The CenTax Business-to-Worker Register
Technical Note

Arun Advani[†*] Arnaud Dyèvre[‡*] Sebastian Gazmuri-Barker[§*]
Helen Hughson[§*] Sanaya Mahajan[§*] Andy Summers[§*]

May 2026 – Version 1.0
(Latest version)

ABSTRACT. This paper introduces the Business-to-Worker Register (BWR), a database of the universe of UK workers linked to the universe of UK businesses and non-businesses from 2002 to 2022. To build the BWR, we combine (i) information on characteristics and outcomes from individual and business tax returns, (ii) administrative links between businesses and workers, and (iii) external data sources including on patents and business financial accounts. Unlike standard Linked Employer-Employee Datasets (LEEDs), the BWR includes all businesses and non-businesses with no restriction on legal form, industrial sector, employer status, or size. On the worker side, non-employment relationships such as partnerships and self-employment are included, there is no restriction on the age of workers or residency status, and all jobs are observed. We develop methods to estimate profits, investments, capital stocks and depreciations from firm tax returns. To benchmark the BWR, we compare worker and business populations to official estimates. We also benchmark our estimates of firm-level variables against external financial accounts.

Keywords: Matched employer-employee data, administrative data, business demography, business performance

JEL codes: C80, C81, D22, D31, J21, L25

[†] Warwick ; [‡] HEC Paris ; [§] LSE ; [*] CenTax.
For helpful comments, we thank Anthony Savagar, Alwyn Young, numerous HMRC analysts, as well as seminar participants at the ADR UK conference, the Bank of England, the Collège de France, the Darlington Economic Campus, and the Low-Pay Commission. Gabriel Lobo de Oliveira provided outstanding research assistance. Dyèvre thanks the Science for Progress Initiative & MIT Sloan, where he did part of this work.

INTRODUCTION

Some of the most important questions in applied economics require observing workers and firms jointly, at scale, and over time. How is GDP growth distributed across firms and workers? How much of workers' pay is determined by the firms they work for? How do firm-level shocks and policies pass through to workers? How do worker-level shocks and policies affect firms? Answering these questions requires a panel dataset that simultaneously covers all forms of workers *and* businesses, with meaningful measures of economic activity for both.

In this paper, we introduce the Business-to-Worker Register (BWR): a population-scale, matched worker-business panel for the UK covering the period 2002-2022. The BWR links the universe of UK workers to the universe of UK businesses and non-businesses. To build it, we combine 21 distinct databases, supplementing microdata on businesses and workers provided by HM Revenue and Customs (HMRC) with non-fiscal information from Companies House, Orbis IP, and the Office for National Statistics (ONS). As a result, the BWR includes extensive demographics and income variables on the worker side, as well as variables on profits, capital stock, depreciation, investment, R&D expenditures and patents on the business side. We provide a detailed description of the construction of the BWR, benchmark its coverage against official statistics, and validate firm-level variables against publicly available financial accounts.

The BWR improves upon existing Linked Employer-Employee Databases (LEEDs) along several dimensions. Conventional LEEDs focus on linking employees to their employers but the BWR expands this focus by (1) including all work relationships between workers and organisations, (2) maintaining a coherent individual spine over time, (3) maintaining a coherent organisation spine over time, (4) adding a rich set of variables to workers and, (5) adding a rich set of variables to businesses.

On the worker side, the BWR includes not only employees but also non-employment work relationships such as partners, sole proprietors, and directors in corporations.[1] These non-employment relationships are economically important. They account for 15% of all workers in the UK in 2021, and 66% of all private businesses have at least one non-employment relationship that year.[2] Beyond their prevalence, non-employment relationships in unincorporated businesses have been shown to be important for understanding the rise in top income shares over the last few decades in the US (Cooper *et al.*, 2016; Smith *et al.*, 2019; Kopczuk and Zwick, 2020), and for measuring

[1] Strictly speaking, UK tax legislation refers to directors as "office holders", of which directors are the predominant category. We use "directors" throughout the paper for readability.

[2] We find that the vast majority of businesses with non-employment relationships are sole proprietorships (4.3 million in 2021, out of 7.1 million private businesses). Partnerships only account for 361,000 businesses in 2021.

top incomes and characterising the tax responses of business owner-managers in the UK (Advani and Summers, 2024; Miller *et al.*, 2024). Secondly, the BWR includes all workers regardless of age or residency status while a lot of LEEDs restrict coverage to the working-age residents. Non-residents working for a UK company or partnership are a substantial minority of the UK workforce and contribute to UK domestic output, so are important to account for in economic analyses even though (by definition) they are not tax resident in the UK. Similarly, workers past the statutory retirement age are a key population of interest when studying retirement decisions and their impacts on firms. Lastly, the BWR includes all jobs held by each worker within a tax year, rather than assigning each worker to a single main job, allowing us to observe multiple job-holding and complex work arrangements. We complement individual records in the BWR with demographics (age, sex, location, migrant status, country of origin), tax-derived income variables (labour income, including the trading profits of sole traders and partners as well as earnings from employment, plus capital income), pensions, and taxes paid.

On the business and employer side, the BWR also extends coverage beyond what is typically covered by LEEDs. It covers businesses that employ workers as well as the many non-employing businesses such as small partnerships and single-person sole proprietorships in the UK. It also covers non-businesses: non-profit organisations, public-sector entities, and household employers.[3] Total counts of businesses and non-businesses observed in the BWR track official estimates (Department for Business, Energy & Industrial Strategy, 2022), which are based on extrapolations from surveys, employer registrations and VAT records. But our data captures a larger number of active businesses than official estimates. We complement business records in the BWR with variables derived from tax returns, including turnover, profits, intermediate inputs expenditure, investment, capital stock, R&D spending, patents and business taxes. Variables such as investment, accounting profits (EBIT/EBITD) and capital stock are not directly observed in tax returns, but we develop methods to recover them from tax data and validate them against external financial accounts. This, in turn, enables us to appropriately characterise active businesses based on observable characteristics rather than formal registrations in tax records. It is the combination of these rich sets of firm and individual outcomes with universal linkages that make the BWR a unique asset for economic analysis.

Three institutional features of the UK tax system and its data management make the BWR possible. First, HMRC maintains a business identifier register (known internally as the Business

[3]For non-profits and public-sector entities, the BWR currently captures only those that are employers. Future extensions via Charity Commission records will broaden coverage to non-employing organisations that nonetheless have workers or owners.

Lookup Table, or BLT) that links key business identifiers over time, including corporation tax identifiers, Pay As You Earn (PAYE) employer references, Value Added Tax (VAT) numbers, and registration numbers (CRNs) from Companies House (CH), the UK's public register of incorporated companies.[4] This spine allows us to build a coherent business panel with many-to-one links between business IDs: corporations can have several payroll numbers and, conversely, several companies can share a single VAT number. Second, all required worker and business tax datasets are accessible within a single secure research environment (the HMRC Datalab), which allows consistent linkage, harmonisation, and enrichment while preserving taxpayer confidentiality. The centralised storage of data within one location at HMRC allows us to get access to UK-wide populations unlike, for instance, in the US where some databases of worker characteristics are scattered around individual States. It also allows us to have access to datasets on different types of businesses in one place. The last key characteristic of the UK tax datasets we use is that they are not heavily pre-processed before being made available to researchers. This matters in the case of the number of jobs individuals have in the data. All jobs are kept in the Datalab tax data, unlike in other countries where only the highest-paying job of a worker in a year is kept.

Relationship to existing databases. To situate the contribution of the BWR relative to existing business-worker datasets, we report in Table 1 a set of often-used datasets in the economics literature, along with their main characteristics such as number and types of workers covered, number and types of businesses covered, the years of coverage and the papers describing them.

Take for example the Swedish linked employer-employee data infrastructure, a standard-setting dataset for economic analysis. The Swedish data is a combination of rich individual demographics (LOUISE, then LISA registers) and business financial accounts (FEK) linked together through a comprehensive worker-employer linkage register (RAMS). The combination of these databases comes close to capturing the universe of workers and businesses in Sweden, with the following caveats. Business accounts exclude financial institutions (Friedrich *et al.*, 2025, p. 1, online appendix), non-resident workers (Statistiska centralbyrån, 2016, p. 10), and workers older than 70 are not covered by the LOUISE database (Engbom *et al.*, 2023, p. 398). The BWR improves on these margins by including all workers regardless of age or residency status, and all businesses regardless of legal form, size, or industrial sector, including financial institutions.

Some datasets with more limited coverage than the Swedish ones or the BWR have a larger number of workers, due to the scale of the underlying population. This is the case of the US

[4]The BLT is built from the Inter-Departmental Business Register (IDBR) created and maintained by the Office for National Statistics, the UK statistical agency.

LEHD and MEF databases which have been extensively used in economic analysis. The LEHD is now a standard US Census product covering all private jobs in the US that are recorded by State unemployment insurance programmes. Importantly, independent contractors, unincorporated self-employed workers and certain farm workers are not covered by the LEHD (Graham *et al.*, 2022). In 2010, the latest year for which we could find numbers, 120 million workers were in the LEHD (Green *et al.*, 2017) while the BWR had "only" 26 million workers that year. The BWR, while having fewer workers per year than LEHD, covers all UK workers in all UK businesses and non-businesses and includes business variables that the LEHD does not.

In the UK, several other LEED-building efforts are currently underway. We are aware of three main projects. First, the Department for Work and Pensions (DWP) are leading a project that involves linking individual-level data and firm-level data held by DWP, via the Responsive Administrative Data Infrastructure ("RAPID") and Employment Characteristics Dataset ("ECD"). Second, the Office for National Statistics (ONS) are leading a project (with the Economic Statistics Centre of Excellence, ESCoE) that involves linking PAYE (RTI) data on employees with firm-level data on employers, via the ONS Longitudinal Business Database (LBD), which is an evolution of the ONS Interdepartmental Business Register (IDBR). Third, an academic project funded by Administrative Data Research UK, known as the Wage and Employment Dynamics project (WED), has linked data on a 1% sample of employees from the Annual Survey of Hours and Earnings (ASHE) with firm-level data from the Business Structure Database (BSD), together with a range of other administrative data sources (Whittard *et al.*, 2022). WED researchers have also proposed a roadmap for a more developed LEED covering the full employee population (Forth *et al.*, 2025).

Relative to these other projects, the BWR makes several important contributions. First, we extend coverage beyond employers and employees to all types of worker and business. This introduces additional data linking challenges at the individual level, the organisation level and between individuals and organisations, but as we have emphasised, this step is essential to provide a complete picture of the UK labour market and wider economy. Second, the BWR incorporates a significantly longer panel back to 2002, whereas (so far as we are aware) only the DWP's project currently extends before 2015. This longer panel is crucial when seeking to use the data for the analysis of government policy and measurement of long-term trends. Third, as we have also emphasised above, an important contribution of the BWR lies not merely in its data linkage, but the rich set of outcome measures that we construct at both the individual and organisation-level, which underpins the database's utility for economic analysis. Finally, as a result of this technical

note, the BWR is now accompanied by detailed publicly accessible information on data sources and methodology, which have not yet been published for either the DWP or ONS projects.

Country	Dataset	Years	Businesses (latest yr)	Business types	Workers (latest yr)	Worker types	Source	Link
UK	**BWR**	2002–2022	7.02m	All (incorporated, partnerships, self-employed, gov. & NPOs)	41.70m	Employees, directors, partners, self-employed	This paper	
Austria	ASSD	1972–2021	267k	Firms with ≥1 social-security employee	3m	Non-farm, non-self-employed, non-civil-servant	Zweimüller *et al.* (2009); Böheim and Pichler (2025)	Link
Brazil	RAIS	2007–2014	3.7m	All businesses with employees	51m	All workers at registered employers	Morchio and Moser (2026)	Link
Canada	CEEDD	2001–2020	5.5m	All unincorp. enterprises & private corporations	8m	Employees in included businesses & owners	Statistics Canada (2018)	Link
Denmark	MEE (IDA+FS)	1980–2013	127k	All businesses (no panel)	3.8m	Employees, self-employed; one job per worker	Bobbio and Bunzel (2018)	Link
Finland	FOLK/ FLEED	1988–2016	—	All businesses	1.2m	All workers aged 15–70	Statistics Finland (2024)	Link
France	DADS	1976–2019	2.9m*	All businesses with employees	54m (cross-section); 4.5m (panel)	Non-agricultural salaried workers	Harrigan *et al.* (2021); Bozio *et al.* (2023)	Link
	DSN	2017–present	2m reports/month	All businesses	20m/month	All employees	INSEE	Link
Germany	LIAB	1975–2021	1.8m*	Businesses with ≥1 social-insurance employee	1.2m*	Workers subject to social security, marginal employment, or in receipt of benefits	Panahian Fard *et al.* (2024)	Link
Italy	INPS	1974–2015	1.6m*	All employers in private non-agricultural sector	22m*	All employees in private non-agricultural sector (excl. domestic & self-employed)	Casarico and Lattanzio (2024)	Link
Netherlands	SSD	2006–2019	1m**	All firms	9m**	All workers	Goos *et al.* (2022)	Link
Norway	*a-ordningen*	2015–present	260k**	All businesses with employees	4.6m**	All employees	Norwegian Tax Administration	Link
Portugal	QP	1983–2018	300k*	All private businesses with ≥1 wage earner	3m*	All workers of reporting firms	Carneiro *et al.* (2023)	Link
Sweden	LISA/RAMS/ FEK	1985–2016	279k*	All businesses	5.97m*	Employees, self-employed	Engbom *et al.* (2023)	Link
	LISA/RAMS/ FEK	—present	1.80m	All businesses	5.11m	—	Statistics Sweden	Link
US	LEHD	1990s–present	6.2m	All employers	120m	All workers covered by unemployment insurance (excl. self-employed & independent contractors)	Green *et al.* (2017); Graham *et al.* (2022)	Link
	MEF	1978–present	6.1m	Employing businesses only	155m	All workers who ever had a SSN	Song *et al.* (2019)	Link
	LBD-NE	—present	31.5m	Non-employing businesses only	33m	Non-employing business owners	Goetz *et al.* (2025)	Link

TABLE 1. Common datasets with matched employer-employee linkages

Notes: The table lists the main databases of workers matched to the entities for which they work that have been used in the Economics literature. *Unique through period. **Average across period. Sources with full bibliographic references are available in the bibliography; data documentation files are available on disk.

While the BWR has some unique strengths compared to other datasets, it also has some limitations. First, the BWR does not include variables such as education or occupational information.[5]

[5]However, detailed industries of businesses are observed. For some applications, this information can be helpful to understand the type of work that workers do, and serve as a substitute for occupation data.

Second, a caveat that is shared by other databases is that informal work, grey and black market activities are also not covered. Lastly, the social security system is operated by a different government department from the tax system. This means that we do not observe social security benefits in HMRC data unless these payments are taxable (as with the state pension, for example). Note however that wages do not need to be taxed to be observed in our data; any wage payment made through the PAYE system will be recorded by the BWR.

Roadmap. The remainder of the paper is organised as follows. Sections 1 and 2 show how the worker and organisation spines are constructed, respectively, with Section 2 ending by benchmarking organisation counts against official statistics. Section 3 details how we link individuals and organisations and ends by benchmarking worker counts against official statistics. Section 4 shows how we construct individual-level variables, while Section 5 deals with business variables and benchmarks them against external sources. Section 6 concludes. Throughout the paper, we use *businesses* to refer to corporations, partnerships, and sole proprietors. Businesses and firms are used interchangeably.[6] Non-profits, public-sector entities, and household employers are referred to as *non-businesses*. Finally, businesses and non-businesses are collectively referred to as *organisations*.

1. Constructing the individual spine

The individual spine is the worker-year backbone of the BWR. Building this spine requires addressing the challenge that no single administrative dataset provides a complete register of workers in the UK. A further complication is that HMRC datasets at the worker level are indexed by one or more (anonymised) identifiers that are not consistent across datasets. Therefore, the first construction step is to homogenise these identifiers. All individual IDs are anonymised by HMRC before being made available to researchers; we never observe the underlying raw identifiers in the Datalab.

Unique Taxpayer Reference numbers (UTR) index tax returns known as 'Self Assessment', while National Insurance Numbers (NINos) are used to administer the Pay As You Earn (PAYE) withholding system. In some circumstances a Temporary Reference Number (TRN) can also be issued to specific workers.[7] An individual may have one, two, or all three of these identifiers, and in some circumstances may have multiple NINos or TRNs.

[6] We default to using the term businesses, but favour the use of "firms" when referring to established concepts such as the "firm size distribution".

[7] TRNs essentially act as a placeholder until a NINo is issued. For example, a TRN can be issued to a newly arrived worker who has not yet been issued a NINo (which are assigned to British residents at age 16).

Datasets	IDs	Coverage
Self Assessment - "Valid Views" Main individual tax return	UTRs, NINos	1997–2022
SA102 Employment pages of Self Assessment	UTRs	1997–2022
SA104 Partnership pages	UTRs	1997–2022
COP 10% annual sample of employer–employee records	NINos/TRNs	2001–2008
NPS 10% annual sample of employer–employee records	NINos/TRNs	2009–2014
PAYE – P14 Annual employer–employee records (tax year 2003 missing)	NINos, TRNs	2002–2014
PAYE – RTI Real-Time Information employer–employee records	NINos/TRNs	2015–2025

TABLE 2. Individual-level inputs used to build the individual spine

Notes: The table lists the individual-level tax filings used to build the individual spine, together with the identifiers each source carries. UTR is the Self Assessment Unique Taxpayer Reference; NINo is the National Insurance Number; TRN is the Temporary Reference Number issued while a permanent NINo is pending. Where multiple identifiers appear in the same record (*e.g.*, NINo and TRN in P14), we use their co-occurrence to build the custom ID described in the text. Coverage years refer to the tax years currently available in the Datalab. In some datasets, NINos and TRNs are combined in one column (COP, NPS, and RTI) but in the P14 data, they appear as separate columns.

