Matching firms engaged in publicly funded training in the Inter Departmental Business Register

Gavan Conlon, Pietro Patrignani, Daniel Herr and Sophie Hedges

Technical Report for Briefing Note 003

March 2017

CENTRE for ECONOMIC PERFORMANCE
ies institute for employment studies
LE London Economics
The University Of Sheffield.

Executive Summary

As part of a recent project estimating the impact of publicly funded training on industry and firm-level outcomes,[1] London Economics undertook a matching exercise aiming to identify firms engaged in publicly funded training (identified in the **Individualised Learner Record** (ILR) and **Employer Data Service** (EDS) data) with the corresponding enterprise level information contained in the **Inter Departmental Business Register** (IDBR), which is the live register of enterprises in the UK.

Although the data matching was relatively successful (successfully matching approximately 70% of entities across the data sets), possible methodological improvements were identified in the exercise, and it was decided to repeat the process incorporating these. This technical report presents the details of the matching strategy; the extent to which the matching strategy was successful; and headline information on the characteristics of the firms identified as engaged in publicly funded training.

Having implemented the various matching stages, the proportion of EDS entities (**256,394** in total) that were matched with the IDBR was estimated to be **216,931** (corresponding to **84.6%**) with **39,463** unmatched (**15.4%**)[2]. This represents a significant improvement on previous attempts, and although smaller training firms are under-represented in the data, ahead of the proposed introduction of the Apprenticeship Levy in May 2017, this dataset will allow for (for instance) a detailed understanding of the baseline incidence of training (especially amongst larger firms). Furthermore, this data set will facilitate researchers and policy makers' understanding and assessment of post-Levy outcomes at both an aggregate and disaggregated level; the impact of the Levy on Levied-enterprises and non-Levied enterprises, as well as the composition of employees receiving training.

[1] "Estimating the impact of publicly funded training on industry and firm-level outcomes" (May 2016), BIS Research Report 177. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/522105/bis-16-255-impact-publicly-funded-training-on-industry-outcomes.pdf

[2] The matching exercise described here was undertaken with partial ILR data for 2014/15. The revised matching exercise on the full 2014/15 ILR data (308,150 EDS entities) achieved an 82.5% match rate.

1. Introduction

As part of the project "Estimating the impact of publicly funded training on industry and firm-level outcomes" (BIS research paper 177[3]), London Economics undertook a matching exercise aiming to identify firms engaged in publicly funded training (identified in the **Individualised Learner Record** (ILR) and **Employer Data Service** (EDS) data) with the corresponding enterprise level information contained in the **Inter Departmental Business Register** (IDBR), which is the live register of enterprises in the UK. Subsequently, the Department commissioned a revision and improvement of the matching process and this technical report presents the details of the matching strategy; the extent to which the matching strategy was successful; and headline information on the characteristics of the firms identified as engaged in publicly funded training.

The matching exercise involved merging the 2014/15 ILR with details on training at the individual level and an employer identifier for training undertaken through the employer. The SFA commissioned a third party provider (Blue Sheep) to collect information on the employers engaging in publicly funded training. This firm level information is based on a variety of sources and provided in a database called "Employer Data Service". Using the firm level characteristics available in the EDS, it is possible to match with the IDBR, which is the official source of information for businesses in the UK and allows for further linking to ONS' surveys. Each stage of the process was subject to a range of quality assurance activities and information on the stage at which the exercise was successful was retained for each matched observation. The final outputs of the matching process consist of a lookup dataset containing the **EDS identifier** linked to the enterprise and, when available, **local unit identifiers** in the IDBR; as well as a dataset with information on ILR training undertaken at the enterprise and local unit level. The unit of observation in this analysis is the firm or enterprise engaged in publicly funded training.

2. Data description

The analysis presented here is based on the following data sets:

- The **Individualised Learner Record (ILR)** contains detailed information on the course and characteristics for publicly funded Further Education (FE) courses, as well as the characteristics of individual learners. The information is supplied by learning providers throughout the Further Education system. The ILR is organised by academic year (1st August – 31st July) and the data specification varies to some extent from year to year. ILR data is collected from providers that are in receipt of funding from the Skills Funding Agency (SFA), the Education Funding Agency (EFA) and co-financed European Social Funds (ESF). For training undertaken through the employer, an employer identifier (**A44**) is attached to the dataset in order to identify the organisation engaging in publicly funded training. The current study focuses on **training undertaken through the employer** (including Apprenticeships) and uses ILR data for

[3] "Estimating the impact of publicly funded training on industry and firm-level outcomes" (May 2016), BIS Research Report 177. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/522105/bis-16-255-impact-publicly-funded-training-on-industry-outcomes.pdf

the academic year 2014/15 (the previous matching exercise focused on the years from 2010/11 to 2013/14).

- The **Employer Data Service (EDS)** (or 'Blue Sheep') data is a composite database containing information on the characteristics of UK firms (or sites within firms, hence 'business entities')[4]. The constituent information is collected from a range of different sources (e.g. Thomson Directories, Companies House, Dun and Bradstreet etc.). However, in the case of evidence gaps, the EDS data may also contain information and data that is directly sourced from the EDS Help Desk (which could be potentially provided by the particular entities engaged in training activities). The EDS contains information on entity characteristics including:

  - Company name and trading name,
  - Entity's postcode,
  - Number of employees[5],
  - Turnover at site and group level,
  - Sector of activity as defined by SIC code,
  - Year of foundation,
  - Company Registration Number (CRN), where available, and
  - A range of other entity-level characteristics.

  The most recent version of the EDS data extract received by London Economics in May 2016 contained information on more than **21 million entities** (not all of them identifying live entities).

- The **Inter Departmental Business Register (IDBR)** is the comprehensive list of UK businesses that is used by Government for statistical purposes and provides the main sampling frame for surveys of businesses carried out by the Office for National Statistics and other Government departments.

