ECONOMIC STATISTICS CENTRE OF EXCELLENCE

A COLLABORATION WITH Office for National Statistics

The UK Longitudinal Business Database

Yohannes Lemma, Silvia Lui, Louise Romaniuk, Jakob Schneebacher and Nick Wolf

ESCoE Technical Report TR-23

September 2023

ISSN 2515-4664

TECHNICAL REPORT

The UK Longitudinal Business Database
Yohannes Lemma, Silvia Lui, Louise Romaniuk, Jakob Schneebacher and Nick Wolf
ESCoE Technical Report No. 23
September 2023

Abstract

This paper introduces the UK Longitudinal Business Database (LBD), a new and experimental business reference database developed by the Office for National Statistics (ONS). It describes the input data, the data linkage and filtering methodology behind the LBD. The LBD is a re-usable longitudinal data spine that is built by using consecutive snapshots of the Inter- Departmental Business Register (IDBR), the UK business register, to capture longitudinality. IDBR unit-specific economic activity criteria are applied to all units of the business structure hierarchy. This process creates the closest measure to date of the active UK business population. The data spine only reflects dynamics embedded in various IDBR snapshots without imposing assumptions to variables that measure business activities. Hence, it provides a flexible, reusable and consistent framework to link firm-level data from administrative and survey sources, via unique identifiers, to create longitudinal linked datasets that suit different user needs. This paper aims to facilitate proper use and understanding of LBD-generated data by ONS-accredited microdata researchers. To aid understanding, it also presents comparisons with existing UK and US business demography datasets, new evidence on firm entry and exit and on the changing distribution of the UK business population at a quarterly frequency, 1999-2022. LBD-created datasets will be accessible to accredited researchers in the ONS Secure Research Service (SRS).

Keywords: Firm-level data infrastructure; data linkage; business demography; firm-dynamics

JEL classification: C80; C81; D22; L25

Yohannes Lemma, Office for National Statistics
yohannes.lemma@ons.gov.uk

Published by:
Economic Statistics Centre of Excellence
King's College London
Strand
London
WC2R 2LS
United Kingdom
www.escoe.ac.uk

The UK Longitudinal Business Database

Yohannes Lemma[1], Silvia Lui[1,2], Louise Romaniuk[1], Jakob Schneebacher[3,4], and Nik Wolf[1]

[1]*Office for National Statistics*
[2]*Economic Statistics Centre of Excellence*
[3]*Competition and Markets Authority*
[4]*King's College London*

September, 2023

Abstract

This paper introduces the UK Longitudinal Business Database (LBD), a new and experimental business reference database developed by the Office for National Statistics (ONS). It describes the input data, the data linkage and filtering methodology behind the LBD. The LBD is a re-usable longitudinal data spine that is built by using consecutive snapshots of the Inter-Departmental Business Register (IDBR), the UK business register, to capture longitudinality. IDBR unit-specific economic activity criteria are applied to all units of the business structure hierarchy. This process creates the closest measure to date of the active UK business population. The data spine only reflects dynamics embedded in various IDBR snapshots without imposing assumptions to variables that measure business activities. Hence, it provides a flexible, reusable and consistent framework to link firm-level data from administrative and survey sources, via unique identifiers, to create longitudinal linked datasets that suit different user needs. This paper aims to facilitate proper use and understanding of LBD-generated data by ONS-accredited microdata researchers. To aid understanding, it also presents comparisons with existing UK and US business demography datasets, new evidence on firm entry and exit and on the changing distribution of the UK business population at a quarterly frequency, 1999-2022. LBD-created datasets will be accessible to accredited researchers in the ONS Secure Research Service (SRS).

JEL codes: C80; C81; D22; L25

Keywords: Firm-level data infrastructure; data linkage; business demography; firm-dynamics

---

Disclaimer: *Any views expressed are solely those of the authors and do not necessarily represent those of the Office for National Statistics, the Competition and Markets Authority, or the wider UK government. This paper uses ONS statistical research datasets. Outputs may not exactly reproduce National Statistics aggregates.*

Corresponding author: Yohannes Lemma (yohannes.lemma@ons.gov.uk)

Acknowledgements: Many other ONS colleagues contributed to the Longitudinal Business Database. We would like to especially thank Zahir Islam, Josefa Lavandero and Geoffrey Megardon. Mika Morgan provided excellent research assistance in the early stages of this project. We are thankful to Arathi Bhaskar, Russell Black, Sumit Dey-Chowdhury, Mirko Draca, Richard Dorsett, Yannis Galanakis, Nathan Goldschlag, Kyle Jones, Alex Montag, Will Shepherd, Lilly Taylor, Tom Wickersham, Mark Williams and seminar participants at the ESCoE Annual Conference 2021 and the Government Statistical Service Methodology Conference 2021 for helpful feedback. Anthony Savagar deserves particular thanks for his help improving this draft.

1 Introduction

Constructing business databases for the UK that can capture genuine longitudinality (and be used to link data from different sources) has been challenging. Attempts have been made to create business databases for UK business analysis purposes, such as the Business Structure Database (BSD). However, these existing products come with certain limitations. Firstly, many of them use single snapshot data from the Interdepartmental Business Register (IDBR) to track continuity and changes across a time period. As a snapshot only reveals what has happened at a point in time, these products provide a more limited view on changes in business structure. Besides, survey and admin data usually cover information relevant to a reference period. Using a product created from a single snapshot for linkage will result in an increased number of unmatched cases and therefore under-utilised data. Secondly, many existing linked business datasets are created for specific analytical purposes or to answer specific research questions. Hence, often specific data cleaning criteria are imposed in the creation of the dataset itself that make it only suitable for the intended use. As the purpose or the research questions change, the source data then needs to be re-linked and cleaned anew. As a result, the UK currently lacks a reusable and consistent longitudinal framework built using data from the business register that enables linkage of data sources for multiple purposes.

Since the introduction of the Digital Economy Act 2017, the type and volume of data the Office for National Statistics (ONS) receives has been increasing. These include administrative data sources such as the Value Added Tax (VAT), the Trade in Goods (TIG) data, Pay As You Earn (PAYE) data and Corporation Tax (CT) records. Linking new and existing pertinent data sources allows researchers to better understand UK business dynamics. Therefore, a consistent framework for data linkage is vital. In view of the limitations of the existing UK business databases, coupled with an increasing need to link different types of data whilst ensuring that data are utilised efficiently, the ONS has produced the Longitudinal Business Database (LBD).

This paper explains the construction of the LBD and presents some new descriptive facts about the UK business population, 1999-2022. The LBD is a new and experimental business reference database at quarterly frequency. The LBD is at its core a re-usable longitudinal data spine, built by using consecutive quarterly snapshots of the IDBR.

The concept of a "data spine" is itself quite new. The LBD makes a clear distinction between the data spine and linked datasets. A data spine is a series of reusable longitudinal identity links constructed using snapshot data from the IDBR that are broken down and stored in different data tables. It contains key identification variables that enable linkages across different datasets tailored for different user needs. No assumptions about business measurements are imposed in its creation, and it captures purely the dynamics of firms and changes in their structure over time as reflected

in the IDBR. It therefore provides users the flexibility to impose assumptions when creating linked datasets that suit their own analytical purposes.

The LBD differs from existing linked IDBR products such as the Business Structure Database (BSD) and the Longitudinal IDBR (LIDBR, built by the then Department for Business, Energy and Industrial Strategy, BEIS) in three major ways: first, it provides information on business demography at quarterly frequency, crucial for economic analysis at business-cycle frequencies. Second, by using information from two adjacent snapshots it establishes longitudinality for all units that comprise the structure of a business. This feature allows the LBD to capture quarterly changes in activity faithfully. Third, it makes explicit the rules used by the ONS to capture economic activity on the IDBR and converts them into activity criteria that give rise to an economically meaningful definition of the business population.

The LBD has two purposes: first, it offers researchers a new business demography dataset that makes better use of the structure of the IDBR to investigate issues of firm entry and exit and the evolving distribution of firms over time. Second, the LBD spines provide a consistent frame for linking together a variety of longitudinal firm-level data from survey, administrative and other alternative data sources. It thus creates a consistent and robust structure for the flexible creation of many linked datasets, which cater to a wide range of research questions. Moreover, the use of data spines allows for data quality improvements of key variables through linkages with administrative or survey data sources. For instance, in future stages of this project, the LBD can be used to assign timely data on turnover and employment to all businesses in the IDBR from administrative sources (such as CT, VAT and PAYE schemes), and to use the LBD to link data from survey and web-scraped sources. This will enable researchers not only to extend the analysis of firm-level labour productivity from firms captured in the Annual Business Survey (Black, 2022) to the whole IDBR business population, but also to investigate the relationship between productivity and a wide range of firm characteristics. An early use case of the LBD is Lui et al. (2020), which used an earlier, internal version of the LBD to link PAYE records to investigate business dynamism trends in the UK economy since 1997.

The present version of the LBD only takes the IDBR as its input. The IDBR was created in 1994 to provide the ONS with a register from which to draw representative samples for its business surveys. The IDBR contains all businesses that at some point in their lifecycle have either paid VAT and therefore at one point have recorded turnover of over £85,000 or have hired employees that required registration for PAYE National Insurance contributions. In any given quarter, this amounts to roughly 3 million businesses, somewhat less than the 4 million incorporated businesses registered at Companies House and the estimate of about 6 million businesses counting all microbusinesses in the UK (ONS, 2022). Nonetheless, businesses on the IDBR represent roughly 97% of turnover and

88% of employment (BEIS, 2022).

The methodology for compiling the LBD spines consists of three steps: first, we use two consecutive snapshots of the IDBR to establish longitudinality for the intervening quarter. Second, we apply economic activity criteria that are tailored to the specific IDBR business unit in the business structure to classify the activity status of that unit. For instance, the economic activity criteria for reporting units takes into account information from the ONS on whether the business is survey-active; the economic activity criteria at the enterprise level on the other hand rely on active tax units (VAT and PAYE) and how recently employment and turnover values were updated. We use these steps to compile spines for each of the business units: enterprise group, enterprise, reporting unit and local unit. In ongoing work, we are building similar spines for the administrative side of the business: VAT unit, PAYE unit, Company Registration Number (CRN) and address information. In a final step, we link the spines together using the relationships on the IDBR.

The LBD is of course not the first attempt to build longitudinally linked firm-level datasets to analyse firm entry and exit, firm lifecycles or their contributions to aggregate productivity. In the UK, the BSD has been widely used to understand issues from productivity (Du and Bonner, 2016) to innovation (Evans and Nadeem Khan, 2009) and labour market power (Abel et al., 2018). Both the BSD and the LIDBR differ somewhat in their input data and methodology from the LBD. The LIDBR in particular does not attempt to improve the business structure aspect of the IDBR, but instead highlights the variable and often substantial lags with which IDBR data are updated from an array of different sources.

Abroad, the US Longitudinal Business Database (Jarmin and Miranda, 2002) is a model example and close relative, but again there exist clear differences. First, the US LBD captures firms at an annual frequency. Second, it is built around the five-yearly Census of Manufacturing, with additional information in non-Census years from administrative sources. As a result, the information the US LBD contains is incredibly comprehensive and accurate in Census years (see section 4), but less so in the intervening period (Jarmin and Miranda, 2002; Chow et al., 2021). The UK LBD, on the other hand, is consistent between quarters. Finally, the primary focus of the US LBD is on establishments (local units in the ONS business structure classification) whereas this is less the case for the UK LBD. The US LBD has proven to be an incredibly important dataset to answer questions around firm dynamics (Acs and Mueller, 2008; Fort et al., 2013), business dynamism (Fort et al., 2013), import-export decisions (Bernard et al., 2011; Antràs et al., 2022) and labour issues (Haltiwanger et al., 2013; Falato et al., 2022), among others. Other countries, for instance New Zealand (Fabling and Sanderson, 2016) and South Africa (Pieterse et al., 2018), have similar data products.

The rest of the paper is organised as follows. Section 2 outlines the differences between the LBD and similar existing data products. We will introduce the concept of a data spine and explain how it differs from traditional linked datasets. We will also explain how the LBD can enable better utilisation of data. Section 3 describes the underlying IDBR snapshots the LBD takes as its inputs. Section 4 explains the construction of the LBD, including the components of the data spines, the definition of reference quarters from adjacent IDBR snapshots, and the economic activity criteria applied for each IDBR unit of the business. Section 5 compares the LBD to existing UK business data linking attempts. Section 6 provides some initial descriptive results, including comparison to other business demography series and evidence on the evolving distribution of UK businesses at quarterly frequency. A final section 7 discusses current shortcomings of the input data and methodology and outlines planned next steps.