To cover all workers across all tax datasets and to homogenise worker IDs, we create a custom, unique identifier per individual. This custom ID is associated with UTRs, NINos and TRNs as they exist. It enables us to link individuals across datasets and over time in a stable way, and to join all relevant records together to build a complete picture of an individual's earnings. HMRC has provided various linked NINo-UTR and TRN-NINo matches, from their central data infrastructure. We also add our own links where we observe identifiers paired in datasets with these fields. Where multiple possible matches exist and conflict, the pair observed most often across datasets and years is chosen, with the caveat that identifiers with years of birth more than one year apart are de-linked where a conflict exists.

Table 2 lists the individual-level tax filings used to build the individual spine and the individual-level variables. With these datasets, we observe employees, unincorporated business owners (partners and sole proprietors), and company directors. We observe unremunerated directors through the Directors Spine (see Section 3) but we omit them from worker totals in the BWR for comparability with official worker totals.

Because we observe individuals through the tax system, an individual enters the panel only in years in which they receive taxable income including earnings, pension income, unincorporated business income, capital income and capital gains. Individuals with income below the personal

allowance are captured but unincorporated business owners whose receipts fall below the self-employment filing threshold (currently £1,000 per year) do not appear in our data. Individuals who never received taxable income or gains, such as children and adults who never worked, are absent by construction. Because HMRC data does not contain payments administered by the Department for Work and Pensions, the BWR also does not observe untaxed benefits such as Universal Credit, Housing Benefit, Employment and Support Allowance, or unemployment benefits; individuals who live exclusively on those benefits and do not engage in paid work are thus not captured. The BWR is therefore not a full register of the working-age population, it is a register of economically active individuals in years in which they receive taxable income or gains.

Territorially, the BWR covers all UK tax residents. It also includes some non-residents: those with a UK employer (observed through the PAYE system) and those with UK-source property income, pension income, self-employment income from work done in the UK, partnership income if the partnership has UK operations, and capital gains from UK land and property.[8]

The number of unique individuals in the individual spine significantly exceeds the taxpayer population in any given year, for two main reasons. First, the spine pools all unique individuals identified over the entire period covered by the data (more than twenty years), rather than reflecting a single year's snapshot. Second, it includes individuals who were not liable to tax but who interacted with HMRC in any of the datasets available to us.

The closest existing comparator to our individuals spine is the population targeted by HMRC's Survey of Personal Incomes (SPI) dataset. The SPI comprises a stratified sample of tax records drawn from SA and PAYE data and is used to produce HMRC's Personal Income Statistics, which covers the population of individuals with an Income Tax liability. The SPI population is somewhat broader than the Personal Income Statistics population because the latter excludes individuals with an SA and/or PAYE record who did not have any Income Tax liability, such as those with taxable incomes below the personal allowance and/or who only received taxable capital gains. The SPI population is (to our knowledge) the broadest population of individuals that has been constructed by HMRC for statistical purposes.

There are several reasons why the population covered by our individuals spine can differ from the SPI population. First, the SPI is constructed using a data source (SA Valid Views Version A) that is extracted from administrative systems approximately three months after the relevant SA filing deadline and therefore excludes some late filers. The SPI incorporates weights to account for these late filers, although recent research by Delestre *et al.* (2025) has documented that for the

[8]For details on the SA non-resident population, see the appendix of Advani *et al.* (2025a).

period up to 2017, this approach led to an overestimate of the true SA population. Our individuals spine is unaffected by this issue, because we use additional SA data sources that directly capture late filers beyond the cut-off for inclusion in SA Valid Views Version A, without the need for weighting.

Second, the SPI is constructed by first excluding from the PAYE population any records where an SA return was issued to the individual, then joining the independent samples from SA records and the (remaining) PAYE records. This entails a different method of de-duplicating individuals across SA and PAYE data compared with the BWR, so could result in different estimates of the total number of individuals across the combined populations.

Third, our individuals spine can include some individuals who do not have either an SA record or PAYE record, if they appear in any of the other data sources available to us. For example, an individual may appear in the Migrant Worker Scan (MWS) but not in SA or PAYE data. Individuals who cannot be linked to any SA or PAYE record are not included in our count of workers in the BWR because by construction they cannot be linked to any organisation. However, for some analytical purposes it can still be useful to retain available information about these individuals and, for example, they may be linked to an SA or PAYE record in future. Consequently, we retain these records within our full population.

The Unique Customer Record (UCR) project, initiated internally by HMRC to link unique individuals across all of their interactions with the department, would be valuable for improving the individual spine, although it is currently unclear which years it covers. The project has subsequently been tendered commercially and is expected to cover business entities as well as individuals.

2. Constructing the organisation spine

The organisation spine is the organisation counterpart to the individual spine. It includes corporations, partnerships, and sole proprietors (collectively referred to as *businesses*), non-profits, public-sector entities, and households (referred to as *non-businesses*) under a single register. There are many potential definitions of an organisation, all suited to different contexts. Company branches might count as one independent organisation in one application and not in another. At the margin, it is hard to know how to treat some businesses according to this definition, even with perfect information on the structure of the business. Consider for example a corporate group comprising subsidiaries that each perform separate trades: each subsidiary is semi-autonomous, yet ultimately controlled by the group parent. Is the group itself the organisation, or is each subsidiary

its own organisation? With the BWR, we target the ONS definition of a business "entity"; broadly, an autonomous decision-making unit. We also offer suggestions on how to aggregate entities up to larger groups based on our data and data we could add to the BWR.

All organisations that interact with HMRC, either through taxes paid on profits or on wages, are observed in the BWR. An organisation is included if it meets any of the following criteria. First, it has filed a corporation tax return (CT600), or a self-employment return (SA103 for sole proprietors and SA104/SA800 for partnerships). Second, it has taxable sales or supplies above the VAT threshold, observed via VAT records. Third, it employs at least one worker, even where it makes no taxable profits or sales, observed via PAYE. In combination, these criteria deliver full coverage of UK organisations engaged in any economic activity.

Table 3 lists the business-level tax filings and external business datasets used to build the firm spine. We keep all unique IDs observed across these datasets but some IDs are grouped and subsumed by others. For instance, a CT600-filing corporation may have multiple PAYE reference numbers if it runs several payrolls. But we aggregate up payrolls at the corporation level, and only use the CT600 ID to identify the corporation in the BWR. As a result, our organisation spine has fewer IDs than the union of CT600, SA800, sole proprietor UTR, and PAYE IDs observed across all datasets. Section 3 provides more information about how IDs referring to the same business are grouped together.

Assigning industry codes to organisations requires to fetch individual-level industry codes and aggregate them up at the organisation level. We pull SIC codes from all individual-level datasets that contain them (NPS, RTI, SA800, and the "Valid Views" extracts), and add any we can infer from other filings related to the same employer PAYE scheme references (using links from SA102 and/or P14), to fill any gaps. We then define an individual's main industry as the first known SIC code attached to their highest total earnings.

2.1. **Territorial scope.** The territorial scope of the BWR is determined by UK tax residence and UK permanent establishment. A company is UK-resident if it is incorporated in the UK, or if it is incorporated abroad but has its "central management and control" in the UK. A non-resident company is treated as having a permanent establishment (PE) in the UK if it has a fixed place of business in the UK from which it carries on its business in whole or in part, such as an office, branch, factory, workshop, or place of extraction of natural resources. Partnerships are required to file a UK tax return (SA800) if they have UK-resident partners or if they have UK operations;

Datasets	IDs	Coverage
Business tax filings		
CT600 Corporation tax returns SME and large-company R&D tax credits	CT600 taxpayers, CRNs	2001–2023
SA800 Partnership returns including supplementary pages (SA800PS, SA801, SA802)	SA800 taxpayers, partners' UTRs	1997–2021
Valid Views (SA103-equivalent) Sole-proprietor self-employment returns ("Valid Views")	UTRs	1997–2022
Lookup tables		
Business Lookup Table Links between organisation IDs	VAT, PAYE, CT600, CRN	∼1990s–2025
Inter Departmental Business Register Links between organisation IDs	VAT, PAYE, CT600, Entref	∼1990s–2019
Industry classifications		
RTI, NPS, "Valid Views", SA800, SA103 SIC codes reported directly or inferred from other records for the same business	UTRs, NINos	1997–2025

TABLE 3. Datasets used to build the organisation spine

Notes: The BWR is constructed inside HMRC's Datalab. Conditional on project justification and approvals, additional boxes on tax returns can be requested and incorporated. Coverage years reflect the data actually used by the BWR construction scripts; not all variables are populated in every year.

in the latter case, even partnerships whose partners are not UK-resident must file, although they report only profits from UK sources.

The BWR does not cover non-resident organisations, even when they employ UK-resident workers. The income earned by those workers is still observed on the individuals side of the BWR, but it cannot be linked to any organisation in the data. Note, however, that a foreign company with a UK permanent establishment will appear in the BWR via the UK tax return that it is required to file in respect of the profits attributable to that establishment.

2.2. **Organisation identifiers and statistical units.** There is no authoritative organisation ID that covers all businesses and non-businesses in the UK. We therefore define our own organisation IDs, separately across business and non-business legal forms. Corporations are identified by their CT600 taxpayer number, partnerships by their SA800 taxpayer number, and sole proprietors by their individual-level Unique Taxpayer Reference number. We identify non-businesses by their employer reference ID. This latter alphanumeric ID is accompanied by a "status" variable in

HMRC's Business Lookup Table that enables us to separately identify public entities, non-profits, and households.[9]

Defining IDs in this way has implications for what we consider an organisation in the BWR. Corporations, partnerships and sole proprietors in the BWR are best thought of as the production units responsible for reporting taxable profits and sales. These units may not align with corporate groups, brands, autonomous subsidiaries, or other levels of aggregation of interest to researchers. Several reasons lead us to define corporations as the legal entities chargeable for corporate tax. First, the CT600 variables we use to construct financial accounts (see Section 5) are reported at the legal entity level. Using these variables at another level of aggregation would require us to make hypotheses about how each variable is distributed across entities in a group, and which group is the most relevant (CT600 corporate group? VAT group? Ultimate owner of firms?).[10] Second, total numbers of corporations in the BWR are broadly in line with official estimates, as we show in subsection 2.3, thus providing some reassurance that our fiscal definition of a corporation is reasonable. Lastly and perhaps most importantly, the ONS Entref identifier, our target identifier for autonomous production units, has virtually a one-to-one mapping with the CT600 identifier of corporations. In the legacy cut of the IDBR that we use, there is between zero and one CT600 ID per Entref in 99% of cases.[11] For the same reasons, our tax-based definitions of partnerships and sole proprietors also seem reasonable.

We identify public and government entities by the industry classification of their workers: we use the modal industry associated with their employer reference number across all their employees. To assign an industry code to an individual, we pull SIC codes from all individual datasets that contain them (SA102, NPS, RTI, P14, SA800, and the "Valid Views" extracts) and define an individual's main industry as the one where their highest total earnings are coming from. We then manually classify five-digit SIC codes as being industries that are predominantly public or government. For instance, SIC 84110 (general public administration activities) to SIC 84300 (compulsory social security activities) denote public entities, as does SIC 64110 (central banking). We

[9]Like individual IDs, all organisation IDs are anonymised by HMRC before being made available to researchers. We never observe the underlying raw identifiers in the Datalab.

[10]Links between CT600 filers under a common corporate group may be available to HMRC, but they are not available to us. One may use ownership links available in the Orbis database to link corporations with a common owner. We have not done so but this is an active area of development of the BWR at CenTax. We have links between CT600 IDs and VAT IDs in the IDBR. Note however that the decision to link CT600 filers under one larger group is made by companies themselves at their own discretion, not by HMRC. Even with the data on corporate groups, ownership and VAT groups, it is unclear that it provides us with the most reliable and homogeneously defined set of autonomous production units.

[11]28% of Entrefs are not associated with a corporation tax number: these are Entrefs that are not corporations and are registered either as a VAT filer or as a PAYE employer, the two criteria to be included in the IDBR.

then aggregate these SIC codes by employer reference numbers across their employee-year observations. We keep the mode of the reported SIC codes to create time-varying an time-invariant industry classifications of public entities.[12] Non-profit organisations[13] and household employers[14] are similarly identified by the modal industry of the employee-year observations.

Several extensions to the firm spine are on our agenda. First, we will include ownership links, sourced from Orbis ownership data, so that tax-filing entities belonging to the same ultimate owner can be aggregated for group-level analysis. Second, we will track changes in legal form over time, so that a business that incorporates, disincorporates, or switches between partnership and LLP is followed as a single economic entity rather than resetting at each transition. Lastly, we will attempt to extend the ONS enterprise reference crosswalk sourced from the legacy IDBR beyond its current 2019 cutoff.

2.3. **Benchmarking organisation counts against official statistics.** Having constructed the organisation spine, we can benchmark organisation counts against the official Business Population Estimates (BPE). We defer worker-count comparisons to Section 3, because those require the worker–organisation links established there. The Business Population Estimates are produced annually by the Department for Business and Trade (DBT). They primarily rely on administrative business registrations (via the IDBR, built from VAT, PAYE and Corporation Tax registrations) which are completed by survey-based extrapolation from the Labour Force Survey (LFS) to estimate business and worker counts (Department for Business and Trade, 2025). A key conceptual difference with the BWR is that the BPE are designed around *registered entities* at a point in time (measured as of January 1st of each year). By contrast, the BWR directly observes economic activity of businesses using tax records: we define an active business as an entity that has revenue and/or non-zero wage workers at any point during the tax year.

Figure 1 provides an aggregate comparison of business counts by legal form, with the overall series shown in panel a. The remaining panels in Figure 1 disaggregate this comparison by business type.

Where the series overlap (2010 onward), our counts track the official BPE series fairly closely in terms of broad trends while extending coverage back to 2002. Overall, we estimate an average of

[12] For all applications, we use the time-invariant classification.

[13] SIC codes such as 88000, 88100 and 88990 which cover "Social work activities without accommodation" excluding child day care, and "Higher education" (SIC codes 85400 to 85500), among others, are used to classify employers into non-profit organisations.

[14] Only two industries capture household employers; "Households as employers of domestic personnel" (SIC code 97000) and "Undifferentiated goods- and services-producing activities of private households for own use" (SIC codes 98000 to 99000).

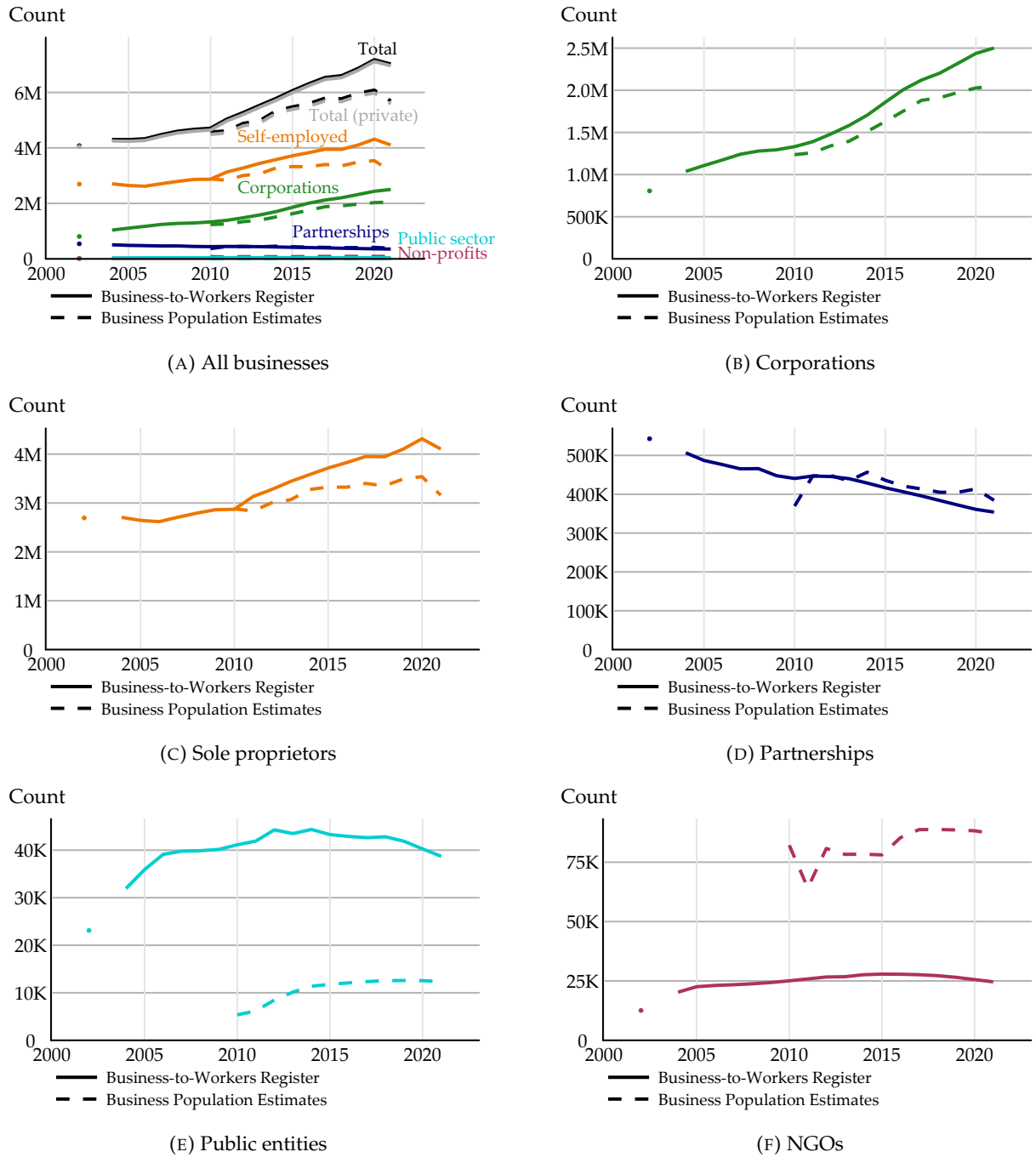


FIGURE 1. Organisation counts comparisons: BWR and official estimates

Notes: The six panels show time series of organisation counts in the UK over 2002–2022, ordered top to bottom and left to right as all businesses, corporations, sole proprietors, partnerships, public entities, and non-profit organisations. Solid lines show BWR counts and dashed lines show the official Business Population Estimates produced annually by the Department for Business and Trade; the official statistics start in 2010 only. The BWR series break in the early years reflects incomplete data availability for parts of the worker-linkage infrastructure of the BWR.

