OECD Statistics Working Papers 2024/07

Towards a better understanding of data-intensive firms in the United Kingdom

Julia Schmidt,

Graham Pilgrim,

Annabelle Mourougane

https://dx.doi.org/10.1787/f8d640cc-en

OECD

Organisation for Economic Co-operation and Development

SDD/DOC(2024)6

Unclassified

English - Or. English

27 August 2024

STATISTICS AND DATA DIRECTORATE

Towards a better understanding of data-intensive firms in the United Kingdom

SDD Working Paper No. 126

Contact: stat.contact@oecd.org.

JT03548560

OECD STATISTICS WORKING PAPER SERIES

The OECD Statistics Working Paper Series – managed by the OECD Statistics and Data Directorate – is designed to make available in a timely fashion and to a wider readership selected studies prepared by OECD staff or by outside consultants working on OECD projects. The papers included are of a technical, methodological or statistical policy nature and relate to statistical work relevant to the Organisation. The Working Papers are generally available only in their original language – English or French – with a summary in the other.

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the authors.

Working Papers describe preliminary results or research in progress by the authors and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed and may be sent to the Statistics and Data Directorate, OECD, 2 rue André Pascal, 75775 Paris Cedex 16, France.

This document, as well as any statistical data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city, or area.

The release of this working paper has been authorised by Steve MacFeely, OECD Chief Statistician and Director of the OECD Statistics and Data Directorate.

www.oecd-ilibrary.org/economics/oecd-statistics-working-papers_18152031

Abstract / Résumé

By combining information from online job postings with firm-level financial data provided by Orbis, as well as firm-level merchandise trade data, this paper seeks to get a deeper understanding of the characteristics and performance of data-intensive firms in the United Kingdom since 2015. Data-intensive firms are defined here as firms which are hiring data-related skills. One key contribution of the analysis is to match in a more efficient way the two data sources, Lightcast and Orbis, which are now used extensively in the economic literature. Both the number and the share of data-intensive firms increased sharply in the United Kingdom from 2015 to 2021, with a peak in 2020. The number of highly data-intensive companies and data-intensive multinationals (MNEs) display the same pattern. A large share of data-intensive firms operate within the information and communication industry and are predominantly located in the Greater London area, especially in London itself. Those firms tend to employ more staff and are more capitalised than non data-intensive firms. They are on average more productive, generate more revenues and trade more in foreign markets. While data-intensive firms can be found in all firm size groups, the firms displaying on average the highest level of data intensity were medium sized in 2015 but are now small sized. In terms of international trade, UK data-intensive firms are, generally, more export intensive than non data-intensive firms, but estimates vary across industries.

Keywords: Data intensity, natural language processing, job advertisements, United Kingdom.

JEL codes : C80, C88, E01, J21.

****

Combinant les informations provenant des offres d'emploi en ligne avec des données d'entreprise, tels qu'Orbis ou la base UK trader, cet article cherche à mieux comprendre les caractéristiques et les performances des entreprises à forte intensité de données au Royaume-Uni depuis 2015. Ces entreprises sont définies comme celles qui recrutent des compétences liées aux données. L'un des principaux résultats de l'analyse est la mise en correspondance de deux sources de données de manière plus efficace : Lightcast et Orbis, qui sont désormais largement utilisées dans la littérature économique. Le nombre d'entreprises à forte intensité de données a augmenté de manière significative au Royaume-Uni entre 2015 et 2021, avec un pic en 2020. Le nombre d'entreprises à forte intensité de données et de multinationales à forte intensité de données présente la même tendance. Une grande partie des entreprises à forte intensité de données opèrent dans le secteur de l'information et de la communication et dans la région du Grand Londres et à Londres en particulier. Ces entreprises ont tendance à employer plus de personnel et à être plus capitalisées que les entreprises qui ne consomment pas beaucoup de données. Elles sont en moyenne plus productives, génèrent plus de revenus et commercent davantage sur les marchés étrangers. Même si les entreprises à forte intensité de données se retrouvent dans tous les groupes de taille d'entreprise, les entreprises affichant en moyenne le niveau d'intensité de données le plus élevé étaient de taille moyenne en 2015, mais sont désormais de petite taille. En termes de commerce international, les entreprises britanniques qui utilisent beaucoup de données exportent généralement plus que les autres, mais les estimations varient selon les secteurs.

Mots-clés : Intensité des données, traitement automatique du langage naturel, offres d'emploi, Royaume-Uni.

Codes JEL : C80, C88, E01, J21.

