upon request and in the NT-SDF. Information is inevitably lost during a concordance process, and if researchers rely on our SIC 5 variables (not our preferred option, as discussed in Section 7), they are advised to browse through these tables, particularly if a focus on particular industries is important for their project.

6 Industry variable review

In this section we present an analysis of the internal and external consistency of the different industry variables, after they have all been converted into the SIC 5 classification system as outlined above. Specifically we examine each variable's completeness (the proportion of firms with non-missing data for this variable), over-time switching (the rate at which firms switch industries over time for each variable), across-variable matching (how well each variable matches the other industry variables at the firm level), and external validity (how aggregate industry statistics for a given variable match similar statistics for external sources).

We perform this analysis over 4 different subsamples of the CIT data. The first sample, which we call the 'CIT Panel', comprises all records found in the SARS-NT Panel. However, a substantial proportion of firms in the SARS-NT data do not seem to be economically active—they either are

marked as dormant, or have zero or missing sales, or zero or missing variable costs.[21] We therefore create a subsample of 'economically active' CIT Panel firms using these criteria, which we call Real CIT firms.[22] When using the Real CIT subsample we have no observations for 2008, since in the SARS extraction we use there is no sales information for firms in this year; hence in the figures below these Real CIT series start in 2009. A further sample restriction common in researcher practice is restricting the firm sample to those that can be matched with IRP5 data, or equivalently those that have employment information. We call this sample without restrictions on firms being economically active the CIT-IRP5 sample. Finally, the CIT-IRP5 sample with economic activity restrictions is the Real CIT-IRP5 sample. In some cases, it is not feasible or useful to present results for all subsamples, but results for all subsamples are available upon request.

6.1 Completeness

We first check the completeness of each industry variable. We calculate the proportion of observations that are not system-missing for each industry variable, out of the total firm count per year. Note that this is an imperfect measure of meaningful industry information per variable, as firms in a category such as 'Agriculture not elsewhere classified (n.e.c.)' do indeed lack detailed sub-sector information, but in our counts here they would be counted as non-missing so long as there is an equivalent SIC 5 category to which they can be assigned.[23]

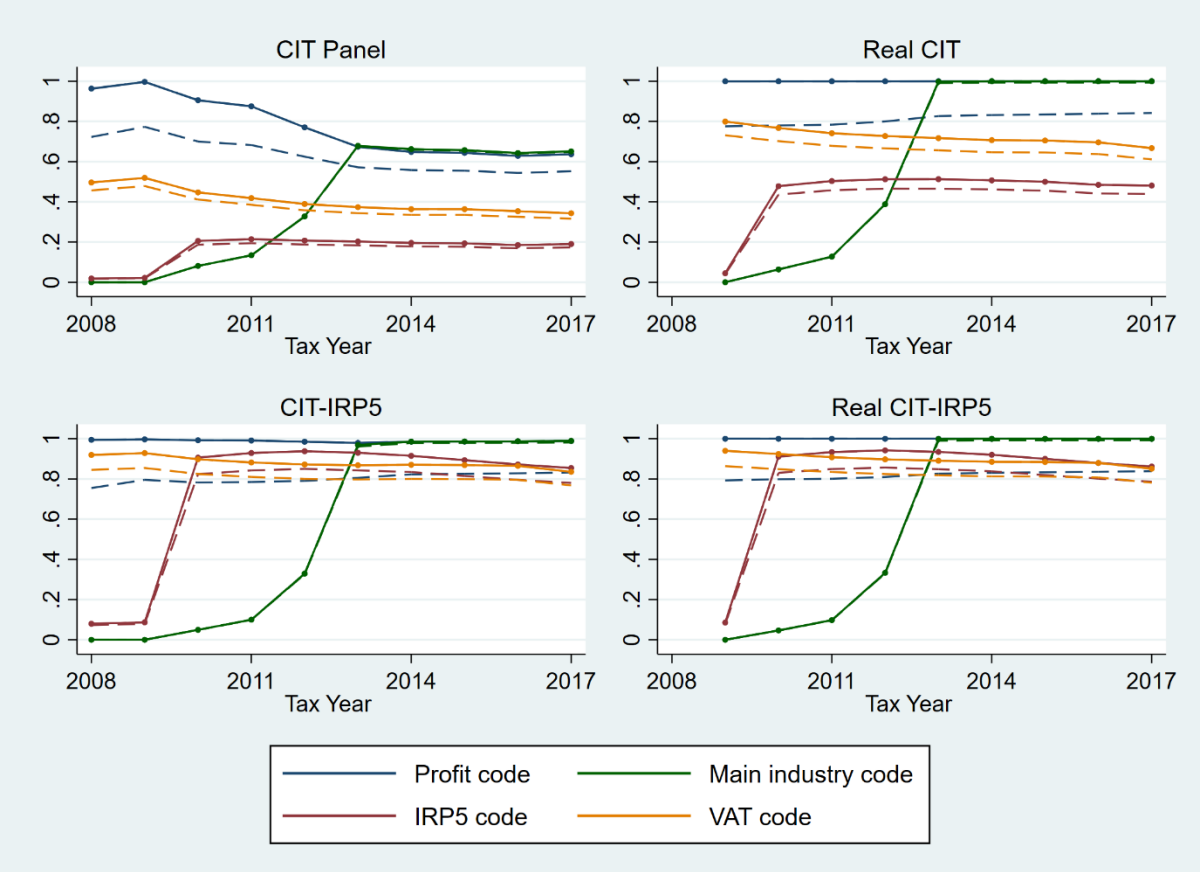
Figure 9 presents completeness proportions for the four subsamples, using three different versions of each variable: The solid line is the proportion of firms with non-missing observations of the industry variable in question when it is converted to the SIC 5 1-digit level, while the dashed line is the equivalent non-missingness when converted to the SIC 5 3-digit level. A further connected line (with points shown at each year) is the non-missingness of the underlying industry category before it is converted to any SIC 5 category. That this last line is indistinguishable from the SIC 5 1-digit category shows that the conversion to SIC 5 loses almost no information at the 1-digit level. This is not true at the SIC 5 3-digit level, with substantially more missing firms at the SIC 5 3-digit level than the SIC 5 1-digit level when converting from the various SARS systems, the Profit Code being particularly affected. As mentioned above, the SIC 7–SIC 5 conversion does not result in much missingness at the 3-digit level, probably because we had Statistics South Africa linking tables to work from.

[21] For our purposes here we define variable costs as the sum of income statement cost of sales and labour expenses.

[22] A potential issue with our approach here is that, subsequent to our data analysis, we were informed that a non-trivial number of firms in financial business services have no sales data but do have data on EBITDA (earnings before interest, taxes, depreciation, and amortization), for example. This may mean that we exclude currently active firms when using this specification. However, this is unlikely to be consequential for our analysis: none of our conclusions rest on the performance of Real CIT FIRE sector firms, and results are frequently so similar for the CIT Panel and Real CIT firms that we report only one set of results.

[23] The extent of this problem should not be overstated. We do have *some* information about a firm in 'Agriculture not elsewhere classified (n.e.c.)': we know that (if the variable is correctly entered) the firm is not in another named Agriculture sub-industry.

Figure 7: Completeness of industry classifications across 4 subsamples



Note: solid lines indicate the proportion of firms with non-missing industry information for the specified variable after the variable has been converted to the SIC 5 1-digit level. Dashed lines indicate the proportion of firms with non-missing industry information for the specified variable after the variable has been converted to the SIC 5 3-digit level.

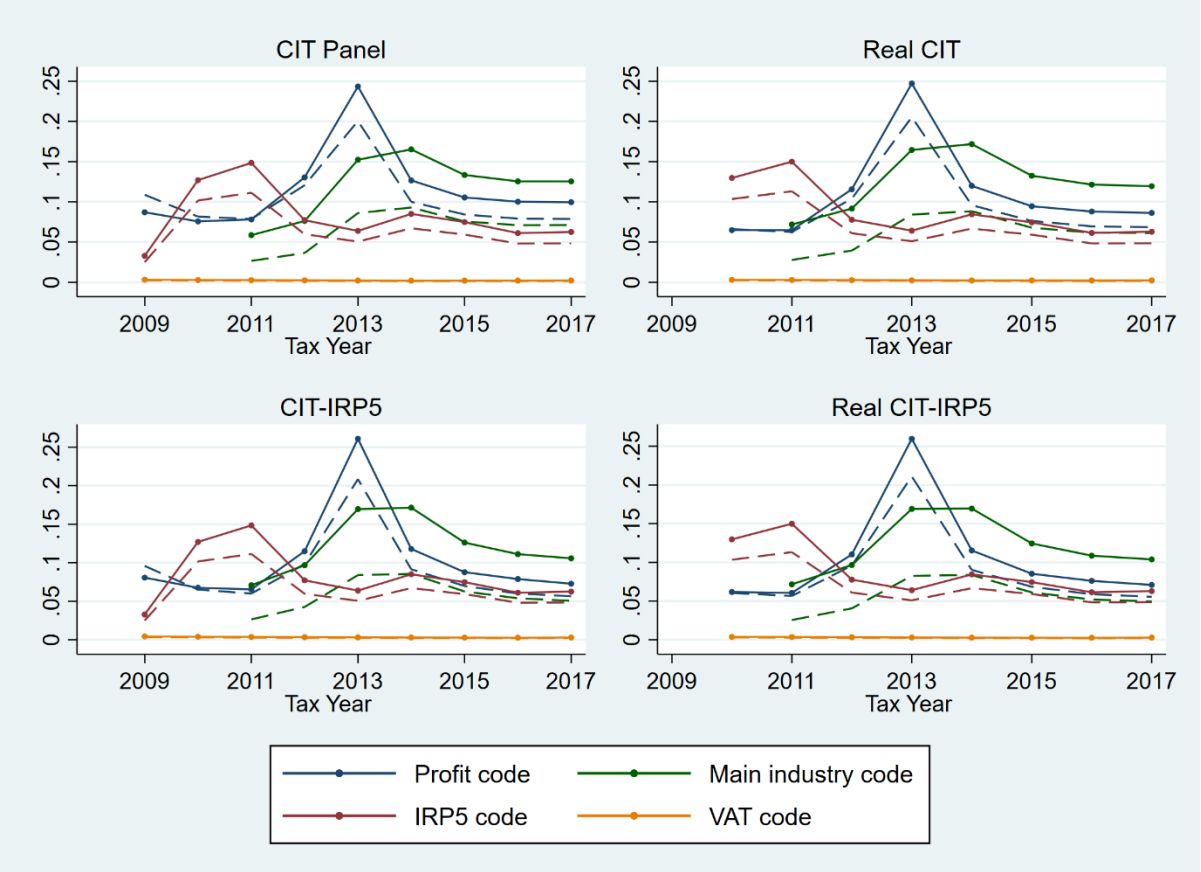
Source: authors' calculations using the SARS-NT data.

Figure 9 also shows that, while industry variable coverage is not particularly good for the entire CIT Panel, when the sample is restricted to economically active firms or firms linked to the IRP5 data, virtually all firms have Profit Code data, and the same is true for the Main Industry Code from tax year 2013 onwards. The lack of Main Industry Code information prior to 2013 is to be expected, as the variable was only introduced onto the forms in 2013. Main Industry Code information on records prior to 2013 will have come from form revisions.

6.2 Switching

As discussed by Newman et al. (2013), some amount of firm industry-switching can be expected over time, as firms change their activities. In Figure 10 we show, for each industry variable, the proportion of firms per year that switch to a non-missing industry category from a different non-missing industry category in the prior year, out of all firms with non-missing industry information in the relevant and prior years. These are thus 'real switches', and are not affected by switches in and out of a system-missing industry variable or the number of firms that stay system-missing in a particular industry variable over time. Two lines are shown per industry variable: switching at the SIC 5 3-digit level (solid line with points) and switching at the SIC 5 1-digit level (dashed line).