12% more businesses relative to the BPE every year, with incorporated businesses and sole proprietorships contributing the most to this difference. Differences in levels and short-run movements

reflect three measurement differences: (i) declining LFS response rates in recent years, (ii) statistical extrapolation versus direct observation, and(iii) definitional differences in what constitutes an "active" business.

Starting with incorporated businesses (panel b of Figure 1), the BWR estimates 250,000 (14%) more firms per year on average compared to the BPE. The main explanation for this discrepancy is that the BPE only consider businesses that are PAYE and/or VAT registered while we count all businesses with a corporate tax return, irrespective of whether they meet the PAYE and/or VAT registration criteria. Both approaches also count different identifiers, which could represent different aggregate units: we count the unique identifier in the corporate tax data while the BPE count "Entref", an identifier created by the ONS and representing a consistent production unit. There is however a close alignment between Entref and CT600 identifiers so it is not likely that this difference of IDs explains much of the discrepancy. Lastly, while we do try to align with the BPE methodology for these comparisons and count only businesses that contain January 1st in their filing period, our consideration of business activity is at any point in the year. This is different from the BPE which will only capture businesses in the IDBR extract of January 1st. In other words, if a business starts its accounting period in July, stops trading in November and closes its accounts in June of the following year, it will appear as an active business in our data (because January 1st is between July and June) but not in the BPE (because it ceased trading before January 1st). Importantly, our definition of business activity makes our estimates more conservative as we only count businesses with some revenue and/or non-zero wage workers (*i.e.* "active" businesses). The BPE do not impose such restrictions on business counts. It is therefore notable that we find more active corporations in every year than what the BPE report.

Amongst unincorporated businesses (partnerships and sole proprietorships), any differences with the BPE can be primarily attributed to the fact that we provide direct counts from the tax data while the BPE rely on a statistical extrapolation of the LFS. Beyond the benefits of using administrative data *vis-à-vis* survey data for these counts, the survey has seen declining response rates in recent years making it a less reliable source for such estimations. Even with these methodological differences, our counts for the total number of partnerships (panel d of Figure 1) are close to the BPE except for 2010. We estimate around 19,000 (5%) fewer partnerships per year on average from 2011. However, we count 71,000 (19%) more partnerships than the BPE in 2010. The relatively larger difference in 2010 is explained by a methodological shift in the BPE from 2011 that allowed for more accurate estimations for both partnerships and sole proprietorships from

2011 onward. Our data on partnerships only contains start and end months of the reporting periods, unlike corporations, for which we have the exact start and end days. To best approximate the definition of active partnerships used by the BPE we impose that the start and end months of partnerships contain the month of January. We further impose that a partnership is counted as active if it generates a non-zero turnover in its reporting period or if a non-zero number of partners are observed.

For sole proprietorships, we estimate an average of 510,000 (16%) more businesses per year than the BPE (panel c of Figure 1). As the 2011 methodology shift affected sole proprietorships as well, we exclude 2010 from these comparisons. While we do restrict counts to active sole proprietorships, we are unable to check (yet) if the filing period of the businesses contains January 1st. However, the fundamental difference in the estimation strategy between the BWR (directly counting observations) and BPE (statistical extrapolation) could also be playing an important role in the gap between our numbers and the BPE's.

Finally, we estimate substantially more entities for central and local government compared to the BPE (panel e of Figure 1). We get 32,000 (330%) more entities per year on average. The opposite is true for non-profit organisations (NPOs) where we get 56,000 (67%) fewer entities per year on average (panel f). We believe our numbers are not necessarily the most reliable here. Two factors are likely at play to explain these discrepancies. First, exactly which entities are classified as central and local government, and NPOs (based on SIC codes) is not the same in the BWR and the BPE. For example, the BPE methodology lists only three SIC codes being counted as central and local government[15] while the BWR uses a much broader, custom classification. In contrast, it is currently unclear which SIC codes are considered as NPOs by the BPE, but it is likely that we are allocating fewer SIC codes to NPOs (and allocating them to central and local government instead). Second, and more importantly, we are relying on a much more disaggregated identifier for these entity counts—the PAYE scheme reference—than what would be considered a consistent organisation ID. As a result, two payroll numbers used by one organisation will be counted as two public entities, while the BPE will (correctly) count one entity. A hospital, for instance, may keep one payroll number for medical staff and another for admin staff. In this case, the hospital would be counted as two public entities instead of one in the BWR. The BPE, on the other hand, use the ONS Entref which would aggregate several PAYE schemes into one organisation. This particularly contributes to the higher entity counts that we get for central and local government.

[15]SIC codes counted as central and local government by the BPE: 841 (administration of the state and the economic and social policy of the community), 842 (provision of services to the community as a whole), and 843 (compulsory social security activities)

3. Linking individuals and organisations

With the individual and organisation spines in place, we proceed to connecting them. Our aim is to link individuals to the organisations for which they work. We use the term *job* to denote a specific worker–organisation relationship, of which we distinguish five types: employee, sole proprietor (where the individual and organisation coincide), partner, director, and owner-manager. We refer to any individual holding at least one job as a *worker*. A given organisation or worker can hold multiple jobs, and multiple job types, simultaneously: a company may have many employees and many directors; a partnership may have both partners and employees; and a single individual may be, at once, an employee of one firm, a director of another, and a sole proprietor in their own right. Links between organisations and workers are therefore many-to-many. Some specific use cases require us to collapse this structure into unique worker-to-organisation links where one worker is associated with one organisation. For this, we designate a *primary job* for each worker. We develop methodologies to include all five types of jobs in the BWR: employment, sole proprietorship, partnership, director (both remunerated and unremunerated), and owner-manager links.

In the current version of the BWR, we do not include off-payroll work relationships as BWR jobs. These are cases where an individual supplies labour to a client through a personal service company or other intermediary. These work arrangements, sometimes described as "disguised employment" are considered to be employer-employee-like links by HMRC (known as "inside IR35"). Due to the lack of data on inside-IR35 relationships in the Datalab, we do not include them in the BWR at this stage. However, we are exploring ways to incorporate them in future versions of the BWR.

3.1. **Employment relationships.**

Employment links between workers and their employers are drawn from two sources: the PAYE returns filed by employers (P14 and RTI), and the SA102 pages attached to individuals' Self Assessment returns. The PAYE employment links connect individuals' *NINos* to their employer PAYE reference number, while SA102 connects individuals' *UTRs* to their employer. Our individual spine described in Section 1 allows us to homogenise these two types of links and define an employer-by-unique ID register. We take the union of employment links reported in PAYE and SA102 as our universe of employer-employee relationships from 2002 to 2022. Unfortunately, P14 data for 2003 is missing at HMRC so we do not have employment links in that year, beyond those reported in SA102. An important step in the pre-processing of employment links consists in removing employer numbers that are occupational pension reference

numbers. Some remain even after filtering out employer-employee links marked as "P" (for pension) in the RTI data. These PAYE schemerefs in the RTI and P14 data is used to pay pensioners rather than ongoing employees, and if left in place would generate spurious worker-firm links in the BWR. We identify and exclude these schemerefs using a combination of HMRC's occupational-pension indicator, the age profile of payees, and a short hand-curated list of known problematic schemes.

Once individual IDs are linked to employer reference numbers, we still need to link the employer numbers to the economic units at the level of which we have outcome variables, *i.e.*, corporations, partnerships and sole proprietors. The IDBR and BLT lookup tables allow us to aggregate employer numbers at the corporation level. In 93% of the cases, corporation identifiers that are linked to at least one employer number have just one employer number. 0.11% have 5 or more.[16]

Linking partnership IDs (from SA800, where the partnership profits, turnover and capital expenditures are reported) to their employer reference numbers is more complicated because neither the BLT nor the IDBR record these links, so the connection between a partnership and the employees on its payroll must be inferred. We do this in two stages. The first stage uses the fact that many partnerships put at least some of their partners on payroll. For every {partnership UTR, employer number} pair we observe a partner appearing on, we count one "vote" toward this linkage. More occurrences of partners in {partnership UTR, employer number} pairs over the years strengthen the statistical link between the two IDs. We keep pairs with at least two votes and restrict to UTRs whose legal status in the BLT is "partnership". Where a schemeref has candidate links to more than one UTR (or *vice versa*), we retain the pair with the highest vote count. The second stage reconstructs links from partners' careers. For a partner observed in a given partnership, we scan their earlier PAYE history for schemerefs they worked for. Schemerefs held by several partners of the same partnership are taken as candidate employer references for that partnership, after filtering by BLT legal status (`== partnership`). Here again, the more partners "graduate" from a PAYE employer number to a partnership, the stronger the link between the two IDs. The two stages are combined and cross-validated against each other. Because the link is statistically inferred rather than administratively recorded, it carries measurement error that users should account for in applications.

Sole proprietors are trivially linked to their business as the reporting unit in Self Assessment. Sole proprietors may also employ workers, beyond the owner. Unfortunately, no link between

[16]Corporations may decide to run several payroll numbers in parallel. For instance, one payroll for directors or senior managers and another for the general workforce. Or maybe to keep two functions separate within the business; marketing and production for example.

the sole proprietors' UTRs (which index the Valid Views datasets from which we derive sole proprietorships' outcomes) and their employer reference number exists. Unlike with partnerships, we cannot rely on sole proprietors appearing on the payroll of their businesses or on the volume of transitions from a previous employer to the business. We can however link employees in sole proprietorships to their co-employees through their common employer reference number, and we have information about the total wage bill of sole proprietorships in the Valid Views extracts. CenTax is actively working on data-driven ways to match sole proprietors' UTRs to their employer reference numbers by making use of validation samples provided by HMRC.

3.2. **Partnership links.** Partnership links are recovered from the SA800 Partnership Statement (PS) returns filed by the partnership itself and listing its partners, and from the SA104 supplementary page that individual partners attach to their own Self Assessment returns. SA800 filings provide us with partners-to-partnership links from 2011 to 2021 and SA104 returns provide us with partners-to-partnership links from 1997 to 2022. There is not a perfect overlap in the SA800PS and SA104 links in the years where they coincide, so we take the union of links in both datasets to minimise the risk of missing links that would not be reported in one dataset or the other, by mistake.

3.3. **Director links.** We identify directorships using two data sources. First, directors are instructed to file the SA102 (Employment) page for each directorship and to indicate on the form that they were a company director. We link directors to their company via the PAYE scheme reference reported on SA102. Where an individual has reported being a director but the scheme reference is missing, we link them to their company using the closest matched RTI record for that individual (based on amount of income reported). However, some directors may not file SA102 if they are unremunerated, or even if remunerated they may report their income on SA102 but fail to tick the box indicating that they are director.

To address these potential gaps, we supplement the SA102-based link with a second data source known as the 'Directors Spine'. This dataset was constructed by HMRC by matching directors in Companies House data with SA records, using name, date of birth and address. In principle this dataset should cover all directors of UK companies who are also SA filers, although in practice only around half of all directors recorded in Companies House are matched to an SA record.[17] For matched individuals, the dataset contains both their individual UTR and the CT UTR (in turn,

[17]For further details see Miller *et al.* (2024), Appendix A5.

matched one-to-one from Companies House CRN), providing the link between the director and their company.

By combining these two data sources, we obtain a dataset of directors links that is more comprehensive than either data source alone, although it is likely that we still miss some links in cases where an individual has failed to indicate that they were a director on SA102 and their SA record (if they have one) is not matched to any Companies House record. It would be possible to extend coverage by improving the quality of the Directors Spine, in particular by linking Companies House records using the full date of birth (instead of only month and year of birth) and by matching to PAYE records as well as SA records.

3.4. **Owner-manager links.** Owner-manager links are constructed from SA102 data, the Directors Spine, and extracts from SA103.[18] We reclassify a director as an owner-manager when two conditions both hold: the director's firm is a close company, and dividends are the director's main source of income. Among directors of corporations, we currently only include remunerated directors in the BWR — those captured via SA102 (with the PAYE schemeref or RTI-fallback link described above) or via a Directors Spine record that coincides with an SA102 or PAYE record at the same firm. Unremunerated directors are observed through the Directors Spine alone, and we have cleaned the full Spine (remunerated and unremunerated) for other applications within the team, but for comparability with official worker totals we do not include them in the BWR.

Table 4 summarises the linkage inputs for each job type, distinguishing administratively observed records from BWR-constructed linkages.

Several extensions to the linkage infrastructure are on our agenda. First, we plan to publish summary statistics and joint distributions of workers and organisations at the job, individual, and organisation levels, which can serve as moments for macro models and as additional benchmarks. Second, umbrella companies and personal service companies (PSCs) complicate the assignment of workers to the appropriate firm: individuals providing services through a PSC are ideally classified as employees of their client(s) if inside IR35 (the "off-payroll working" rules) and as owner-managers otherwise, and temporary workers supplied via an umbrella agency should be linked

[18] HMRC makes available to researchers two important datasets built from Self Assessment returns: the "Valid Views" and the "Invalid Views" extracts. Both Valid and Invalid Views contain subsets of all the variables derived from Self Assessment returns. They cover different population though: individuals in Valid Views are considered to have a valid UK filing address (hence the name), and individuals in Invalid Views are taxpayers without a valid UK filing address. We use both datasets to build owner-manager links.

Datasets	IDs	Coverage
Employee ↔ employer linkages		
PAYE – P14 Annual employer–employee records (tax year 2003 missing)	NINo ↔ PAYE schemeref	2002–2014
PAYE – RTI Real-Time Information employer–employee records	NINo ↔ PAYE schemeref	2015–2022
SA102 Employment pages of Self Assessment	UTR ↔ employer name/ref	1997–2022
Partner ↔ partnership linkages		
SA104 / SA800 Union of individual partnership pages (SA104) and partnership-filed partner lists (SA800)	UTR ↔ partnership UTR	1997–2022
Partnership ↔ PAYE schemeref links BWR-constructed; inferred from partners' payroll records and career trajectories	partnership UTR ↔ PAYE schemeref	1997–2022
Director ↔ company linkages		
SA102 director flag Self Assessment filers who tick the director box; link to company via PAYE schemeref (RTI fallback)	UTR ↔ PAYE schemeref	1997–2022
Directors Spine HMRC-constructed; tax-to-CH match on name, date of birth, and address	UTR ↔ CT UTR	2002–2022
Firm-ID crosswalks		
Business Lookup Table Links between organisation IDs	VAT, PAYE, CT600, CRN	∼1990s–2025
Inter Departmental Business Register Links between organisation IDs	VAT, PAYE, CT600, Entref	∼1990s–2019

TABLE 4. Administrative linkage tables used to connect individuals to firms

Notes: The IDs column lists only the identifiers through which each row links individuals to organisations; each dataset contains additional variables. The Partnership ↔ PAYE schemeref row is BWR-constructed and carries measurement error; the Directors Spine row is HMRC-constructed by matching Companies House data with tax records and inherits coverage gaps from that match (see Section 3). Applications should treat both links probabilistically. Owner-manager links are not listed separately: they are directors from the third block reclassified using close-company status and SA103 dividend information (see Tables 2 and 3).

to their client firms rather than to the umbrella. We do not currently have solutions to these issues but plan to engage with HMRC to scope the magnitude of affected workers and firms and to develop a strategy for further statistical work in this area.

3.5. **Benchmarking worker counts against official statistics.** Section 2.3 compared organisation counts in the BWR to the official Business Population Estimates (BPE) (Department for Business and Trade, 2025). Having now established the worker–organisation links, we can benchmark worker counts as well. The key conceptual difference is that the BPE are designed around jobs

at a point in time (measured as of January 1st of each year), whereas the BWR counts individual workers over the full tax year and explicitly includes non-employee work relationships.