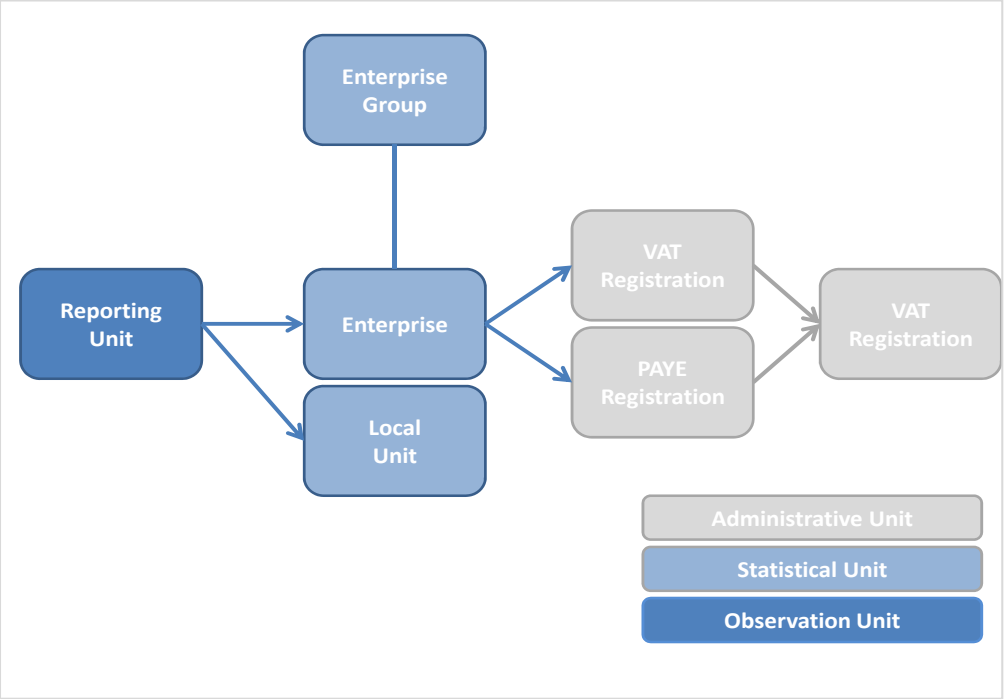
  The IDBR is organised in different datasets, and information is updated **each quarter**. The **enterprise unit** is at the centre of the IDBR classification, and all other units (i.e. local unit) can be linked to the enterprise. Using this database structure (see Figure 1), it is possible to distinguish between:

  - Statistical units (Enterprise, Local Units and Enterprise Group);
  - Administrative units (VAT and PAYE units also containing the Company Registration Number);
  - Observation unit (Reporting Unit);

[4] In general, to distinguish EDS entities from the various IBDR units (enterprises, local units etc.), we refer to "entities" throughout the document to identify EDS 'firms', and refer to "enterprises" (or other relevant aggregation) when discussing IDBR units. When comparing information for matched entities across data sources, IDBR data may refer to EDS entities (labelled "IDBR – entity level"), which involves duplication on the IDBR side, as multiple EDS entities may correspond to the same enterprise or to the enterprise unit (labelled "IDBR – enterprise level"), with no duplication on the IDBR side. See Figure 1

[5] Last available year and two prior years

Figure 1: Structure of the IDBR



Source: http://www.ons.gov.uk/ons/about-ons/products-and-services/idbr/index.html

- The IDBR is a key data source for analyses of business activity. It covers around **2.7 million live enterprises** in all sectors of the UK economy, other than some very small businesses (those without employees, and with turnover below the relevant tax threshold) and some non-profit making organisations. The IDBR also reports information on over 5.7 million non-live enterprises. The information used in the IDBR is obtained from the following main sources:

  - HMRC VAT - Traders registered for VAT purposes with HMRC;
  - HMRC PAYE - Employers operating a PAYE scheme, registered with the HMRC;
  - Companies House - Incorporated businesses registered at Companies House;
  - Dun and Bradstreet for Enterprise Group information;
  - The Business Register and Employment Survey (BRES) and other ONS surveys.

  Information contained in the IDBR includes:

  - Company name and trading name;
  - Address including postcode;
  - Unit (e.g. enterprise) 'birth' date;
  - Unit (e.g. enterprise) 'death' date;
  - Standard Industrial Classification (UK SIC 2007 and UK SIC 2003);
  - Employment and employees (updated from administrative sources (PAYE and VAT records) and ONS Surveys (Business Register Employment Survey))[6];
  - Turnover (updated via administrative sources (HMRC VAT and PAYE records) and ONS Business Surveys (Annual Business Survey));

[6] For more information on data sources for employment and turnover and the updating frequency see https://www.ons.gov.uk/ons/guide-method/method-quality/specific/business-and-energy/business-population/further-information-about-idbr-sources.pdf

- Legal status (company, sole proprietor, partnership, public corporation/nationalised body, Local Authority or non-profit body);
- Enterprise group links;
- Country of ownership;
- Company Registration Number.

Current data set (2014/15)

We attempted to match the 2014/15 ILR/EDS data[7] with the IDBR for companies engaging in publicly funded training. The match was undertaken between the 2014/15 ILR/EDS data and the IDBR 2015 Q3 (third quarter).

Overall, there were **256,394** entities undertaking publicly funded training in 2014/15[8]. Around **85%** of these entities (**216,931**) were matched in the IDBR. This is a significant improvement on previous attempts to match the data sets (the original matching attempt by the ONS had a match rate of around 50%[9], while a more recent matching exercise by London Economics (2016) achieved a match rate around 70%[10]).

As planned in the preliminary stages of the work, we also undertook the following steps to understand the quality of the match, assess the extent of mismatches and identify possible adjustments to the approach:

- A manual review of a sample of matched cases for each stage of the process;
- Additional tweaks in the matching algorithms to capture extra matches;
- Using a further data source on company information (FAME database) to see whether it was possible to identify some of the entities left unmatched; and
- A manual review of cases left unmatched to understand the possible reasons

The improved approach was then also applied to earlier years to generate a consistent time series for the period 2010/11 to 2014/2015. Compared to the previous matching approach, all original steps meeting the quality criteria were retained, and a number of new stages were introduced (based on name similarity, SIC code and postcode, and involving probabilistic matching).

[7] For firms to receive the relevant public funding relating to training undertaken in the firm, they must provide a number of details relating to the basic characteristics of the firm and where the training takes place (predominantly). Therefore, for all incidences of publicly funded training, there is information on the entity involved; however, information is regularly partial and not fully accurate (for instance the provision of a trading name rather than registered company name), thereby making the matching exercise between the ILR/EDS and IDBR the challenging component of the analysis.

[8] These figures refer to partial 14/15 data. The full 14/15 ILR dataset contains 308,150 EDS entities undertaking publicly funded training. 254,000 of those have been identified in the IDBR (82.5%).

[9] See page 23 of "Estimating the Impact of Training on Productivity using Firm-level Data", (May 2012), BIS Research Paper 72. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32303/12-766-estimating-impact-training-on-productivity.pdf

[10] See page 13 of "Estimating the impact of publicly funded training on industry and firm-level outcomes" (May 2016), BIS Research Report 177. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/522105/bis-16-255-impact-publicly-funded-training-on-industry-outcomes.pdf

3. Overview of methodology

The matching strategy was based on developing an approach to identify those entities engaging in publicly funded training (as defined by the EDS) within the IDBR (at the enterprise and local level). The matching approach can be split into four different macro phases, as summarised in **Table 1**. Each phase consisted of two or more stages and the validity of each stage was assessed in the quality assurance process. Below we outline the matching approach. All figures refer to entities with at least one training record in the 2014/15 ILR matched to the September 2015 IDBR extract.