Table of contents

Towards a better understanding of data-intensive firms in the United Kingdom 5
1. Introduction 5
2. Selected literature review 6
3. Data 7
4. Methodology and empirical strategy 9
5. Key features of data-intensive firms 13
6. Performance of data-intensive firms 21
7. Conclusion 27
References 28

Annex A. Explaining the Natural Language Processing approach 32

FIGURES

Figure 1. Overview of the approach 10
Figure 2. Number and share of data intensive firms over time, United Kingdom 14
Figure 3. Data intensity by company size, United Kingdom 15
Figure 4. Share of firms across industries in the United Kingdom 16
Figure 5. Regional concentration of data intensive jobs, United Kingdom 17
Figure 6. UK cities with the highest demand for data skills 18
Figure 7. Number of highly data-intensive firms in the United Kingdom 19
Figure 8. Share of data vs. non data-intensive MNEs, United Kingdom 20
Figure 9. Comparison between data and non data-intensive firms, United Kingdom 21
Figure 10. Average multifactor productivity levels by firm type, United Kingdom 22
Figure 11. Engagement in exporting of data and non data-intensive firms, United Kingdom 23
Figure 12. Export intensity by industry, United Kingdom 24
Figure 13. Average number of products traded, United Kingdom 24
Figure 14. Share of data-intensive firms trading by product categories, United Kingdom 25
Figure 15. Data-intensive firms received a higher part of their total revenues from exports 26

Figure A.1. A natural language processing pipeline 32
Figure A.2. Example of a tokenisation process 33

TABLES

Table 1. Share of missing values in Lightcast data 8
Table 2. Main advantages and disadvantages of using Lightcast data 8
Table 3. Share of companies retained compared to the Lightcast dataset 12
Table 4. Descriptive statistics of key firm indicators in the panel dataset 13

Towards a better understanding of data-intensive firms in the United Kingdom

By Julia Schmidt, Graham Pilgrim and Annabelle Mourougane[1]

1. Introduction

1. The rapid pace of digitalisation, and particularly the increasing reliance on data, over the past decades has triggered substantial analysis on the behavioural changes of firms, and how economic policy needs to adapt to these changes. Given the heterogeneity of firms in digital technology uptake across industries, economic research has shifted somewhat to examine directly how individual firms have adopted those technologies and how this has affected their performance, with a lot attention directed to productivity and engagement in international trade (Andrews, Criscuolo and Gal, 2016[1]; Bernard et al., 2018[2]).

2. Data-intensive firms are expected to have a different business model, given their strong reliance on a non-rival input, which allows usage by multiple parties simultaneously without exhausting the asset, even though technical and legal restrictions can sometimes make access to data excludable (OECD, 2022[3]). While there is no official definition of what constitutes a data-intensive firm, the term is often assimilated to firms that have invested heavily in data or with a substantial endowment in data skills (e.g. data analytics). They can be identified using either firm or household micro-data or alternative data, such as online job postings. By providing timely and granular information, the latter have for instance been increasingly exploited to quantify the data content of jobs at country and industry levels using natural language processing (Schmidt, Pilgrim and Mourougane, 2023[4]). These data can also provide information at the firm level of data contents of jobs but fail to inform on firm performance.

3. By combining information from online job postings with firm-level datasets, such as Orbis or HM Revenue and Customs (HMRC)'s UK trader dataset, this paper seeks to get a deeper understanding of the characteristics and performance of data-intensive firms in the United Kingdom since 2015. Those firms are defined here as firms which are advertising jobs with data-related (data entry, database and data analytics) skills. One key by-product of the analysis is to match two data sources in an efficient way, Lightcast and Orbis, which are now used extensively in the economic literature, looking at the data-intensity of firms or industries and at firm-level productivity developments.

[1] The financial support for the research presented in this paper was provided by the Department for Business and Trade (DBT), United Kingdom. This paper benefited greatly from comments, guidance, and support by Asa Johansson, Josh De Lyon, Peter Gal and Valentine Millot (all OECD), Nikos Tsotros and Dylan West (DBT). The authors would like to thank Virginie Elgrably for excellent support in formatting the document.

4. The key insights from the empirical analysis are as follows:

- The number of data-intensive firms, registered in the United Kingdom, increased sharply from 2015 to 2021, with a peak in 2020. The number of highly data-intensive firms with data intensity above 50% and of data-intensive multinationals (MNEs), defined as companies with at least two subsidiaries in other countries and data intensity above 10% display the same pattern.
- A large share of data-intensive firms operate in the information and communication industry and are located in the Greater London area, especially in London itself.
- Those firms tend to employ more staff and be more capitalised than non data-intensive firms. They are on average more productive, generate more revenues and trade more in foreign markets.
- Small firms have, on average, accelerated their hiring rate of data-related skills more than medium-sized and large firms. While data-intensive firms can be found in all size groups, firms which had the highest demand for data-related skills were generally medium sized in 2015 but were small sized in 2021.
- Turning to international trade, the share of export volume, including goods and services, in gross output is higher among data-intensive firms. When zooming into merchandise trade data, there is no discernible difference in terms of number of products traded between data-intensive and non data-intensive firms. Data-intensive firms export and import essentially machinery and equipment, electrical machinery and optical, and photographic medical instruments.