Figure 2 provides the worker-count benchmarking panels, with the overall comparison shown in panel a. The remaining panels disaggregate this comparison by organisation type and show that, once the links are established, the BWR broadly reproduces the main time-series patterns of the official statistics. For workers in incorporated businesses (panel b of Figure 2), we count those who work for incorporations identified as "active". However, because we estimate more incorporated businesses to begin with, our worker counts for these are also higher: 5,200,000 (25%) more per year on average. In addition, we do not restrict our counts to workers active on January 1st unlike the BPE which use an IDBR snapshot from January 1st and take employment counts as of that date.

For worker counts within unincorporated businesses, the use of the LFS has an additional limitation. The BPE only use a respondent's first and second jobs as reported in the LFS. The tax data is not subject to this limitation, and once the worker–organisation links are established we can count all jobs associated with an individual. In terms of the number of workers in partnerships (which include both partners and employees), panel d of Figure 2 shows that we get around 160,000 (10%) more per year on average excluding 2010 (as the BPE methodology was less accurate then). We count workers in "active" partnerships only, but unlike the BPE we do not restrict the counts to workers active on January 1st, so we get higher estimates.

For workers in sole proprietorships, we estimate an average of 1,500,000 (36%) more workers (sole proprietors themselves and their employees) per year than the BPE (panel c of Figure 2). As the 2011 methodology shift affected sole proprietorships as well, we exclude 2010 from these comparisons. While we do restrict counts to workers within active sole proprietorships, we are unable to count workers active on January 1st. However, the fundamental difference in the estimation strategy between the BWR (directly counting observations) and BPE (statistical extrapolation) could also be playing an important role in the gap between our numbers and the BPE's.

For worker counts, we estimate 1,300,000 (25%) more workers per year on average in central and local government and 130,000 (6%) fewer workers per year on average in non-profit organisations relative to the BPE (panels e and f of Figure 2). As discussed in subsection 2.3, differences in public/non-profit classification and the use of PAYE scheme references rather than ONS Entrefs also shape these worker-count gaps.

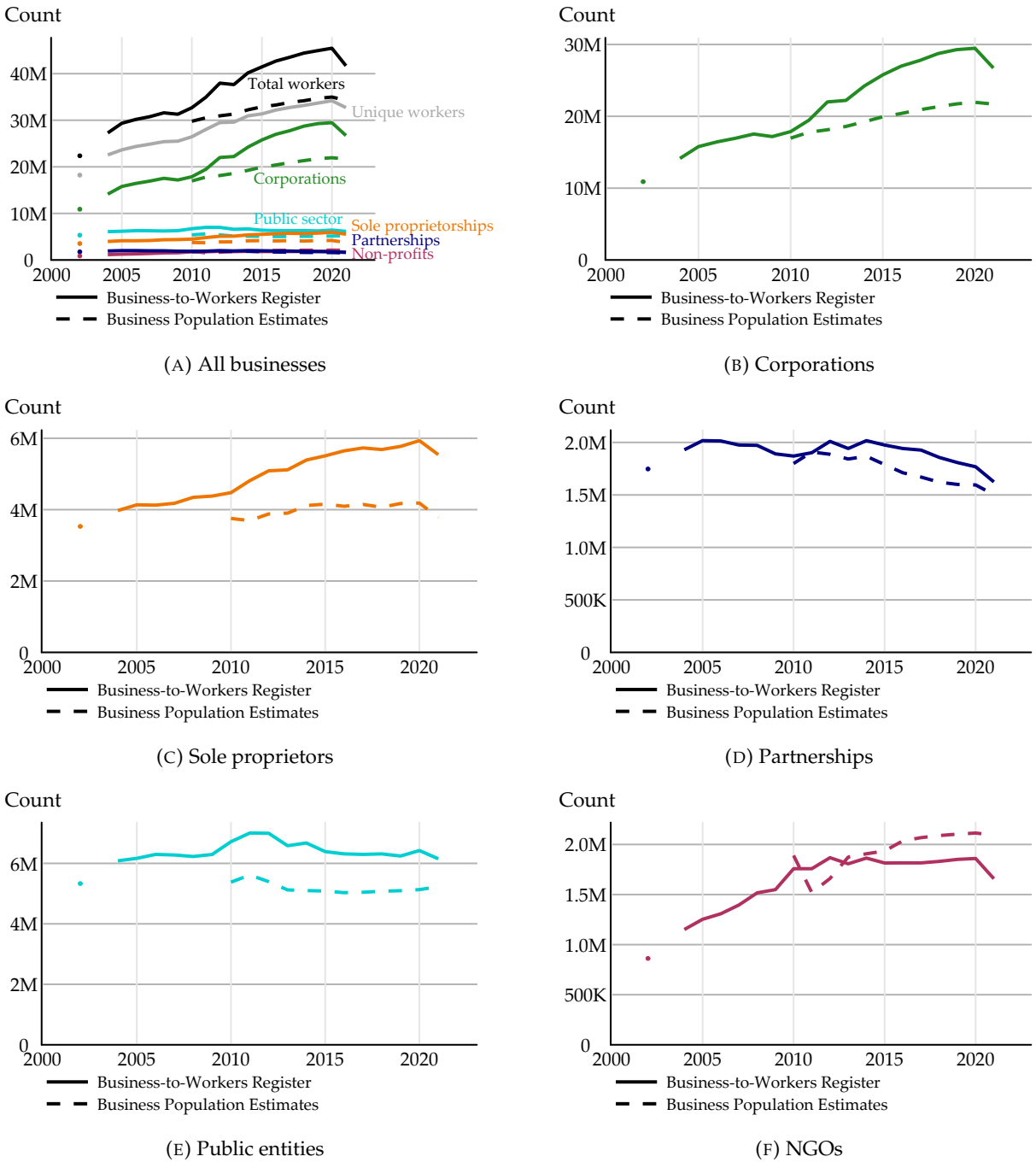


FIGURE 2. Worker counts comparisons: BWR and official estimates

Notes: The six panels show time series of worker counts in the UK over 2002–2022, ordered top to bottom and left to right as all businesses, corporations, sole proprietors, partnerships, public entities, and non-profit organisations. The first panel also reports total and unique-worker counts. Solid lines show BWR counts and dashed lines show the official Business Population Estimates produced annually by the Department for Business and Trade where comparable series are available; the official statistics start in 2010 only. The BWR series break in the early years reflects incomplete data availability for parts of the worker-linkage infrastructure of the BWR.

4. Constructing individual variables

We construct measures of worker incomes and characteristics using several tax data sources that we link using our individuals spine. Most information comes from Self-Assessment returns and PAYE records and spans the universe of individual taxpayers for the period 2002-2022. We observe income at the individual×tax-year level, allowing us to study annualised changes over time. Employment and pensions income is observed monthly from 2015. Tax data also contains several demographic characteristics, most of which are observed at the individual×tax-year level, although we process this information to render it time-invariant where appropriate. We also observe time-varying information about individuals' reported home address.

4.1. **Data sources.** We utilise two main data sources: Pay As You Earn (PAYE) and Self-Assessment (SA). Collectively these cover the full population of individuals with taxable income and/or capital gains, including all workers. Although commonly referred to as the "taxpayer population", this includes individuals with no final tax liability (*i.e.* beyond strictly "taxpayers") provided that they have income or gains that are within the scope of Income Tax and/or Capital Gains Tax. However, it does not include individuals who have no chargeable income or gains in the relevant year, for example those who are exclusively in receipt of non-taxable social security payments.

PAYE data covers all employees and remunerated directors working for UK employers whether or not the individual is UK tax resident,[19] plus all individuals in receipt of a UK registered pension. Information is reported to HMRC directly by employers and pension providers and covers all cash earnings (*e.g.*, basic salary, overtime, bonuses, etc.) paid by employers, plus all taxable pension income paid by UK registered pension schemes.[20] From 2015, the data also includes information about the taxable value of benefits-in-kind provided to employees.

PAYE data comes from two sources, reflecting changes in PAYE administrative systems over time. For 2002-2014 inclusive, we use P14 data extracted from the end-of-year statement provided by employers and pension providers, covering the full PAYE population. This includes annualised information on total taxable income received and Income Tax paid for each employment/pension, plus the PAYE scheme reference.[21] For 2015 onwards, we use RTI data, which similarly covers the full PAYE population but includes a much wider range of variables. This includes annualised

[19]Provided that at least one employee is paid at or above the lower earnings limit for Class 1 NICs, which was £72 per week in 2002, rising to £120 per week in 2022.

[20]Since only taxable pension income is reported, this excludes the tax-free lump sum paid at commencement of the pension, valued at up to 25% of the total pension pot.

[21]COP and NPS data provides some limited additional information for a 10% random sample of this population.

information about individuals' employment characteristics, plus monthly information on taxable income, again with the PAYE scheme reference.

SA data covers all individuals who filed a Self Assessment tax return, subject to some missing coverage for some late filers. This includes non-workers with taxable income and/or gains. Specific filing criteria have changed over time, but broadly are based on three factors: (1) the receipt of any taxable income or gains for which no withholding tax has been applied (*i.e.*, income other than employment income, pensions income, or some savings income); (2) a threshold for total taxable income above which all individuals must file regardless of income source (set at £100,000 for most of the period covered by our data); and (3) certain additional status-based criteria, such as being a partner or company director. Non-residents are also required to file if they have received income from UK property.

SA data comes from two sources. The first is known as "SA Valid View" and its counterpart "SA Invalid View".[22] Together these two datasets cover the full population of SA filers who filed within approximately nine months after the relevant filing deadline. Valid View covers a wide range of key variables from the full SA return, typically including all of the variables that are direct inputs to the final tax calculation for Income Tax and CGT. It also includes derived variables defined by HMRC that aggregate individual variables from the SA return for statistical and analytical purposes, for example for inclusion within the Survey of Personal Incomes (SPI) and associated official statistics.

The second type of data source is known as an "SA custom extract". These datasets correspond to specific pages from the SA return (*e.g.*, SA102 Employment, or SA103 Self-Employment, etc.) and other data tables extracted from CESA, the Self Assessment administrative system. They include all individuals who filed a version of the relevant SA page at the time the extract was taken, including some late filers who are missing from SA Valid View and Invalid View. Using an extract of the SA table corresponding to the final tax calculation, we obtain a dataset that includes all SA filers (including late filers, up to the date of extract) irrespective of which specific SA page(s) they filled out. Variables correspond to individual boxes on the SA return.

Finally, in addition to PAYE and SA data, we use a variety of additional tax data sources that we link at individual-level (based on the tax identifier) using our individuals spine. These include, for example data from the Migrant Workers Scan (MWS), Individual Savings Account (ISA), pension schemes giving relief at source ("COM100"), and Inheritance Tax (IHT). These datasets are not

[22]The difference between these is whether the individual reported a valid postcode in their address field. The version of Valid View that we use contains significantly more variables than Invalid View, although both can be used to construct broad measures of income and provides information on key demographic characteristics.

currently fully integrated within the BWR data infrastructure but we are able to use them to produce additional analyses of individuals that complement the primary measures of characteristics and incomes that we describe in the following sections.

4.2. **Income: terminology and coverage.** The data collected by HMRC focuses on income that is within the scope of Income Tax, meaning the income is chargeable to Income Tax, rather than that Income Tax is due.[23] We refer to this as "fiscal income". Fiscal income is widely used for statistical purposes as well as for tax administration and policy modelling. However, it differs from other measures of income commonly used for economic statistics, such as the definition of household sector income used in the National Accounts. Advani *et al.* (2023) provide a detailed comparison of the coverage of fiscal income versus the National Accounts definitions. These measures of income differ along various dimensions; overall National Accounts definitions of income are broader than fiscal income although the latter is not simply a subset of the former.

In general, fiscal income is reported to HMRC even if no Income Tax is due. For example, employment income that is below an individual's non-taxable allowance (the "Personal Allowance") is still reported by their employer via PAYE. There are some circumstances where fiscal income is not reported to HMRC and hence is not observed in tax data. For example, some fiscal income is only reportable if the individual has been required to file an SA return for other reasons (*e.g.* savings income below savings allowance; "disregarded income" of non-residents). There are also rare instances where fiscal income is not reportable even if individual is filing anyway (*e.g.* trading income below trading allowance), although by construction these exceptions tend to be quantitatively small as they are designed as *de minimis* easements.

Non-fiscal income—*i.e.* income that is not within the scope of Income Tax—is typically excluded from PAYE and SA data, although in several cases it is nevertheless possible to measure them using other linked tax data sources. For example, income from savings held within an ISA is omitted from SA returns because it is not taxable; however, it is nevertheless reported to HMRC by ISA providers. The retained profits of UK companies are not reported on SA returns until they are distributed to shareholders as dividends; however, these profits are observable using CT600 data and could be allocated to individuals based on their directorships (Miller *et al.*, 2024). We have not yet included these non-fiscal income sources within the BWR data infrastructure but this is an agenda for further work.

[23] We use "taxable income" interchangeably with "fiscal income", although the former is sometimes by others to denote income on which Income Tax is due.

Some sources of non-fiscal income cannot be observed directly from any other linked tax data sources. For example: employment income that is transferred directly by the employer into a pension scheme via a "salary sacrifice" or "net pay" arrangement; the tax-free lump sum that is paid to an individual on commencement of their pension; and social security payments that are exempt from Income Tax, which includes most means-tested benefits. The significance of these missing sources of income will depend on the specific purpose of the analysis: for example, the exclusion of means-tested benefits will more significant if seeking to study dynamics at the lower end of the labour market than when focused on impacts at the top.

4.3. **Income: variable construction and decomposition.** We define total income (TI) using the same definition adopted in HMRC's Survey of Personal Incomes (SPI) and associated official statistics.[24] Broadly this equals the total amount of income that is chargeable to Income Tax in the relevant tax year, prior to any deductions, allowances or reliefs. For individuals who only have a PAYE record for the year ("PAYE-only cases"), we treat their reported employment and/or pension income as equal to their total income. For individuals who have a record in SA Valid View (whether or not they also have a PAYE record), we use the variables in SA Valid View to construct total income, which includes their employment and/or pension income. For individuals such as late filers who are in an SA custom extract but not in SA Valid View, we use total income from the tax calculation table.[25]

We are able to break down total income into its component sources and have constructed variables at three levels of granularity. At each level, income types are mutually exclusive and collectively capture all of total income.

At level 1, we decompose total income into two "major" types: earned income (TEI) and investment income (TII). Our definitions of TEI and TII broadly follow the definitions used in the SPI, but we make a choice to diverge from HMRC's classification in some cases. For example, whereas "other income" is classified within TEI by HMRC, we assign it to TII because quantitatively its component parts are predominantly miscellaneous forms of income from capital. Nevertheless, our division into earned and investment income still tracks the legal form of the specific income source: for example, dividends are assigned to TII even though in some cases (*e.g.* personal service companies) they may reflect a return on labour.

[24] In previous work we have verified the correspondence of our total income measure by comparing across in the PAYE and SA microdata who we match with the SPI-internal dataset.
[25] This step is a work in progress.

At level 2, we further decompose TEI and TII into a total of nine "minor" types. TEI comprises employment income; sole proprietor income; partnership income; pensions income; and social security income. TII comprises savings income; property income; dividends; other investment income. Again, these minor types are closely based on the definitions used in the SPI, but we deliberately diverge in a small number of cases. For example, our definition of partnership income only includes trading profits (assigned to TEI), whereas the SPI definition of partnership income additionally includes investment income arising to partnerships, which we instead assign to the underlying investment type and include within TII.

At level 3, we are able to further decompose each minor type of income into the most granular units as reported on the SA return. For example, "interest from gilts" is reported separately on the SA return from interest from bank accounts. In turn, these two micro types (and others) contribute to savings income, which in turn contribute to total investment income (TII). Accordingly, our variable construction hierarchy enables us to produce a fully flexible decomposition of total income flowing all the way from individual boxes on the tax return, via minor and major income types, up to total income, in a way that is mutually exclusive and collectively comprehensive.

Finally, although not strictly fiscal income, we are also able to measure fiscal capital gains using data from the SA108 (Capital Gains) page and link this information to the individual's other income. Since CGT is broadly only charged on the disposal of assets, our measure of capital gains is effectively a (partial) measure of realised gains: we do not observe the accrued gains on assets that have not yet been sold or otherwise disposed of. Fiscal capital gains also exclude realised gains that are outside the scope of Capital Gains Tax, for example transfers between spouses, gains on assets held until death or after emigration, and gains on main homes. As with income, we decompose observed gains into their major types, although this taxonomy has changed over time due to changes in the structure of the SA108 form.

4.4. **Characteristics.** Although not as comprehensive as other data sources such as census data, tax data provide an extensive range of information about individuals' demographic characteristics, including age, sex, migrant status, and (partially) whether they have children. We also observe each individuals' geographic location based on their reported home address, converted to hyperlocal area.

We obtain information on each individual's tax year of birth (from which we construct their age) and sex from PAYE, SA, and other sources. We interpret any changes over time in reported tax

year of birth and sex as most likely attributable to errors in reporting, so construct time-invariant values for each individual based on the modal reported value across all data sources.

Tax data also enables us to identify who is a "migrant", which we define as an individual who arrived in the UK after the age of 18. We construct our migrant indicator from two main sources. First, we develop a novel indicator based on the structure of each individual's National Insurance Number (NINo) combined with their tax year of birth.[26] Second, we supplement this using data from the Migrant Workers Scan (MWS). It is important to use both of these indicators together as the MWS does not identify migrants who arrived in the UK before 2002. For individuals who are present in MWS, we also observe the primary nationality that they report upon arrival to the UK, and their date of arrival.