Figure 8: Industry-switching over time across 4 subsamples



Note: solid lines indicate the proportion of firms switching industry at the SIC 5 3-digit level, while dashed lines indicate industry-switching at the SIC 5 1-digit level. Switches (or staying) in and out of system-missing industry variables are not considered.

Source: authors' calculations using the SARS-NT data.

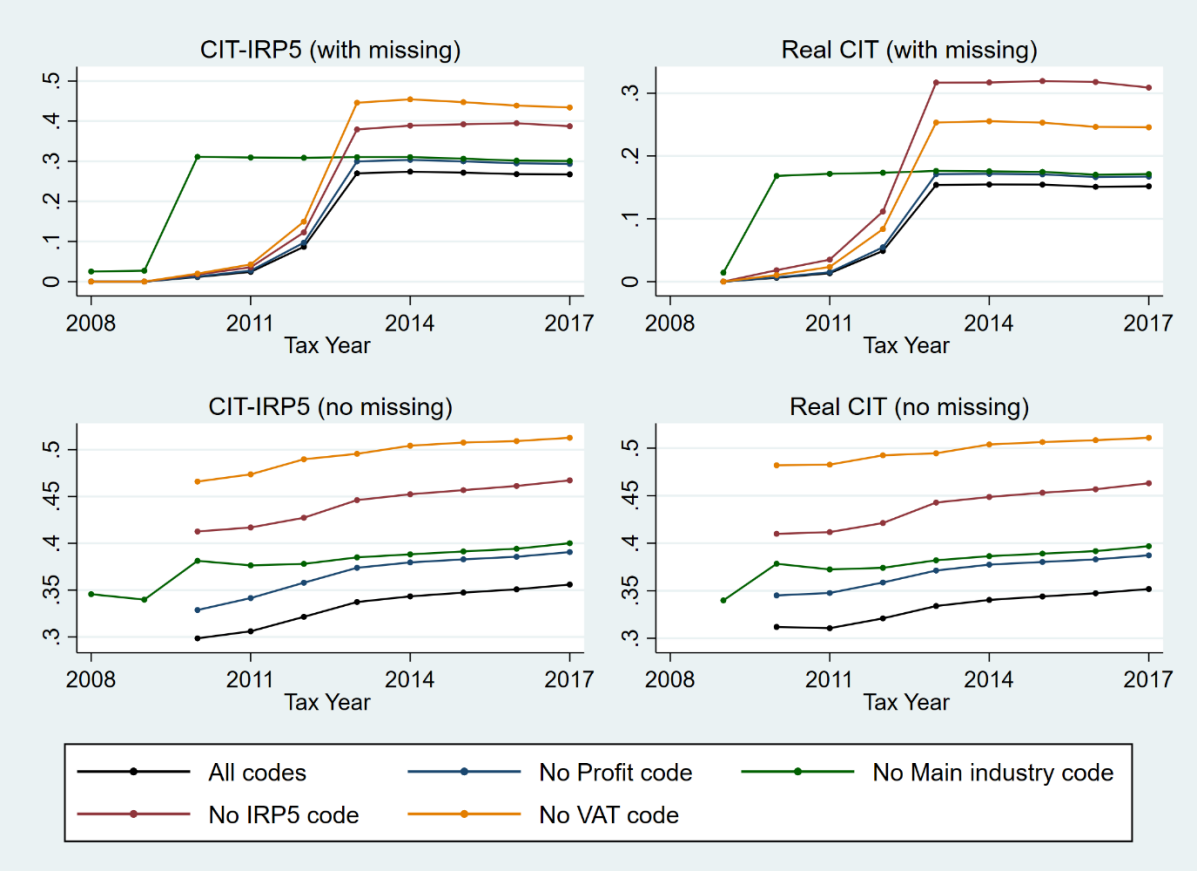
Figure 10 shows that there is virtually no switching over time in the VAT Activity Codes, which supports our suggestion in Section 4 that these codes likely come from the firm VAT registration process. There is, however, meaningful switching in the IRP5 codes, which would seem to cast doubt on a similar explanation for the origin of the IRP5 Activity Codes. There is a dramatic spike in switching in Profit Codes in 2013, likely explained by this being the year the ITR14 form was introduced and the IT14 was phased out. We cannot tell whether the switch in reported industry is due to increased prevalence of e-filing or different documentation of the SARS Profit Codes, but the switch is concerning insofar as the continuity of the Profit Code series is concerned. In the post-2013 years there is noticeably more switching in the Main Industry Code at the SIC-5 3-digit level than there is in the other codes, but switching levels are similar at the SIC 5 1-digit level. It is possible that the difference in switching at the 3-digit level is driven by the greater completeness of the Main Industry Code at the SIC 5 3-digit level than is the case for other variables.

6.3 Across-variable matching

Plausible over-time switching is one measure of internal consistency. Another is how well a given industry variable matches other industry variables in the dataset: for example, if some firm is in Manufacturing according to industry variable X, but in Mining according to the other three industry variables, this is cause for some scepticism about how well industry variable X captures the firm's industry characteristics. In Figure 11 we compare industry variables at the SIC 5 1-digit level and report how often industry categories agree with each other for firms in the SARS-NT

panel. We start by using all industry variables, and then examine how the rate of industry agreement increases as we take out particular industry variables. A large increase in the matching rate due to excluding industry variable X would suggest that industry variable X is frequently the 'odd one out' in how it assigns industries.

Figure 9: Across-variable agreement for different combinations of industry variables



Note: each line shows the proportion of firms that have matched industry codes at the SIC 5 1-digit level across industry variables, for the subset of industry variables indicated in the graph legend. Graphs in the top row do not adjust for missing values in industry variables (treating them as any other category), while graphs in the bottom row exclude firm-observations where any of the industry codes is missing. A higher line suggests that dropping the relevant variable improves matching more.

Source: authors' calculations using the SARS-NT data.

Figure 11 shows figures for two subsamples: the Real CIT and CIT-IRP5 subsamples.[24] We show two graphs for each subsample. In the top row, we do not adjust for missing values in industry variables (treating them as any other category). Graphs in the bottom row exclude firm-observations where any of the industry codes are missing.

The bottom row of Figure 11 suggests that when firms with missing industry categories are excluded, the worst-matched industry variables are the VAT Activity Codes, then the IRP5 Activity Codes, then the Main Industry Codes, and then the Profit Codes (i.e. the biggest improvement in matching is obtained by dropping the VAT Activity Codes, then the IRP5 Activity Codes, and so on). However, the Profit Codes and Main Industry Codes are only negligibly different from each other with respect to how well they match other industry variables, and they are clearly the two

[24] The pictures are extremely similar for the CIT Panel and Real CIT-IRP5 subsamples, hence there being no point in showing them. They are available from the authors on request.

best-matched industry variables. This conclusion is not immediately self-evident when looking at the top row of Figure 11, which includes firms with system-missing industry information, but differences are easily explained. Dropping the Main Industry dramatically increases the rate of matching across variables before 2013, but this is because it is generally missing in these years—recall that it was only introduced in the 2013 tax year. Once the Main Industry Code is introduced, dropping it adds a very similar benefit to dropping the Profit Code when it comes to matching—that is, only a negligible improvement. A similar phenomenon explains why dropping the IRP5 Activity Codes improves matching more than dropping the VAT Activity Codes for the Real CIT subsample in the top row. This subsample includes many firms not matched to the IRP5 data, meaning that a high proportion of IRP5 Activity Codes are missing. When restricting to the CIT-IPR5 panel, the VAT Activity Codes are again the worst-matched industry variables. The overall conclusion therefore is that when 'real' industry information exists, the Profit Codes and IRP5 Activity Codes are most consistent across industry classifications.

This conclusion requires further testing before it is robust, however. For example, it is possible that the Profit Codes and Main Industry Codes match only each other very well, and that this drives the results—but they could both be wrong. While in the interest of space we do not report the results here, we perform additional tests by directly examining how often each variable assigns 'odd-one-out' industry classifications to firms, for all combinations of three industry classifications. This shows that even when excluding the Profit Code, for example, the Main Industry Code is still less frequently the odd-one-out than the VAT or IRP5 Activity Codes. These unreported results robustly support the conclusion that the Profit Code and Main Industry Code are most likely to match other industry variables, followed by the IRP5 Activity Codes and then the VAT Activity Codes.

6.4 External comparisons

Here we compare aggregate time series for the four SARS-NT industry variables—VAT Activity Codes, IRP5 Activity Codes, Main Industry Codes, and Profit Codes—with those generated by four external sources of South African data: the Tax Statistics, Quarterly Employment Statistics, Quarterly Labour Force Surveys, and Quarterly Financial Statistics.[25] Sample restrictions are generally required to make our time series comparable to the external data—this and the specific variables compared are discussed in each sub-section. We present only the 6 largest (out of 9) industries per figure for clarity of presentation.[26] The quarterly data from Statistics South Africa are aggregated to the tax-year level. The overarching conclusion we draw from this exercise is that the time series of the Main Industry Code variable best matches the external data.

Tax Statistics

The Tax Statistics used here are a collation of individual CIT spreadsheets, which are available for each year from SARS, and specifically Table A3.4.2 for each year (National Treasury and South African Revenue Services 2008–2019). We use these data to construct a count from SARS of the number of firms in each industry, and compare this with what we find in our data using the CIT Panel subsample, which is the sample most directly comparable to the CIT statistics.

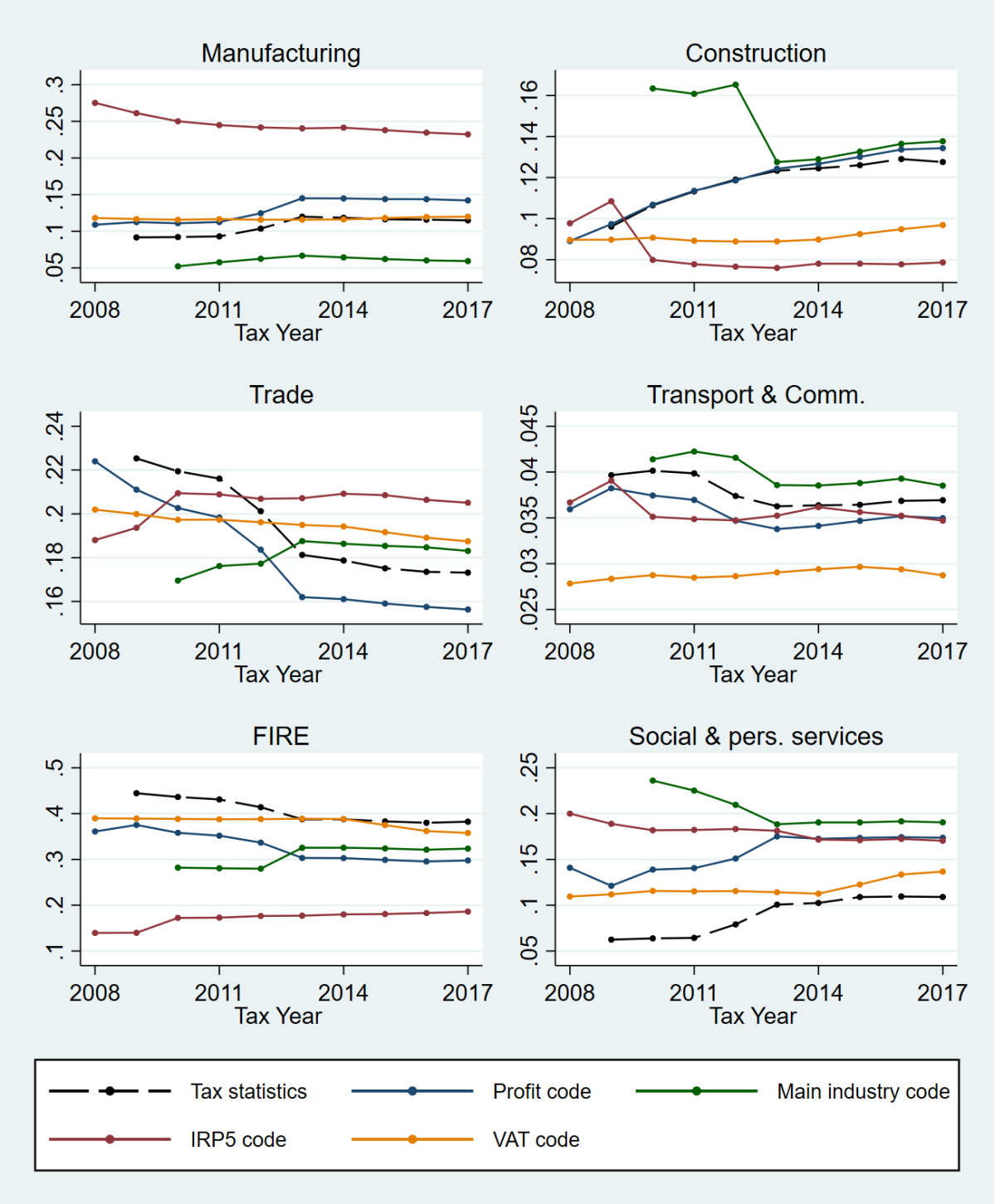
Figure 12 shows the proportion of all firms accounted for by each SIC 5 1-digit industry, per variable. It is evident that the Tax Statistics time series is in general most closely tracked by the

[25] The Tax Statistics are produced by SARS; all other data are produced by Statistics South Africa.