Table 1: Overview of methodology

Phase	Description	Stages
1	Company CRN; Parent CRN	1 to 2
2	Company name / trading name and postcode/postcode district	3 to 6
3	Fuzzy matching	7 to 20
4	Quality assurance	Applied to all stages

1) **Matching the ILR-EDS to the IDBR using Company Registration Number** and parent/ultimate CRN (CRN is available for around 35% of the EDS/ILR entities and in the IDBR);

2) **Matching the ILR-EDS to the IDBR using company/trading name and various variations of address and postcode** (using information on company details available in the EDS dataset and the IDBR "ADDRESS" file). This stage consisted of several steps and was undertaken after standardising company names in both datasets (e.g. ensuring 'Ltd' recoded as 'Limited').

3) **Fuzzy matching the ILR-EDS to the IDBR using entity characteristics** including combinations of

   a) Company name[11]/ trading name (with and without vowels)
   b) First seven letters of company name/trading name
   c) First word of company name/trading name
   d) Full Postcode, Postcode District, and Postcode Area
   e) SIC Code (2003) and/or SIC Code (2007)
   f) Name similarity[12]
   g) Third or fourth word of company name[13]
   h) Address similarity[14]
   i) Legal status[15]
   j) *Reverse* company name (e.g. first seven letters of the reverse name)
   k) Name of EDS company and address of IDBR entities

[11] Match based on company name only when only one *entref* is available in the IDBR (or one *entref* covering at least 80% of cases) - and company is not a partnership or unincorporated business (based on the definition in the EDS).
[12] The tolerance (i.e. number of different characters) was based on the overall number of characters in the name: no differences in names up to 7 characters; 1 digit for strings of 8-10 characters; 2 digits for strings of 11-14 characters and 3 digits for longer strings (string length is always computed on the shorter string).
[13] Whether at least one other word (excluding the first two) matched across the two strings.
[14] Address similarity is based on building number and whether the other strings of the EDS address (sub_building_name, building_name, primary_thoroughfare_name) match the IDBR address.
[15] Legal status checks exclude from the match any EDS entities matched to IDBR entities classified as "Sole Proprietors" or "Partnership".

4) **Quality assurance:** This stage consisted of evaluating the quality of matched EDS entities for each stage of the process. Random samples of 60 matched entities were checked for the accuracy of the match for each stage of the matching process. The procedure also identified and evaluated unmatched entities (i.e. those that could belong to a specific category, but were then left unmatched due to validity checks). Quality assurance was undertaken as an iterative process until the final matching methodology was defined. For cases left unmatched at the end of the process we undertook a manual review of 100 unmatched cases, looking for possible matches in the FAME database and for potential matches in the IDBR (not retained during the matching process).

3.1 Matching process

This section describes the matching procedure in greater detail. All matching stages[16] were undertaken on both live and non-live IDBR units.

Phase 1: Match using Company Registration Number

Around 89,200 EDS-ILR entities had a valid Company Registration Number (corresponding to approximately **35%** of the total **256,400** EDS-ILR entries)[17]. In the IDBR, information on CRN is available at the enterprise-level through the links with the PAYE and VAT units.

Thus, the first step consisted of undertaking a direct match based on Company Registration Number for the EDS-ILR entries with a valid CRN. The EDS contains information (when available) not only on Company Registration Number, but also on parent CRN and ultimate CRN (i.e. the CRN of the parent and ultimate parent company) and we used the different levels of CRN information in the first matching stage. **Overall, in the IDBR, we managed to identify slightly more than 79,000 EDS-ILR entities based on Company Registration Number** (the vast majority through the entities' own CRN, while the parent or ultimate CRN accounted for around **800** matches), corresponding to about **32%** of all EDS-ILR entries (and **89%** of entries with a valid CRN). After the first stage, there were around **175,000** EDS-ILR entities still unmatched (approximately **69%** of the total).

Before undertaking the second and third phases, company and trading names across both the EDS and IDBR datasets were standardised as follows:

- Punctuations, special characters, company type (e.g. ltd, plc) were removed, and various replacements ("&" with "and", "United Kingdom" with "UK" etc.) were implemented;
- Company names were then reformatted as one lower case string concatenating all words in the name and removing all spaces;
- Postcodes were also reformatted and postcode area was identified

[16] With the exception of stage 19, which was undertaken on live IDBR records only. Stage 19 is based on a probabilistic approach and is highly computationally intensive and only yielded slightly more than 100 matches.
[17] This proportion drops to 30% in the final 14/15 ILR-EDS dataset.

Phase 2: Match using full company/trading name and postcode

In this part, we used information on company details to match EDS-ILR entities in the IDBR. The first step in this stage was based on matching using **full company** or **trading name** and various expressions of postcode: **full postcode** (e.g. AB1 1BA), and **postcode district** (e.g. AB1)[18].

Overall, we were able to identify a further **78,800 EDS-ILR entities** in this step (**31%** of the total), with the vast majority using the full postcode along with full company name or full trading name:

- **56,075** (**22%** of the total) using full company name and various postcode aggregations (**48,450** of these using the full postcode);
- A further **22,750** (**9%**) using the full trading name and various postcode aggregations (**19,930** of these using the full postcode);

After this procedure, slightly less than **100,000** EDS-ILR entities (**38%** of the total) were still unmatched.

Phase 3: Fuzzy match using expressions of company name and postcode

With the fuzzy matching approach, we tried to identify EDS entities in the IDBR based on expressions of company name (i.e. to match companies having a certain degree of similarity) *and* full postcode[19]. The process involved several combinations of these expressions.

The objective was to simultaneously maximise the number of valid matches and to minimise the number of false matches, as there was a clear trade-off between these two goals and mismatches were likely to occur.

In general, when using fuzzy matching approaches, the main issues are typically associated with:

- Companies having different names in the two sources: for example “Langdon Care Home” and “Langdon Residential Care” or “Shelton Infant School” and “Shelton Nursery”. The strings based on the full name may be quite different;
- Companies having part of the name identifying the geographical area: for example “Somerset Café” and “Somerset Care” share 11 of 12 characters, but are clearly different companies;
- Small businesses bearing the owner’s name may be difficult to identify (i.e. multiple matches may be possible or it may be difficult to understand whether two businesses are actually the same).

The fuzzy matching strategy stages can be grouped in the following categories:

[18] When matching on postcode district and postcode area, we set restrictions when multiple IDBR matches with different enterprise reference numbers were available in the same postcode aggregation.
[19] While we also attempted to use postcode district in the fuzzy matching process, the resulting matches did not meet the quality criteria.