There is no universal requirement for individuals to report to HMRC that they have children. Consequently, tax data does not contain any widely applicable indicator for whether an individual has children, let alone a link between the parent and child. However, children are relevant for various tax and social security purposes, and we are able to use some of these to obtain a partial indicator for individuals who have children. We have so far developed this indicator using two sources: claims for parental leave by employees in RTI data, and claims for free childcare hours and/or tax-free childcare via the Childcare Account, which HMRC administers. Collectively these sources provide extensive (but not comprehensive) coverage for children born since 2015.

Finally, SA and PAYE data contain information on the geographic location of individuals, based on their reported home address. Since full address information would be directly disclosive, this is converted to Output Area (a hyperlocal area equivalent to around five postcodes) by HMRC staff in the data that we use. Output Area information can be missing either where it was not reported or the address is abroad. We treat geography obtained from SA data as likely to be reasonably up to date, whereas information from PAYE relies on the employee or pensioner having proactively updated their employer or pension provider on any change in address. We are able to convert Output Area into any higher level of ONS geography (from LSOA to NUTS1) and utilise a lookup table to assign higher level geographies for Scotland and Northern Ireland.

5. Constructing firm-level variables

We now turn to business-side characteristics, inputs and outcomes. A comprehensive array of firm-level characteristics like the number of employees, industries, or geographic areas of businesses can be assembled by fetching variables scattered across Datalab datasets, thus building key

[26] See further Advani *et al.* (2025b).

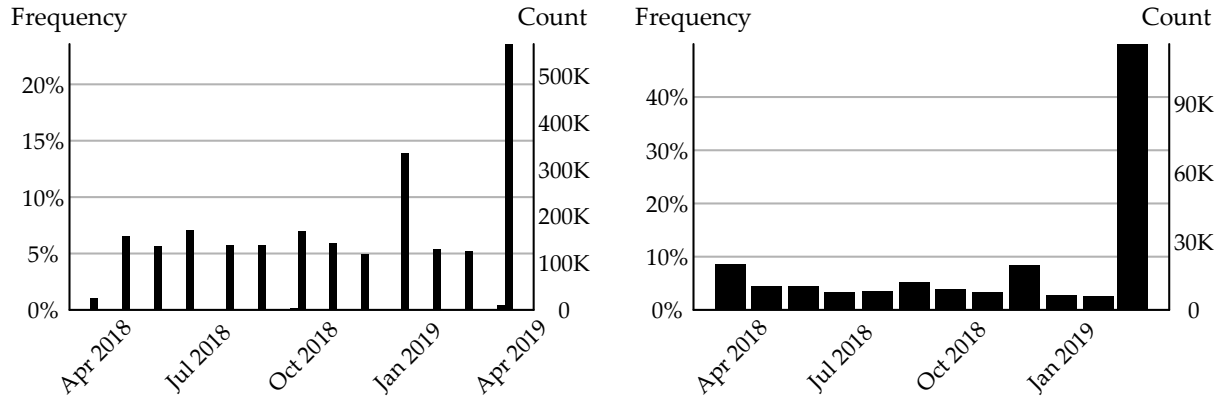
observables that can be used as outcome measures, controls, or even sources of exogenous variation (see for instance the R&D tax credit discontinuity used in Dechezleprêtre *et al.* (2023)). We complement Datalab tax datasets with external data on patents. Variables about economic performance such as profits, turnover, capital assets and investment can also be created from Datalab tax datasets but they require some additional processing to convert tax-accounting values into quantities that economists or financial analysts would consider economically meaningful. We describe this processing in this section. We then turn to benchmarking our estimates. Finally, we detail how we construct other business characteristics in the third part of this section.

5.1. **Building financial accounting variables from tax returns.** Profits reported by firms in their tax returns may differ from those reported in their accounting statements, sometimes significantly (Bilicka, 2019), posing challenges for researchers interested in reliable measures of business performance. Tax returns and accounting statements serve different purposes: financial accounts aim to measure economic performance consistently across firms and over time, while tax accounts implement tax law and policy objectives (*e.g.* investment incentives), which mechanically affects taxable income.

Three features of tax return data is first-order for our tax-to-financial accounts conversion exercise. First, reporting periods are heterogeneous across firms, requiring us to translate firm-specific accounting periods into a consistent tax-year panel. Second, some unincorporated businesses can report on a cash basis for tax purposes, while financial accounts are typically accrual-based, generating timing differences between book and tax measures. Third, the tax code departs from financial accounting in its treatment of investment and depreciation: tax returns generally report capital allowances and balancing charges rather than investment and proceeds from the sales of assets. To construct harmonised, economically interpretable measures, we reverse-engineer π_{ft} (operating profits), I_{ft} (investment), K_{ft} (capital stock) and δK_{ft} (depreciation) from tax returns in three steps. We summarise the methodology here. Appendix A provides more details.

Step 1: Harmonising reporting periods. Business tax returns often use different accounting periods, even within a given legal form. These reporting periods may not coincide with the UK tax year. To build our yearly panel of businesses and workers, we need to make sure that business- and worker-level variables are defined over the same period. We therefore express each business-year

at the tax-year frequency (April of $t-1$ to March of t).[27] Figure 3 shows the distribution of end-of-period dates for corporations (left panel) and partnerships (right panel) in tax year 2020. About one quarter of corporations file in line with the tax year; most do not. The misalignment is smaller for partnerships, but a majority still close their accounts on a different date.[28]



(A) Corporations (B) Partnerships

FIGURE 3. Distributions of end-of-period dates in tax returns (tax year 2020)

Notes: The unit of analysis is a corporation or a partnership observed in tax year 2020 (April 2019 to March 2020; the last full fiscal year before the first Covid lockdown). We keep only surviving businesses that report a full fiscal year of activity in their tax returns. The histogram shows the counts of businesses for each end-of-period reporting date. Our data has exact days of reporting for corporations but just months for partnerships, so the partnership histogram is at the monthly level while the corporation one is at a daily frequency. Sole proprietorships file business income through Self Assessment and are therefore not included in these two panels.

To map returns onto tax years, we apportion each reported variable by the days or months of overlap between the business's accounting period and the UK tax year. A corporation starting its books on 1 January of year t and closing them on 31 December of the same year—a common choice—has three months of its accounting period (January to March of t) in tax year t and nine months (April to December of t) in tax year $t+1$. In such a case, we would split its variables 3/12–9/12. This harmonisation is needed to compare firms to each other, link workers to businesses, and to match business-level values to external financial accounts data.

Step 2: Inferring investment and capital from capital allowances. Tax returns do not report investment (I_{ft}) and capital stock (K_{ft}) directly.[29] Instead, they report capital allowances and balancing

[27] This minimises data transformation. We could have chosen calendar years instead, but this would have forced us to apportion business-level *and* worker-level variables at the calendar year frequency.

[28] Sole proprietorships report business income through the individual's Self Assessment, so those accounts are already aligned with the tax year by construction.

[29] Note that the SA800 returns do have some balance sheet data for a subset of partnerships, the medium-sized ones with annual turnover between £85,000 and £15 million. We are using this data for benchmarking.

charges, which reflect tax depreciation schedules and tax adjustments on disposal of capital assets. We use these components to infer investment flows and to construct a capital stock time series via a perpetual inventory approach, using asset-type-specific depreciation assumptions where the tax forms permit a breakdown by asset category. The key methodological insight here is to estimate the capital stock that a firm is allowed to tax-depreciate from year $t-1$ to t on its year-t tax return to extract the value of new investment in year t. The intuition behind our methodology is the following. If a firm buys a piece of equipment for £100,000 in 2019, it is allowed to report a capital allowance of £18,000 (applying an 18% tax depreciation rate) in 2019, and £14,760 in 2020 (18% of (£100,000 - £18,000)). If we observe a capital allowance of £24,760 in 2020, we infer that a new investment has been made by the firm, and that the firm was allowed to tax-depreciate £10,000 (£24,760 - £14,760) of this new investment. The value of the new investment in 2020 is therefore $I_{f,2020} = \frac{10,000}{18\%} = 55,555.55$. The resulting I_{ft} and K_{ft} measures are designed to be interpretable as economic investment and productive capital inputs rather than tax concepts.

For firms that start investing within the time period of our panel (2001-2023 for firms, 1997-2022 for partnerships and sole proprietors), we initialise the capital pool $K_{f,0}$ at $\frac{\text{Capital allowance}_{f,0}}{\delta_t^{TAX}}$, where δ_t^{TAX} is the tax code depreciation rate that applies to this type of capital. We then capitalise the capital stock by the perpetual inventory method. We use the estimated values of I_{ft}, the initial value of K_{f0} and *financial accounting* depreciation rates (the rates at which capital really depreciates, in an economic sense) taken from the UK Generally Accepted Accounting Principles. For firms already present in the panel in the first available year, we initialise capital in the first observed year to its long-run value, taking the average value of investment in subsequent years as the steady-state investment average. Appendix A provides a detailed description of our methodology.

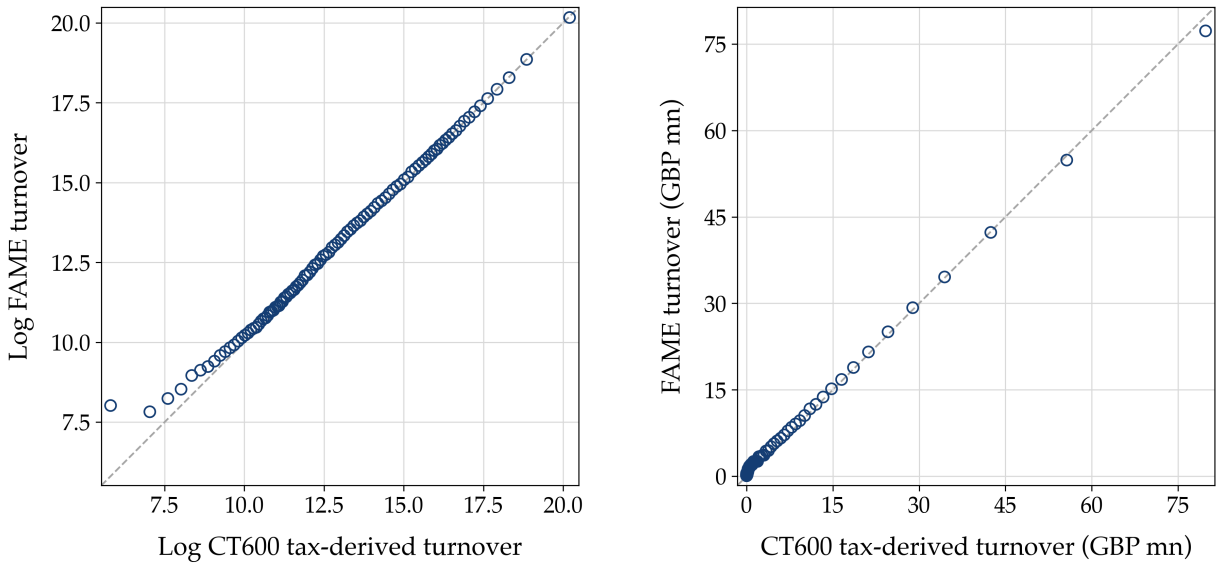
Step 3: Calculating operating profits (EBIT/EBITDA-style measures). We then adjust taxable profits by removing tax-specific capital allowance and balancing charges. We then add our own estimate of depreciation, yielding measures close in spirit to EBIT and EBITD. In practice, this involves (i) purging the profit measure from capital allowances and balancing charges and (ii) applying the depreciation implied by our reconstructed capital stock. We can then make further adjustments to get to the profit measures we are interested in, by adding back interest deductions (EBIT) and our estimations of depreciation (EBITD). We have not yet constructed measured of amortisation

(depreciation of intangible assets) so we have no estimates of EBITDA.[30] We assess the quality of these constructed outcomes by comparing them to external accounts data (FAME/ORBIS) for the subset of corporations where such validation is feasible. SA800 partnerships also offer validation opportunities because medium-sized partnerships report balance sheet and depreciation accounts.

5.2. **Benchmarking with FAME and SA800 data.** To assess how closely our reconstructed measures track their financial-accounting counterparts, we compare them to balance-sheet information for the subset of firms that can be matched across sources. For corporations, the reference data comes from FAME/ORBIS, which is built from the statutory accounts that UK companies file at Companies House; for partnerships, it comes from the balance-sheet fields observed in SA800 returns. This is best interpreted as a *quality-assessment exercise* rather than as a population benchmark: the overlap samples are selective, whereas the BWR is designed to cover the full economy. The benchmark datasets themselves are also incomplete in ways that vary systematically with firm size and reporting period: the Companies House filing requirements that FAME inherits differ by size category and have been revised several times over our window. Appendix B summarises the size thresholds and the set of financial items that each size category is required to file in each regime, which is useful context for the gaps between the tax-derived and FAME series reported below. Similarly, not all partnerships need to report balance-sheet information in SA800.

Operationally, we sort firms into 100 equally sized cells and compare the tax-derived variable on the horizontal axis to the reference variable on the vertical axis, so the 45-degree line provides a direct visual benchmark for alignment. Figures 4–10 show the 2012 comparisons, with log and level versions displayed side by side for each outcome. Logs have the advantage of facilitating the comparison of unevenly distributed quantities across firms but remove zero (and negative, for profits) observations in our data or in the benchmark datasets. Removing zeros can be problematic for variables that typically have a mass point there, which is the case for investment, capital, profits and to some extent, turnover. We report turnover, capital, depreciation, EBIT, and investment for corporations, and capital and depreciation for partnerships. We chose 2012 as the middle year of our panel, giving enough time for our capitalised estimates of the capital stock to reach their long-term values. Three features emerge from these graphs. First, across outcomes, the relationship is generally close to the 45-degree line, with especially tight alignment in logs. Second, the most striking divergence between our estimates and the benchmark datasets is related to capital. We often underestimate capital stock compared to FAME (for corporations) and SA800 (for

[30] We could use our data on R&D investments and patents to estimate the R&D component of intangible capital. An area for future improvement.



(A) Log scale

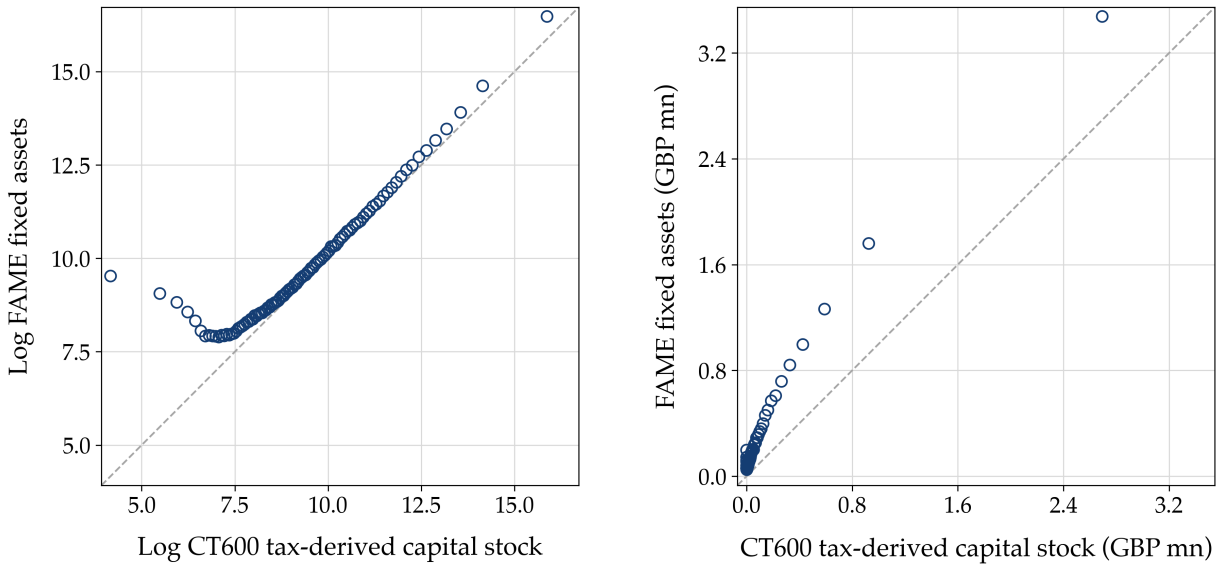
(B) Level scale

FIGURE 4. Corporation turnover: tax-derived and FAME measures (2012)

Notes: Binscatter comparing the BWR tax-derived turnover measure (CT600, horizontal axis) to the FAME benchmark turnover measure (vertical axis) for CT600-linked corporations in 2012. Firms are sorted into 100 equally sized bins, and each dot plots the average of the horizontal- and vertical-axis variables across the firms in that bin. The 45-degree line provides a direct visual benchmark for alignment. Panel (a) uses log scales (dropping zeros and non-positive values); panel (b) uses levels.

partnerships). Relatedly, we overestimate depreciation for partnerships and for corporations with large values. This could indicate that our (financial accounting) depreciation rates are too high.[31] These depreciation rates were chosen by us based on our understanding of Generally Accepted Accounting Practices (GAAP) in the UK. Refining these rates and our measure of capital is an active area of future research at CenTax. Third, there is often a divergence between our estimates and the benchmark for very small values of our estimates. These divergences suggest that the benchmark overestimates (or we underestimate) the value of interest. See for instance the log graphs of capital, depreciation, EBIT, investment, and, to some extent, turnover. A likely explanation is that threshold effects introduce some selection bias in the type of firms that make it into the benchmarks (FAME or SA800). In Companies House, not all firms have to report profits, turnover and balance-sheet variables. In particular, small firms are exempt from these reporting requirements for half of our panel period. As a result, the benchmark-BWR pairs of firms that appear in our comparison charts for small values of estimated capital, investment or turnover link large firms in the benchmark (above the reporting threshold) to small firms in the BWR. Taking the benchmark

[31] See Table A.1.