2 The LBD as a data spine

This section provides an overview of the differences between the LBD and existing UK data products of a similar kind. We discuss how the structure and construction of the LBD naturally lends itself to re-usability and flexibility, as well as offering a data linkage framework that enables better utilisation of data. We also discuss how the LBD differs from existing data products in its approach to business activity criteria.

2.1 Re-usability and flexibility

A single firm is represented as different business units in the IDBR. These units are enterprise group, enterprise, reporting unit, local unit, VAT unit and PAYE unit. The identification and information of each of these units are stored in separate data tables (snapshots). As a result, the identity links between these business units are not readily observed. The creation of the LBD therefore requires re-establishing the "lost" identity links to gather information about the whole firm.

The re-usability and flexibility of a data spine makes the LBD different from existing products , which are primarily linked datasets created to answer particular research questions. These linked datasets are created by imposing assumptions that suit their custom purpose. The cleaning criteria established to create these datasets also vary according to the purpose of the study. When the purpose or research questions change, underlying source data needs to be re-linked to create new datasets. The LBD data spine only reflects the dynamics readily observed from the business register. We do not impose further assumptions or restrictions during the construction of the LBD spine that may affect one's interpretation of dynamism. As such, the LBD data spine is very flexible and can be re-used to create many other linked datasets. Figure 1 below highlights the difference between a data spine and linked dataset in this respect. The spine contains only identification information

but not measurement variables. Hence it provides flexibility as users can define cleaning criteria that suit their needs when using it to create linked datasets.

Figure 1: Data spine vs linked dataset

Data spine	Linked dataset
• It is a consistent framework to produce multiple linked datasets using information from the IDBR.	• It is constructed by linking data to the data spine.
• It captures genuine longitudinality and continuity.	• It can be constructed for longitudinal or cross-sectional analysis, depending on the research questions.
• It contains identification information only, including reference numbers, industry markers, location markers, address references, (estimated) birth date, etc.	• It contains identification information and variables that measure firms' economic performance that comes from the data linked to the spine.
• Criteria established to construct a data spine capture only the dynamics of firms as reflected in the IDBR and do not vary with research questions.	• Its format is subject to cleaning criteria and assumptions imposed by users according to the research questions or problems in hand.
• A data spine is used to produce different linked dataset for multiple uses. It is not tailored for a single use.	• It is tailored for one type of use (according to the research question of interest).
• A data spine is re-usable.	• It needs to be re-constructed if the research questions or problems in hand have changed.

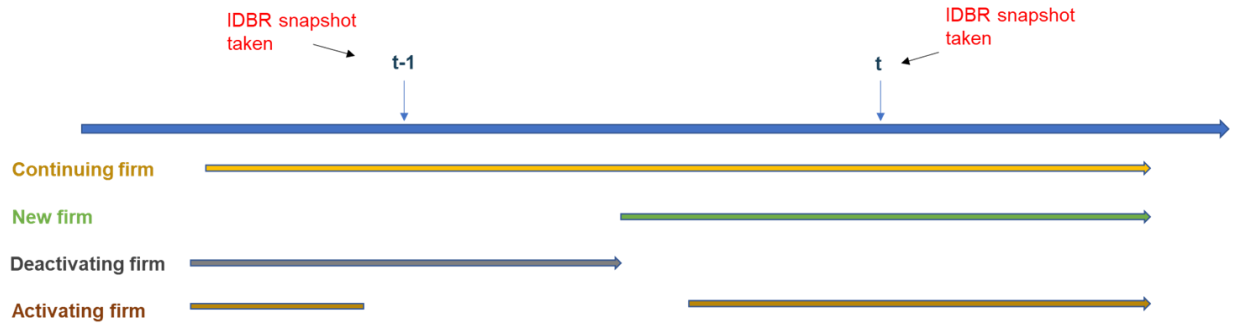
2.2 Better data utilisation

A key novel feature of the LBD is its construction using quarterly IDBR snapshots, to ensure changes over a reference period are captured appropriately.[1] The problem of using individual snapshots to construct database can be illustrated with a simple example as shown in figure 2. To fix ideas consider two points in time, $t-1$ and t when the IDBR snapshots are taken. We can think of four types of firms: (1.) a continuing firm that has started trading activity before $t-1$ and continue trading activity until sometime after time t; (2.) a new firm that has started trading activity between $t-1$ and t; (3.) a deactivating firm that has started trading activity sometime before $t-1$ but ceases trading activity sometime between $t-1$ and t (note that it may or may not resume activity after t); (4.) an activating firm that has stopped trading activity some time before $t-1$ but resumes trading activity sometime between $t-1$ and t.

If taking individual snapshot as an entry of a database for a reference period, the activity status of the four firms will appear differently in each database as illustrated in Figure 3. The IDBR is

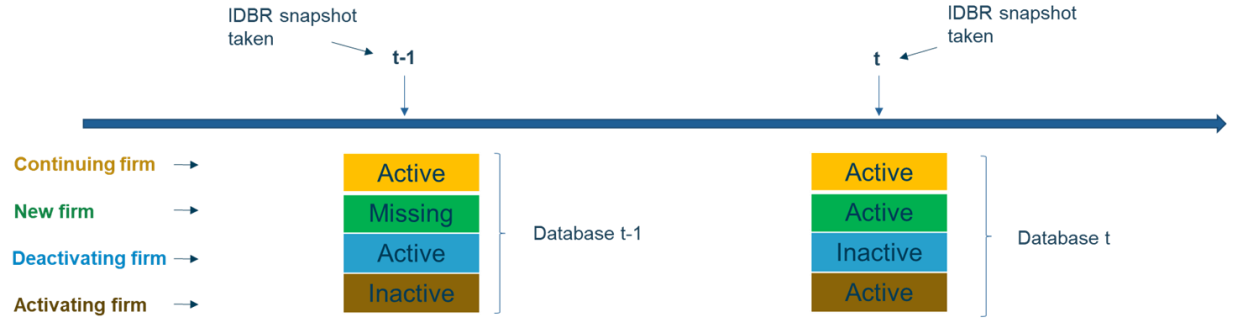
[1]Advantages of using consecutive snapshot for longitudinal analysis are highlighted in previous work by Lui et al. (2020) and Black et al. (2019), which used an internal version of the LBD datasets held at the ONS.

Figure 2: Firms with different activity status (Source: Black et al. (2019))



a survey sampling frame. Those firms that are inactive on the register at the point in time when the sample is taken are not chosen. Suppose we want to link a survey that is sampled between $t-1$ and t to the set of firms on the database built using individual snapshots, but there is no IDBR snapshot taken between these two points. Linking the survey to database t will lead to unmatched response from deactivating firms; while linking it to database $t-1$ would lead to unmatched responses from the new firm and the activating firms. This problem causes data to be under-utilised.

Figure 3: Activity status of firms shown in single snapshot (Source: Black et al. (2019))

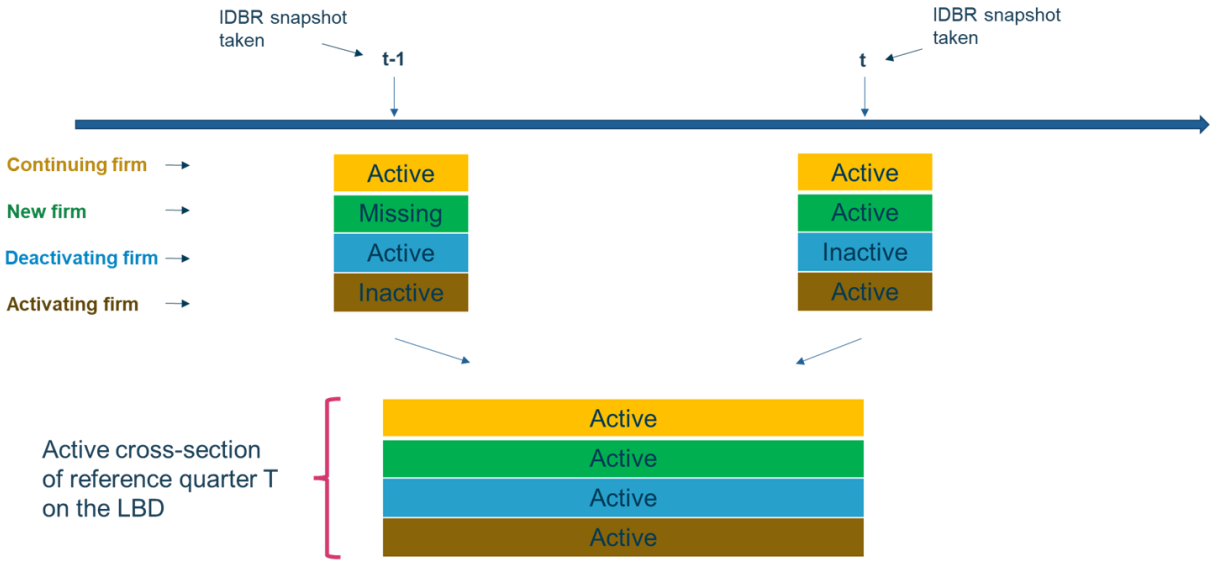


The data spine makes use of consecutive snapshots from the register to deduce any signs of activity over each reference quarter. Using the above example, all four types of firms will be captured in the active cross-section of reference quarter T on the LBD as shown in Figure 4. Snapshot $t-1$ and t sit at the two ends of a reference period. This resolves the problem of unmatched responses.

2.3 Inactivity vs death

When constructing the LBD data spine, we define businesses as "active" or "inactive", rather than "alive" or "dead". Drawing information about actual birth and death of firms from the business register is challenging. When the IDBR receives information about birth and death of a specific IDBR unit of a firm, usually lower-level administrative business units i.e., VAT and PAYE, this

Figure 4: Reference quarter *T* LBD active cross section (Source: Black et al. (2019))



information is then updated on the IDBR. A snapshot produced after this information is recorded will capture such incidents. However, since birth and death information of IDBR units could come at a lag, there are occasions when the IDBR could receive new information about the birth and death of a unit after the creation of a snapshot. As the IDBR snapshot data are never revised, such new information is updated on a new snapshot but not the old ones, making it difficult to work out the actual birth and death of an (administrative) unit. It is even more challenging for firms that have complex business structures that contain multiple lower-level administrative units.

Besides, death of an IDBR business unit within an enterprise does not always imply death of the whole enterprise. We observe birth and death of different business units for the same firms due to events such as mergers, acquisitions and takeovers; or merely changes in firms' business structures not necessarily due to those events. Creating a death variable implies making certain assumptions about what counts as business "death". This could affect the flexibility of the spine. Hence, when constructing the LBD data spine, we only ensure it reflects the dynamics and changes of different IDBR units of the same enterprise over time, thereby allowing the user to choose their own definition of business death in the creation of linked datasets.

3 Input Data

3.1 The Inter-Departmental Business Register

The construction of the LBD relies on the Inter-Departmental Business Register (IDBR) as its primary input source. Launched in 1994, the IDBR holds information on much of the UK business

population and is used by government for surveying and analytical purposes. It follows the regulations set out in the Eurostat Business Registers recommendations manual (Eurostat, 2010). The register is located and maintained at the Office for National Statistics (ONS) site in Newport.

Business information is primarily gathered from the following administrative sources:

- HM Revenue and Customs (HMRC) Value Added Tax (VAT)
- HMRC Pay-As-You-Earn (PAYE)
- Companies House
- Department for Environment, Food and Rural Affairs (DEFRA)
- Department of Finance and Personal, Northern Ireland (DFPNI)

Additionally, data are gathered from Dunn & Bradstreet, an external commercial data provider, as well as the ONS Business Register and Employment Survey (BRES) and other ONS surveys. Introducing these additional data sources allow for a more detailed industry and local region breakdown of UK business information than the primary administrative sources.

3.2 Which businesses are captured (and which are not)?

The main administrative sources for the IDBR are HMRC VAT and HMRC PAYE as well as Companies House. The administrative nature of the sources negates the need for estimation or imputation. The VAT and PAYE information is shared with ONS under the Value Added Tax Act 1994 and the Finance Act 1969, respectively. Incorporated business information is extracted from Companies House. Through these sources, the IDBR has a direct feed to all new businesses registered with these government departments, allowing the IDBR to capture approximately 2.8 million registered UK businesses.

The IDBR does not capture very small businesses which are not VAT registered (businesses are not required to register if their annual turnover falls below the VAT threshold, currently £85,000) or have no employees. Therefore many self-employed workers are excluded. Our most current estimates of the 'unregistered' population provided by the then Department for Business, Energy and Industrial Strategy (BEIS) amount to around 2.9 million. These unregistered businesses are very small in terms of turnover and employment, representing around 7% of estimated total employment and 1% of economic activity. However, the coverage and scope of these businesses can vary between industries.