5. The paper is organised as followed. After a short review of the economic literature in Section 2, Section 3 and 4 present the data and the approach used to identify data-intensive firms. Section 5 describes the main features of data-intensive firms and Section 6 discusses their performance. A last section concludes.

2. Selected literature review

6. The main focus of this paper is on gaining insights on data-intensive firms, where they operate and whether they perform better than non data-intensive firms. While data-intensive firms are generally known to be a subset of firms that rely heavily on data collection, processing and analysis to drive their business decisions, their precise definition varies across papers. For some, those are firms which heavily invest in digital technology, for others those are firms whose manpower is specialised in data production skills (e.g. data entry, database or data analytics skills). The overlap between the two groups is likely to be large.

7. In this context, several strands of the literature are of relevance to the paper. The first strand deals with measurement issues and seeks to make digitalisation "visible" in official statistics (Ahmad and Schreyer, 2016[5]). A number of improvements have been put forward in the context of the 2025 revision of the System of National Accounts and of the Balance of Payments international standards. This includes broadening the scope of intangible assets which will need to be explicitly accounted in official statistics (see SNA update site).

8. Furthermore, several attempts have been made to measure the data-intensity of jobs, and in turn of an industry or an economy (Corrado et al., 2022[6]; Schmidt, Pilgrim and Mourougane, 2023[7]; Statistics Canada, 2019[8]). One main innovation in this field is to rely on online job advertisements, which can provide timely and very granular information, although so far, the information has been exploited mostly at the economy-wide and industry level.

9. The second strand of the literature is related to the characteristics and performance of firms which have invested in new technologies and the heterogeneity of firm-level productivity performance across industries, building on micro-data (Andrews, Criscuolo and Gal, 2016[1]; Berlingieri, Blanchenay and

Criscuolo, 2017[9]; Berlingieri et al., 2020[10]; Gal et al., 2019[11]; Cette, Corde and Lecat, 2017[12]). A number of studies have stressed the differences between small and large firms in adopting new technology (Coyle et al., 2022[13]; Bricongne, Delpleuch and Lopez Forero, 2021[14]). Coyle and Nguyen (2018[15]) found that cloud-using firms tend either to be large businesses or digitally native start-ups. Focusing on the United Kingdom, Calvino et al. (2022[16]) found that a significant share of AI adopters can be found in Information and Communication Technologies and Professional Services, and is located in the South of the United Kingdom, particularly around London.

10. The complementarity between human capital and technology has been put forward as a key explanation of differences in technology adoption. The idea would be that a sufficient technological skill endowment is a pre-requisite for firms to make the most of new technology (Sorbe et al., 2019[17]). Bloom, Sadun and Reenen (2012[18]) found that technology is used more productively by better managed firms, while Grundke et al. (2018[19]) underlined the importance of having a bundle of skills, including cognitive and non-cognitive skills to harness the potential of digital technologies. Similarly, there is evidence of complementarity between technology adoption and investment in intangibles such as data and software (Corrado et al., 2021[20]).

11. Data-intensive firms are found to perform better than non data-intensive firms along a number of dimensions. Firms that exhibit on average a higher share of investment in digital technologies are also characterised by a faster rate of multifactor productivity growth, but the difference is small and not all firms and industries experience significant productivity gains from digitalisation (Anderton, Botelho and Reimers, 2023[21]). Defining digital firms as those which invest in digital assets, Coyle et al. (2022[13]) reckoned that digital adopters have higher multifactor productivity than non-adopters, when the former are endowed with in-house capabilities.

12. The objective and main novelty of this paper is to bring together those two strands and get further insights on data-intensive firms, including their economic performance and their engagement in international trade. Given the absence of a readily available comprehensive micro-dataset, the approach is to combine information from Lightcast, a database of online job advertisement, with Orbis, a micro-database widely used in the literature, and the HMRC's UK trader database, to add on information on international merchandise trade.

3. Data

13. The analysis combines three different datasets to enable analysis on the geographical, financial, and trade characteristics of data-intensive firms in the United Kingdom.

Online job advertisement data

14. The online job advertisements data are provided by Lightcast, a private data provider previously known as Emsi Burning Glass or Burning Glass Technologies. They gather job postings from close to 40 000 online sources, such as job boards, employer sites, newspapers, and public agencies using web scraping techniques. Lightcast initially covered Australia, Canada, New Zealand, Singapore, the United Kingdom and the United States, and its geographical coverage has been expanded in the most recent vintages to include EU, African and Latin American countries.