[26] All graphs are available from the authors on request.

time series of our Profit Code variable. This is more reassuring than it is enlightening: we use the same base data, and the Tax Statistics originate in the Profit Codes, so one would hope that the results match.

Figure 10: Proportion of firms accounted for by each industry, comparing with Tax Statistics



Note: each solid line shows the proportion of firms in the CIT Panel sample accounted for by each industry using the specified industry variable. The black dashed line shows the same proportion from the SARS Tax Statistics.

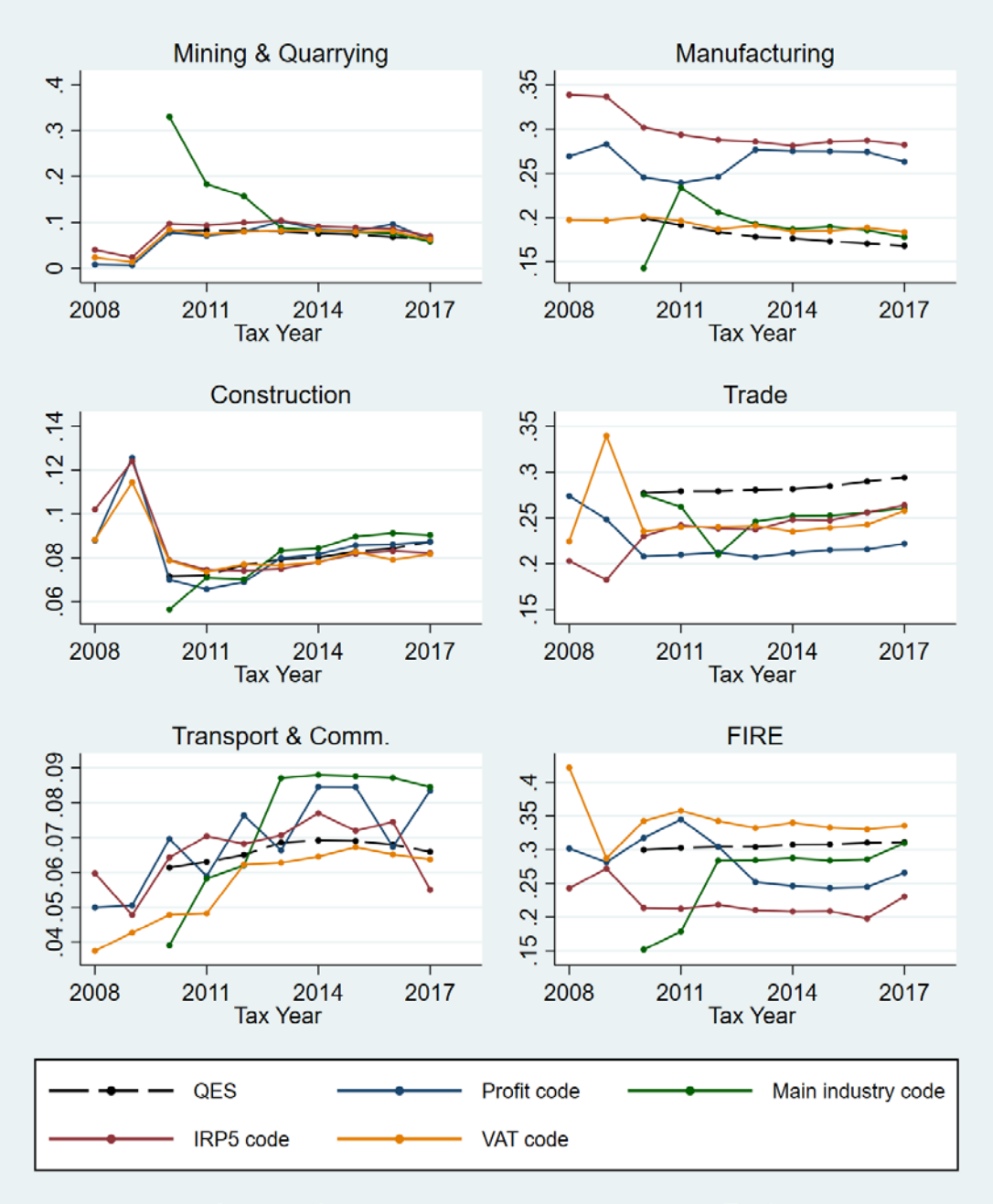
Source: authors' calculations using the SARS-NT data and SARS Tax Statistics.

Quarterly Employment Statistics (QES)

The QES are drawn from an enterprise-based survey conducted by Statistics South Africa, from the universe of VAT-registered business. Using payroll information, employment and earnings are reported. The survey does not cover the agricultural sector, private households, or public sector entities. An Excel file with all quarterly results for the last 10 years is available from Statistics South Africa's website.

In Figure 13 we present total employment in the CIT-IRP5 subsample accounted for by each industry, compared with the analogous time series from the QES.

Figure 11: Proportion of employees accounted for by each industry, comparing with QES



Note: each solid line shows the proportion of employees in the CIT-IRP5 sample accounted for by each industry using the specified industry variable. The black dashed line shows the same proportion from the QES.

Source: authors' calculations using the SARS-NT data and Statistics South Africa QES.

For each industry variable we drop SARS-NT records in the SIC 5 1-digit 'Agriculture, hunting, forestry, and fishing', because the QES explicitly does not include agricultural firms. The SIC 5 1-digit sector of 'Community, social and personal services' presents a challenge, as many but not all industries classified in this sector provide public services or are otherwise not-for-profit and may not be comparable to QES firms. Because delineating between these types of firms would be difficult, we drop firms in this sector from the SARS-NT sample as well as the QES statistics, to create more comparable samples.

We use the CIT-IRP5 subsample because we only want to look at firms with employment information. They are not reported here but conclusions are qualitatively the same if total remuneration is used instead of employment, or if the Real CIT-IRP5 sample is used. It is apparent from the graphs that in general the VAT Activity Codes and (post-2012) Main Industry Codes best match the QES time series, the former matching better than the latter. The surprisingly good performance of the VAT Activity Codes may be due to the QES data originating in VAT-registered firms.

Quarterly Labour Force Survey (QLFS)

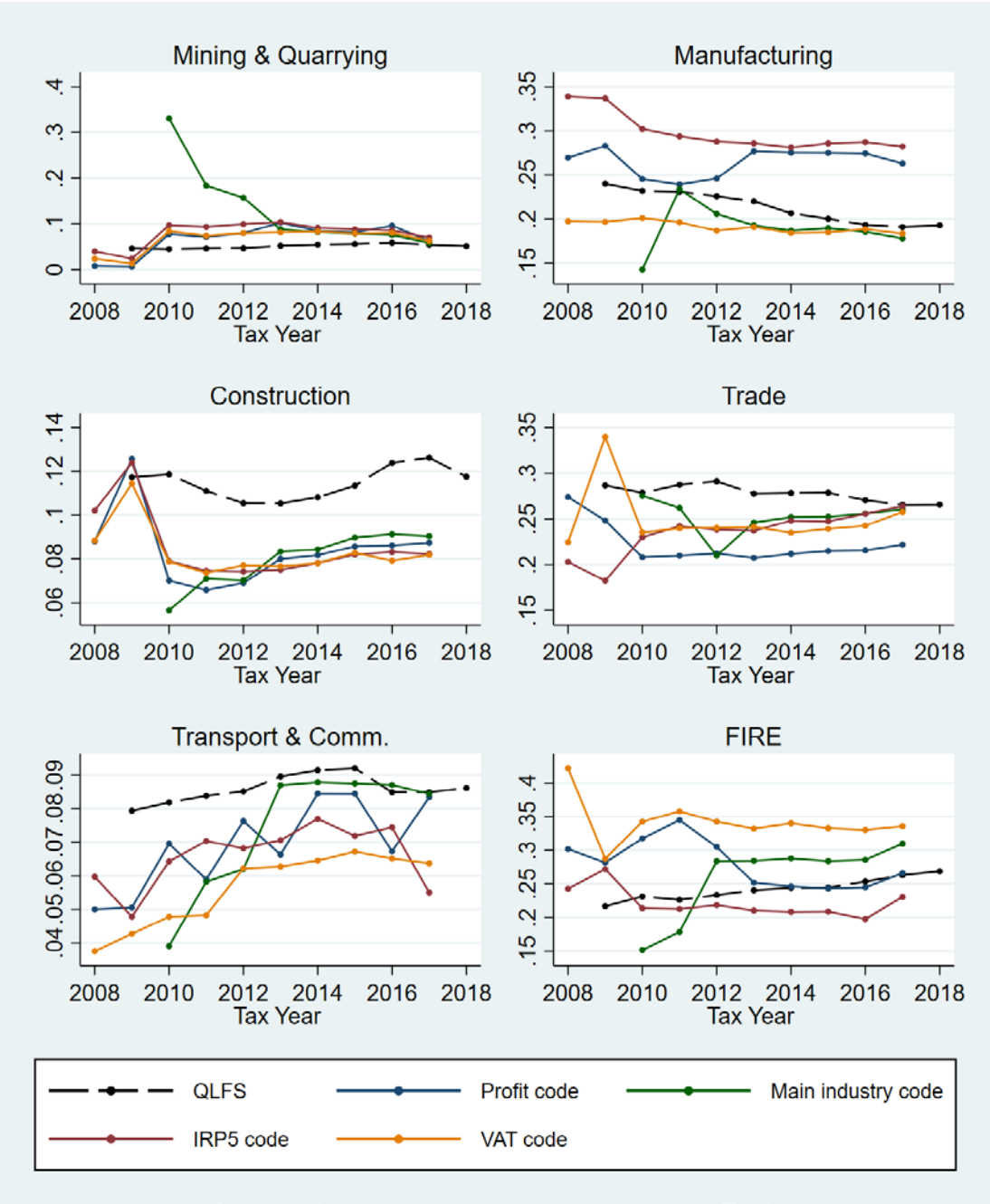
The QLFS is a household survey conducted by Statistics South Africa. As such, it is not limited to recording labour market outcomes in particular establishments. The survey microdata provide rich detail on many aspects of the South African labour market, but we restrict our attention to the employment information available in industry-disaggregated QLFS time series made available by Statistics South Africa (Statistics South Africa 2019b). We focus on table 3.3 from this resource, and only examine employment in formal sector industries, which in the QLFS excludes the agricultural sector.