- **Stages 7-10**: these stages are based on expressions of company and trading name (first seven letters of company/trading name, first word of company name and company name after all vowels have been removed[20]) with postcode;
- **Stage 11** matches EDS company/trading name with IDBR address (based on the first seven letters): many entities can be recorded in the EDS with the name of the location (e.g. Shelton Road Nursery) and in the IDBR under a different name (e.g. Little Angels Nursery). All matches were conditioned on the two entries also sharing the same postcode and 2 digit SIC code;
- **Stages 12-15** are based on the 5-digit and 3-digit SIC 2007 and SIC 2003 codes and postcode, with checks on overall name similarity;
- **Stages 16-17** are based on the last seven letters of company name (starting from the last letter) and the last word of company name with postcode. As usual checks on SIC code and name similarity were carried out;
- **Stages 18** groups all entries based on postcode, then sorts on company name and looks for matches in nearby entries, applying the usual checks on name and SIC code;
- In **Stage 19** we performed **probabilistic matching** based on company name and postcode. This procedure involved a STATA-written package, *reclink2*[21] designed to produce a similarity score based on the selected variables. Use of such algorithm is highly computationally intensive and not entirely feasible in the context of large datasets such as the EDS-IDBR. For this reason, we resorted to using the command towards the end of the matching procedure. Postcode equality was required as a pre-condition to obtain a high quality of matching and to limit the computation time. This stage was only implemented when matching on live IDBR units and yielded a small number of matches due to the high threshold set on the similarity score.

Throughout the fuzzy matching process, the matched expressions were combined with an assessment of the overall similarity across company names (i.e. the number of different characters allowed to identify a valid match) and other validity checks to assess the quality of the match. In particular:

- We set a validity check based on the different number of characters across the two strings (depending on the overall length of the string);
- For longer strings, we looked at whether the third and fourth words were contained in the matching string: for example, “Langdon Care Home” and “Langdon Residential Care” would be matched thanks to the presence of the word “care” (when sharing the same postcode);
- When using the last 7 letters of the name, we removed the most common words (e.g. ‘services’, ‘consulting’) to ensure that the match would not be based on frequently used (and potentially misleading) names;
- We noticed that many schools (especially primary schools) in the EDS had no corresponding match based on company name at the same address but had a

[20] This was undertaken so that words such as “company” and “compeny” are identified by the same string.
[21] The module reclink2 was developed by Wasi and Flaaen (see The Stata Journal (2015) 15, Number 3, pp. 672–697 “Record linkage using Stata: Preprocessing, linking, and reviewing utilities”).

corresponding IDBR local unit entry identifying the local council with the same address and exact SIC code. Thus, it appears that schools may be recorded as "branches" of the local council in the IDBR, rather than with the actual school name. That raised the question of how to correctly identify these entries, especially as there are often cases where a nursery is located within the school complex (and has a similar name), but does not constitute part of the same "enterprise". Thus, when possible, entries in the EDS with SIC code 85200 (primary schools) or 85310 (secondary schools) were matched to IDBR entries with the same SIC code identifying the local authority.

- We used the availability of SIC 2007 and SIC 2003 codes in both the EDS and IDBR datasets to refine the match:
  - Strings such as "Shelton Infant School" and "Shelton Nursery" share a low overall level of similarity, but may have the same SIC code across the two data sources. Thus, we used SIC codes (mainly at the two digit level) to identify entries across the IDBR and EDS dataset. If "Shelton Infant School" and "Shelton Nursery" had the same postcode and both share the same two digit SIC code (e.g. 85), they were considered a valid match;
  - The limitations of this approach are missing SIC codes, SIC codes recorded differently in the two data sources, the presence of multiple SIC codes and the possibility that two companies sharing part of the name, postcode (or postcode district) and SIC Code are in fact different.

A further **57,000** EDS-ILR entities (slightly less than **23%** of the total) were matched based on the fuzzy matching strategy.

The last step concerned matching of enterprises belonging to groups:

- **Stage 20:** *Direct match based on company name alone (but not postcode);*

  Registered company names cannot be exactly the same as another registered company's name (or even too similar) unless they are part of the same company. We tried to exploit this rule and match directly on full company name (excluding postcode). However, the match was only undertaken when all (or the vast majority of) units in the IDBR having the same name also shared the same *entref.* For example, if there are 5 entries in the IDBR with full name "A1Telecom", and there is an unmatched EDS entry with the name "A1Telecom", the EDS entry will be associated to the IDBR *entref* for "A1Telecom" as long as all (or at least 80% of[22]) IDBR units with name "A1Telecom" have the same *entref.* We were able to identify a further **2,500** EDS-ILR entities with this approach (**1%** of the total).

At the end of the previous matching phases, there were around **39,000** EDS-ILR entities left unmatched (approximately **15%** of the total).

A summary of the matching stages – updated data set (2014/15)

Having implemented the various matching stages, Table 1 below illustrates the number of matches that were achieved in each stage, as well as the proportion of the EDS entities that were matched with the IDBR. The analysis indicates that of the **256,394** entities, **216,931** were

[22] To allow for the presence of other units (e.g. non-trading units) with same name but different *entref.*

matched (**85%**) with **39,463** unmatched (**15%**). Stages based on fuzzy matching also had additional checks as previously described.

Table 2: Summary of matching stages

Description	Stage	Number of Obs.	% Matched
	2014/15		
Company CRN	Stage 1	79,866	31.1%
Parent CRN	Stage 2	802	0.3%
Company name and postcode	Stage 3	48,448	18.9%
Trading name and postcode	Stage 4	19,931	7.8%
Company name and postcode district	Stage 5	7,627	3.0%
Trading name and postcode district	Stage 6	2,819	1.1%
First 7 letters of company/trading name, postcode and SIC code	Stage 7	21,862	8.5%
First 7 letters of company name/ trading name, and postcode	Stage 8	7,921	3.1%
Company name without vowels and postcode	Stage 9	291	0.1%
First word of company name/trading name and postcode	Stage 10	7,177	2.8%
First 7 letters of address (IDBR), company name postcode and SIC code	Stage 11	6,241	2.4%
Full SIC 2007 and postcode	Stage 12	4,572	1.8%
Full SIC 2003 and postcode	Stage 13	457	0.2%
3-digit SIC 2007 and postcode	Stage 14	894	0.3%
3-digit SIC 2003 and postcode	Stage 15	1,165	0.5%
Reverse first 7 letters of company name and postcode	Stage 16	1,734	0.7%
Last word of company name and postcode	Stage 17	733	0.3%
Postcode and company name similarity	Stage 18	1,797	0.7%
Probabilistic matching based on company name and postcode	Stage 19	118	0.0%
Company name (groups sharing same enterprise reference number)	Stage 20	2,476	1.0%
	Unmatched	39,463	15.4%
	Matched	216,931	84.6%
	Total	256,394	100.0%

Note: Matching to the IDBR. Stages based on fuzzy matching also have additional checks on company name, frequent expressions in names (e.g. 'services'). Stages in green are "new" stages compared to the previous approach adopted in the earlier matching attempt (BIS (2016) (here)), and stages shaded in blue are stages where significant changes/improvements have been made to the methodology adopted previously.