Over-coverage in the IDBR can also occur, where businesses are no longer trading but are yet to be removed from the register, or duplicate business information has been gathered from multiple

sources. These cases account for a small proportion of the IDBR population.

The IDBR is updated daily, bi-monthly and quarterly with VAT, PAYE and other information. Snapshots of the IDBR – archives of the register taken at a given point in time - are taken on an annual, quarterly and monthly basis, with quarterly snapshots being taken near the end of every quarter. These are used for constructing the LBD spines. Section 4 describes the construction of these spines in more detail.

3.3 Business structure on the IDBR

A business on the IDBR can be decomposed into three types of business units:

1. **Administrative Units**: VAT trader and PAYE employer information supplemented with incorporated business data from Companies House.
2. **Statistical Units**: A group of legal units under common ownership is called an *enterprise group*; an *enterprise* is the smallest combination of legal units (generally based on VAT and/or PAYE records), which has a certain degree of autonomy within an enterprise group. An individual site (for example, a factory or shop) in an enterprise is called a *Local Unit*.
3. **Observation Units**: Reporting Units hold the mailing address to which the survey questionnaires are sent; the questionnaire can cover the enterprise, or parts of the enterprise identified by lists of local units.[2]

Figure 5 provides an example of the structure of a business found on the IDBR. These different units provide the basis for joining the LBD reference tables together for the LBD spine. A single reporting unit usually corresponds to one enterprise, except in the case where multiple reporting units are required for custom surveying purposes for enterprises with locations in different regions of the UK.

Furthermore, a business structure can be defined as being either 'simple' or 'complex'. A complex business is one that has multiple lower level units that span industries (as defined by their Standard Industrial Classifications, or SIC, code) or regions.

Figures 6 and 7 compare the structure of a simple and complex business, where 'ENTGRP', 'ENT' and 'LU' represent enterprise group, enterprise and local statistical units, respectively, 'RU' represents reporting observation units and 'VAT' and 'PAYE' represent administrative units.

[2]Businesses may find it easier to provide survey responses based on neither administrative or statistical units. For instance, a business might have separate accounting and management structures for its regional divisions, or its manufacturing and retail divisions.

Figure 5: IDBR Business Structure

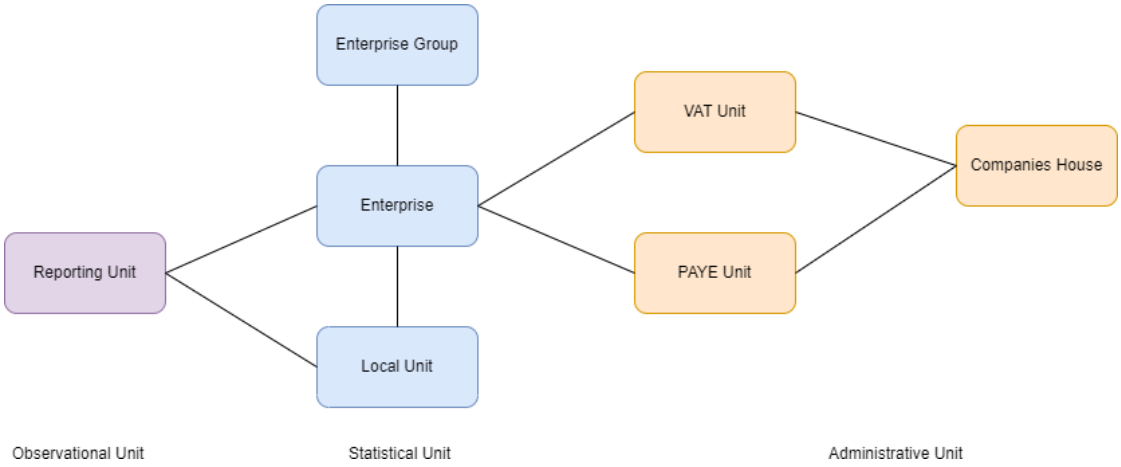
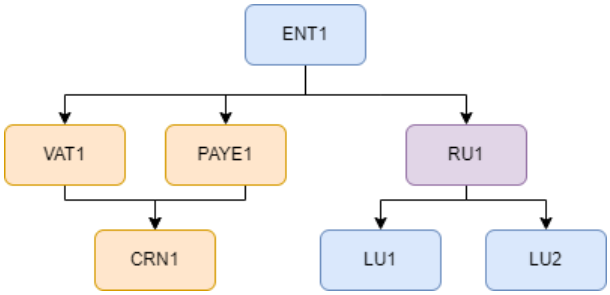


Figure 6: Simple IDBR Business Structure



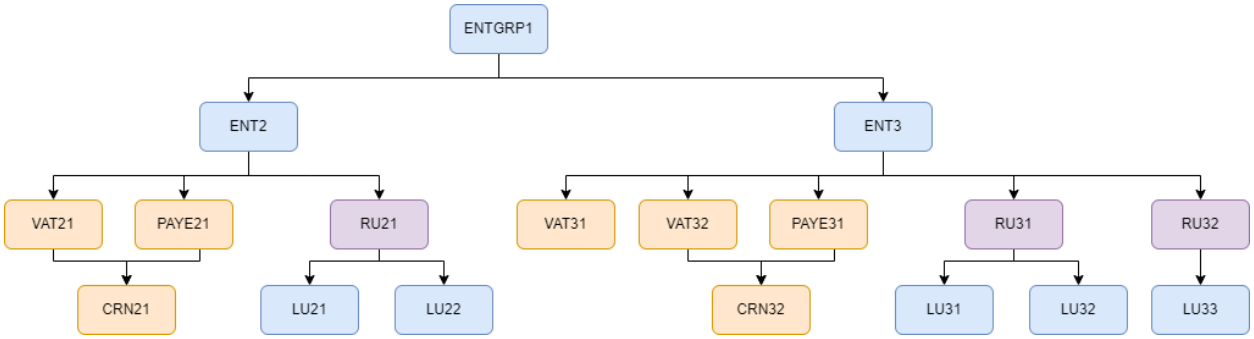
3.4 Variables and timing

The IDBR (snapshot) tables hold information on business names, addresses, standard industrial classification (SIC), employment, turnover, ownership and legal status, among other things. Additionally, every type of business unit is automatically given a unique reference number when it enters the IDBR. This allows us to link snapshots of IDBR tables. Also, this method provides us with the framework for construction of the LBD longitudinal spine.

Table 1 provides the list of variables in the first LBD dataset release, and records the source IDBR table. Definitions of these variables are given below:

- **Entref, wowref, crn, payeref, ruref, vatref, luref** – unique reference numbers assigned to the enterprise, enterprise group (who owns what), company registration number, PAYE, reporting unit, VAT and local unit respectively.
- **Birthdate, deathdate** – the birth date and death date of the enterprise.
- **Employment and turnover** - employment totals and turnover derived from administrative data like PAYE returns and ONS surveys such as the Business Register and Employment

Figure 7: Complex IDBR Business Structure



Survey (BRES).

- **Standard Industrial Classification** – the standard industrial classification code of the business unit based on either the SIC2003 or SIC2007 classification.
- **Legal status** – the legal status of the firm (limited company, limited liability partnership, sole trader, partnership, etc.).

Table 1: Key variables from IDBR tables

Variable	Enterprise	Entgroup	Payeunit	Repunit	Vatunit	Localunit	Crnunit
entref	x		x	x	x	x	
wowref		x					
crn			x		x		x
payeref			x				
ruref				x			
vatref					x		
luref						x	
birthdate	x	x	x	x	x	x	
deathdate	x	x	x	x	x	x	
employment		x		x		x	
turnover	x	x		x	x		
Sic2007	x	x				x	
legalstatus	x		x	x	x		

4 Construction of the LBD

This section explains the construction of the Longitudinal Business Database (LBD) from the IDBR quarterly snapshot data described in Section 3. We proceed in three steps: first, we construct different components of the longitudinal spine. Each of these components is specific to a business unit of a firm on the IDBR. Second, we combine these components to re-establish the business structure of

businesses over time. This is not otherwise observed since information about each business unit of the same business is stored in different data tables (snapshots). Finally, we classify active businesses.

4.1 The LBD data spines

The LBD data spine consists of five components, which correspond to identity links between business units on the IDBR. The five components are:

1. **L-WE**: Longitudinal link between Enterprise Group and Enterprise
2. **L-WERL**: Longitudinal link between Enterprise Group, Enterprise, Reporting Unit and Local Unit
3. **L-WEP**: Longitudinal link between Enterprise Group, Enterprise and PAYE
4. **L-WEV**: Longitudinal link between Enterprise Group, Enterprise and VAT
5. **L-WEC**: Longitudinal link between Enterprise Group, Enterprise and Company Registration Number

Figure 8: Construction of LBD spine

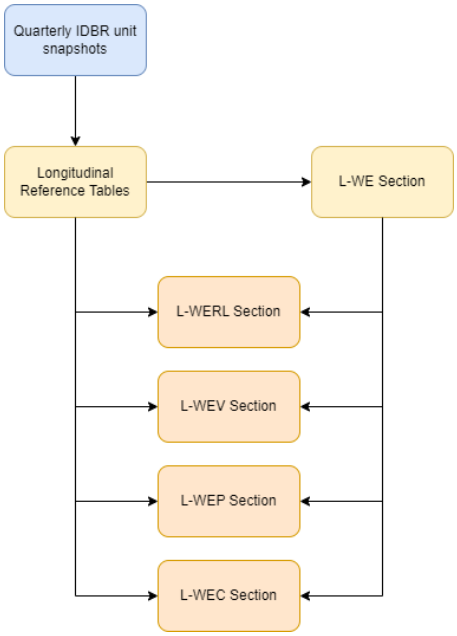


Figure 8 shows the different components of the LBD data spine. The L-WEP, L-WEV and L-WEC components are currently under construction.

4.2 IDBR quarterly files

To construct the components of the LBD data spine, we use quarterly snapshots from the IDBR. Information on the statistical, observational, and administrative units are returned in individual files. These are shown in Table 2.

Table 2: IDBR snapshots (data tables)

File Name	Unit References	Unit Variables	Unit Type	LBD Spine
Entgroup	wowref	Enterprise Group	Statistical Unit	L-WE, L-WERL, L-WEP, L-WEV, L-WEC
Enterprise	entref	Enterprise	Statistical Unit	L-WE, L-WERL, L-WEP, L-WEV, L-WEC
Wow	wowref, entref	NA	Linking Table	L-WE, L-WERL, L-WEP, L-WEV, L-WEC
Repunit	entref, ruref	Reporting Unit	Observational Unit	L-WERL
Localunit	entref, luref	Local Unit	Statistical Unit	L-WERL
Lulist	ruref, luref	NA	Linking Table	L-WERL
Paye	entref, payeref	Paye Unit	Administrative Unit	L-WEP
Vat	entref, vatref	Vat Unit	Administrative Unit	L-WEV
CRN	crnunit	NA	Linking Table	L-WEC
Irh Chsic92	crnunit	CRN Unit	Administrative Unit	L-WEC

4.3 Building longitudinality from adjacent snapshots

To build the components of the spine, we first construct longitudinal reference tables for each IDBR file. We complete this in two stages: we first create quarterly reference tables using two adjacent quarterly snapshots. Quarters in the reference tables include records from the two adjacent quarters of snapshots, current and lagged. These reference tables are then merged chronologically.

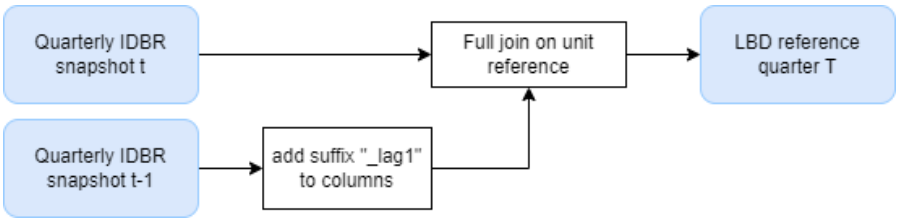
Existing IDBR-based datasets use annual snapshots directly. However, this captures outcomes at a particular point in time rather than over a period.[3] A quarterly reference table is built using two consecutive IDBR quarterly snapshots (t and $t-1$) to create one reference quarter, T. We use a full join to keep all the unit references from both quarters. Each row will contain the unit reference, unit variables (values from the more recent quarter, t) and lagged unit variables (values from the previous quarter, $t-1$).