(A) Log scale

(B) Level scale

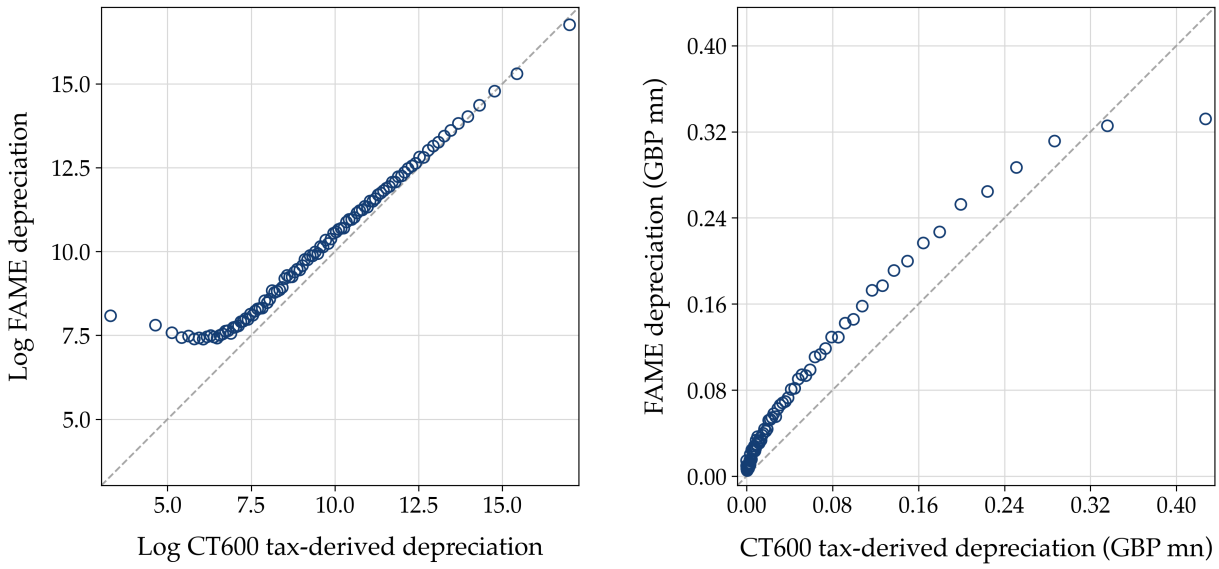
FIGURE 5. Corporation capital: tax-derived capital stock and FAME fixed assets (2012)

Notes: Binscatter comparing the BWR tax-derived capital stock (capitalised from CT600 investment flows, horizontal axis) to the FAME fixed-assets benchmark (vertical axis) for CT600-linked corporations in 2012. Firms are sorted into 100 equally sized bins, and each dot plots the average of the horizontal- and vertical-axis variables across the firms in that bin. The 45-degree line provides a direct visual benchmark for alignment. Panel (a) uses log scales (dropping zeros and non-positive values); panel (b) uses levels.

values as the ground truth, and assuming that the measurement error from our estimates is symmetrically distributed around the true value, the reporting threshold filters out below-threshold values associated with even lower estimates of ours, thus leaving only the above-threshold-low-estimate pairs to appear in the graph.[32] Appendix B provides a detailed timeline of reporting requirements in Companies House, for context.

5.3. **Other firm characteristics.** Beyond the accounting-style measures reconstructed above, the organisation side of the BWR (the Business Register) attaches a set of firm-level characteristics that support applications in industry dynamics, local labour markets, firm entry and exit, and

[32]To make this situation concrete, consider the following example. A corporation that starts trading in 2000, before our panel starts, with an initial investment in fixed assets of £1 million. It then only invests £10,000 a year from 2001 onward. Because we only observe the stream of £10,000 yearly investments in the BWR, we underestimate the total capital stock, and this pair of observations contributes to the uptick in capital values at the bottom of the 45-degree-line graph; *i.e.*, we incorrectly underestimate the total capital stock of this firm. Conversely, consider another firm that invests £10,000 a year from 2001 onward, with no initial capital stock before 2001. This is a small firm that falls below the Companies House reporting threshold. We estimate the capital stock in our data but this pair of observations is dropped because of missing data in FAME. Therefore, it does not contribute to bringing the scatterplot closer to the 45-degree line; nor do any of the observations where our estimates are below the true value and the Companies House value is below the threshold. A related but similar argument applies to estimates of turnover where the measurement error on our end comes from the apportioning of values to fiscal years.



(A) Log scale

(B) Level scale

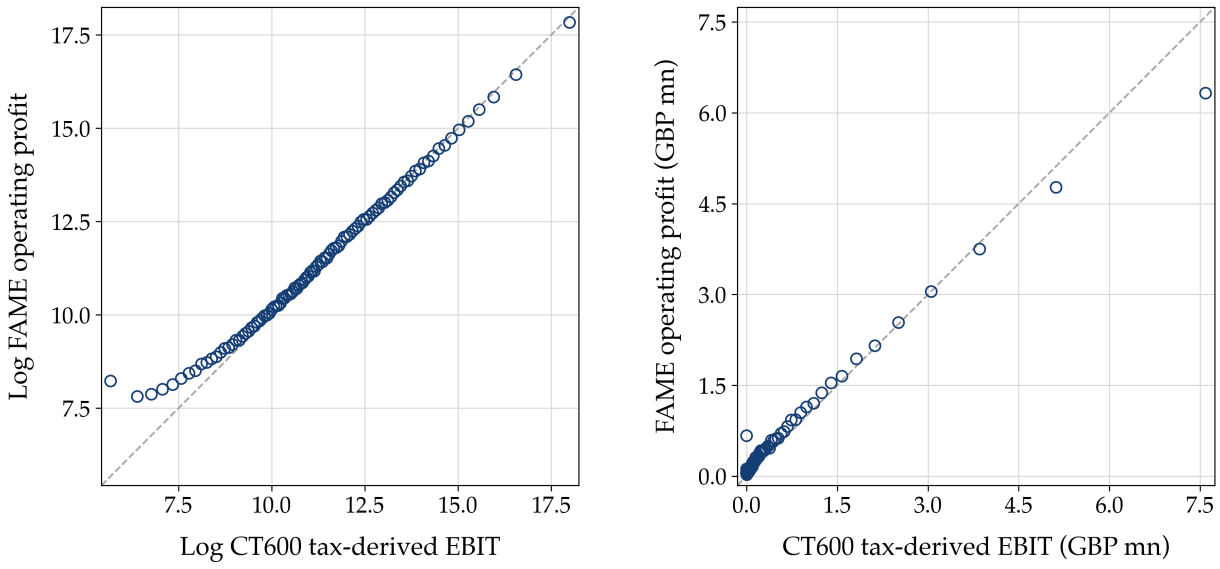
FIGURE 6. Corporation depreciation: tax-derived and FAME measures (2012)

Notes: Binscatter comparing the BWR tax-derived depreciation measure (horizontal axis) to the FAME depreciation benchmark (vertical axis) for CT600-linked corporations in 2012. Firms are sorted into 100 equally sized bins, and each dot plots the average of the horizontal- and vertical-axis variables across the firms in that bin. The 45-degree line provides a direct visual benchmark for alignment. Panel (a) uses log scales (dropping zeros and non-positive values); panel (b) uses levels.

innovation. Table 5 lists the input datasets used to construct them; the paragraphs below describe each variable family, and its coverage by legal form.

Industry. Every business-year is assigned one or more SIC2007 codes. For corporations, the Trade Classification Number (TCN) reported on the CT600 is mapped to SIC2007 using an ONS-provided weighted many-to-many crosswalk, with the maximum-weight match taken as the primary code and lower-weight matches retained in wide form. For PAYE schemes and partnerships, the industry is the modal SIC2007 across the scheme's workers, taken from the self-reported industry field on SA102 and PAYE data. To avoid relying on the statistical linkage between partnership IDs and schemerefs we created, we only use industry information from *partners*, not employees, to determine the industry of partnerships. Government, non-profit, and public-administration employers are given the SIC codes of their modal employee-year observations.

Geography based on workers' Output Areas. Each business-year is located by aggregating its workers' Output Areas (OAs) up to the firm level. Worker-level OAs come from PAYE Geography files (2002–2023) for employees and Self Assessment location extracts (1997–2023) for partners and sole



(A) Log scale

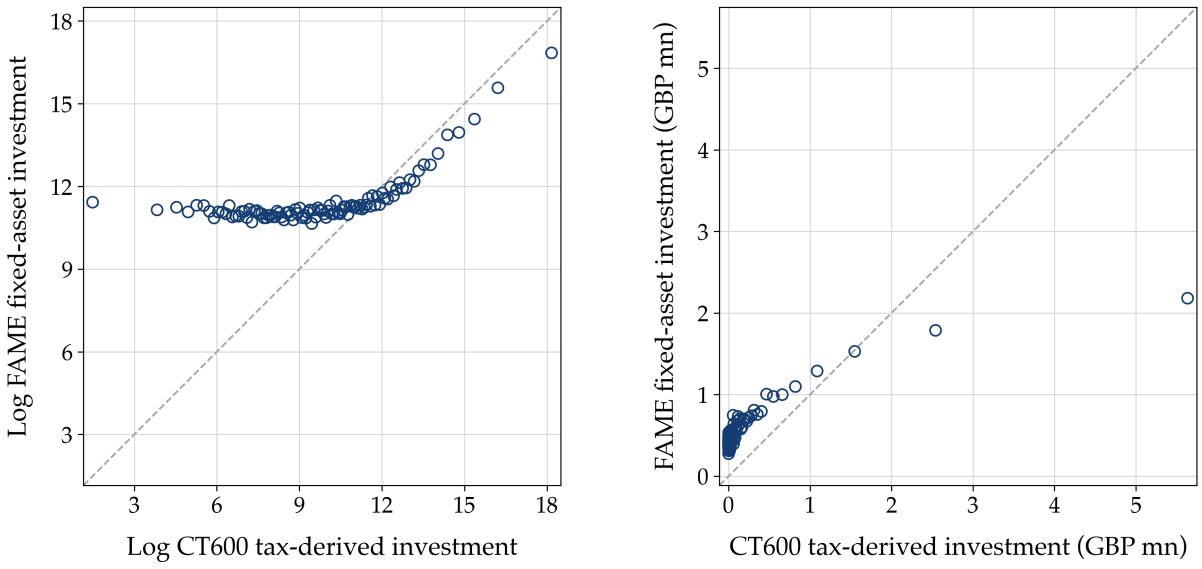
(B) Level scale

FIGURE 7. Corporation EBIT: tax-derived and FAME operating-profit measures (2012)

Notes: Binscatter comparing the BWR tax-derived earnings before interest and tax (EBIT) measure (horizontal axis) to the FAME operating-profit benchmark (vertical axis) for CT600-linked corporations in 2012. Firms are sorted into 100 equally sized bins, and each dot plots the average of the horizontal- and vertical-axis variables across the firms in that bin. The 45-degree line provides a direct visual benchmark for alignment. Panel (a) uses log scales (dropping zeros and non-positive values, hence excluding loss-making firms); panel (b) uses levels.

proprietors. The 2011 reclassification of OA codes is harmonised via correspondence tables provided by the ONS. At the firm level, the BWR retains the top-5 most-common OAs by worker count, a time-varying Travel-to-Work Area, and a time-invariant postcode. For partnerships, the same geographical information is retained, but we only use geography data from partner files for the same reason as for industry information.

Births and deaths. The BWR does not collapse firm entry and exit into a single birth/death definition. Instead, each business-year carries a panel of binary activity indicators: non-zero wage bill in PAYE, non-zero turnover or input value on a filed VAT return, IDBR-recorded VAT birth/death window, non-zero CT600 turnover or pre-tax earnings, FAME "Active or Live" status, and analogous partnership and self-employment flags. Users can construct alternative data-driven birth/death definitions as their application requires. For example, birth as the first year any flag fires and death as the last, or a definition restricted to the subset of flags relevant to a given legal form. For instance, in our organisation counts presented in section 2, we consider a business active if it reports a non-zero wage bill in PAYE, or a non-zero turnover.



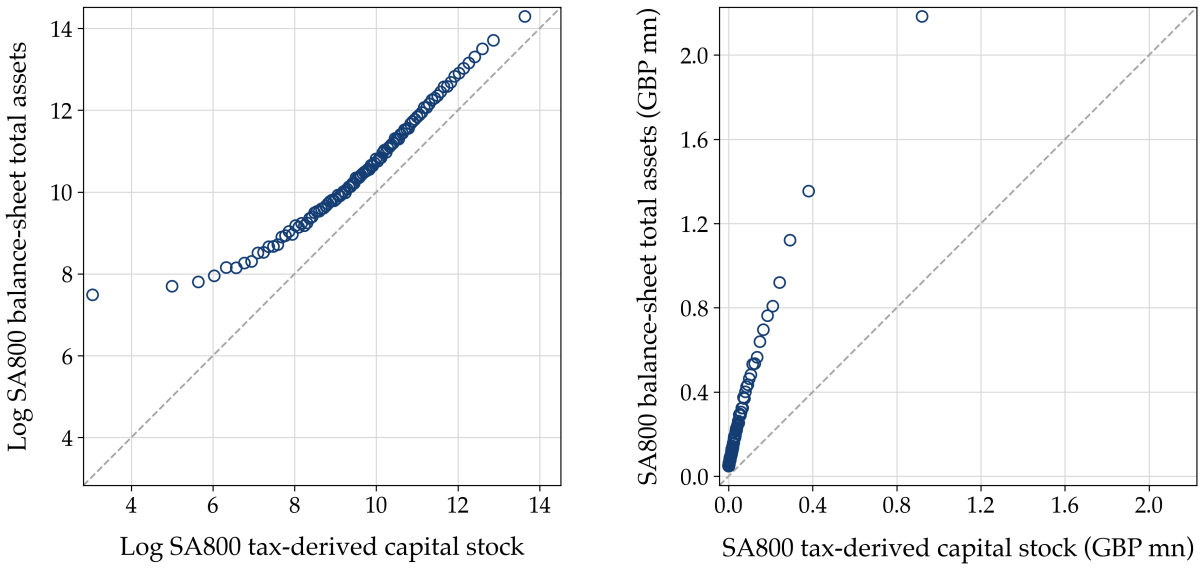
(A) Log scale (B) Level scale

FIGURE 8. Corporation investment: tax-derived and FAME fixed-asset investment measures (2012)

Notes: Binscatter comparing the BWR tax-derived fixed-asset investment measure (horizontal axis) to the FAME fixed-asset additions benchmark (vertical axis) for CT600-linked corporations in 2012. Firms are sorted into 100 equally sized bins, and each dot plots the average of the horizontal- and vertical-axis variables across the firms in that bin. The 45-degree line provides a direct visual benchmark for alignment. Panel (a) uses log scales (dropping zeros and non-positive values); panel (b) uses levels.

Patents. Patent records from the European Patent Office (EPO), the UK Intellectual Property Office (IPO) and patents identified as being filed by British entities are linked to corporations by Company Registration Number, via a CRN-to-CT600 ID crosswalk. We obtained the patent data from Orbis IP, a proprietary database. Patent-year assignment uses the earliest of the application filing date across patents of a same patent family (patents protecting the same innovation but filed in different jurisdictions). We build three variables at the firm×year level: raw patent count, sum of claims of patents,[33] and the sum of forward citations at various time horizons. Some patents appear in the EPO, IPO and British companies patent extract and we deduplicate patents using information on patent families. Coverage is 1990–2022 for all three sources. Patents can only be linked to corporations because we do not have lookup tables between CRNs and partnerships IDs or sole proprietors' UTRs.

[33] A patent claim is a short description that defines the precise boundaries of the invention for which exclusive IP rights are granted. Patent may contain one or multiple claims. The number of claims on a patent is often used to measure the technological scope of an invention.