As with the QES, we exclude the SIC 5 1-digit 'Community, social and personal services' category from both our SARS-NT and QLFS time series, but in this case because it is the QLFS that likely includes many government sector activities that will not be found in our SARS-NT panel. We again use the CIT-IRP5 subsample of the SARS-NT data, and the results are even more supportive of our overarching conclusions if the Real CIT-IRP5 subsample is used.

Our results, shown in Figure 14, are not as clear-cut as in the QES, but again suggest that the (post-2012) Main Industry Code is the code best matched to the external data. In the one important case where the Profit Code is better matched than Main Industry Code—the 'Financial intermediation, insurance, real estate and business services' sector—the conclusion is reversed if the Real CIT-IRP5 subsample is used, suggesting that this particular exception is not very robust.[27] It is encouraging that results from a household survey (the QLFS) are roughly consistent with what we find from the firm-based statistics (QES and QFS, as discussed below).

[27] This may be related to the Real CIT definition issue discussed in footnote 22.

Figure 12: Proportion of employees accounted for by each industry, comparing with QLFS



Note: each solid line shows the proportion of employees in the CIT-IRP5 sample accounted for by each industry using the specified industry variable. The black dashed line shows the same proportion from the QLFS.

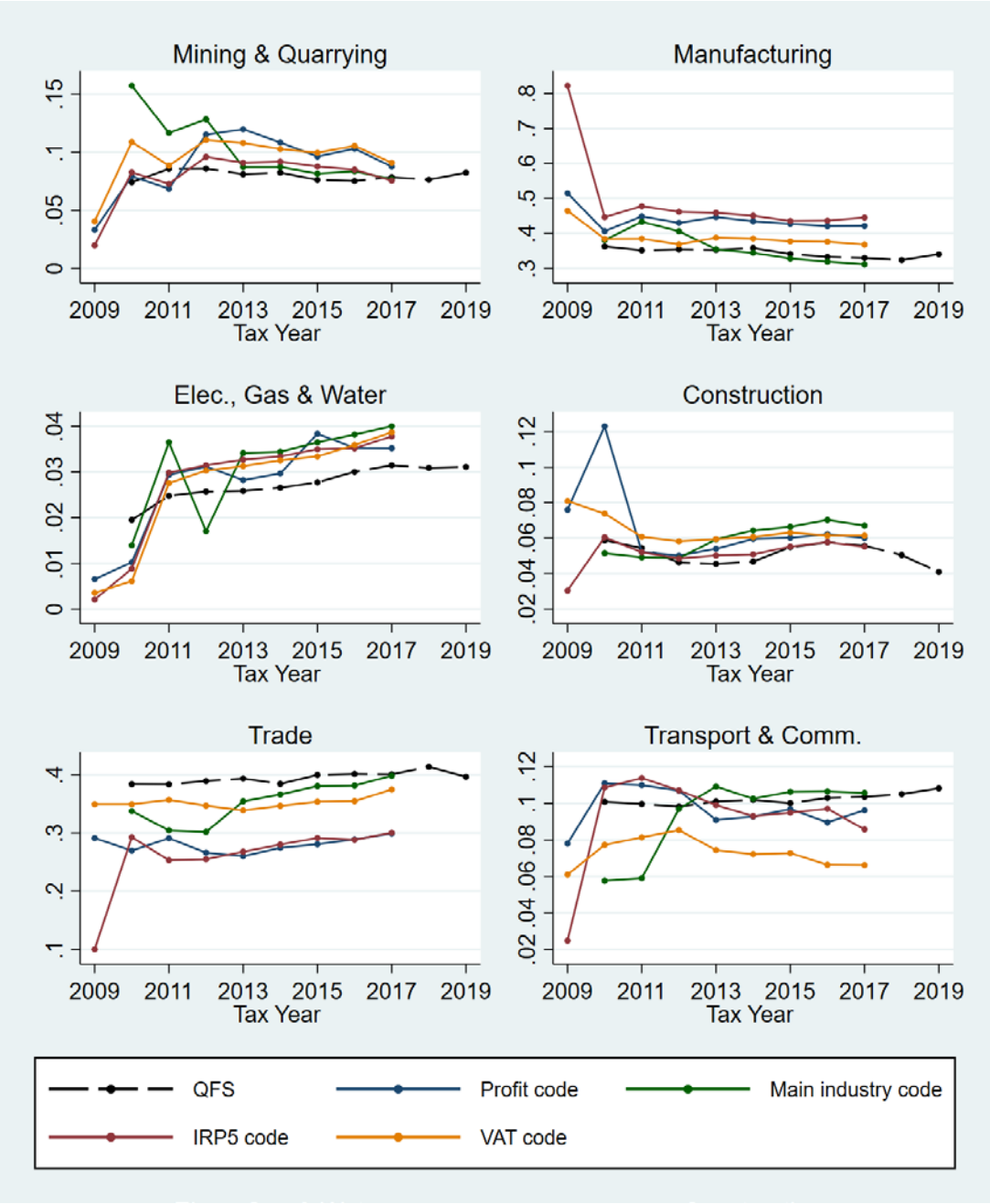
Source: authors' calculations using the SARS-NT data and Statistics South Africa QLFS.

Quarterly Financial Statistics (QFS)

The QFS come from a Statistics South Africa survey of formal-sector enterprises, excluding firms in agriculture, the FIRE sector, government, and educational institutions. Firms of various sizes are sampled, though firms with VAT turnover of less than R2 million are excluded. Sampling weights are applied so that the results are representative of all enterprises. As far as we are aware, there is no readily available time series of QFS statistics; we therefore create a time series by scraping quarterly-reported spreadsheets from the Statistics South Africa website (Statistics South

Africa 2009–2019).[28] The particular value of the QFS for our purposes is that it has balance sheet information, so that we can compare the external validity of the industry codes outside of the subsamples with employment information used for the QES and QLFS comparisons. In Figure 15 we compare turnover time series from the QFS with gross sales time series using the SARS-NT 'Real CIT' subsample, but the results are extremely similar across all 4 subsample options.

Figure 13: Proportion of sales accounted for by each industry, comparing with QFS (turnover)



Note: solid lines show the proportion of gross sales in the Real CIT sample accounted for by each industry using the specified industry variable. Black dashed lines show the same proportion from the QFS, using turnover.

Source: authors' calculations using the SARS-NT data and Statistics South Africa QFS.

[28] Excel tables are reported for the vast majority of quarters at the Statistics South Africa landing page for that quarter's reports.

We exclude the SIC 5 1-digit 'Agriculture, hunting, forestry, and fishing' and 'Financial intermediation, insurance, real estate and business services' sectors from our SARS-NT data as firms in these industries are excluded from the QFS, and again exclude the 'Community, social and personal services' sector from both our SARS-NT and QFS time series, due to the comparability issues raised previously.

From Figure 15 it can be seen that the (post-2012) Main Industry Code is again generally the variable best matched to the external data source, especially when attention is focused on the largest sectors.

7 'Best' industry variable

The discussion in Section 6 suggests that the most reliable industry variable in the raw data is the Main Industry Code. It is as complete as the Profit Code after 2012; it matches the other industry variables at about the same rate as the Profit Code; it matches the external data better than any other variable; in its raw form it comes in the highly granular and modern SIC 7 system; and it can be relatively completely concorded to the standard South African system of SIC 5. Industry-switching over time is non-negligible but relatively constant over time.

The major drawback of the variable is that it is only well populated after the switch to ITR14 forms in 2013. This is not an issue to take lightly, as discussed below. While the Main Industry Code is therefore the basis of our preferred 'best' industry variable, we also create an additional variable, based primarily on the Profit Code, for researchers needing a longer industry time series based on raw reported variables. We discuss these two variables below, and then reproduce some of the analysis of Section 6 to compare their performance.

7.2 Imputed Main Industry Code

The major issue with the Main Industry Code is its sparse coverage in the pre-2013 tax years. The way to address this that causes minimal violence to the underlying codes is to iteratively impute missing industry codes with time-neighbouring non-missing observations for the same firm. This process is the industry imputation implemented by the old panel creators.

The process is as follows[29]:

1. Replace each firm-year observation system-missing in the Main Industry Code with the non-system-missing Main Industry Code for the same firm in the immediately **following** period, if such a non-system-missing **following** firm-year Main Industry Code exists.
2. Replace each firm-year observation **still** system-missing in the Main Industry Code with the non-system-missing Main Industry Code for the same firm in the immediately **prior** period, if such a non-system-missing **prior** firm-year Main Industry Code exists.
3. Repeat steps (1) and (2) until no replacements are made at either step.

The process is therefore a kind of over-time 'nearest neighbour' imputation, with a preference given to backwards rather than forwards imputation. It still allows some industry-switching over time, provided underlying raw records include more than one industry per firm, but switching will artificially be zero for the majority of firms pre-2013, and the timing of switching that does exist

[29] Stata code implementing this procedure is available from the authors on request and in the NT-SDF.

in this period will frequently be incorrect. The method does not impute industries across gaps in the data—that is, when firms exit the panel for at least a year and then re-enter. It seems a particularly strong assumption that there is no industry-switching when this occurs.[30]

The major downside of this approach is that firms with Main Industry Code information prior to 2013 will disproportionately be firms that survive into 2013, rendering the pre-2013 sample an implicit 'balanced panel' and increasingly unrepresentative going back in time. This issue is not easily addressed. One option would be to assign to firms missing Main Industry Code data the industry categories of other industry variables in the dataset. Though we implement something similar for the Composite Profit Code discussed below, this has two major downsides when applied to the Main Industry Codes.

First, this cross-variable industry assignment requires that industry variables be converted to one classification system—in this paper we use the SIC 5 system as the base category. However, in the concordance process significant detail is lost and error introduced as systems are converted. The detail and modernity of the SIC 7 system is a major benefit of the Main Industry Code and we are averse to losing this. Second, cross-variable assignment brings in its own biases. Consider a process where non-missing Profit Codes (the variable with the longest time series and comparable internal consistency to the Main Industry Code) are used to assign industries to pre-2013 firm-year observations missing Main Industry Codes after imputations. As Section 6.4 makes clear, the Profit Codes tend to over-assign firms to Manufacturing and under-assign to Trade. A naïve examination of firm entry and exit using this hypothetical composite industry variable would therefore over-estimate the extent to which Manufacturing firms fail compared with Trade firms, because it would be disproportionately firms that do not survive until 2013 that are assigned to Profit Code industries.

There is value in a simpler and more transparent measure, and we do not use cross-industry assignment in our Imputed Main Industry Code.

7.3 Composite Profit Code

The major concern with the Profit Code variable is that it does not match very well with the external data sources examined in Section 6.4. Its major advantage is that it has a high rate of record coverage across the time periods of the CIT data, while like the Main Industry Code (and unlike the IRP5 and VAT Activity Codes) it is generally internally consistent.

To create a 'best' Profit Code we therefore focus on addressing the external validity issue by using cross-variable industry assignment to replace the Profit Code when it does not match other classifications. After all industry variables have been converted to the SIC 5 system, we create a Composite Profit Code according to the following algorithm[31]:

1. Create Imputed Main Industry Code variables as in Section 7.2, separately for the SIC 5 1-, 2-, 3-, 4-, and 5-digit levels.
2. Generate a new SIC 5 1-digit variable that is equal to the raw non-missing SIC 5 1-digit Profit Code for a given firm if the Profit Code and Imputed Main Industry Code match.

[30] If required, individual researchers can of course implement this cross-gap imputation themselves.

[31] Stata code implementing this procedure is available from the authors on request and in the NT-SDF.