Combining two consecutive snapshots means our quarterly reference table T, contains informa-

[3]The ONS annual business demography methodology uses historical information at different points in the year to produce a file with some longitudinality elements which underpins published official statistics.

tion from the start and end of the quarter. We can compare the current variable values with the lagged values to understand the unit's activity during the quarter. Figure 9 illustrates the construction of these quarterly reference tables.

Figure 9: Construction of quarterly reference tables



Using the quarterly reference outputs, we create a longitudinal reference table for each unit. We append the quarterly reference tables in one large table with each row containing the unit reference, unit variables , lagged unit variables (such as employment and turnover) and year-quarter. At this point we have recovered the longitudinal information for individual IDBR units. We can then link these longitudinal reference tables to construct the components of our LBD spine which recover the longitudinal links between units.

Consider the hypothetical example of the two snapshots and four enterprises shown in Tables 3 to 5. In this hypothetical example, our final longitudinal reference table across the three reference quarters and four enterprises would look like Table 6.

Table 3: 2018 Q1 Enterprise IDBR snapshot (t-1)

entref	enterprise variable
ENT1	100
ENT2	7
ENT3	1020

Table 4: 2018 Q2 Enterprise IDBR snapshot (t)

entref	enterprise variable
ENT1	120
ENT2	0
ENT3	1700
ENT4	50

Table 5: 2018 Q2 Enterprise Quarterly Reference Table (t)

entref	enterprise variable	lagged enterprise variable
ENT1	120	100
ENT2	0	7
ENT3	1700	1020
ENT4	50	NA

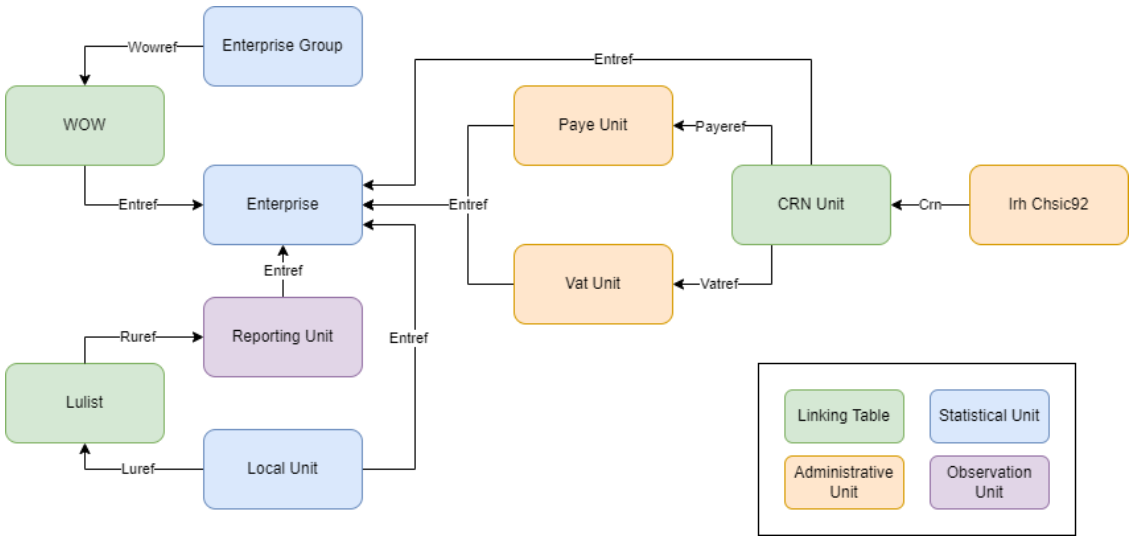
Table 6: Example Longitudinal Reference Table

entref	enterprise variable	lagged enterprise variable	year	quarter
ENT1	100	NA	2018	1
ENT2	7	NA	2018	1
ENT3	1020	NA	2018	1
ENT1	120	100	2018	2
ENT2	0	7	2018	2
ENT3	1700	1020	2018	2
ENT4	50	NA	2018	2
ENT1	NA	120	2018	3
ENT2	NA	0	2018	3
ENT3	NA	1700	2018	3
ENT4	NA	50	2018	3

4.4 Linking IDBR units

To recover the longitudinal link between units that are not readily observed in individual snapshots, we then link the longitudinal reference tables (reftable). Figure 10 illustrates how different identifiers of different IDBR business units are used to link the IDBR tables.

Figure 10: Unit References to Link IDBR Tables



To create the L-WE component of the spine we join three longitudinal reference tables: the ent-group reftable, which contains wowref and entgroup values; the enterprise reftable, which contains entref and enterprise values; and the WOW reftable, which contains the entref-wowref link.

To join the tables we take the following two steps:

1. We left join the entgroup reftable to the wow reftable on wowref, year, quarter. This keeps all values from the wow table and only values from the entgroup table that have a match in that year-quarter with a wowref in the wow reftable.

2. We then left join this table on entref, year, quarter to the enterprise reftable. This keeps all values from the enterprise table and only values from the joined wow-entgroup table that have a match in that year-quarter with an entref in the enterprise reftable.

To assemble the L-WERL part of the spine requires the most complicated joining strategy. This is because we need to use four tables to accurately link enterprise groups, enterprises, reporting units and local units. From these four tables we end up with just one longitudinal table, the L-WERL component, with the longitudinal link recovered between each unit.

The four tables required are the L-WE component, which contains the entref-wowref link and enterprise and entgroup values; the Reporting Unit (RU) Reftable, which contains the entref-ruref

link and RU values; the Local Unit (LU) Reftable, which contains the entref-luref link and LU values; and finally the Lulist Reftable, which contains the ruref-luref link for active so-called 500 RUs. Note that there are two types of RUs in the IDBR, the 499 RUs and the 500 RUs. 500 RUs are reporting units in an enterprise containing multiple reporting units. Unique reporting units within an enterprise are always denoted as 499 RUs. The types of RUs firms have on the IDBR convey meaningful information about the structure of a firm. Firms with simple business structure such as the one in 6 have only one RU and will be given a 499 RU. A single firm can only have one 499 RU, not multiple. If at some point during the lifetime a business expands to become one with a complex business structure such as the one in 7, its 499 RU will be replaced by multiple 500 RUs. According to the IDBR rules, firms cannot have both active 499 RU and 500 RUs. To recover the entref-ruref-luref link for these unique 499 RUs, we can link the L-WE spine to the repunit reftable (on entref, year, quarter) and then the joined L-WE-repunit table to the localunit reftable (on entref, year, quarter).

For 500 RUs, because multiple RUs are linked to the same enterprise, we cannot link reporting units to local units via the entref. Instead, we need the direct ruref-luref link from the lulist table. In other words, we link the L-WE spine to the repunit reftable (on entref, year, quarter) and then join the L-WE-repunit table to the lulist reftable (on ruref, year, quarter). The lulist only contains the ruref-luref link, and not any of the local unit values. To retrieve the local unit values, we must then link the joined L-WE-repunit-lulist table to the localunit reftable on (on luref, year, quarter).

4.5 Activity criteria

Having established reference quarters from adjacent snapshots, and linked the tables to reconstitute the business structure longitudinally, in a final step we apply a set of unit-specific economic activity criteria. As the L-WEP, L-WEV and L-WEC components of the spine are still under construction, our discussion in this section focuses only on active enterprises, active RUs and active LUs.

The economic activity criteria we establish to identify active enterprises fulfil two purposes: filter out inactive enterprises, and classify active enterprises into informative categories. We apply the activity criteria in a specific order. First, we identify active enterprises and remove any rows that contain inactive enterprises in that year-quarter. This leaves us with the subset of the IDBR enterprise population that is currently economically active according to our measure of economic activity.

After filtering for only active enterprises, we mark reporting units as active or inactive. In cases where an enterprise is linked to multiple reporting units and at least one is currently active, we remove all inactive reporting units linked to the enterprise. If the enterprise is linked to no reporting units or only inactive reporting units, we do not remove these rows. This is because we do not want to remove any currently active enterprises.

We repeat a similar process for local units and only remove inactive local units where the reporting unit is linked to a currently active local unit. If the reporting unit is linked to only inactive or null local units, we do not remove any rows. All units we identify as inactive but do not remove will be marked as dead in the unit activity type. To identify active enterprises, we classify all enterprises for each quarter into three categories: deleted, inactive or active. An enterprise is marked as deleted if it has been removed from the IDBR during the reference quarter so is present in snapshot $t-1$ but not in t.

We have three conditions to mark enterprises as inactive based on the Annual Business Demography criteria. All enterprises are marked as inactive if their turnover and employment values are zero for two consecutive snapshots. If the enterprise has at any point been linked to a live administrative unit we mark the enterprise as inactive when its live VAT and PAYE count is zero for two consecutive snapshots. However, if the enterprise has never been linked to a live administrative unit during its history on the register, we cannot use this as a marker. We instead will mark the enterprise as inactive when it has a deathdate value for two consecutive snapshots. Before enterprises are deactivated they will have a missing value in the *ent deathdate* column. The majority of active enterprises will have an associated administrative unit, however a small proportion of enterprises are currently trading but do not meet the threshold for requiring a PAYE or VAT unit.

All enterprises that have not been marked as deleted or inactive will be recorded as active. We filter the dataset to only keep active units. This will include enterprises in the quarter they first deactivate. Because each enterprise must meet the criteria for being inactive for two consecutive snapshots all enterprises that have been active at any point during a quarter will be kept in the LBD spine. For example, if an enterprise deactivates in a quarter and its turnover and employment values are set to zero we will mark it as active until the following quarter where the turnover and employment have been zero for two consecutive snapshots. This is because we do not know at which point during the quarter the enterprise deactivated. We only remove enterprises when they have been inactive for an entire quarter.

The statistical function of a reporting unit is to respond to official surveys. The reporting units of firms on the IDBR provide a sampling frame. When a sample is drawn from the IDBR, reporting units with death dates are excluded. We call such RUs as survey inactive at that point in time. A challenge in establishing criteria to identify active RUs is that a RU that is found to have a death date (survey inactive) recorded at some point in time can get back to “live” (survey active) due to changes of the firm’s business structure. It therefore makes the value of the *ru deathdate* variable unreliable.

We identify if the reporting unit is currently active in a quarter using the *ru deathdate* variable. Because the RU death date values in the IDBR are unreliable, rather than looking at the date, we only check if the column has any values. Before the reporting unit is marked as inactive on the IDBR, it will have a missing value in the death date variable. When the reporting unit has a death date value for two consecutive snapshots, we mark the reporting unit as inactive. Those that have no death date values in in one of the two consecutive snapshots sitting on the two ends of a reference quarter, we define it as an active RU in that reference quarter.

We also apply an additional check for identifying inactive 500 RUs. All currently active 500 RUs should appear in the lulist table. In cases where the 500 RU is in the repunit table but not in the lulist table we will mark the RU as inactive.

LUs on the IDBR are assigned "deathcodes" to describe their activity status. For the purposes of our WERL spine, for an LU to be considered active its "deathcode" in either snapshots, t or $t-1$, must be 0 or 6. Value 0 indicates the LU is "live". Value 6 are usually assigned to an LU while clerical investigation are undertaken to determine whether the LU is actually "live" or be marked as"dead" on the IDBR. Any other"deathcode" value corresponds to an inactive LU unit.

4.6 Active types

The ENT activity marker classifies all active enterprises. It records changes to an enterprise's activity status within the reference quarter so we can track the enterprise's activity through time. To decide the enterprise's activity marker we look at the activity status of the enterprise in the previous, current and following snapshot.[4] Table 7 describes the enterprise unit active types. Figure 11 shows how the active types of a enterprise in reference quarter T appears across three reference quarters.

The RU activity marker defines the reporting unit's active type. It records how the entref-ruref links have changed within the quarter. Table 8 describes the reporting unit active types. A new reporting unit could have been created and linked to a new or existing enterprise, or an existing reporting unit may have moved within the quarter to a new or existing enterprise. Although in each snapshot a reporting unit can only be linked to one enterprise, a reporting unit can move between complex enterprises over time. For the active types we consider two types of enterprises:

- **Simple enterprise**: Always linked to only one RU - these RUs have a ruref starting with 499

[4]A special case of enterprise status are so-called dummy enterprises, which are described in more detail in the appendix.

Table 7: Enterprise Unit active types

Active Type	Description
continuing_active	Enterprise is active and does not change activity status.
birth	Enterprise is created and appears on the register for the first time.
death	Enterprise deactivates.
reactivation	Enterprise has appeared before on the register and becomes active again after a period of inactivity.
birth_and_death	Enterprise is created and deactivated in the same quarter.
death_and_deleted	Enterprise is deactivated in this quarter and is removed from the register in the next quarter.
birth_and_deleted	Enterprise is created in this quarter and is removed from the register in the next quarter.