Phase 4: Quality Assurance

Table 3 illustrates the number of matches that were achieved in each stage, as well as the proportion of the EDS entities that were matched with the IDBR. At the end of the matching process, 216,931 EDS entities were matched (**84.6%**) with 39,463 unmatched (**15.4%**). This represents a significant improvement (approximately 9 percentage points) on previous matching attempts.

However, to understand the quality of the matching process (including those that were matched based on CRN), 60 matched cases were selected at random from each of the stages and a manual check of the accuracy of the match was undertaken. The manual check suggests that approximately **216,542** of these matches are 'good' matches (**99.8%**), with just **389** being mistakenly matched (**0.2%**). These estimates are based on random samples of 60 entries and it is possible that the extent of the mismatch is greater than presented here, however, the very low incidence of false positives provides some evidence of the validity of the matching process.

Table 3: Assessment of matching progress by stage

Stage	Number of Obs.	% Matched	Manual Check			Match rate (Correct + Probably/ All)	Good quality matches	Poor quality matches
			Definitely Correct	Probably correct	Incorrect			
Stage 1	79,866	31.1%	60	0	0	100.0%	79,866	-
Stage 2	802	0.3%	60	0	0	100.0%	802	-
Stage 3	48,448	18.9%	60	0	0	100.0%	48,448	-
Stage 4	19,931	7.8%	60	0	0	100.0%	19,931	-
Stage 5	7,627	3.0%	58	2	0	100.0%	7,627	-
Stage 6	2,819	1.1%	56	4	0	100.0%	2,819	-
Stage 7	21,862	8.5%	60	0	0	100.0%	21,862	-
Stage 8	7,921	3.1%	55	5	0	100.0%	7,921	-
Stage 9	291	0.1%	58	2	0	100.0%	291	-
Stage 10	7,177	2.8%	57	3	0	100.0%	7,177	-
Stage 11	6,241	2.4%	54	6	0	100.0%	6,241	-
Stage 12	4,572	1.8%	57	2	1	98.3%	4,496	76
Stage 13	457	0.2%	47	10	3	95.0%	362	95
Stage 14	894	0.3%	59	1	0	100.0%	894	-
Stage 15	1,165	0.5%	57	2	1	98.3%	1,146	19
Stage 16	1,734	0.7%	57	2	1	98.3%	1,705	29
Stage 17	733	0.3%	47	11	2	96.7%	709	24
Stage 18	1,797	0.7%	51	6	3	95.0%	1,707	90
Stage 19	118	0.0%	54	4	2	96.7%	114	4
Stage 20	2,476	1.0%	54	3	3	95.0%	2,352	124
Unmatched	**39,463**	**15.4%**	**Results from Manual check**			**Results from Manual check (weighted by number of stage cases)**		
Matched	**216,931**	**84.6%**	**1,120**	**63**	**17**	**99.2%**	**216,542**	**389**
Total	**256,394**	**100.0%**	**93.3%**	**5.3%**	**1.4%**			

Note: Stage 1 and 2 are based on Company Registration Number and parent/ultimate CRN and were undertaken before all other stages. Note that as part of the robustness checks undertaken, we varied the order to the stages undertaken.

In more detail, the analysis indicates that the initial stages based on firm characteristics including name and detailed postal address were assessed to be of extremely high quality. However, the quality of the match remained high also for later stages of the process, given that potential stages not meeting the quality threshold had not been retained (each stage has at least **95%** of good matches).

Assessing a sample of unmatched entries

The unmatched units were also assessed in terms of quality, to understand whether 'looser' requirements could help increase the matching rate without decreasing the quality of results. This process was reiterated with minor adjustments to refine the matching stages and strengthen the validity of the code.

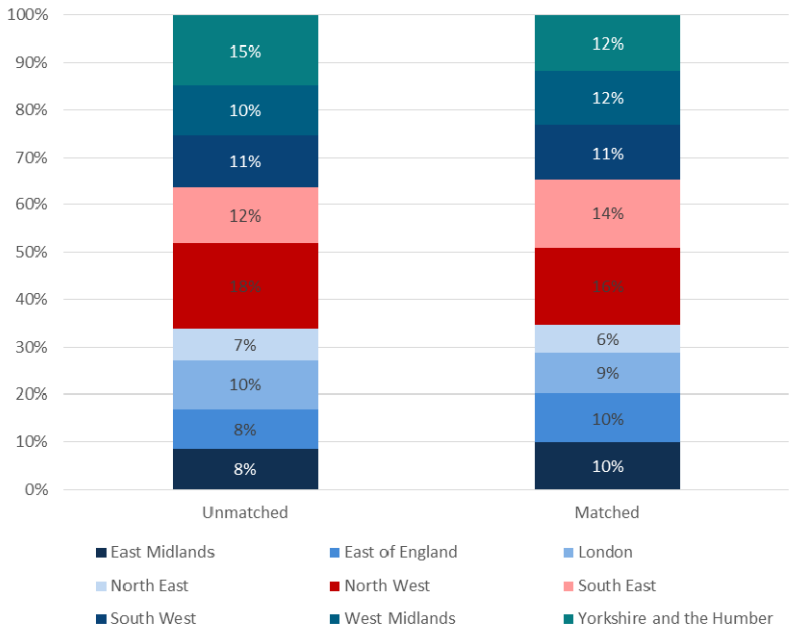
We extracted a sample of 100 unmatched EDS firms and searched them on the Orbis database (which includes FAME)[23]. In the vast majority of cases, no match could be found. Out of the 100 units, only 14 could be matched to entities on Orbis. However, out of these 14, only 9 were correct matches. We then proceeded with a careful review of IDBR records which contained either a similar company name or postcode as the records that were found on Orbis. A manual search on the IDBR suggested that 2 out of these 9 companies were available in the IDBR, however, the postcode did not match the one reported in the EDS (explaining why the entries could not be matched).

Finally we manually checked a sample of 60 unmatched EDS entities in the IDBR (by postcode or similar company name) and found that 6 (10%) had a potential match, but could not be matched due to different postcodes. About 58% of these EDS entries identified as "Sole Traders", explaining why no valid match was found in the IDBR. 27% of these entries had either 'Other' or a missing legal form.

3.2 Comparing matched and unmatched firms

Matched and unmatched firms are evenly distributed across all regions of England (Figure 2). Northern regions are slightly more prominent in the unmatched group, although the shares never differ by more than **2%** or **3%** between matched and unmatched.