3. If the new 1-digit SIC 5 variable is still missing for a firm, and the non-missing SIC 5 1-digit Imputed Main Industry Code matches the VAT Code and IRP5 Code, assign this SIC 5 1-digit value to the firm.
4. If the new 1-digit SIC 5 variable is still missing, assign the raw non-missing SIC 5 1-digit Profit Code to the firm.
5. Repeat the above process for, sequentially, the SIC 5 1-, 2-, 3-, 4-, and 5-digit levels, but only create a non-missing record for these lower levels of disaggregation if it nests into the immediately higher level of industry aggregation for that firm-year observation.
6. Impute across time as in Section 7.2.

The logic of this algorithm is that if the Profit Code matches the imputed Main Industry Code, we can be quite sure that the correct industry is assigned. If the Profit Code does not match the Imputed Main Industry Code, but all other categories agree with the Imputed Main Industry Code, then it is likely that the Main Industry Code assignment is correct. Only if none of these options is possible do we use the raw non-matching Profit Code.

With the Profit Code we do not need to be worried about the decrease in granularity that a composite industry variable requires when all industry variables are converted to SIC 5, since the raw SARS Profit Code system is not suitable as an industry categorization anyway. However, the issues of a selected sample and various biases as discussed in the previous section do need to be considered, and the complex construction of the Composite Profit Code makes such issues difficult to fully anticipate. This variable therefore comes with a warning for users, though we present it as an alternative nonetheless, since a complete industry time series based on consistent underlying raw records will be important for some users. Both options come with drawbacks.

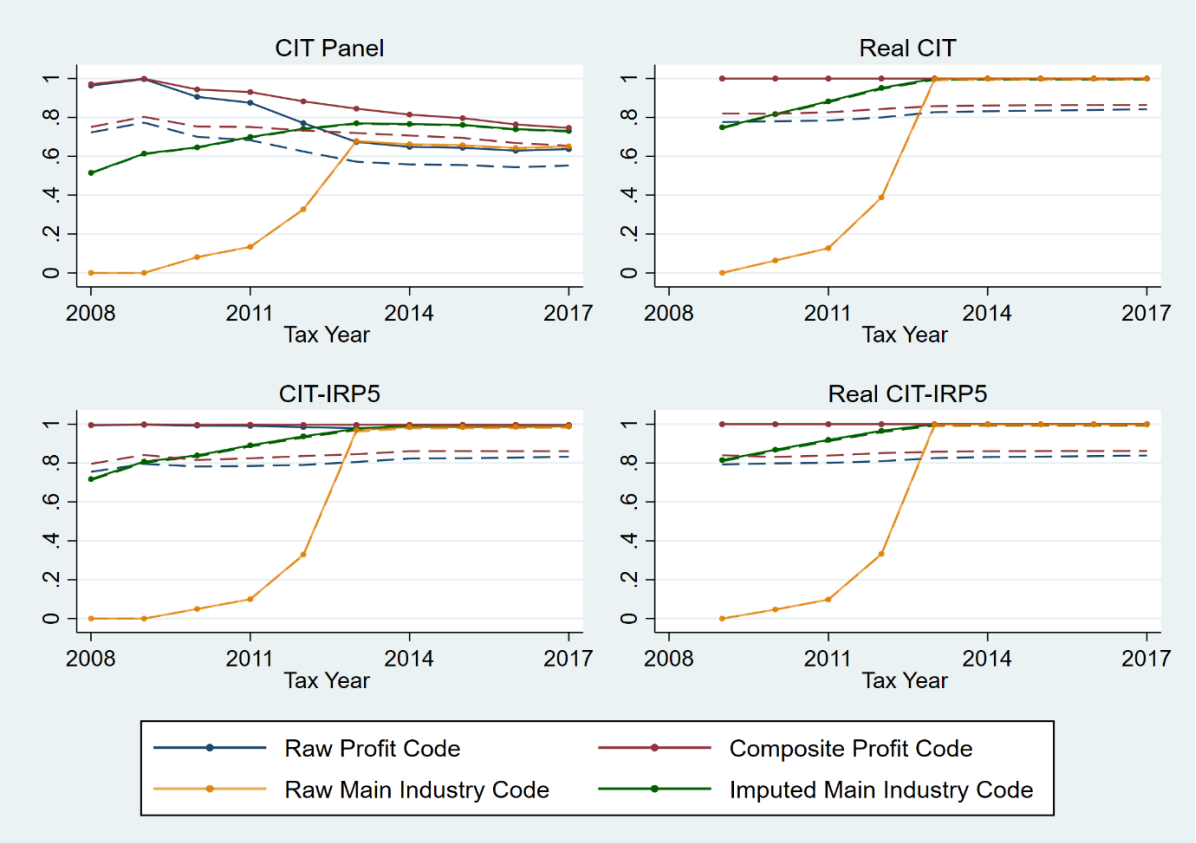
7.3 Internal and external validity

In this section we present a subset of the figures presented in Section 6.4, for the Imputed Main Industry Code, Composite Profit Code, Raw Main Industry Code, and Raw Profit Code. The exact same procedures for generating figures is used as in Section 6.4, but with different industry variables.[32] The overarching conclusion is that the Imputed Main Industry Code and Composite Profit Code clearly improve upon the raw variables across all dimensions, though issues noted above (such as artificially low switching in the Main Industry Codes prior to 2012) are also evident.[33] The Composite Profit Code almost always improves upon the matching of the Raw Profit Code to the external data, while the Imputed Main Industry Code vastly improves upon the coverage of the Raw Main Industry Code and extends consistently back in time.

[32] All figures use industry variables in SIC 5 format for purposes of comparability, though as mentioned our suggested Imputed Main Industry Code for the data would retain its SIC 7 format.

[33] In unreported figures we also check how the Imputed Main Industry Code performs against a Composite Main Industry Code, and the Composite Profit Code against an Imputed Profit Code. While our main reasons for suggesting the Imputed Main Industry Code and Composite Profit Code are the arguments in Sections 7.1 and 7.2, it is reassuring that these unreported figures seem to support our recommendation.

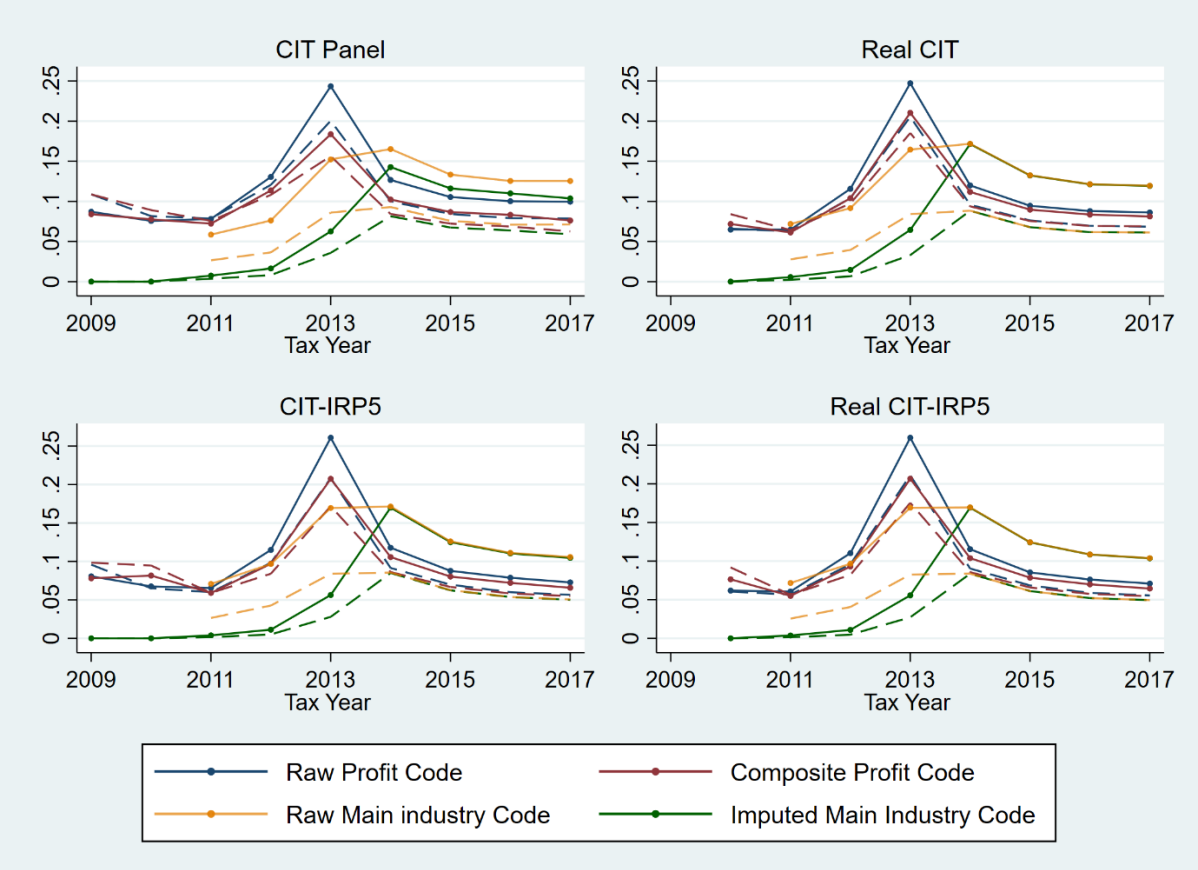
Figure 14: Completeness of industry classifications across 4 subsamples



Note: solid lines indicate the proportion of firms with non-missing industry information for the specified variable after the variable has been converted to the SIC 5 1-digit level. Dashed lines indicate the proportion of firms with non-missing industry information for the specified variable after the variable has been converted to the SIC 5 3-digit level.

Source: authors' calculations using the SARS-NT data.

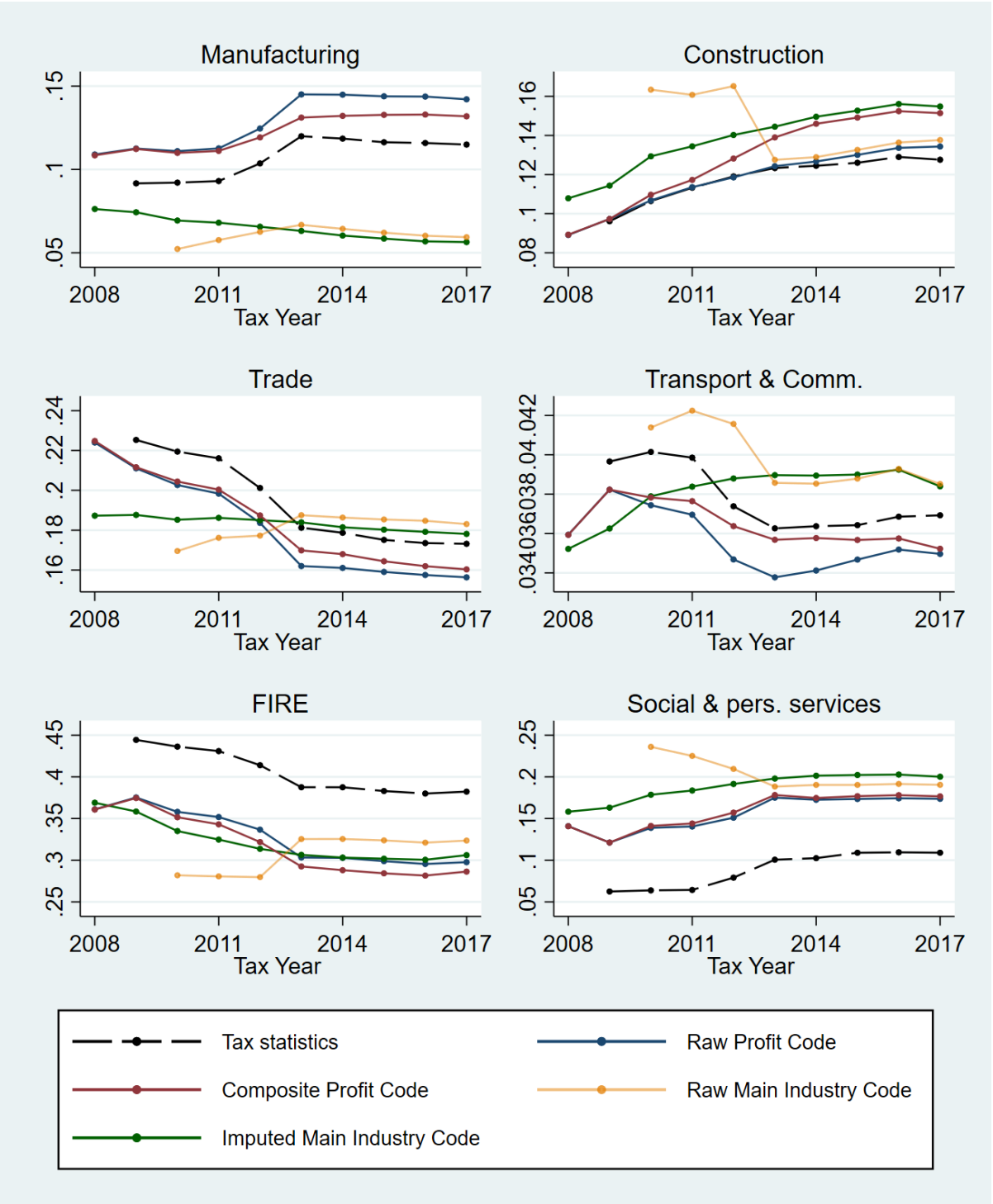
Figure 15: Industry switching over time across 4 subsamples



Note: solid lines indicate the proportion of firms switching industry at the SIC 5 3-digit level, while dashed lines indicate switching industry at the SIC 5 1-digit level. Switches between (or staying in) system-missing industry variables are not considered.