- **Complex enterprise**: Linked to at least two RUs - these RUs have a ruref starting with 500

499 RUs do not move between enterprises, one 499 RU is created per enterprise and will remain with the same enterprise until the enterprise becomes inactive or the RU is replaced with multiple 500 RUs. Therefore, we never see cases of 499 RUs moving between enterprises. These impossible types are listed in Table 9.

Figure 11: Activity type of an enterprise (in reference quarter T)

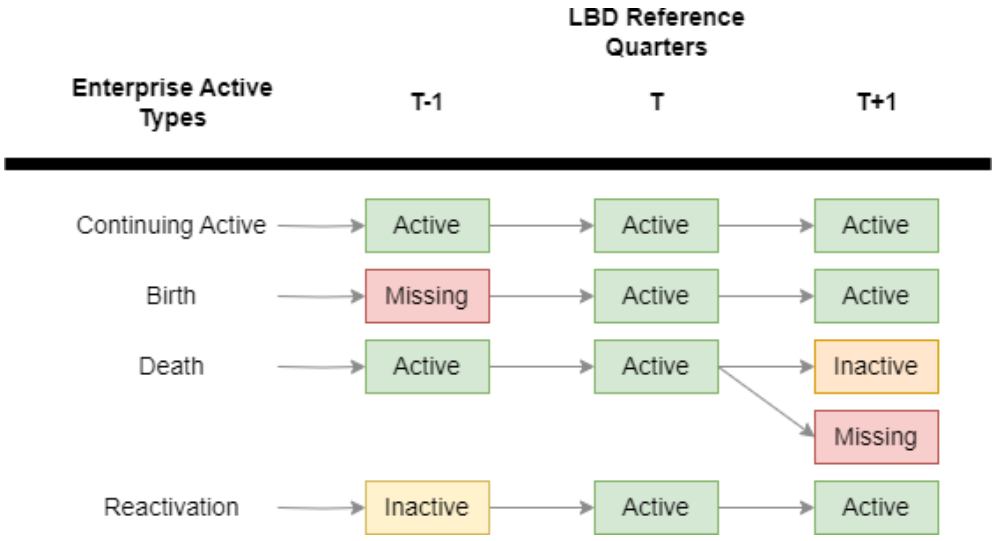


Table 8: Reporting Unit active types

Active Type	Description
Type A	Active new RU belonging to an existing simple enterprise.
Type B	Active new RU belonging to a new simple enterprise.
Type E	Active existing RU belonging to the same simple enterprise in both quarter snapshots.
Type F	Active new RU belonging to a new complex enterprise.
Type G	Active new RU belonging to an existing complex enterprise.
Type H	Active existing RU belonging to an enterprise in one snapshot and moving to a new complex enterprise in the next snapshot.
Type I	Active existing RU belonging to an enterprise in one snapshot and moving to another existing complex enterprise in the next snapshot.
Type J	Active existing RU belonging to the same complex enterprise in both snapshots.
Type K	Dead RUs linked to active ENT without any other active RUs (we keep these rows so we don't remove any active enterprises).
Type L	Null RUs linked to an active ENT - this category shouldn't exist. Live enterprise should always have at least one live RU. However, due to missing IDBR repunit tables for some quarterly snapshots, some enterprises have no linked RU.

Table 9: Impossible Reporting Unit active types

Active Type	Description
Type C	Active existing RU belonging to an enterprise in one quarter snapshot moving to another new simple enterprise in another quarter snapshot.
Type D	Active existing RU belonging to an enterprise in one quarter snapshot moving to another existing simple enterprise in another quarter snapshot.

The LU activity marker defines the local unit's active type. Where the local unit is an existing active LU the marker records how its enterprise and reporting unit links have changed within the quarter. Although in each snapshot a local unit will only be linked to one enterprise and reporting unit, a local unit can move between enterprises and reporting units over time. The local unit may be re-linked to a new entref, ruref or both. Since 499 RUs do not move between enterprises, any LUs linked to a 499 RU must either have moved to a new entref and ruref or remained with the same entref and ruref. A summary of LU active types is provided in Table 10.

Table 10: Local Unit active types

Active Type	**Description**
Type A	Active new LU.
Type B	Active existing LU relinked to an RU and ENT: Active existing LU that was relinked to a different RU and ENT during that reference quarter.
Type C	Active existing LU relinked to an RU with the same ENT: Active existing LU that was relinked to a different RU, but within the same ENT during that reference quarter.
Type D	Active existing LU with the same RU relinked to an ENT: Active existing LU linked to the same RU, but relinked to a different ENT during that reference quarter.
Type E	Active existing LU linked to the same RU and ENT: Active existing LU that is linked to the same RU and ENT at the beginning and end of the reference quarter.
Type K	Dead LU: Dead LUs linked to active RUs without any other active LU.
Type L	No Linked LU: RU without a linked LU (we keep these rows so we don't remove any active enterprises or reporting units).

5 Comparisons with similar datasets

The UK LBD is of course not the first attempt to build a comprehensive dataset of longitudinally linked, firm-level microdata for the UK. This section compares the UK LBD to two other such endeavours. First, we discuss differences and similarities with the Business Structure Database (BSD), a previous data product from the Office for National Statistics (ONS) to rebuild the business structure from the IDBR. Secondly, we compare the UK LBD with the Longitudinal IDBR (LIDBR), an experimental dataset built by data scientists in the then Department for Business, Energy, and Industrial Strategy (BEIS) which focuses on classifying the quality of input sources on the IDBR. The appendix also provides a comparison with the LBD's older and much better known US counterpart.

5.1 The Business Structure Database

The Business Structure Database (ONS, 2006; Evans and Welpton, 2009) represents a previous attempt to build a linked, firm-level business database for the UK. The BSD is constructed from annual snapshots of the IDBR and includes data on enterprise and local units. The purpose of the BSD is to create a version of the IDBR for research use, taking into account changes in ownership and the restructuring of businesses. Demographic events at enterprise and local unit level, included in the BSD, are in accordance with recommendation from Eurostat. At the enterprise and enterprise group level, demographic events comprise of:

- Birth and death
- Mergers and takeovers
- Break-ups and splitoffs
- Change of ownership and trade sale

Births, deaths and transfers are also included at the local unit level.

Like for the LBD, a set of rules help classify these demographic events and their dynamics. The demographic events markers are based upon the continuity of reference numbers and the links between reference numbers between successive annual cross sections of the IDBR. The input data for the BSD comes from the same annual snapshot, taken in March, as the annual ONS publication on business activity, size and location. However, for its annual business demography publication the ONS takes extracts straight from the IDBR and uses a slightly different methodology to classify activity status, births and deaths of businesses over a particular year.

Widely used (Evans and Nadeem Khan, 2009; Du and Bonner, 2016; Abel et al., 2018), the BSD nonetheless suffers from three substantial drawbacks: first, the Reporting Unit (RU) is not part of the business structure in the BSD, making it difficult to link to ONS survey data.[5] Second, the BSD only uses annual snapshots and is therefore of limited usefulness for business-cycle analysis. Third, the BSD does not provide true longitudinality through the simultaneous use of multiple snapshots.

5.2 The Longitudinal IDBR

The Longitudinal IDBR (LIDBR) is an experimental business dataset constructed by the then Department of Business, Energy and Industrial Strategy (BEIS) that also uses the IDBR as its input data. Like the Business Structure Database, it has an annual frequency. Unlike the LBD and the BSD, the primary focus of the LIDBR to date has been to investigate data input sources and data input lags, and to construct a series of flags that allow researchers to make their own choices on what input sources to prioritise.

Employment and turnover information on the IDBR are updated from administrative data sources as well as ONS surveys. As a result, there are timing and data quality issues to contend with when using this information from the IDBR. Large firms are surveyed every year and information pertaining to their employment and turnover is mainly sourced from ONS survey data. It will therefore generally be timely and of good quality. For smaller firms, since they are not included in the surveys every year, information will be derived from administrative sources, mainly VAT and PAYE. Unlike information from ONS surveys, data from the administrative sources may be measured at higher levels of aggregation or may not correspond directly to the variables of interest. To tackle these issues, the LIDBR cross-references information on employment and turnover against alternative data sources to identify inaccurate instances and apply adjustments dictated by pertinent business rules. The LIDBR currently covers the period from 2007 to 2017. There are currently no plans to update the LIDBR for future years.

6 Descriptive results

This section presents initial descriptive results from the LBD, focusing on total active enterprises and employment from the LBD between 1999 Q1 and 2022 Q2. These results have been derived from the LWE-spine, and are entirely focused at the enterprise level.

6.1 LBD active enterprises

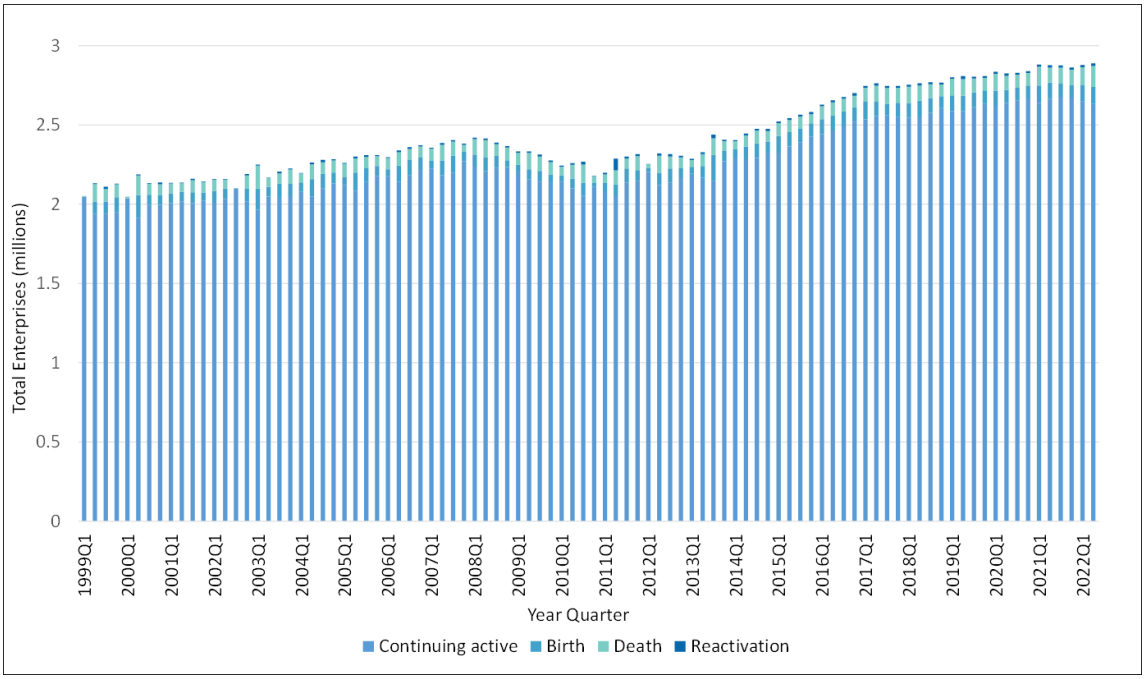
All our analysis in this paper excludes enterprises with the activity types *birth and death* and *birth and deleted* from statistics on active enterprises (definitions for these activity types outlined in Ta-

[5] RU-LU and RU-enterprise lookup tables are provided to researchers that want to bring survey information to an establishment or enterprise. Issues of business structure changes are side-stepped this way.

ble 7). This is because they are most likely related to IDBR administrative changes rather than representing real economic activity. We also combine activity types *death* and *death and deleted* into one category, as they both capture active enterprises that cease to be active before the next quarter.

Figure 12 below shows the number of enterprises by active types. On average 93% of enterprises are continuing active enterprises, firms that are active in the previous, current and following IDBR snapshots. Roughly 3% of enterprises are born and 3% die in each quarter with 0.5% reactivating. Reactivating firms are enterprises that have previously appeared on the IDBR and are registering economic activity for the first time after a period of inactivity.