Figure 2: Comparison of matched and unmatched firms by region



Source: ILR/EDS enterprises matched in the IDBR and unmatched;

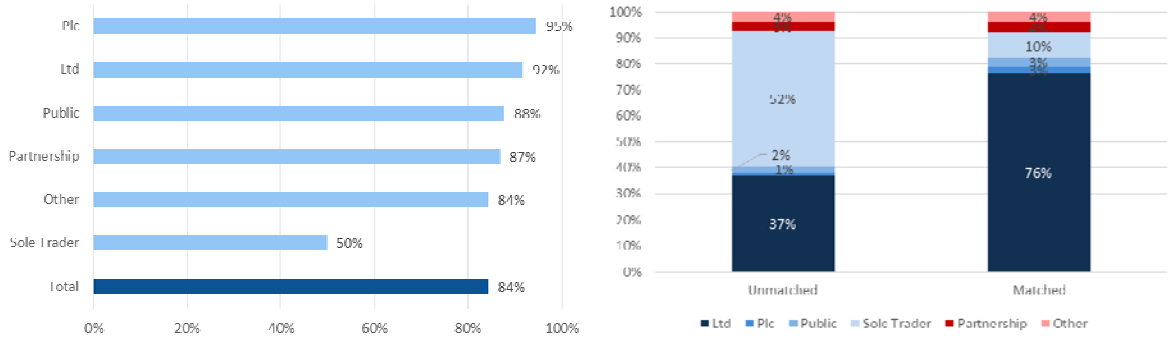
Matched and unmatched firms differ more markedly in terms of legal status and in terms of size. Looking at the match rates by legal form (Figure 3 left panel), the sole trader category is largely unmatched (**50%**). The matching rate for all other legal categories was satisfactory (at least **84%**). A large proportion (**52%**) of unmatched firms is represented by sole traders (Figure

[23] The Fame database is administered by Bureau van Dijk and contains comprehensive information on 9 million companies in the UK and Ireland (including non-active companies) http://www.bvdinfo.com/en-us/our-products/company-information/national-products/fame.

3 right panel). In contrast, this category takes up only **10%** of matched entities. Whilst a large share (**37%**) of unmatched units turns out to be limited companies, this group is the most frequent form among matched firms (**76%**).

A key explanation for this concerns the requirements by which enterprises are included in the IDBR, namely, enterprises in all sectors of the UK economy that are registered for VAT and PAYE. Hence, the Inter Departmental Business Register excludes very small businesses (i.e. those without employees, and those with revenues below the relevant VAT threshold (£82,000 in 2016)), as well as some non-profit organisations. As a result, it would be expected that relatively small sole traders will not be included in the IDBR, resulting in relatively small match rates for these types of companies.

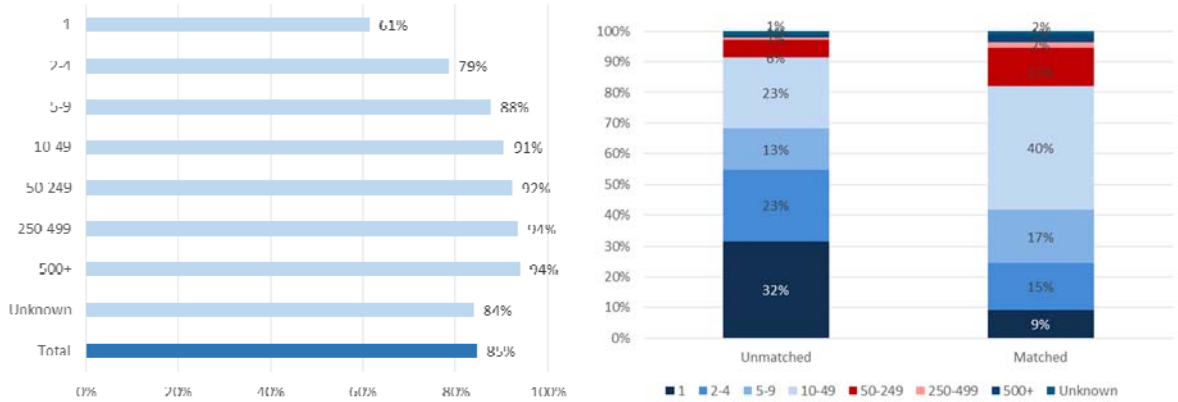
Figure 3: Comparison of matched and unmatched firms by legal status



Source: ILR/EDS enterprises matched in the IDBR and unmatched;

In relation to number of employees, the matching approach was relatively successful at achieving a match for medium sized or large entities, with a match achieved for approximately **94%** of training entities with more than 250 site employees and approximately **92%** of training entities with between 50 and 249 site employees. In contrast, the match rate for small training entities with a single site employee amounted to only **61%**. As a result, while entities with one (site) employee make up **32%** of unmatched training entities, they account for only **9%** of matched entities.

Figure 4: Comparison of matched and unmatched firms by size



4. Descriptive analysis – characteristics of firms engaging in publicly funded training

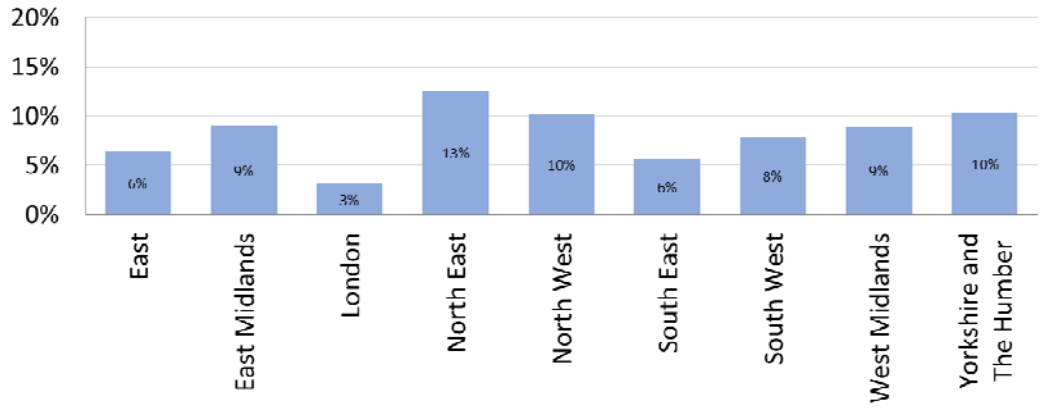
Below we present an analysis of enterprise characteristics: the charts describe the proportion of (matched) enterprises undertaking publicly funded training in 2014/15 by company size, region, sector of activity and legal status. Attention is restricted to England only as the ILR covers English learning providers only (a small number of enterprises are located in the other Home Nations).