Source: authors' calculations using the SARS-NT data.

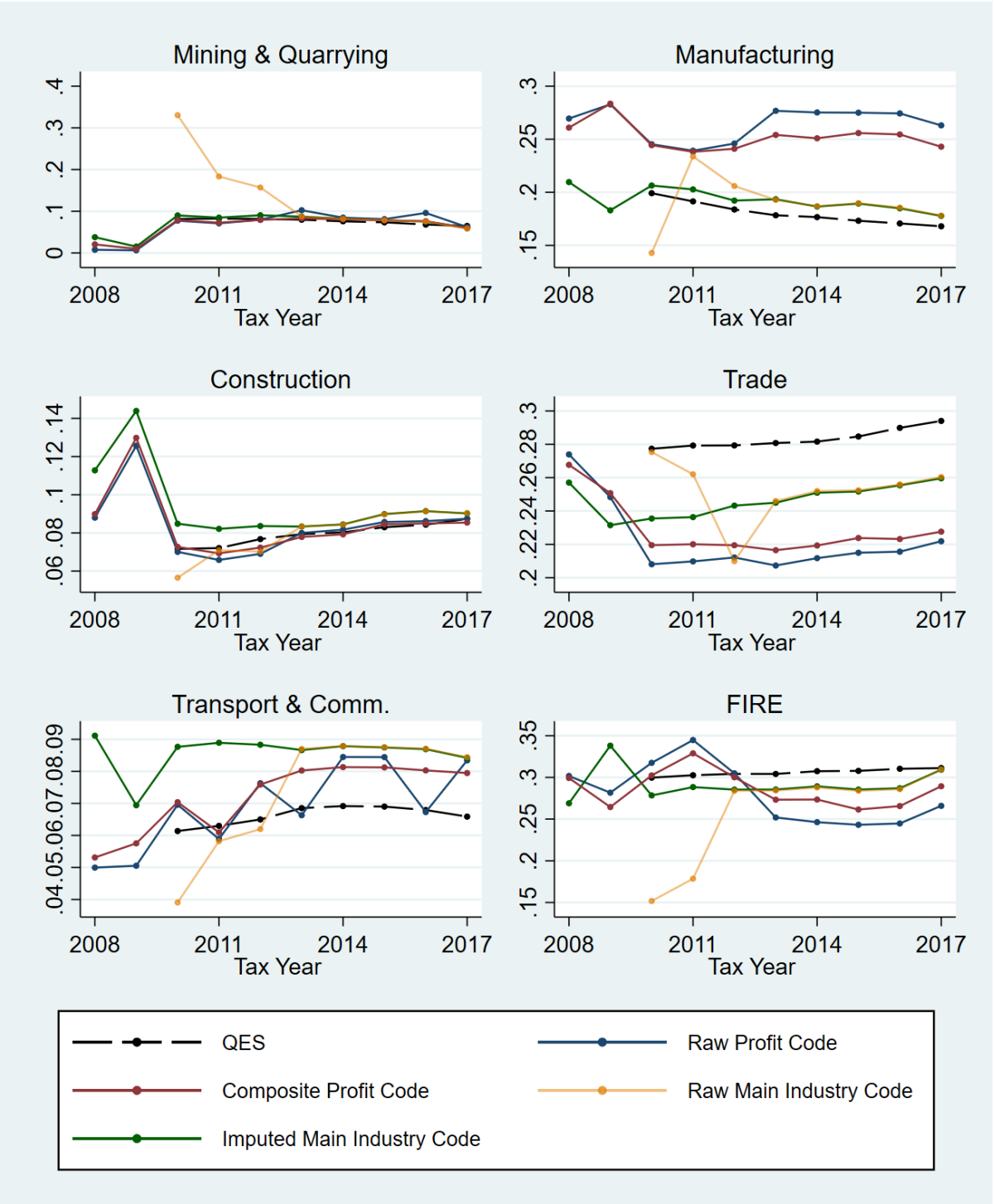
Figure 16: Proportion of firms accounted for by each industry, comparing with Tax Statistics



Note: each solid line shows the proportion of firms in the CIT Panel sample accounted for by each industry using the specified industry variable. The black dashed line shows the same proportion from the SARS Tax Statistics.

Source: authors' calculations using the SARS-NT data and SARS Tax Statistics.

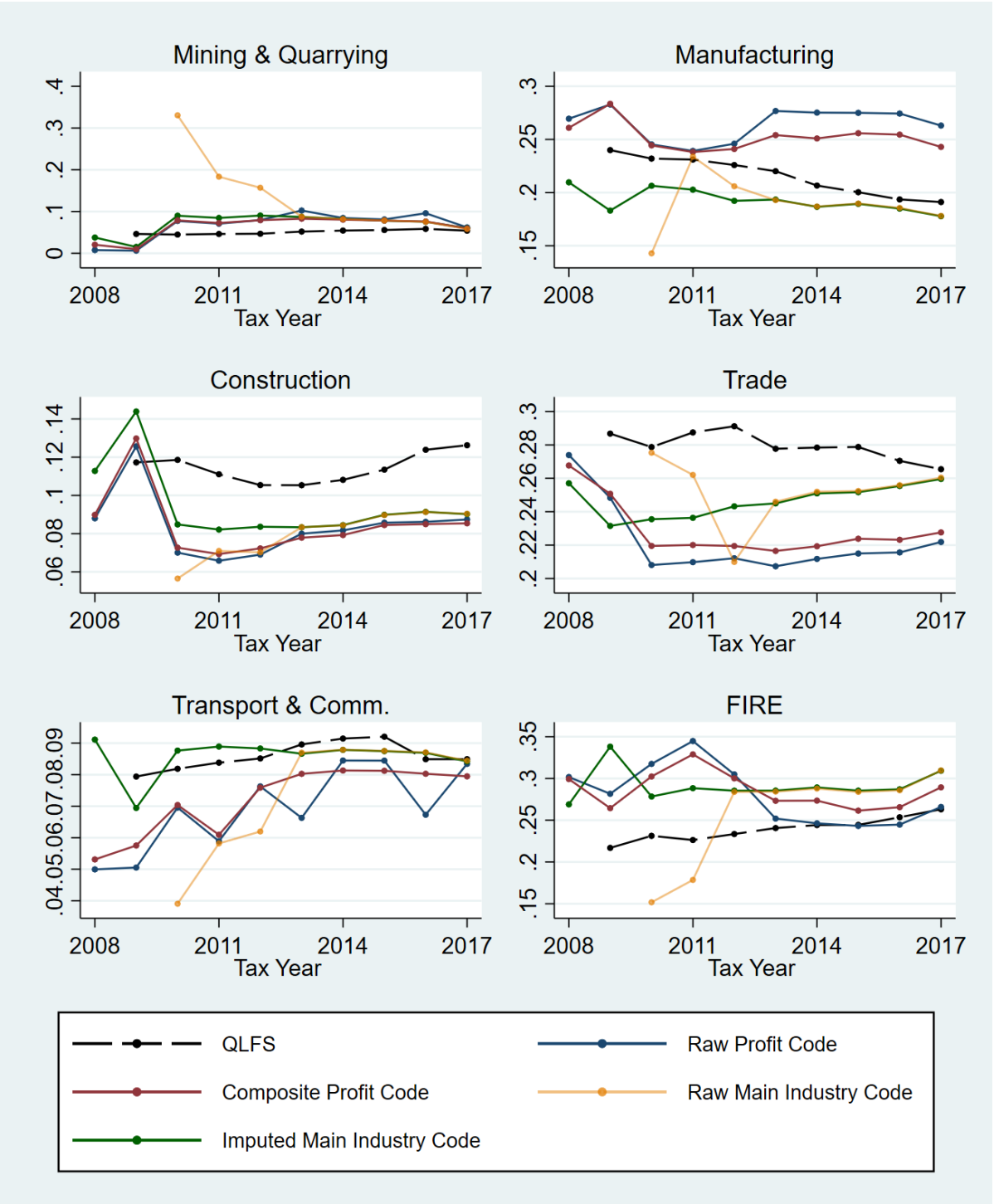
Figure 17: Proportion of employees accounted for by each industry, comparing with QES



Note: each solid line shows the proportion of employees in the CIT-IRP5 sample accounted for by each industry using the specified industry variable. The black dashed line shows the same proportion from the QES.

Source: authors' calculations using the SARS-NT data and Statistics South Africa QES.