Figure 12: LBD enterprises by active type

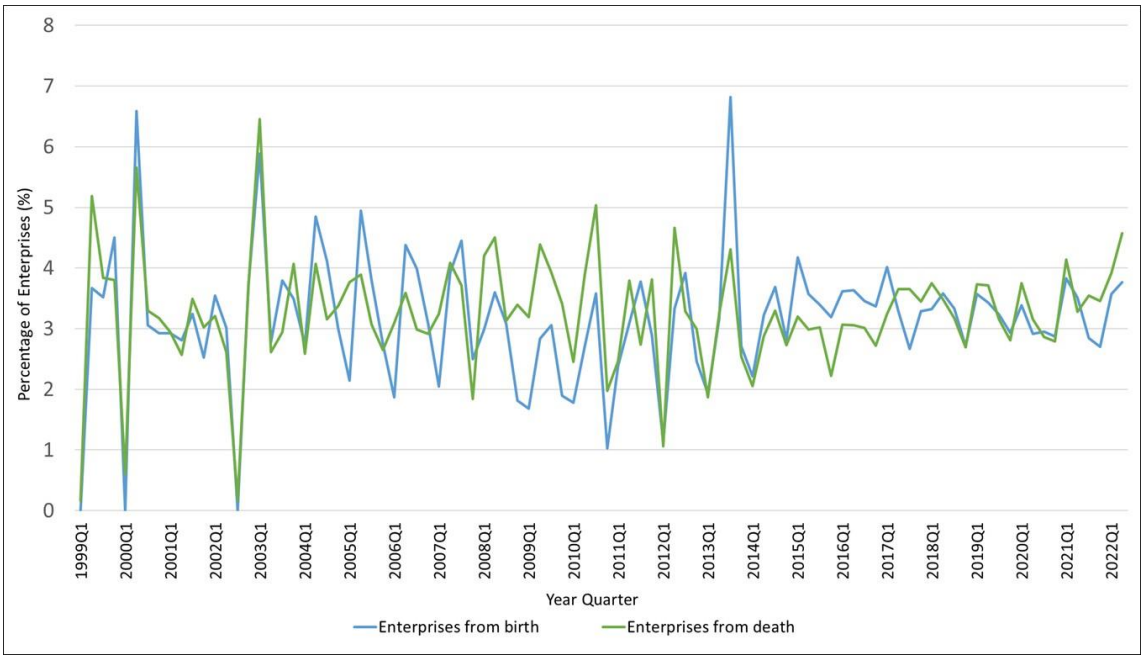


The total number of active enterprises on the LBD has increased from 2 million in 1999 Q1 to 2.9 million in 2022 Q1. This can be seen in Figure 12. Over time, active enterprises have gradually increased with the notable exception of the Great Financial Crisis (GFC). After the GFC, enterprises fell from 2.4 million in 2008 Q3 to their lowest point at 2.2 million in 2010 Q4. Total active enterprises did not recover to pre-GFC levels until 2013 Q4.

Figure 13 shows the quarterly birth rate and death rate of enterprises. The birth rate is the number of births as a percentage of the total number of enterprises, and the death rate is defined analogously. Quarters where birth or death rates are zero can be from missing IDBR snapshots or due to delayed data updates. Zero values are followed by unusually high levels in the following quarter which suggests a catching-up effect in the next quarter. This implies there are data quality

issues, at least in the early years of our data series. Following the GFC, death rates remain relatively constant while births drop, leading to a fall in active enterprises. Birth rates recover around 2011 and overtake deaths in 2014, leading to a rise in active enterprises on the LBD.

Figure 13: Quarterly enterprise birth and death rate



6.2 LBD active workforce

The LBD workforce has increased from 26 million in 1999 Q1 to 33 million in 2022 Q1. On average, the LBD workforce increases 0.26% each quarter, although with a lot of variation between quarters, as shown in Figure 14. Early quarters in the LBD are especially variable. This could be due to changes in the true nature of the economy or due to data quality issues in the IDBR. During the GFC, average employment growth between 2007 Q4 and 2011 Q1 fell below zero. Employment growth gradually recovered before slowly declining from 2017 onwards. Figure 15 provides additional detail on the drivers behind these changes. Both the employment change contributions of firm births and deaths have become lower and less variables in recent years, but the employment change associated with firm deaths in particular has fallen since the GFC.

6.3 LBD firm size and industry

Figure 16 shows the number of active enterprises in each quarter by firm size. Most enterprises on the LBD are micro firms with between zero and nine employees. Averaged across all years, 89% of active enterprises are micro firms with only 0.38% of active enterprises employing over 250 employees. Large firms however contribute disproportionally to total economic activity on the LBD. The

Figure 14: Quarter on quarter employment change

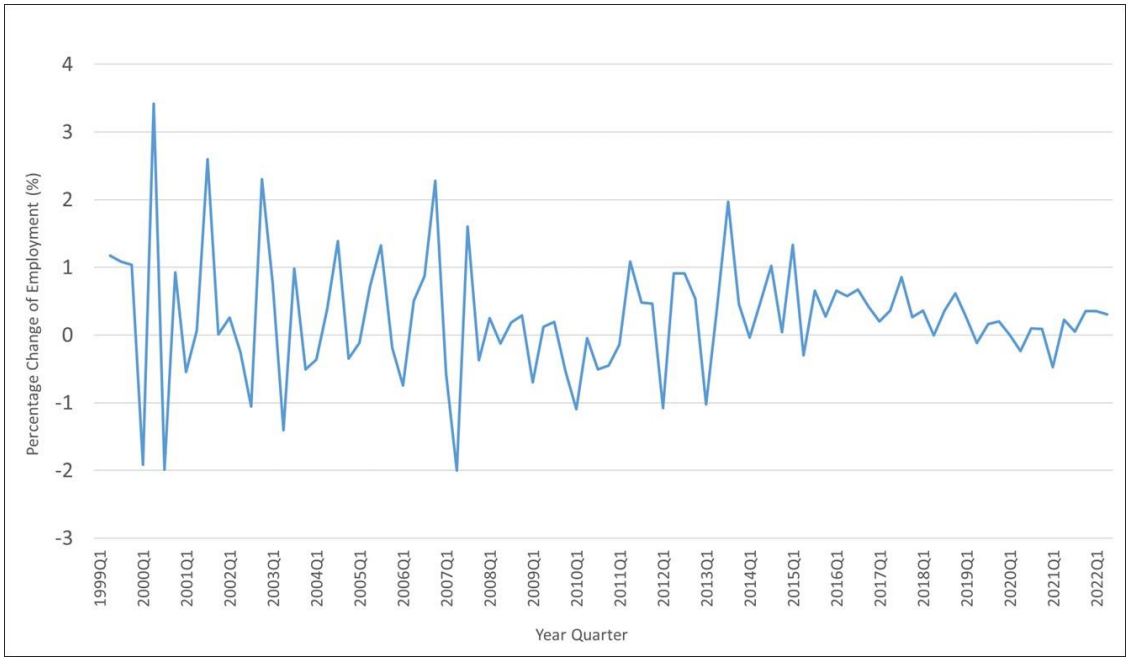
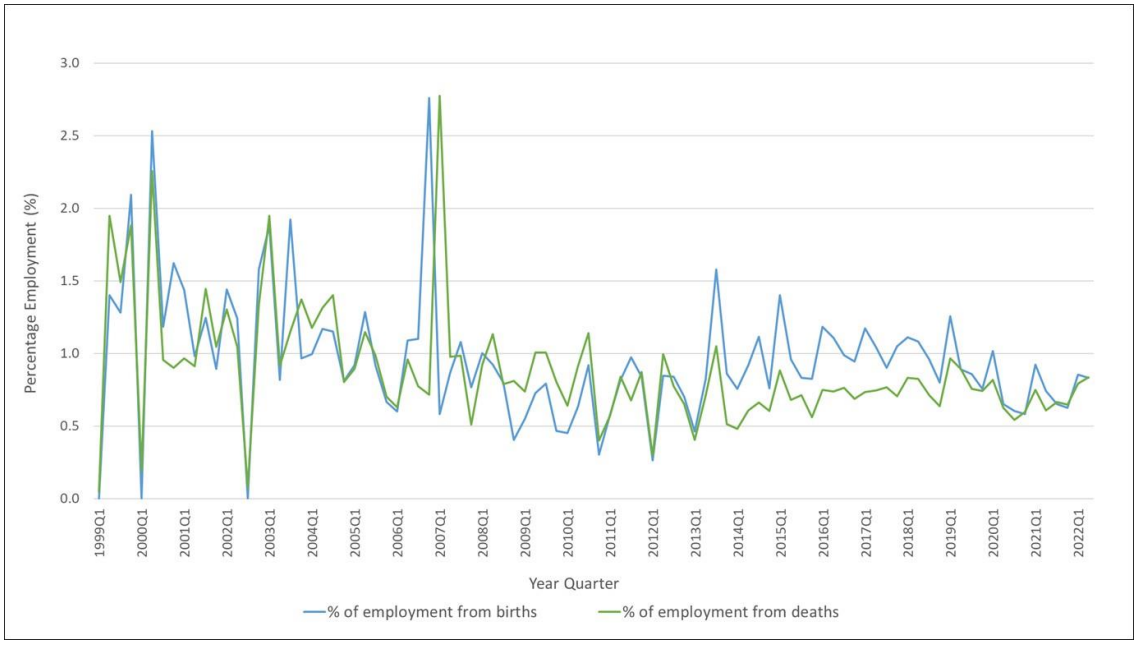


Figure 15: Quarterly percentage of employment from births and deaths



average large firm employs 1,700 workers, whereas the mean micro firm employs just over 2 workers. Looking at Figure 17, large firms employ 55% of the total LBD active workforce with micro firms employing only 18%.