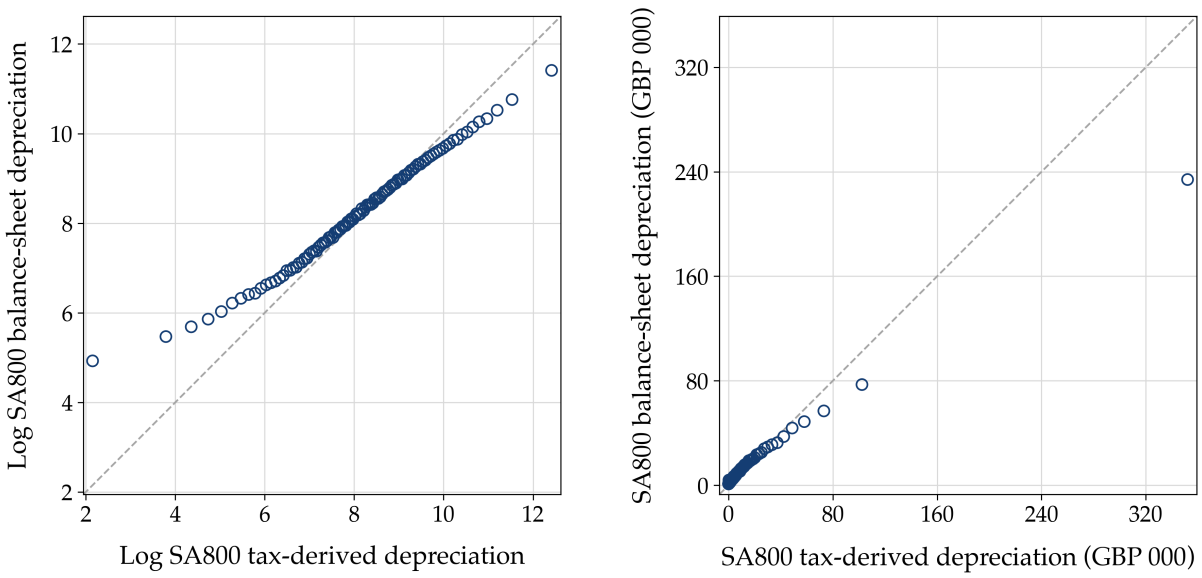
(A) Log scale (B) Level scale

FIGURE 9. Partnership capital: tax-derived capital stock and balance-sheet total assets (2012)

Notes: Binscatter comparing the BWR tax-derived capital stock for partnerships (capitalised from SA800 investment flows, horizontal axis) to the SA800 balance-sheet total-assets benchmark (vertical axis) in 2012. Firms are sorted into 100 equally sized bins, and each dot plots the average of the horizontal- and vertical-axis variables across the firms in that bin. The 45-degree line provides a direct visual benchmark for alignment. Panel (a) uses log scales (dropping zeros and non-positive values); panel (b) uses levels.

R&D expenditure. R&D variables are built from HMRC's administrative R&D tax-credit records, which are subsets of the CT600 filings, and cover the SME R&D relief scheme (2000–2021) and the large-company RDEC scheme (2003–2021). The final business register carries values of R&D tax credits, qualifying expenditure, and common-cost deductions. These variables are not available for partnerships or sole proprietors, whose R&D activity does not enter the tax-credit administrative files.

Intermediate inputs from VAT. We measure intermediate inputs volumes from VAT returns. In particular, we use box 7 ("Total value of purchases and all other inputs, excluding VAT"). VAT reporting is at the VRN level (VAT registration number), which may group up to several hundred subsidiaries. In the case of corporations, for which we have at CT600 Id-to-VAT ID mapping from the IDBR, the group total is apportioned equally across the CT600 IDs linked to the VRN. Coverage is 2005–2023. Box 4 (input VAT recovered) is carried alongside Box 7 for reference.



(A) Log scale (B) Level scale

FIGURE 10. Partnership depreciation: tax-derived and balance-sheet measures (2012)

Notes: Binscatter comparing the BWR tax-derived depreciation measure for partnerships (horizontal axis) to the SA800 balance-sheet depreciation benchmark (vertical axis) in 2012. Firms are sorted into 100 equally sized bins, and each dot plots the average of the horizontal- and vertical-axis variables across the firms in that bin. The 45-degree line provides a direct visual benchmark for alignment. Panel (a) uses log scales (dropping zeros and non-positive values); panel (b) uses levels.

6. CONCLUSION

This paper introduces the Business-to-Worker Register (BWR), a population-scale matched panel of workers and businesses for the UK covering 2002–2022. The BWR links the universe of UK workers (employees, directors, partners, and sole proprietors) to the universe of UK businesses and non-businesses, including non-profits, public-sector entities, and household employers. It reconstructs firm-level accounting measures (turnover, operating profits, investment, capital stock, and depreciation) from tax returns where no balance-sheet data exist, and validates them against external financial accounts. The resulting dataset covers roughly 42 million workers and 7 million businesses per year, including the large population of non-employing sole proprietorships and small partnerships that most existing linked employer-employee databases miss.

The BWR is a first step in a broader effort to map the full network of economic relationships between individuals and businesses in the UK. Two links are still missing from this picture: ownership ties connecting individuals to the businesses they hold equity in, and household ties connecting individuals who share resources and make joint economic decisions. Adding shareholder-to-business links would allow researchers to trace how profits, investment, and shocks flow from

Datasets	IDs	Coverage
NSPL National Statistics Postcode Lookup; postcodes mapped to output areas and travel-to-work areas	postcode	rolling
PAYE Geography files annual worker-level output-area assignments	NINo	2002–2023
Self Assessment Location extracts travel-to-work-area assignments for SA filers, partnerships, and employers	UTR	1997–2023
ONS TCN–SIC2007 weighted crosswalk historical Trade Classification Number sectors mapped to SIC2007		static
HMRC R&D tax-credit administrative files SME R&D relief and RDEC large-company schemes	taxpayer_anon	2000–2021
EPO patent records European Patent Office filings, linked via consolidated CRN	CRN	1990–2022
IPO patent records UK Intellectual Property Office filings, linked via consolidated CRN	CRN	1990–2022
VAT returns Box 7 intermediate inputs; VRN-to-taxpayer apportionment	VRN	2005–2023

TABLE 5. Input datasets for non-financial firm characteristics

Notes: The BWR is constructed inside HMRC's Datalab. Coverage years reflect the data actually used by the BWR construction scripts; not all variables are populated in every year.

businesses to their beneficial owners, and back. Adding intra-household links, identified from co-location and shared savings accounts, would make it possible to study how labour supply, income risk, and wealth accumulate within households rather than in isolation. Building these extensions is an active area of research at CenTax, and the necessary administrative datasets are available within the HMRC Datalab. Taken together, these three layers (worker-to-business, owner-to-business, and intra-household) would provide a unified empirical infrastructure for studying the distribution and growth of economic activity in the UK.

REFERENCES

ADVANI, A., D. BURGHERR, AND A. SUMMERS (2025a): "Taxation and Migration by the Super-Rich," CESifo Working Paper No. 11870. [*Cited on page 9.*]

ADVANI, A., F. KOENIG, L. PESSINA, AND A. SUMMERS (2025b): "Immigration and the Top 1 Percent," *The Review of Economics and Statistics*, 107, 1123–1135. [*Cited on page 30.*]

ADVANI, A. AND A. SUMMERS (2024): "Measuring and taxing top incomes and wealth," *Oxford Open Economics*, 3, i1113–i1129. [*Cited on page 3.*]

ADVANI, A., A. SUMMERS, AND H. TARRANT (2023): "Measuring top income shares in the UK," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186, 241–258. [*Cited on page 27.*]

BILICKA, K. A. (2019): "Comparing UK tax returns of foreign multinationals to matched domestic firms," *American Economic Review*, 109, 2921–2953. [*Cited on page 31.*]

BOBBIO, E. AND H. BUNZEL (2018): "The Danish Matched Employer-Employee Data," Economics Working Paper 2018-03, Department of Economics and Business Economics, Aarhus University. [*Cited on page 6.*]

BOZIO, A., T. BREDA, AND M. GUILLOT (2023): "Using Payroll Taxes as a Redistribution Tool," *Journal of Public Economics*, 226, 104986. [*Cited on page 6.*]

BÖHEIM, R. AND D. PICHLER (2025): "Earnings Volatility in Austria," Economics Working Paper 2025-04, Department of Economics, Johannes Kepler University Linz. [*Cited on page 6.*]

CARNEIRO, A., P. PORTUGAL, P. RAPOSO, AND P. M. M. RODRIGUES (2023): "The Persistence of Wages," *Journal of Econometrics*, 233, 596–611. [*Cited on page 6.*]

CASARICO, A. AND S. LATTANZIO (2024): "What Firms Do: Gender Inequality in Linked Employer-Employee Data," *Journal of Labor Economics*, 42, 325–355. [*Cited on page 6.*]

COOPER, M., J. MCCLELLAND, J. PEARCE, R. PRISINZANO, J. SULLIVAN, D. YAGAN, O. ZIDAR, AND E. ZWICK (2016): "Business in the United States: Who Owns It, and How Much Tax Do They Pay?" *Tax Policy and the Economy*, 30, 91–128. [*Cited on page 2.*]

DECHEZLEPRÊTRE, A., E. EINIÖ, R. MARTIN, K.-T. NGUYEN, AND J. VAN REENEN (2023): "Do Tax Incentives Increase Firm Innovation? An RD Design for R&D, Patents, and Spillovers," *American Economic Journal: Economic Policy*, 15, 486–521. [*Cited on page 31.*]

DELESTRE, I., B. JACOBS-STROM, H. MILLER, AND K. SMITH (2025): "New trends in self-employment and top incomes," Ifs report, Institute for Fiscal Studies, ISBN 978-1-80103-233-9. Available at https://ifs.org.uk/publications/new-trends-self-employment-and-top-incomes. [*Cited on page 9.*]

DEPARTMENT FOR BUSINESS AND TRADE (2025): "Business Population Estimates for the UK and Regions 2025: Methodology Note," https://www.gov.uk/government/statistics/business-population-estimates-2025/business-population-estimates-for-the-uk-and-regions-2025-methodology-note, published 2 October 2025. [*Cited on pages 14 and 22.*]

DEPARTMENT FOR BUSINESS, ENERGY & INDUSTRIAL STRATEGY (2022): "Business Population Estimates 2022," https://www.gov.uk/government/statistics/business-population-estimates-2022, accessed 2026. [*Cited on page 3.*]

ENGBOM, N., C. MOSER, AND J. SAUERMANN (2023): "Firm Pay Dynamics," *Journal of Econometrics*, 233, 396–423. [*Cited on pages 4 and 6.*]

FORTH, J., A. BRYSON, AND C. PALMOU (2025): "A Roadmap for Developing a New LEED Infrastructure for the UK," ESCoE Technical Report TR-28, Economic Statistics Centre of Excellence. [*Cited on page 5.*]

FRIEDRICH, B., L. LAUN, C. MEGHIR, AND L. PISTAFERRI (2025): "Earnings Dynamics and Firm-Level Shocks," Manuscript, November 24, 2025. [*Cited on page 4.*]

GOETZ, C., H. HYATT, Z. KROFF, K. SANDUSKY, AND M. STINSON (2025): "Business Owners and the Self-Employed: 33 Million (and Counting!)," CES Working Paper 25-60, Center for Economic Studies, U.S. Census Bureau. [*Cited on page 6.*]

GOOS, M., A. MANNING, A. SALOMONS, B. SCHEER, AND W. VAN DEN BERGE (2022): "Alternative Work Arrangements and Worker Outcomes: Evidence from Payrolling," CPB Discussion Paper 435, CPB Netherlands Bureau for Economic Policy Analysis. [*Cited on page 6.*]

GRAHAM, M., E. MCENTARFER, K. MCKINNEY, S. TIBBETS, AND L. TUCKER (2022): "LEHD Snapshot Documentation, Release S2021_R2022Q4," CES Working Paper 22-51, Center for Economic Studies, U.S. Census Bureau. [*Cited on pages 5 and 6.*]

GREEN, A. S., M. J. KUTZBACH, AND L. VILHUBER (2017): "Two Perspectives on Commuting: A Comparison of Home to Work Flows Across Job-Linked Survey and Administrative Files," Census Working Paper CES 17-34, U.S. Census Bureau, Center for Economic Studies. [*Cited on pages 5 and 6.*]

HARRIGAN, J., A. RESHEF, AND F. TOUBAL (2021): "The March of the Techies: Job Polarization Within and Between Firms," *Research Policy*, 50, 104008. [*Cited on page 6.*]

KOPCZUK, W. AND E. ZWICK (2020): "Business Incomes at the Top," *Journal of Economic Perspectives*, 34, 27–51. [*Cited on page 2.*]

MILLER, H., T. POPE, AND K. SMITH (2024): "Intertemporal Income Shifting and the Taxation of Business Owner-Managers," *The Review of Economics and Statistics*, 106, 184–201. [*Cited on pages 3, 20, and 27.*]

MORCHIO, I. AND C. MOSER (2026): "The Gender Pay Gap: Micro Sources and Macro Consequences," *American Economic Review*, 116. [*Cited on page 6.*]

PANAHIAN FARD, D., A. SCHMUCKER, S. SETH, M. UMKEHRER, AND F. ZIMMERMANN (2024): "Linked-Employer-Employee-Data of the IAB: LIAB Longitudinal Model (LIAB LM) 1975–2021," FDZ-Datenreport 04/2024, Institut für Arbeitsmarkt- und Berufsforschung (IAB). [*Cited on page 6.*]

SMITH, M., D. YAGAN, O. ZIDAR, AND E. ZWICK (2019): "Capitalists in the twenty-first century," *The Quarterly Journal of Economics*, 134, 1675–1745. [*Cited on page 2.*]

SONG, J., D. J. PRICE, F. GUVENEN, N. BLOOM, AND T. VON WACHTER (2019): "Firming Up Inequality," *The Quarterly Journal of Economics*, 134, 1–50. [*Cited on page 6.*]

STATISTICS CANADA (2018): "The Measurement of Business Ownership by Gender in the Canadian Employer–Employee Dynamics Database," Analytical Study 11-633-X2018017, Statistics Canada. [*Cited on page 6.*]

STATISTICS FINLAND (2024): "FOLK – Finnish Longitudinal Employer-Employee Data," Research microdata, Statistics Finland. [*Cited on page 6.*]

STATISTISKA CENTRALBYRÅN (2016): "Registerbaserad arbetsmarknadsstatistik (RAMS) 2015," SCBDOK, Produktkod AM0207. [*Cited on page 4.*]

WHITTARD, D., F. RITCHIE, V. PHAN, J. FORTH, A. BRYSON, L. STOKES, C. SINGLETON, AND A. MCKENZIE (2022): "Exploring the Workplace Location Problem in the ASHE," Methodology paper, Wage and Employment Dynamics (WED) Project. [*Cited on page 5.*]

ZWEIMÜLLER, J., R. WINTER-EBMER, R. LALIVE, A. KUHN, J.-P. WUELLRICH, O. RUFF, AND S. BÜCHI (2009): "Austrian Social Security Database," Working Paper 0903, NRN – The Austrian Center for Labor Economics and the Analysis of the Welfare State. [*Cited on page 6.*]

APPENDIX A. ESTIMATING ACCOUNTING VARIABLES FROM THE TAX DATA

This Appendix describes how we recover economically meaningful, financial-accounting-style variables from tax data for corporations filing Corporation Tax returns (CT600), partnerships filing partnership returns (SA800), and sole proprietors filing the self-employment pages of the Self Assessment return (SA103, accessed through the "Valid Views" extracts). The objective is to construct firm-year variables that correspond to the objects used in applied economic work and in financial accounting: operating profits and cash-flow proxies π_{ft} (EBIT and EBITDA-style), investment I_{ft}, capital stocks K_{ft}, and depreciation D_{ft}. Numbers reported by firms in their tax forms do not correspond to the quantities that would be calculated by accountants adhering to the UK Generally Accepted Accounting Principles (UK GAAP) or the International Financial Reporting Standards (IFRS). The central measurement challenge is to undo the tax-specific treatment of capital expenditure (*capital allowances*) and asset disposals (*balancing charges*), and replace it with an accounting-style treatment based on depreciation of productive capital over its useful life. We treat capital pools (*e.g.* "structure and buildings", or "electric vehicle charging points") equivalently across corporations, partnerships and sole proprietors, applying identical depreciation rates throughout.

A.1. **From taxable profits to accounting-style variables.** Our approach exploits the detailed capital-allowance fields reported on each tax return to (i) reconstruct investment and the evolution of the underlying capital stock, (ii) impute accounting depreciation using transparent assumptions on asset lives, and (iii) transform tax-reported profit measures into accounting-style profit measures by adding back Capital Allowances (CAs), netting out Balancing Charges (BCs), and subtracting imputed depreciation.

Tax-return items used. For each firm-year, we extract a tax-profit base and a detailed breakdown of capital allowances and balancing charges. We describe these items here using CT600 box numbers as the concrete reference. The SA800 (partnership) and SA103 (self-employment) returns carry the analogous capital-allowance and balancing-charge fields, which map to the same economic capital pools (see Table A.1).

- **Tax-profit base.** The taxable trading-profit base, net of current-year trading losses is called "Profits before other deductions and reliefs" (in the CT600 form, it is Box 235; the SA800 and SA103 returns report the analogous trading-profit measure).
- **Capital allowances (CAs).** Corporation, partnership and sole proprietorship returns report (subsets of) the following capital allowances. Coverage differ slightly across legal forms and years.

  - Main pool writing-down allowance (MR) (CT600 Box 705)
  - Special rate pool writing-down allowance (SR) (CT600 Box 695)
  - Annual Investment Allowance (AIA) (CT600 Box 690)
  - Full expensing allowance (CT600 Box 691)
  - Super-deduction allowance (CT600 Box 692)
  - Special rate first-year allowance (CT600 Box 693)
  - Structures and Buildings Allowance (SBA) (CT600 Box 711)
  - Business Premises Renovation Allowance (BPRA) (CT600 Box 715)
  - Electric vehicle charge points allowance (CT600 Box 713)
  - Zero-emission goods vehicles allowance (CT600 Box 723)
  - Zero-emission cars allowance (CT600 Box 726)
  - Other allowances (CT600 Box 725 and related boxes)
- **Balancing charges (BCs).** Asset disposal adjustments reported in the pool-specific balancing-charge boxes (*e.g.* for SR pool, MR pool, SBA/BPRA and other categories; box numbers vary by CT600 vintage).