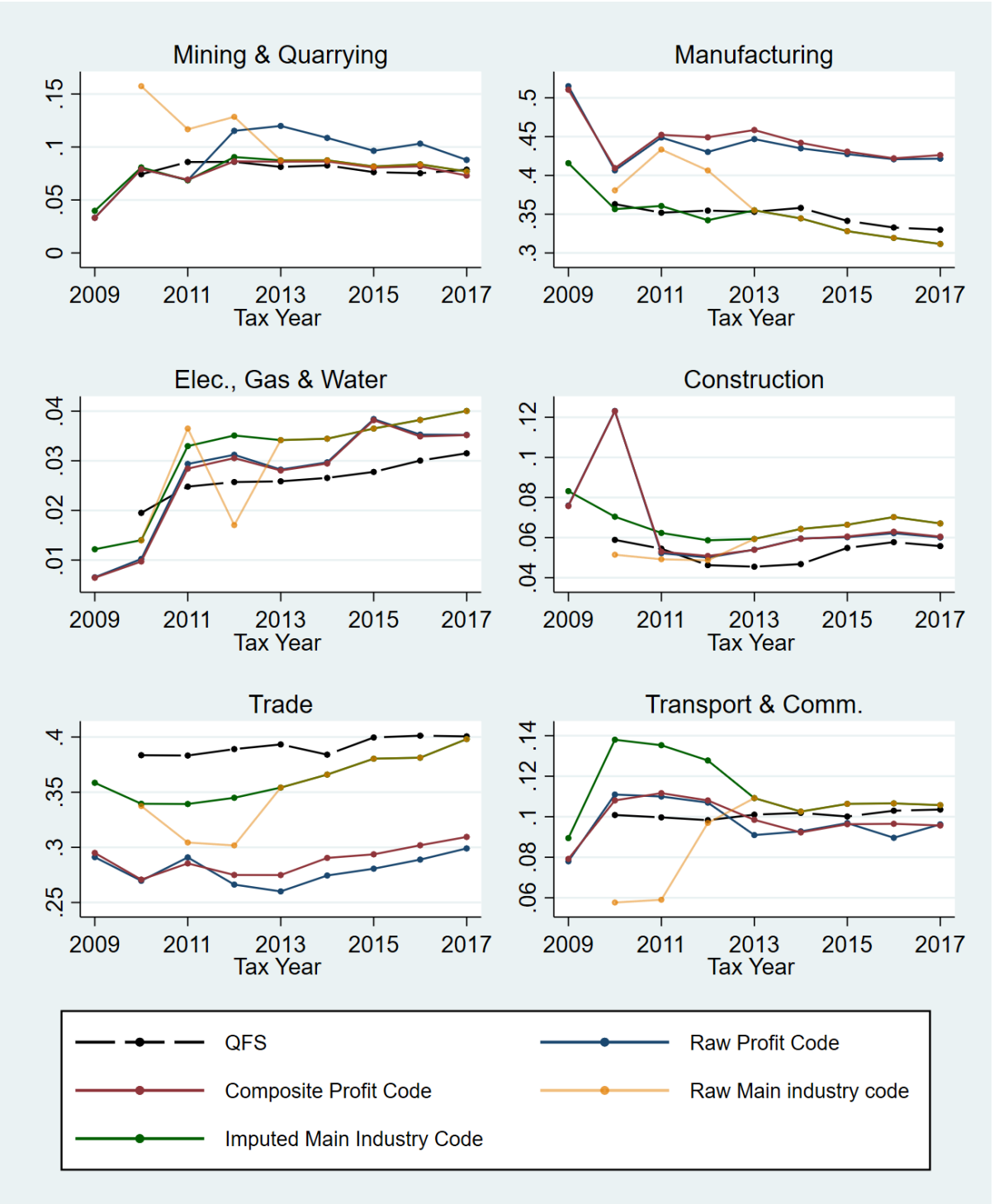
Figure 18: Proportion of employees accounted for by each industry, comparing with QLFS



Note: each solid line shows the proportion of employees in the CIT-IRP5 sample accounted for by each industry using the specified industry variable. The black dashed line shows the same proportion from the QLFS.

Source: authors' calculations using the SARS-NT data and Statistics South Africa QLFS.

Figure 19: Proportion of sales accounted for by each industry, comparing with QFS (turnover)



Note: each solid line shows the proportion of gross sales in the Real CIT sample accounted for by each industry using the specified industry variable. The black dashed line shows the same proportion from the QFS, using turnover.

Source: authors' calculations using the SARS-NT data and Statistics South Africa QFS.

8 Conclusion

The most important task of this paper is to propose a 'best' industry variable for researchers using the SARS-NT dataset. To this end, we suggest using the Imputed Main Industry Code discussed above. We are, however, transparent about the weaknesses of this variable and note that for some purposes the Composite Profit Code will be superior. We have written Stata Code that generates these variables from the raw extractions for the NT-SDF panel-updating team, and this code is available to researchers.

We should acknowledge upfront that the SIC 7 Imputed Main Industry Code is not a new product of this paper: it did exist in the old panel, along with various other industry variables.[34] However, we hope that our documentation and analysis of the different industry variables provides a clearer justification for use of the Imputed Main Industry Code than has existed before. The raw SIC 7 Main Industry Code best matches the other South African data available to us, it is highly internally consistent, and the SIC 7 classification system is granular, modern, and internationally comparable. The major downside of the raw Main Industry Code is that it is very poorly populated before 2013: hence the over-time imputation used for our suggested SIC 7 Imputed Main Industry Code. This imputation does not come without costs, however, and in particular we are concerned about creating a pre-2013 sample that is biased towards firms that survive into 2013. Researchers for whom this a problem may prefer to use our Composite Profit Code or raw Profit Code.

This project has also produced gains beyond these 'best variables', some of them unintentional as we dealt with the lack of systematic industry classification information and concordance for the variables in the SARS-NT data. We have produced new, machine-readable industry classification tables for the SIC 5, SIC 7, SARS Profit Codes, and SARS Activity Codes systems and we have created concordance tables for the conversion of the last three classifications to SIC 5. We have made strides towards understanding the origins of the various SARS-NT industry variables, though not all are fully understood. And we have for the first time developed clear and documented code for the construction of industry variables from the raw SARS extractions, which can be used for panel updating in the future—and which hopefully will facilitate researcher scrutiny of our methods.

[34] To pre-empt possible confusion: while in the analysis of Sections 6 and 7 of this paper we use the Main Industry Code *converted to SIC 5 format*—for the purposes of comparison to the other industry variables—the Main Industry Code we recommend for the panel is in its *original SIC 7 format*. As discussed in Section 7.2, the granularity and modernity of the SIC 7 system is a major benefit of the Main Industry Code, and it is also good to avoid the concordance process where this is possible.

References

Budlender, J. (2019). 'Markups and Market Structure in South Africa: What Can Be Learnt from New Administrative Data?'. Working Paper 2019/58. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2019/692-0

Kerr, A., and M. Wittenberg (2019). 'A Guide to version 3.3 of the Post-Apartheid Labour Market Series (PALMS)'. Cape Town: DataFirst.

National Treasury and South African Revenue Services (2008–2019, multiple years). 'Tax Statistics: Company Income Tax Tables (2008–2019)'. Pretoria: National Treasury and South African Revenue Services. Available at: https://www.sars.gov.za/About/SATaxSystem/Pages/Tax-Statistics.aspx (accessed January 2020).

Newman, C., J. Rand, and F. Tarp (2013). 'Industry Switching in Developing Countries'. *The World Bank Economic Review*, 27(2): 357–88. https://doi.org/10.1093/wber/lhs030

Pieterse, D., E. Gavin, and C.F. Kreuser (2018). 'Introduction to the South African Revenue Service and National Treasury Firm-Level Panel'. *South African Journal of Economics*, 86(S1): 6–39. https://doi.org/10.1111/saje.12156

SARS (2013). *Comprehensive Guide to the ITR14 Return for Companies*. Pretoria: South African Revenue Services.

SARS (2020). 'Find a Source Code' [online]. Pretoria: South African Revenue Services. Available at: https://www.sars.gov.za/TaxTypes/PIT/Tax-Season/Pages/Find-a-Source-Code.aspx (accessed May 2020).

SARS (no date a). *VAT/EMP 403: Vendors and Employers Trade Classification Guide*. Pretoria: South African Revenue Services. Available at: https://www.sars.gov.za/AllDocs/OpsDocs/Guides/LAPD-VAT-G01%20-%20VAT%20403%20Vendors%20and%20Employers%20Trade%20Classification%20Guide.pdf (accessed September 2019).

SARS (no date b). *External Guide: Guide for completion of VAT Registration application* forms (5th rev., VAT-REG-02-G01). Pretoria: South African Revenue Services.

SARS (no date c). *External Guide: Guide for completion of Employer Registration application* (5th rev., EMP-REG-03-G01). Pretoria: South African Revenue Services.

Statistics South Africa (1993). *Standard Industrial Classification of all Economic Activities (SIC) Fifth Edition*. Pretoria: Statistics South Africa.

Statistics South Africa (2012). *Standard Industrial Classification of all Economic Activities (SIC) Seventh Edition*. Pretoria: Statistics South Africa.

Statistics South Africa (2009–2019, multiple years). 'Quarterly Financial Statistics'. Pretoria: Statistics South Africa. Available at: www.statssa.gov.za/?page_id=1866&PPN=P0044&SCH=7616 (accessed January 2020).

TaxTim (2020). 'Business Code Table' [online]. Cape Town: TaxTim. Available at: https://www.taxtim.com/za/tax-guides/definitions/business-code-table (accessed May 2020).

Datasets

National Treasury and UNU-WIDER. CIT-IRP5 Firm 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.

South African Revenue Service. IRP5 2008–2017 [dataset]. Pretoria: South African Revenue Service [producer of the original data], 2019.

South African Revenue Service. IT14 2008–2013 [dataset]. Pretoria: South African Revenue Service [producer of the original data], 2015.

South African Revenue Service. ITR14 2013–2017 [dataset]. Pretoria: South African Revenue Service [producer of the original data], 2019.

South African Revenue Service. Value Added Tax 2008–2017 [dataset]. Pretoria: South African Revenue Service [producer of the original data], 2019.

Statistics South Africa. Quarterly Employment Statistics Breakdown 2009/09–2019/09 [dataset]. Pretoria: Statistics South Africa [producer of the original data], 2019a. Available at: www.statssa.gov.za/?page_id=1854&PPN=P0277 (accessed January 2020).

Statistics South Africa. Quarterly Labour Force Survey Trends 2008–2019Q3 [dataset]. Pretoria: Statistics South Africa [producer of the original data], 2019. Available at: www.statssa.gov.za/?page_id=1854&PPN=P0211 (accessed January 2020).

Appendix A: The general structure of industry classification systems

An industry classification system is a means to assign an (alpha)numeric code to each industry category description. It is typically made up of a few levels of aggregation, each level containing some number of industrial categories, where the number of categories per level generally increases (and never decreases) with each sub-level. While a super-category may be associated with a number of sub-categories, each sub-category should be associated with just one super-category per level. This rather abstract description becomes clearer with an example, which is given in Table A1.[35]

As is apparent from Table A1, each category in each level has both a code (usually numeric) and a label. Where the codes are numeric, as in the example, the N-digit code of a Level N category can be broken up along its digits to indicate all of the other super-categories to which it belongs. This allows aggregation to the 1-digit level when given codes from the 5-digit level, for example.

Note also from Table A1 that sub-categories are not always more detailed than their super-categories. In general, researchers should be aware that 'lower' levels do not always indicate a narrower category, though they should never indicate a broader category. A last point is that industry classification schemes should be exhaustive. Every firm should be classified in some category at every level. In the example in Table A1, fisheries do not fit into any category below the 1-digit level. In reality, the SIC 5 system does have categories for fisheries. If it did not, these firms would be in some category for firms 'not elsewhere classified' (n.e.c.). These n.e.c. categories can be at any level, so there may for example be a 5-digit category for firms which are in the 4-digit 'Farming of equestrian animals' category but farm neither donkeys nor horses, or there may be a general 1-digit n.e.c. category. In general, one can expect a variety of n.e.c. categories at different levels.

[35] Note that the classification scheme in Table A1, while loosely based on the SIC 5 system, is modified and simplified for the purposes of illustration. It is *not* the genuine SIC 5 industry classification scheme even for the subset of industries it represents and should not be used as such.