Figure 16: Active LBD enterprises by firm size

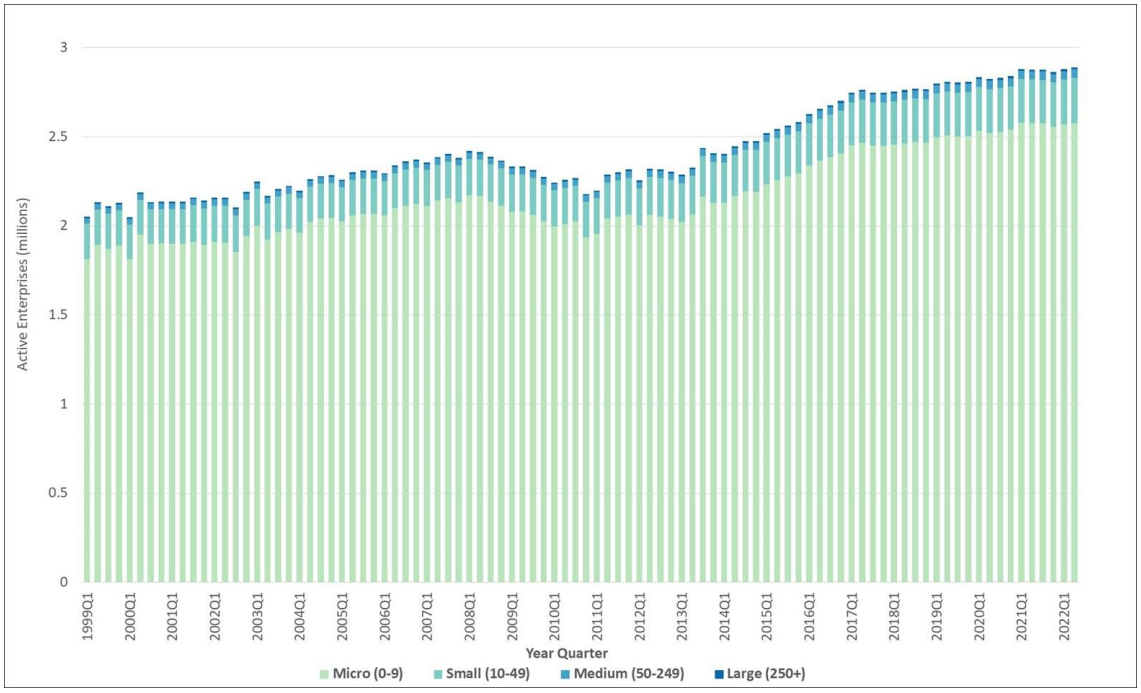


Figure 17: Active LBD workforce by firm size

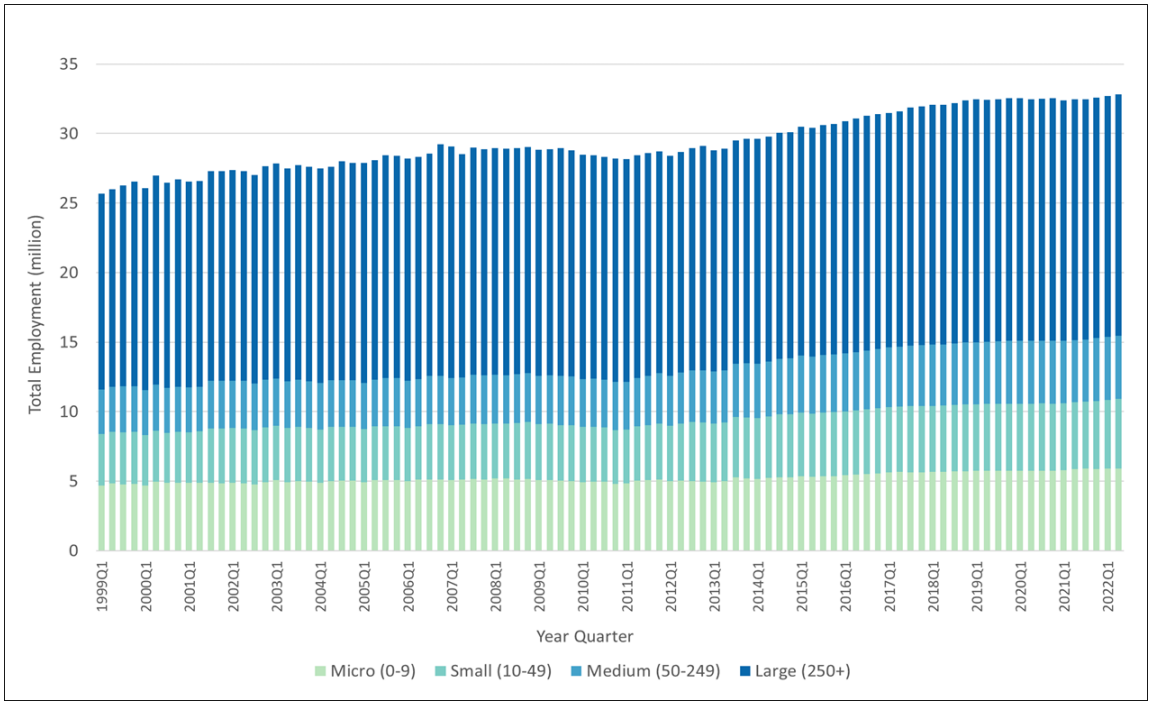
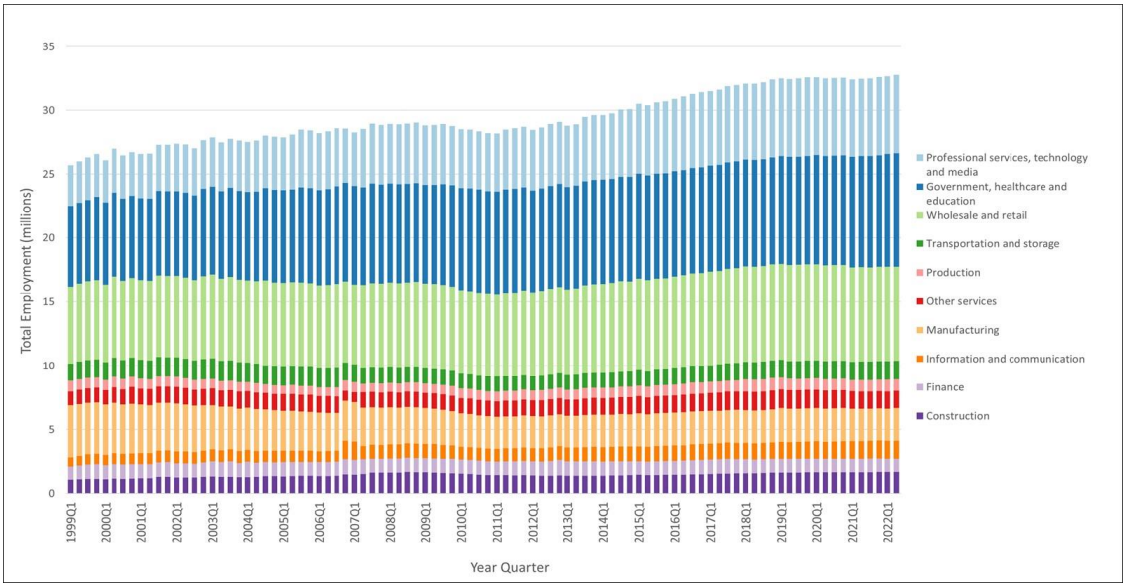


Figure 18 decomposes employment in the LBD by industry. The largest industry in employment is government healthcare and education, employing 26% of the workforce. This is followed closely by wholesale and retail, employing 23%. The smallest industries by employment are production,

finance, and information & communication each employing less than 4% of the total active workforce from the LBD. Between 1999 Q1 and 2022 Q1 professional services, technology & media has experienced the greatest growth with employees increasing 94% over the period. By contrast, manufacturing is the only industry that has seen a decrease in employees, shedding 38% of its workforce since 1999. There is a small inconsistency between quarters before 2006 Q3 and after, which is caused by a change in industry classifications from the older SIC92 standard to the newer SIC2007.

Figure 18: Active LBD workforce by industry



6.4 Comparisons to regular ONS business demography

The ONS already publishes annual and quarterly Business Demography (BD) statistics. The annual series is designated a National Statistic. Like the LBD, the Business Demography statistics use the IDBR as a data source. While both datasets use the same data source, the LBD and BD have methodological differences when categorising businesses as active, a birth or a death. The LBD uses a combination of key variables and administrative units to categorises businesses, whereas BD only uses administrative units. For this reason, we expect to see differences between the two datasets. This subsection explores these differences in more detail.

Figure 19 compares unique enterprise births and deaths between the LBD and BD. The two datasets produce similar trends, with similar peaks and troughs. However, one of the major differences is that the LBD sees consistently more enterprise deaths per year than the BD, whereas births are very similar throughout the five-year time series. This difference in enterprise deaths occurs due to reactivations after death. To account for this, BD releases estimate the number of reactivations and subtracts this estimate from the death counts (ONS, 2021, 2022). The last two

years of the publication are thus considered provisional and subject to revision as more enterprises reactivate. The LBD does not remove reactivations from the death count. Instead it treats death and reactivation as different activity types of the enterprise. To check, we have compared BD deaths unadjusted for reactivation with LBD business deaths. The two series are similar.

Figure 19: Birth and death comparison for LBD and BD

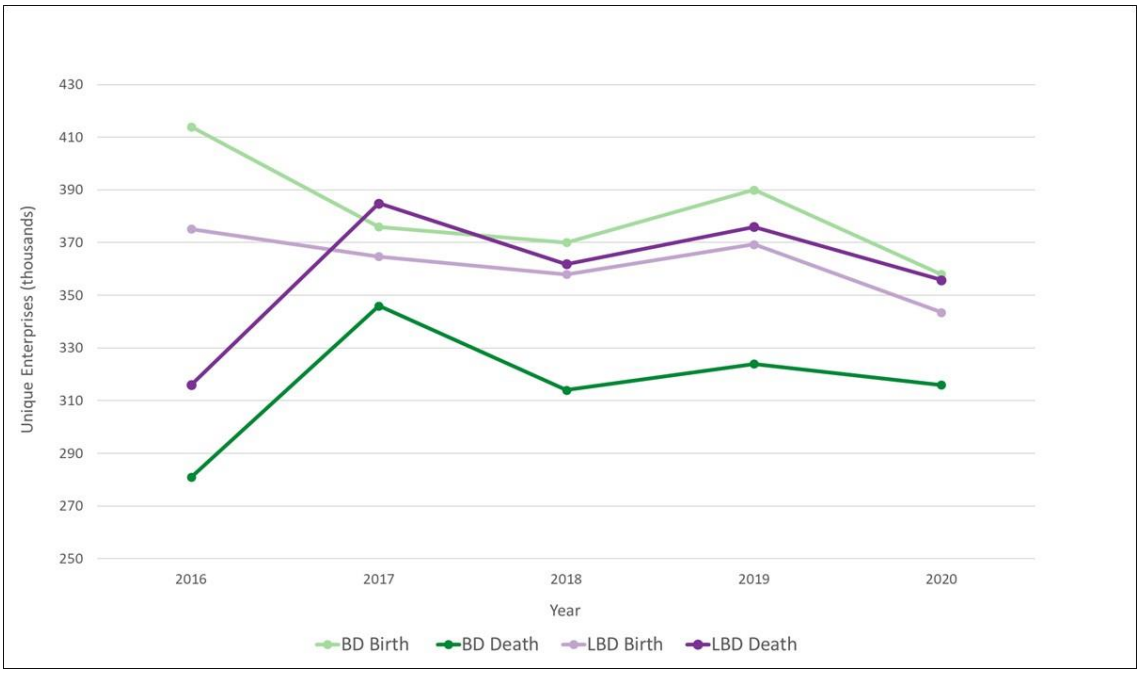


Figure 20 shows an industry breakdown of births and deaths at the enterprise level in 2020, for the LBD and BD data. Once again, the lighter colours represent the LBD data and darker colours represent the BD data. Figure 20 shows that the LBD and BD have very similar birth and death counts for the majority of industries. We see the largest differences in wholesale and retail, where the LBD has substantially more births and deaths.

7 Discussion

High-quality, comprehensive microdata is increasingly crucial for furthering our understanding of economic phenomena. The UK Longitudinal Business Database is an attempt to transform the existing UK firm-level data infrastructure to provide researchers with exactly that.

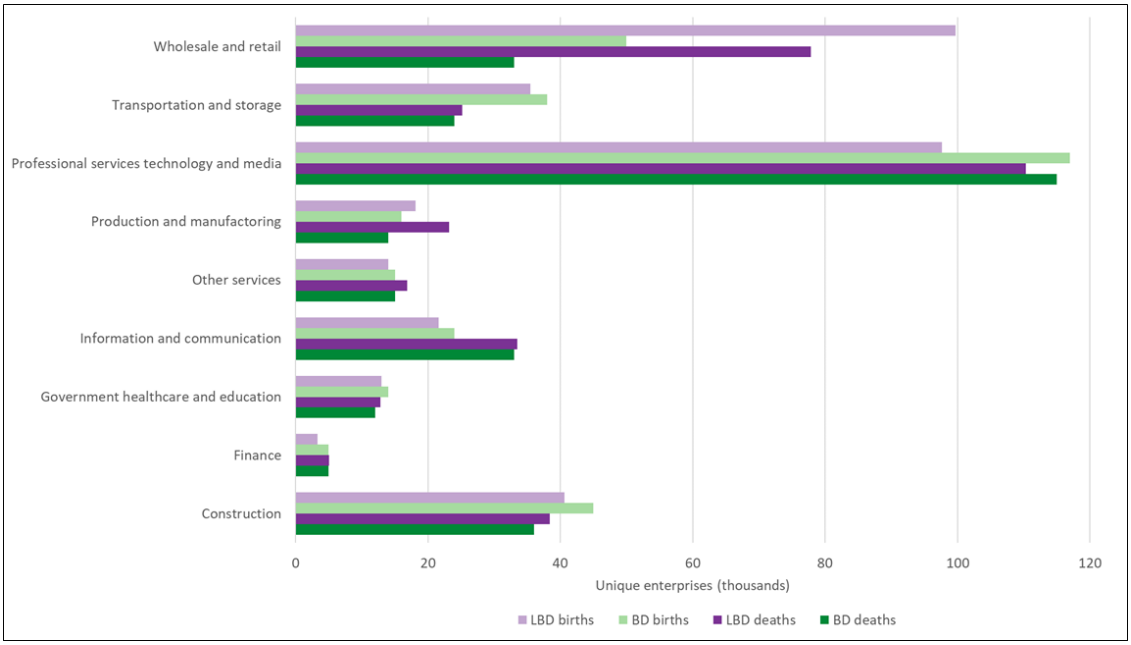
In this paper, we have explained the underlying Inter-Departmental Business Register data that serves as input for our core longitudinal spines. We have described the coverage of the data, the variables it provides and some of the drawbacks. We have then provided a description of the construction of the spines. We showed how we used two consecutive snapshots to build each quarter,

how we linked each unit, how we reconstructed the business structure from these longitudinally linked units and finally which economic activity criteria we use to filter active units.

There are obvious similarities to existing IDBR-based datasets, but we highlight several distinct advantages of the LBD: first, we use two adjacent snapshots to establish true longitudinality; second, we use unit-specific filters to build a clearer picture of the active business population for all units of the business structure; finally, we offer linked firm-level data for the first time at quarterly intervals, an ideal interval to study firm entry, growth and exit at business-cycle frequencies.

The LBD of course also inherits some drawbacks from its input data, the IDBR. First, snapshots are missing for some of the earlier quarters (and do not exist before 1997). Second, some of the economic activity criteria at present rely on timeliness and accuracy of the IDBR employment and turnover data. Where there are lags, we expect some revisions to the LBD as better data becomes available for use.

Figure 20: Industry births and deaths comparisons for LBD and BD



This paper also provides some initial descriptive results using the LWERL spine of the LBD: we show that overall patterns track the ONS annual business demography results quite closely and are in line with the quarterly business demography results (which rely on a different, less accurate methodology). We also present breakdowns of firm size at quarterly frequencies. We show the impact of the Great Financial Crisis on employment and firm entry and exit, and some differences in firm churn between industries.