In this section, we present information on the incidence of firms providing training as proportion of official statistics on business counts. To develop this part of the analysis, we first combined data from EDS-ILR and IDBR; then we calculated the proportion of these firms on business count data from the ONS.

Overall, the **216,931** EDS entities were matched to around **151,700** enterprises, as more than one entity may correspond to the same enterprise (for example a retailer with several stores would appear multiple times in the EDS, but only once at the enterprise level in the IDBR). The overall number of live enterprises for 2015 in England was **2,116,300**, implying that the overall proportion of matched EDS entities undertaking some publicly funded training in 2014/15 was just above **7%**.

Looking at the regional distribution, the largest concentration of firms undertaking publicly funded training is found in the North East (**13%**), North West (**10%**) and Yorkshire (**10%**), while the lowest rate is in the London area (**3%**).

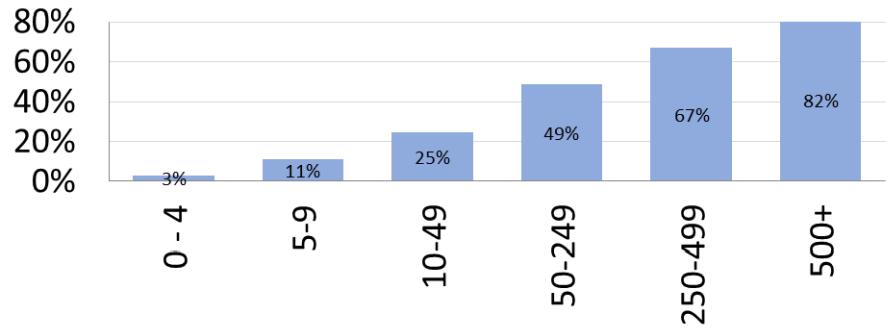
Figure 5: Proportion of matched ILR/EDS firms providing publicly funded training by English region



Source: ILR/EDS enterprises matched in the IDBR; ONS NOMIS Business Count data Note: England only

Looking at enterprise size, unsurprisingly, the largest incidence of firms engaged in publicly funded training is found among large enterprises. For those companies employing more than 500 employees, the coverage is higher than 80% (meaning that at least one employee in these enterprises undertook some form of publicly funded training in 2014/15). Rates are still high for medium-sized enterprises (between 50 and 499 employees) and decline substantially for small and micro businesses (only **3%** of enterprises with fewer than 4 employees engaged in publicly funded training).

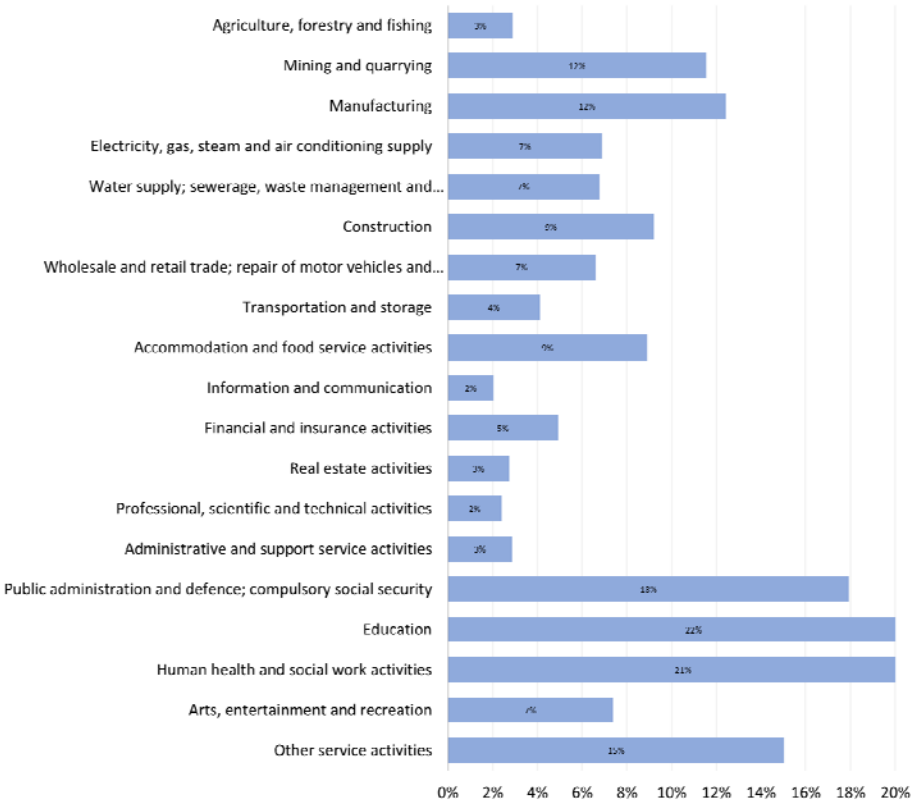
Figure 6: Proportion of matched ILR/EDS firms providing publicly funded training by size



Source: ILR/EDS enterprises matched in the IDBR; ONS NOMIS Business Count data
Note: England only

There is significant variation also in terms of sector of activity, with the highest incidence of engagement in publicly funded training found in the Education, Human Health and Public Administration sectors (between **18%** and **22%**, as shown Figure 7 below).

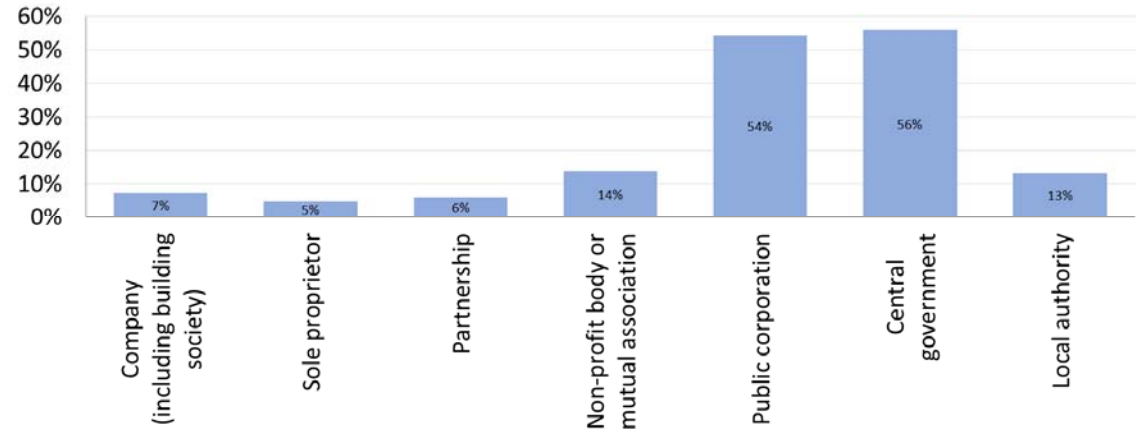
Figure 7: Proportion of matched ILR/EDS firms providing publicly funded training, by sector



Source: ILR/EDS enterprises matched in the IDBR; ONS NOMIS Business Count data
Note: England only; Sectors are classified by SIC code 2007

Reflecting the figure reported for size and sector, there is also a large incidence of matched firms in both the public corporation category, and the central government category (**54%** and **56%** respectively). Conversely, in other types of legal forms, the proportions are found to be relatively low, with sole proprietors showing the lowest incidence (**5%**).