Table A1: Example industry classification

Code (1-digit) Description (level 1)	**1 Agriculture, hunting, forestry and fishing**							**3 Manufacturing**								
Code (2-digit) Description (level 2)	11 Agriculture, hunting, and related services					12 Forestry and logging		30 Manufacture of food, beverage, and tobacco products					31 Manufacturing of textiles, clothing, and leather goods			
Code (3-digit) Description (level 3)	111 Growing of crops		112 Farming of Animals			121 Forestry	122 Logging	302 Manufacture of dairy products		305 Manufacture of beverages			311 Spinning, weaving, finishing of textiles			312 Manufacture of footwear
Code (4-digit) Description (level 4)	1111 Growing of cereals	1113 Growing of vegetables	1121 Farming of equestrian animals		1123 Farming of dairy animals	1210 Forestry	1220 Logging	3020 Manufacture of dairy products		3052 Manufacture of beer		3053 Manufacture of soft drinks	3111 Weaving		3112 Finishing of textiles	3120 Manufacture of footwear
Code (5-digit) Description (level 5)	11110 Growing of cereals	11130 Growing of vegetables	11211 Farming of horses	11212 Farming of donkeys	11230 Farming of dairy animals	12100 Forestry	12200 Logging	30201 Processing of fresh milk	30202 Manufacture of butter and cheese	30521 Non-Sorghum beer	30522 Sorghum beer	30530 Manufacture of soft drinks	31113 Weaving of animal fibres	31114 Weaving of vegetable fibres	31120 Finishing of textiles	31200 Manufacture of footwear

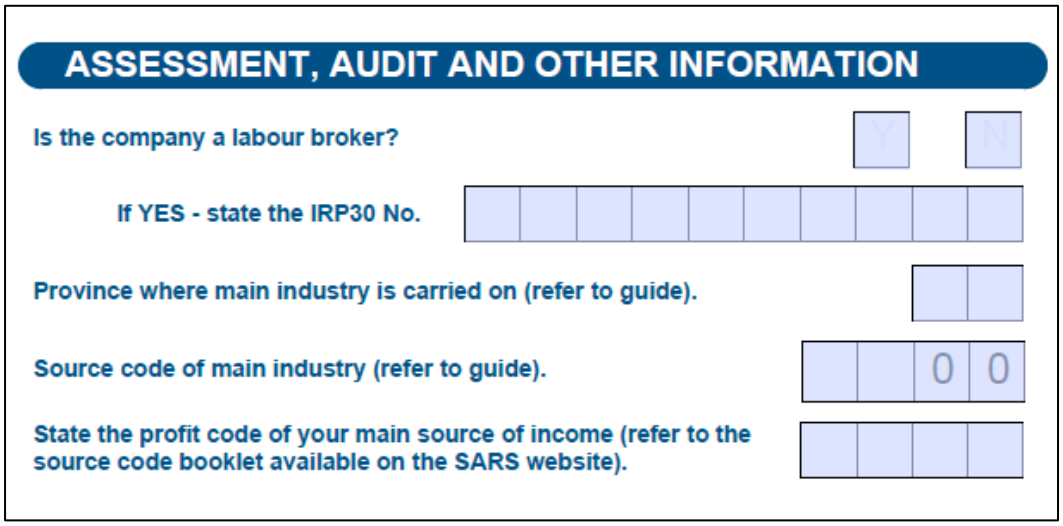
Note: this example classification scheme, while loosely based on the SIC 5 system, is modified and simplified for purposes of illustration. It is *not* the genuine SIC 5 industry classification scheme even for the subset of industries it represents and should not be used as such.

Source: authors' construction.

Appendix B: Company Income Tax (CIT) data—IT14

The IT14 form, like the ITR14 form, makes space for a 4-digit 'profit code', but then requires a 2-digit 'source code of main industry', as indicated in Figure. B1.

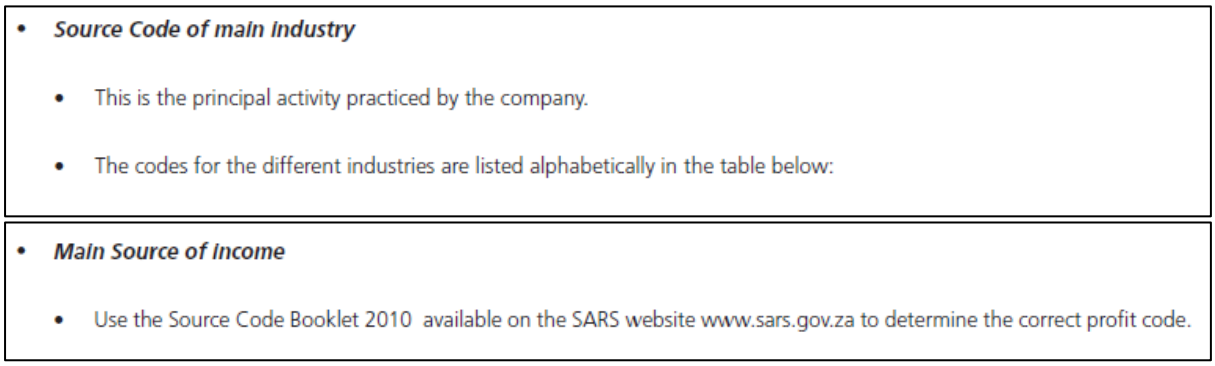
Figure B1: Income Tax Return for Companies form industry fields (IT14)



Source: IT14 v2009.0.0.1 form, available from SARS.

The 4-digit profit code uses the same SARS Profit Code system as in the ITR14, and the same data-cleaning activities are applied to this record. However, the system used for the 2-digit 'Source code of main industry' is unknown to us, or it is perhaps inconsistently applied. While the source codes mentioned in the IT14 user guide (Figure B2) and reproduced in Table B1 are clearly a version of the 2-digit profit codes, actually existing 2-digit records from this field in the IT14 extraction data do not seem to always follow this system, with large numbers of records indicating codes greater than 35. Ultimately, however, the coding system for the 2-digit field is somewhat moot, as IT14 forms are no longer used, a 2-digit variable lacks the granularity to be very useful, the 2-digit system does not maintain continuity with the post-2013 ITR14 SIC 7 system, and current SARS extractions of pre-2013 data (used by SARS-NT panel administrators to make revisions and adjustments from SARS) no longer even include entries from this field. We thus note the existence of this entry but do not attempt to use it in our project.

Figure B2: Guide text to complete the IT14 industry fields



Source: How to Complete the IT14 Return.

Table B1: IT14 2-digit code for ‘Main source of income’

Source Code	Description
0100	Agriculture, forestry and fishing
0200	Mining and quarrying
0300	Food, drink and tobacco
0400	Textiles
0500	Clothing and footwear
0600	Leather, leather goods and fur (excluding footwear and clothing)
0700	Wood, wood—products and furniture
0800	Paper, printing and publishing
0900	Chemicals and chemical, rubber and plastic products
1000	Coal and petroleum products
1100	Bricks, ceramic, glass, cement and similar products
1200	Meta
1300	Metal products (except machinery and equipment)
1400	Machinery and related items
1500	Vehicles, part and accessories
1600	Transport equipment (except vehicles, parts and accessories)
1700	Scientific, optical and similar equipment
1800	Other manufacturing industries
1900	Electricity, gas and water
2000	Construction
2100	Wholesale trade
2200	Retail trade (including mail order)
2300	Catering and accommodation
2400	Transport, storage and communication
2500	Financing, insurance, real estate and business services
2600	Long-term insurers
2700	Educational services
2800	Research and scientific institutes
2900	Medical, dental and other health and veterinary services
3000	Social and related community services
3100	Recreation and cultural services
3200	Personal and household services
3300	Specialised repair services
3400	Agencies and other services
3500	Employment (salary)

Source: How to Complete the IT14 Return.

Appendix C: Miscellaneous Profit and Activity Codes

Table C1: 3501–3534 Activity Codes and category descriptions

Code	Description
3501	Agriculture, Forestry & Fishing
3502	Mining & Stone Quarrying Works
3503	Food, Drink & Tobacco
3504	Textile
3505	Clothing & Footwear
3506	Leather, Leather Goods & Fur (Excluding Footwear & Clothing)
3507	Wood, Wood Products & Furniture
3508	Paper, Printing & Publishing
3509	Chemicals & Chemical, Rubber & Plastic Products
3510	Coal & Petroleum Products
3511	Bricks, Ceramics, Glass, Cement & Similar Products
3512	Metal
3513	Metal Products (Except Machinery & Equipment)
3514	Machinery & Related Items
3515	Vehicle, Parts & Accessories
3516	Transport Equipment (Except Vehicle, Parts & Accessories)
3517	Scientific, Optical & Similar Equipment
3518	Other Manufacturing Industries
3519	Electricity, Gas & Water
3520	Construction
3521	Wholesale Trade
3522	Retail Trade
3523	Catering & Accommodation
3524	Transport, Storage & Communication
3525	Finance, Insurance, Real Estate & Business Services
3526	Public Administration
3527	Educational Services
3528	Research & Scientific Institute
3529	Medical, Dental, Other Health & Veterinary Services
3530	Social & Related Community Services
3531	Recreational & Cultural Services
3532	Personal & Household Services
3533	Specialised Repair Services
3534	Agencies & Other Services

Source: Find a Source Code (SARS 2020).

Table C2: 2-digit Profit Codes and (main activity) categories

2-digit code	Description
01	Agriculture, Forestry and Fishing
02	Mining and Quarrying
03	Food, Drink and Tobacco
04	Textiles
05	Clothing and Footwear
06	Leather, Leather Goods + Fur(Excluding Footwear + And Clothing)
07	Wood, Wood Products and Furniture
08	Paper, Printing and Publishing
09	Chemicals and Chemical, Rubber and Plastic Products
10	Coal and Petroleum Products
11	Bricks, Ceramics, Glass, Cement and Similar Products
12	Metal
13	Metal Products (Except Machinery and Equipment)
14	Machinery and Related Items
15	Vehicles, Parts and Accessories
16	Transport Equipment(Except Vehicles, Parts and Accessories)
17	Scientific, Optical and Similar Equipment
18	Other Manufacturing Industries
19	Electricity, Gas and Water
20	Construction
21	Wholesale Trade
22	Retail Trade
23	Catering and Accommodation
24	Transport, Storage and Communication
25	Financing, Insurance, Real Estate and Business Services
26	Long Term Insurers
27	Educational Services
28	Research and Scientific Institutes
29	Medical, Dental and Other Health and Veterinary Services
30	Social and Related Community Services (Exempt Organisation(s))
31	Recreational and Cultural Services
32	Personal and Household Services
33	Specialised Repair Services
34	Agencies and Other Services

Source: Business Code Table (TaxTim 2020) and VAT/EMP 403 Vendors and Employers Trade Classification Guide (SARS no date a).

Appendix D: VAT Micro Sector

The second VAT industry variable in the SARS extraction is called the VAT Micro Sector, which comes with a 13-category coding system and associated value labels. We have been unable to determine where this variable comes from. It is somewhat unusual, in that despite being very aggregated (having only 13 sectors), it separately distinguishes gold mining and coal mining from 'other mining'. At first glance it looks like a modified SIC 5 1-digit category, but it also separately distinguishes 'Communications' from 'Transport and Storage', which was uncommon in South African classifications until the SIC 7 was published. Given our uncertainty regarding its coding and origin, and its highly aggregated nature, we do not discuss this code further for this project.

Appendix E: Variables

Table E1 provides basic information on the variables used and created in this project, for ease of reference by the researcher interested. These are variables created by the authors from the raw extractions. We provide the variable name, our description of the variable (for details, see main body of this paper), its industry classification system, and its source. The suffix 'X' indicates the digit-level; for example there are in fact five 'itr14_sic7_X' variables in the data, one for each digit-level: 'itr14_sic7_1', 'itr14_sic7_2', 'itr14_sic7_3', 'itr14_sic7_4', and 'itr14_sic7_5'. Below we simply indicate the variable once with an 'X' placeholder for the specific digit-level. A selection of these variables is available in the CIT-IRP5 panel version 4.0, and the remaining variables are available to researchers in the NT-SDF.

Table E1: Variable names, descriptions, and sources

Variable name	Description	System	Source
imp_mic_sic7_Xd	Imputed Main Industry Code	SIC 7	ITR14
comp_prof_sic5_Xd	Composite Profit Code	SIC 5	CIT-IRP5 panel
mic_sic7_Xd	Main Industry Code	SIC 7	ITR14
prof_profcode_Xd	Profit Code	SARS Profit Code	ITR14 & IT14
mic_sic5_Xd	Main Industry Code	SIC 5	ITR14
prof_sic5_Xd	Profit Code	SIC 5	ITR14 & IT14
vat_actcode_Xd	VAT Code	SARS Activity Code	VAT
vat_sic5_Xd	VAT Code	SIC 5	VAT
irp5_actcode_Xd	IRP5 Code	SARS Activity Code	IRP5
irp5_sic5_Xd	IRP5 Code	SIC 5	IRP5

Source: authors' created variables from the raw SARS extractions.