While the current spines represent a significant step forward for tracking and analysing decisions and outcomes over a firm's lifecycle, the LBD remains work in progress. Efforts are ongoing to complete the spines relating to the administrative units (LWEP, LWEV and LWEC in our terminology). Other plans in the short term include improvements to the existing address data.

Beyond the short term, we believe the LBD forms an ideal platform to bring administrative data (starting with VAT, PAYE and corporation tax) to the firm level, using survey data to calibrate the administrative data where possible. This would enable researchers to compute firm-level time series of labour productivity, employment and investment for all businesses on the IDBR. Similarly, the LWEC spine currently under construction is an ideal conduit for bringing balance sheet data from Companies House to the active business population. In the meantime, we hope our first dataset built from the UK Longitudinal Business Database provides researchers with a new tool to answer open questions on firm entry and exit, business lifecycles and determinants of productivity and economic growth, 1999-2022.

References

W. Abel, S. Tenreyro, and G. Thwaites. Monopsony in the uk. *Available at SSRN 3270944*, 2018.

Z. J. Acs and P. Mueller. Employment effects of business dynamics: Mice, gazelles and elephants. *Small Business Economics*, 30(1):85–100, 2008.

P. Antràs, E. Fadeev, T. C. Fort, and F. Tintelnot. Global sourcing and multinational activity: A unified approach. Technical report, National Bureau of Economic Research, 2022.

BEIS. Business population estimates for the uk and regions 2022: statistical release. Technical report, 2022.

A. B. Bernard, S. J. Redding, and P. K. Schott. Multiproduct firms and trade liberalization. *The Quarterly journal of economics*, 126(3):1271–1318, 2011.

R. Black. Total factor productivity from the annual business survey. *Economic Statistics Centre of Excellence Discussion Papers*, TBD, 2022.

R. Black, S. Lui, and P. Wales. The potential of administrative data linkage for economic policy in the uk: Experimental results. Presentation at the ESCoE Conference on Economic Measurement, May 2019.

M. C. Chow, T. C. Fort, C. Goetz, N. Goldschlag, J. Lawrence, E. R. Perlman, M. Stinson, and T. K. White. Redesigning the longitudinal business database. Technical report, National Bureau of Economic Research, 2021.

J. Du and K. Bonner. Decomposing uk aggregate labour productivity and growth:: 1998-2013 using the ons business structure database data. 2016.

Eurostat. Business registers recommendations manual. Technical report, 2010.

P. Evans and M. K. Nadeem Khan. The characteristics of patenters. *Economic & Labour Market Review*, 3(12):20–27, 2009.

P. Evans and R. Welpton. Methods explained-business structure database. *Economic & Labour Market Review*, 3(6):71–75, 2009.

R. Fabling and L. Sanderson. A rough guide to new zealand's longitudinal business database. 2016.

A. Falato, H. Kim, and T. M. von Wachter. Shareholder power and the decline of labor. Technical report, National Bureau of Economic Research, 2022.

T. C. Fort, J. Haltiwanger, R. S. Jarmin, and J. Miranda. How firms respond to business cycles: The role of firm age and firm size. *IMF Economic Review*, 61(3):520–559, 2013.

J. Haltiwanger, R. S. Jarmin, and J. Miranda. Who creates jobs? small versus large versus young. *Review of Economics and Statistics*, 95(2):347–361, 2013.

R. S. Jarmin and J. Miranda. The longitudinal business database. *Available at SSRN 2128793*, 2002.

S. Lui, R. Black, J. Lavandero-Masson, and M. Shafat. Business dynamism in the uk: New findings using a novel dataset. *Economic Statistics Centre of Excellence Discussion Papers*, DP-2020-14, 2020.

ONS. Business structure database user guide. Technical report, 2006.

ONS. Business demography, uk: 2021. Technical report, 2021.

ONS. Uk business; activity, size and location qmi. Technical report, 2022.

D. Pieterse, E. Gavin, and C. F. Kreuser. Introduction to the south african revenue service and national treasury firm-level panel. *South African Journal of Economics*, 86:6–39, 2018.

Appendix

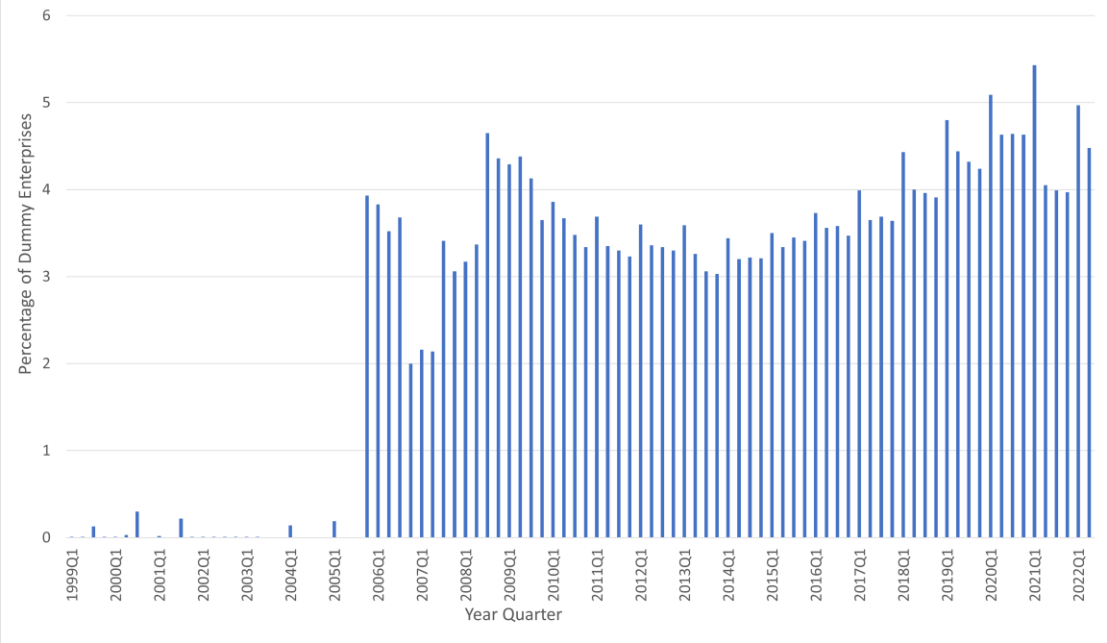
A. Dummy Enterprises

Dummy enterprises are created when there are no live administrative units, VAT and/or PAYE references linked to an enterprise, but the company associated with the enterprise continues to operate under an active enterprise group. These enterprises are created to maintain the links between the different units of a business in the IDBR. A marker *dummyent* has been created in the LWERL prototype, which will enable researchers to decide on whether to include or exclude such enterprises from their analyses.

The employment and turnover for these enterprises are populated based on the Standard Industrial Classification code (SIC) associated with the dummy enterprise. For holding companies (SIC codes 64201 – 64209), employment and turnover are set to 0. For trading enterprises (all other SIC codes), employment is set to 1. The turnover is calculated using a turnover per head ratio across the SICs using businesses that hold true turnover and employment. The ratio is updated annually.

Figure 16 shows the proportion of dummy enterprises in the LWERL. The above figure shows a spike in the proportion of dummy enterprises from 2005 Q4. This spike was due to the introduction of the DEFRA data take-on. This began the creation of dummy enterprises where DEFRA had indicated farming activity, but the IDBR did not have a matching legal unit for the site. This process became standard from November 2005; therefore, we see more stability in dummy enterprise numbers from 2005 Q4 onwards.

Figure 21: Percentage of dummy enterprises in the LWERL spine, by quarter



B. Enterprise Groups

As stated in ECC No.696/93, an enterprise group represents the highest level of the statistical unit hierarchy and can be used to identify enterprises connected through legal and/or financial links. These links often refer to business activity linked by ownership, controlling interest or management.

As linkages often extend outside of the UK, knowledge of international ownership trees is needed to inform ultimate ownership and therefore connected ownership of lower-level activity. The ONS utilise data from Dunn and Bradstreet (D & B), an external provider, who provide information on UK subtrees.

Information on UK subtrees is received from D & B every January and uses the Companies House identifier which the ONS then link to enterprise reference and other statistical units using administrative units such as VAT or PAYE references. The ONS receive information more regularly from HMRC and Companies House which is used to further inform linkages between business activity outside of the yearly D & B update.

Given the source, enterprise groups only accurately reflect ownership linkages for incorporated businesses. Additionally, the IDBR does not allow tracking of quarterly changes to ownership structures in the same way it does for lower levels of the business structure. Further revisions of the Longitudinal Business Database may be able to address these limitations.

C. List of variables

Variable Name	Variable Description
year	Year of IDBR snapshot
quarter	Quarter of IDBR snapshot
entref	IDBR enterprise reference
ruref	IDBR reporting unit reference
wowref	IDBR enterprise group reference
luref	IDBR local unit reference. 8 digits + checkletter
wow_ultfoc	Ultimate Parent reference provided by Dun & Bradstreet
ent_employment	Enterprise employment
ent_legalstatus	1-digit code representing the legal status of the business
ent_sic2007	Enterprise SIC 2007 code
ent_sic92	Enterprise SIC 1992 code
ent_turnover	Enterprise turnover reported for a business in thousands
ent_active_type	LBD activity marker for enterprises
ru_employment	Reporting unit employment
ru_region	Reporting unit Region
ru_rusic2007	Reporting unit SIC 2007 code
ru_rusic92	Reporting unit SIC 2003 code
ru_turnover	Reporting unit turnover
lu_employment	Local unit Employment
lu_region	Local unit region
lu_sic2007	Local unit SIC 2007 code
lu_sic92	Local unit SIC 2003 code
ru_active_type	LBD activity marker for reporting units
lu_active_type	LBD activity marker for local units
dummy_ent	A dummy enterprise indicator variable

D. The US Longitudinal Business Database

The US Longitudinal Business Database, initially described in Jarmin and Miranda (2002) and more recently in Chow et al. (2021), is rightfully one of the most used business datasets in history. The US LBD is a major improvement of its predecessor, the Longitudinal Research Database (LRD), which used plant level data from the Census and Annual Surveys of Manufacturers as its input. As a result, the LRD only covered the manufacturing sector. The US LBD on the other hand covers nearly all non-farm private economy as well as public-sector activities.

The US LBD contains information going back to 1975, but its annual frequency limits its usefulness for research at business-cycle frequencies. Since the US LBD covers roughly 50 years and multiple data collection and storage systems to date, the production process involves a series of steps to prepare the input data, perform data linkage and apply corrections to the linked dataset.

Different components of the US LBD are sourced from a variety of sources, including the Standard Statistical Establishment List (SSEL – also known as the Business Register), Economic Censuses and surveys. The SSEL provides a mail-out frame for the quinquennial Economic Census and is also used as sampling frame for economic surveys. The construction of longitudinal linkages is therefore underpinned by the SSEL.

The first step in the production process is to standardize variables to ensure consistency of data over time. This involves using annual surveys to apply edits to important variables such as geography, industry, and employment. In addition to standardising variables, this step also entails integrating accurate historical files. Given the long time span, it is not uncommon to face missingness and data quality issues, especially in earlier years. In the case of the SSEL, these issues are especially pertinent during the 1970s to 1980s. To fix these issues, the US LBD relies on a combination of newly discovered (at the time) survey microdata to fill gaps and imputations based on prior years where this is not possible.

The second step in the production process is year-to-year matching. This involves using cross-sectional snapshots of the business register and county patterns microdata (CBPBR) to apply edits to important variables such as geography, industry, and employment. Subsequently, longitudinal links are created. This process starts with the 1976-77 year pair and adds subsequent year pairs chronologically thereafter. This process does not distinguish between birth and reactivation of an establishment and so checks for activity going back up to seven years are conducted to correct for this.

The SSEL is continuously updated with administrative and survey data. As a result, it covers different statistical units, which brings about alignment challenges. This is especially acute for the

accurate classification of new entrants, either as a stand-alone entity or part of multi-unit establishments. Increased resource is allocated to updating the SSEL before and after the Economic Census. As a result, data quality declines in the intervening periods. The US Census Bureau conducts annual surveys collecting data on organisational structure of firms, to supplement the information in the intervening years. The survey, however, only covers relatively large firms. As a result, many firms fall outside the scope and any structural changes as well as other critical variables for these firms are only updated in census years.

This updating pattern introduces a spurious pattern of spikes and troughs in the number of births and deaths of establishments. Despite the use of imputation models to retime some of the birth and death, researchers using the US LBD will have to contend with inconsistent breakdown and the attendant complexities. Given its input sources, the UK LBD on the other hand does not face this issue.