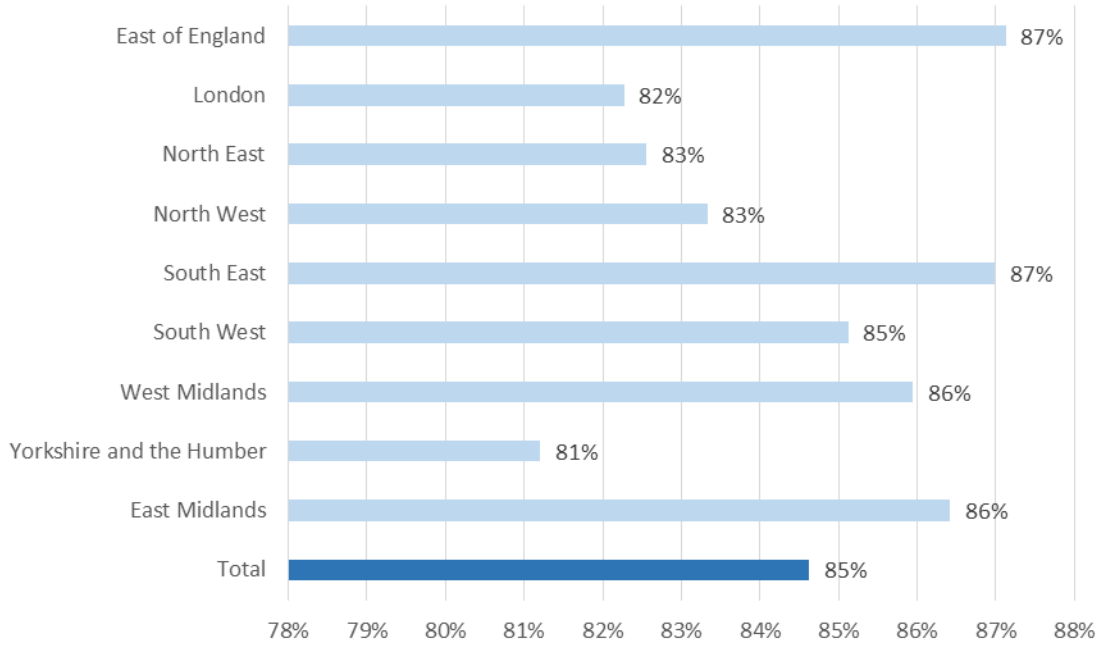
Figure 8: Proportion of matched ILR/EDS firms providing training, by legal status



Source: ILR/EDS enterprises matched in the IDBR; ONS NOMIS Business Count data.
Note: England only

The matching procedure achieved satisfactory rates across all regions, albeit with some variation (Figure 9). The lowest match rates are registered for Yorkshire (**81%**) and London (**82%**). Conversely, the East, the South East, and East Midlands are best represented with match rates of **87%** and **86%** respectively.

Figure 9: Match rates by region



Source: ILR/EDS enterprises matched in the IDBR

5. Conclusions

This matching exercise yielded a highly satisfactory match rate (**84.6%**). The code was progressively refined to attain the best performance for each step of the procedure, while maintaining a high level of quality of the matches. In previous analysis undertaken for the Department, after removing those individual stages of the matching process that resulted in relatively poor quality matches, the match rate stood at approximately **69%**.

The method we developed was notably computationally intensive but also tailored to the specific context of ILR-EDS-IDBR matching. This had the obvious advantage of delivering a very high quality matched dataset. Although sophisticated probabilistic routines are available (e.g. the *reclink* command we used for Stage 19), these are not sufficiently powerful to deal with such large amount of data and variables. Furthermore, these techniques are, just like their user-written counterparts, prone to error; they also require user intervention in the quality assurance phase.

The quality assurance process helped identify ways to further refine the code, and also showed that the routine was successfully identifying a large quantity of correct matches. In addition, an additional search on Orbis (which integrates a number of data sources including FAME) revealed that for those last unmatched units, a very small number could be correctly identified online (**9%**) but were not mapped into the IDBR. This gives us further confidence that the procedure has been successful.

As shown by the descriptive analysis, a relatively large amount (**58%**) of unmatched entities are sole traders. By construction, and due to the less strict requirements, this legal category is subject to significant under-reporting in the IDBR. For this reason, as well as the fact that there may be a lower incidence of training (especially publicly funded training) in this category, a significant proportion falling under this category cannot be fully captured by the matching procedure.

CVER PUBLICATIONS

CVER Briefing Note 003

The incidence of publicly funded training in England

Gavan Conlon, Sophie Hedges, Daniel Herr and Pietro Patrignani

March 2017

CVER Briefing Note 002

The Decision to Undertake an Apprenticeship: A Case Study

Steven McIntosh

March 2017

CVER Research Paper 004

Young people in low level vocational education: characteristics, trajectories and labour market outcomes

Sophie Hedges, Vahé Nafilyan, Stefan Speckesser and Augustin de Coulon

March 2017

CVER Briefing Note 001

Further Education in England: Learners and Institutions

Claudia Hupkau and Guglielmo Ventura

February 2017

CVER Research Paper 003

Vocational vs. General Education and Employment over the Life-Cycle: New Evidence from PIAAC

Franziska Hampf and Ludger Woessmann

November 2016

CVER Research Paper 002

Labour Market Returns to Vocational Qualifications in the Labour Force Survey

Steven McIntosh and Damon Morris

October 2016

CVER Research Paper 001

Post-Compulsory Education in England: Choices and Implications

Claudia Hupkau, Sandra McNally, Jenifer Ruiz-Valenzuela and Guglielmo Ventura

July 2016

The Centre for Vocational Education Research (CVER) is an independent research centre funded by the UK Department for Education. CVER brings together four partners: the LSE Centre for Economic Performance; University of Sheffield; Institute for Employment Studies and London Economics.

Any views expressed are those of the authors, and do not represent the views of DfE. For more details on the Centre, go to cver.lse.ac.uk

Corresponding author: Gavan Conlon, London Economics, and CVER
Email: gconlon@londoneconomics.co.uk

Published by:
Centre for Vocational Educational Research
London School of Economics & Political Science
Houghton Street
London WC2A 2AE