Step 1: Mapping tax pools to economic pools. We build a firm-year panel of capital stocks and investment flows by tracking all capital pools reported by firms. Under the tax code, writing-down allowances are a fixed percentage of the pool's *tax written-down value*, so we can "capitalise" observed allowances to recover the implied pool size.

Two layers of pools are involved. We first work with the *tax pools* defined by the tax return such as the main rate pool, the special rate pool, the Annual Investment Allowance, the Structures and Buildings Allowance, the super-deduction, and so on. We use these to back out an investment flow from the observed allowances and balancing charges, using a methodology we detail below. We then map these tax pools into six *economic pools*; (i) short-lived plant & machinery, (ii) "middle-of-the-road" plant & machinery, (iii) long-lived plant & machinery, (iv) buildings and structures, (v) electric vehicles, and (vi) internal-combustion-engine cars. These economic pools aggregate capital assets (reported in the *tax pools*) that have similar economic depreciation rates. We accumulate capital and impute depreciation in these economic pools.

The mapping from tax pools to economic pools, together with the tax-return inputs used for each legal form, is given in Table A.1; it is identical across corporations, partnerships, and sole proprietors. We determine this mapping between the tax pools and economic pools based on the economic characteristics of the underlying assets, and our ability to identify the type of capital

accumulated by firms in each tax pool. For instance, we make the assumption that Zero-Emission Cars and Zero-Emission Goods Vehicles–two capital pools reported in tax returns–depreciate *economically* at the same rate, which set at 25% based on our understanding of best practices in corporate accounting. Some tax pools like Annual Investment Allowances are used for both short-lived and long-lived capital, so we map them to a bespoke "middel-of-the-road" accounting pool. The derivations below use the main rate and special rate pools as the worked illustration; the other tax pools are handled analogously.

Step 2: Recovering initial capital stocks. Let $t = 1$ denote the first year a business is observed. Define net allowances as allowances net of balancing charges within each pool. For the main rate pool:

$$\widetilde{CA}_1^{MR} \equiv \max\{CA_1^{MR} - BC_1^{MR},\, 0\}.$$

The main pool writing-down allowance is granted at 18% of the tax written-down value, so we initialise the implied main-pool capital as

$$K_1^{MR} = \frac{\widetilde{CA}_1^{MR}}{0.18}. \tag{A.1}$$

Similarly, for the special rate pool, where the writing-down allowance is 6%,

$$K_1^{SR} = \frac{\widetilde{CA}_1^{SR}}{0.06}, \qquad \widetilde{CA}_1^{SR} \equiv \max\{CA_1^{SR} - BC_1^{SR},\, 0\}. \tag{A.2}$$

The rates 0.18 and 0.06 are the statutory writing-down rates in force at the end of our sample; in practice we apply the rate prevailing in each year (for the main pool, 25% up to 2009, 20% in 2010–2012, and 18% thereafter; for the special rate pool, 10%, then 8%, and 6% from 2020).

For allowances that are effectively 100% expensed for tax purposes in the year of purchase (AIA, full expensing, certain first-year allowances, and other 100% categories), we treat the net claimed amount as current-year investment:

$$I_1^a = \max\{CA_1^a - BC_1^a,\, 0\}, \qquad a \in \mathcal{A}_{100\%}. \tag{A.3}$$

The super-deduction (when applicable) allows a 130% deduction for qualifying main-rate investment, so we recover the underlying investment by scaling down the allowance:

$$I_1^{SD} = \max\{(CA_1^{SD} - BC_1^{SD}) \times \tfrac{10}{13},\, 0\}. \tag{A.4}$$

For structures and buildings, the SBA is a straight-line allowance. In 2020 onward, the statutory rate is 3% per year (2% before the rate change). We therefore recover the underlying qualifying

expenditure (and thus the corresponding capitalised value) by reversing the rate:

$$I_1^{SBA} = \begin{cases} \max\{(CA_1^{SBA} - BC_1^{SBA}) \times \frac{1}{0.03},\, 0\}, & \text{for post-2020 qualifying expenditure,} \\ \max\{(CA_1^{SBA} - BC_1^{SBA}) \times \frac{1}{0.02},\, 0\}, & \text{for pre-change qualifying expenditure.} \end{cases} \tag{A.5}$$

Step 3: Inferring new investment from changes in allowances. For pools that are not fully expensed (MR and SR), the annual allowance reflects both (i) depreciation of the existing pool and (ii) additions/disposals. We infer net new investment by comparing the current net allowance to the allowance implied by the previous-year pool absent new investment. take for example the MR pool, with a writing-down rate of 18%, the tax-written-down value rolls over at $1 - 0.18 = 0.82$. Therefore, absent new investment, net allowances would fall by 18% on a shrinking pool:

$$\mathbb{E}_{t-1}\left[\widetilde{CA}_t^{MR} \mid I_t^{MR} = 0\right] = 0.82 \cdot \widetilde{CA}_{t-1}^{MR}.$$

Note that we are working with flows here: the capital allowance at time t is 82% of what it was at time $t-1$. We define net new investment allowance as

$$NNIA_t^{MR} \equiv \widetilde{CA}_t^{MR} - 0.82 \cdot \widetilde{CA}_{t-1}^{MR}, \tag{A.6}$$

that is, the difference between what is expected under no new investment and what is reported in the tax return, and convert it into an implied investment flow by reversing the 18% writing-down rate:

$$I_t^{MR} = \frac{NNIA_t^{MR}}{0.18}. \tag{A.7}$$

Analogously for the SR pool (6% writing-down; rollover factor 0.94):

$$NNIA_t^{SR} \equiv \widetilde{CA}_t^{SR} - 0.94 \cdot \widetilde{CA}_{t-1}^{SR}, \qquad I_t^{SR} = \frac{NNIA_t^{SR}}{0.06}. \tag{A.8}$$

In practice, $NNIA$ can be negative (net disposals); we retain these values for investment-flow accounting, and we impose non-negativity only where required for stock initialisation.

For 100% (or more) allowances, investment is observed directly each year:

$$I_t^a = \max\{CA_t^a - BC_t^a,\, 0\}, \qquad a \in \mathcal{A}_{100\%}, \tag{A.9}$$

with the super-deduction and SBA handled as in (A.4) and (A.5).

Step 4: Updating accounting capital stocks. With the values of investments thus defined across *tax* capital pools, we can capitalise these investments using the standard perpetual-inventory-style recursion with appropriate *economic* depreciation rates. We map tax pools to economic pools and

define the following recursion for each economic pool c,

$$K^c_{t+1} = \max\{0,\ (1-\delta^c)K^c_t + I^c_t\}, \tag{A.10}$$

where δ^c is the economic depreciation rate for pool c (Table A.1). Disposals do not enter as a separate term: they are already embedded in the investment flow I^c_t, which is built net of balancing charges through the net-new-investment-allowance construction above. The $\max\{0,\cdot\}$ floor prevents the reconstructed stock from turning negative after a run of net disposals. A situation that almost never happens in our data.

Step 5: Calculating total accounting depreciation. Let $\mathcal{C}$ denote the set of six economic pools introduced above. Annual, total depreciation for a firm is simply

$$D_t = \sum_{c\in\mathcal{C}} \delta^c K^c_t. \tag{A.11}$$

Table A.1 provides a transparent summary of the rate assumptions used in the current implementation. The same six rates are applied identically to corporations (CT600), partnerships (SA800) and sole proprietors (SA103); only the tax-allowance inputs that feed each pool differ by legal form.

Step 6: Constructing accounting-style profit measures (EBT, EBIT, EBITDA-style). Turning to profit measures, we construct the same variables for all three legal forms–trading profit, earnings before tax (EBT), EBIT, and an EBITDA-style measure–alongside the investment, capital, and depreciation series above. The core accounting reconciliation is:

$$\widehat{\text{EBT}}_t = \text{Tax-profit base}_t + \underbrace{\sum_a CA^a_t}_{\text{add back tax allowances}} - \underbrace{\sum_a BC^a_t}_{\text{subtract tax disposal adjustments}} - \underbrace{D_t}_{\text{imputed accounting depreciation}}, \tag{A.12}$$

where D_t is defined in (A.11). The $\widehat{\cdot}$ symbol on EBT stresses that this quantity is estimated. Intuitively, we (i) remove the tax system's accelerated expensing (add back CAs), (ii) remove the tax system's disposal adjustment (net out BCs), and (iii) impose the accounting expensing rule (subtract imputed depreciation).

To obtain an operating-profit measure, we adjust EBT for non-operating income and expenses where these are observable on the return: we add back financing costs and net out financing income to isolate EBIT. We then define an EBITDA-style measure by adding imputed depreciation

Accounting pool	δ	Tax-allowance inputs
Short-lived plant & machinery `K_sl`	20%	Main Rate Pool (CT600 Box 705); Super-Deduction (Box 692); other allowances and charges (Boxes 725/750). SA800 and SA103 analogues: main-rate plant & machinery and "other" capital allowances.
"Middle" plant & machinery `K_middle`	12.5%	Annual Investment Allowance (CT600 Box 690); Electric Vehicle Charge Points (Box 713). SA800 and SA103 analogues: AIA and EV charge-point allowance.
Long-lived plant & machinery `K_ll`	7.5%	Special Rate Pool (CT600 Box 695); Special Rate First Year Allowance (Box 693); Business Premises Renovation (Box 715). SA800 analogue: special-rate equipment pool. SA103 analogue: special-rate pool plus business premises renovation.
Buildings and structures `K_ibs`	4%	Structures and Buildings Allowance (CT600 Box 711). SA800 analogue: agricultural/industrial buildings allowance plus SBA. SA103 analogue: the same, augmented by the freeport SBA.
Electric vehicles `K_elec`	25%	Zero-Emission Cars (CT600 Box 726); Zero-Emission Goods Vehicles (Box 723). SA800 and SA103 analogues: the corresponding zero-emission allowances.
Internal-combustion-engine cars `K_cars`	10%	Cars not included in the main pool, tracked separately from Box 705 on CT600 and via the equivalent vehicle allowance lines on SA800 and SA103.

TABLE A.1. Accounting depreciation rates used for capital-stock and depreciation imputation

Notes: The table reports the single accounting depreciation rate applied to each economic capital pool in the BWR's reconstruction of firm-level capital stock and depreciation. The same six rates are applied identically to corporations (CT600), partnerships (SA800) and sole proprietors (SA103). Rates correspond to the following assumed useful lives: 5 years for short-lived plant & machinery (office equipment, general machinery), 8 years for the "middle" pool (mixed plant & machinery covered by AIA and related fully-expensed allowances), 13 years for long-lived plant & machinery (integral features, long-life assets, building renovations), 25 years for buildings and structures, 4 years for electric vehicles, and 10 years for internal-combustion-engine cars. Under the UK statutory Full Expensing regime (CT600 Box 691, in force from April 2023) new plant & machinery is expensed in the year of acquisition; Box 691 is not currently in the BWR extract and is therefore not a separate input to the pools above. CT600 box numbers follow the form in use at the end of the sample; the SA800 and SA103 analogues differ in variable naming but map to the same economic categories.

D_t back to EBIT. This measure adds back depreciation but not amortisation: because the tax returns provide limited information on amortisation of intangible assets, it is strictly an EBIT-plus-depreciation (EBITD) concept. We leave amortisation measurement to future work using richer sources on intangible assets. One area of future refinement we are envisioning is to use R&D tax credit data and patent information to estimate intangible capital stocks and impute amortisation.

APPENDIX B. COMPANIES HOUSE ACCOUNTS FILING REQUIREMENTS

Statutory accounts filed at Companies House depend on the size category into which each company falls. Companies House assigns companies to four size categories (micro, small, medium, and large) using a two-of-three test on turnover, balance-sheet total, and employee headcount. A company is initially assigned to the category whose thresholds it satisfies on at least two of the three criteria. Thereafter, moves between categories are governed by a two-consecutive-years rule, so a company's category can change over time but is not re-set mechanically each year. The thresholds and the filing obligations attached to each category have been revised several times over the BWR's coverage window.

Three features of this regime matter for the FAME benchmarking in section 5. First, the micro-entity category did not exist before December 2013; before that date the smallest companies fell under the small-company regime. Second, small companies have never been required to file a profit-and-loss account at Companies House, so turnover and operating profit are never observed in FAME for this size group. Third, the 2016 reforms abolished *abbreviated* accounts and introduced *abridged* accounts for small companies. Before 2016, companies typically prepared full accounts for shareholders and HMRC and then filed a separate abbreviated version at Companies House. After 2016, a small company could instead prepare abridged accounts, which had to be the same accounts sent to shareholders, HMRC, and Companies House and required member approval. *Filleting*, by contrast, means removing items such as the P&L from the public filing, and it was possible both before and after 2016. For our purposes, the key implication is unchanged: the Companies House filing for small companies may omit the P&L throughout the sample. The fixed-asset note attached to the balance sheet, however, is required across all size categories that file a balance sheet, so capital expenditure and depreciation are recoverable for every size category except micro-entities.

Table B.2 reports the size-category thresholds in force over the BWR's coverage window; Table B.3 summarises the financial items that each size category is required to file in each regime.

Category	Criterion	Pre-Jan 2004	Jan 2004 (SI 2004/16)	Apr 2008 (SI 2008/393)	Dec 2013 (SI 2013/3008)	Jan 2016 (SI 2015/980)
Micro	Turnover	—	—	—	$\leq$ £632k	$\leq$ £632k
	Balance sheet	—	—	—	$\leq$ £316k	$\leq$ £316k
	Employees	—	—	—	≤ 10	≤ 10
Small	Turnover	$\leq$ £2.8m	$\leq$ £5.6m	$\leq$ £6.5m	$\leq$ £6.5m	$\leq$ £10.2m
	Balance sheet	$\leq$ £1.4m	$\leq$ £2.8m	$\leq$ £3.26m	$\leq$ £3.26m	$\leq$ £5.1m
	Employees	≤ 50	≤ 50	≤ 50	≤ 50	≤ 50
Medium	Turnover	$\leq$ £11.2m	$\leq$ £22.8m	$\leq$ £25.9m	$\leq$ £25.9m	$\leq$ £36m
	Balance sheet	$\leq$ £5.6m	$\leq$ £11.4m	$\leq$ £12.9m	$\leq$ £12.9m	$\leq$ £18m
	Employees	≤ 250	≤ 250	≤ 250	≤ 250	≤ 250
Large		*Exceeds the medium-company thresholds on at least two of the three criteria*				

TABLE B.2. Companies House size-category thresholds

Notes: A company is initially assigned to the size category whose thresholds it satisfies on at least two of the three criteria. Thereafter, a change in category generally requires the relevant thresholds to be met or breached for two consecutive financial years. The micro-entity category was introduced in December 2013; before that date the smallest companies fell under the small-company regime. "—" indicates that the category did not exist in the relevant period. Column headers show the month in which each revision took effect; the smaller line beneath gives the Statutory Instrument that brought the revision into force.

Financial item	**Micro** (from 2013)	**Small, pre-2016** (abbreviated)	**Small, post-2016**	**Medium, pre-2008** (abbreviated)	**Medium, 2008–2015** (abbreviated)	**Medium, post-2016** (full)	**Large** (all periods)
Turnover	×	×	×	× (combined)	✓ (shown)	✓	✓
Operating profit	×	×	×	Partial (from gross profit)	Partial (from gross profit)	✓	✓
Fixed assets (B/S)	Aggregate only	By type (I/T/Inv)	Full or by type	Full detail	Full detail	Full detail	Full detail
Capital expenditure	×	✓ (FA note)	✓ (FA note)	✓ (FA note)	✓ (FA note)	✓ (FA note)	✓ (FA note)
Depreciation	×	✓ (FA note)	✓ (FA note)	✓ (FA note)	✓ (FA note)	✓ (FA note + P&L notes)	✓ (FA note + P&L)

TABLE B.3. Availability of financial items in Companies House filings

Notes: ✓ denotes that the item is reported in the filed accounts; × denotes that it is not. "FA note" refers to the fixed-asset movement note attached to the balance sheet, which decomposes changes in each fixed-asset category into additions (capital expenditure), disposals, and depreciation charged. "Abbreviated" accounts (pre-2016) were separate Companies House filings derived from fuller accounts prepared for members and HMRC. From 2016, abbreviated accounts were abolished and small companies could instead prepare "abridged" accounts, which had to be the same accounts sent to members, HMRC, and Companies House and required member approval. The "Small, post-2016" column summarises the public filing under that regime; the filed version may also be "filleted", meaning that the director's report and/or the P&L are removed before filing at Companies House. Filleting was possible both before and after 2016. Operating profit is not a prescribed line in UK statutory formats, so its availability depends on whether the filed P&L uses a format from which it can be read directly or inferred from gross profit. "I / T / Inv" abbreviates the Intangible, Tangible, and Investment fixed-asset categories.

The combined reading of Tables B.2 and B.3 accounts for a large share of the FAME–tax discrepancies documented in section 5. Turnover is absent from Companies House filings for small and micro companies throughout the panel, and for medium companies before 2008; this is a filing-requirement gap rather than a data-processing artefact. Operating profit is subject to the same gap, compounded by the fact that it is not a mandated line item in UK statutory formats. Capital expenditure and depreciation, in contrast, are recoverable from the fixed-asset note for every size category except micro-entities, so the FAME coverage gap for these two variables is confined to the smallest firms and to years in which a company falls into the micro category. These regularities help explain why the tax-derived series in Figures 4–10 track FAME most closely for large companies and depart from FAME most visibly in the lower tail of the size distribution.