15. The UK data are available annually for 2012-present (January 2024 at the time of writing). In 2021, the last available year in Orbis and the reference year for this study, the dataset includes 9.9 million job advertisements for the United Kingdom. As one of the key contributions, this study relies on the raw text format rather than utilising the pre-processed structured data provided by Lightcast to estimate data intensity at firm level.

16. The data provide good coverage of the labour demand in the United Kingdom (flow as opposed to labour stock) and allow for insights into the skills and task requirements for each job advertisement. At the occupational level, there is a very limited number of missing values (Table 1). The level of missingness is higher at the industry and regional levels, although with some variability across industries and regions.

17. Comparing LC data to official vacancy data collected by the ONS, Tsvetkova et al. (2024[22]) show that education and health are the most over-represented industries, while the most under-represented industries are accommodation, foods, arts and recreation, and transport and warehousing. For digital sectors such as information, media, and telecommunication as well as professional, scientific and technical activities and administrative and other services the differences between the data sources remain small. The representativeness of United Kingdom industrial data is very stable over time.

Table 1. Share of missing values in Lightcast data

Per cent

Variable of interest	2015	2016	2017	2018	2019	2020	2021
Job ID	0	0	0	0	0	0	0
Lightcast occupational group	6	6	6	6	7	7	8
Industry code - two digits	51	51	50	47	43	43	45
Company name	69	69	68	62	54	54	50
Region	18	16	15	18	19	19	22

Source: Authors' calculations based on the Lightcast data.

18. The main advantage of those data are their timeliness and high level of granularity (Table 2). They have been used in several studies, including in Tsvetkova et al. (2024[22]) and Cameraat and Squicciarini (2021[23]). The main limitation is the decreasing quality when analysing earlier periods, most notably 2012-14 due to the improvements in the Lightcast algorithm. Another limitation is the lack of representativeness when using the firm-level information contained in the job advertisements (for an extended discussion of its quality see Schmidt, Pilgrim and Mourougane (2023[7])).

Table 2. Main advantages and disadvantages of using Lightcast data

Advantages	Disadvantages
Timely data (2012-present)	Decreasing quality of the data the further back in the time (e.g. 2012 data are of worse quality than 2023 data)
Extended country coverage (United Kingdom, Canada, United States, New Zealand, Australia, Singapore, EU countries, selected African and Latin American countries at the time of writing)	Partially harmonised classification for occupation/sectors across countries; strong variance in coverage of certain variables (e.g. regions, firms)
Allow linkage to firm-level and regional data	Limited coverage depending on year and country; no insight on how firms hire (e.g. churn rate)
Standardised occupation and industry classifications within a country	Representativeness is heterogeneous (depending on country, industry, region)
Identify skill demands beyond standard labour market statistics	

Source: Authors' compilation based on Lightcast data.

Micro-level firm data

19. The Orbis database provides harmonised financial and ownership information at the firm level across countries. It covers more than 100 countries and over 400 million firms up to end- 2021 and remains one of the largest global database that combines firms' financial statements and their activity in terms of sales, employment and investment. It includes both private and publicly listed firms. The financial and balance sheet information comes from national business registers, governed by country-specific legal and administrative filing requirements.

20. For the United Kingdom, financial data vary in availability and are commonly less available for micro and SMEs. With adequate treatment (see Section 4 for a detailed description), the Orbis sample for the United Kingdom is representative in terms of coverage of small and large firms over time and by industry and no reweighting is needed (Kalemli-Oezcan et al., 2023[24]; Bajgar et al., 2020[25]).

21. The key variables of interest to the analysis include firm name, firm ID, industry, number of employees, gross output, capital, productivity measures, export revenue and information on the ownership structure (see below).

Product-level merchandise trade data

22. The product level merchandise trade information is published by the UK HM Revenue and Customs (HMRC) (HM Revenue and Customs, 2023[24]). The dataset contains monthly information about the imported and exported goods by firm disaggregated by eight-digit HS codes, starting in 2016 up to January 2022. The files also contain the corresponding Standard International Trade Classification (SITC) heading, that allows to get information on the different categories of goods traded. Further, information on the name of the company, the address, and its postal code is included. Overall, the dataset allows for unique, detailed and timely insights into the type of goods exported and imported by UK companies.

23. There are certain shortcomings that are important to keep in mind when analysing insights from this database. First, the dataset covers only transactions in goods. Second, HMRC applies automated disclosure control to importers and exporters names and address details if a specific commodity segment includes less than three active importers/exporters. In addition, traders can request exclusion from the exporter/importer information up to the 15th day of the month.

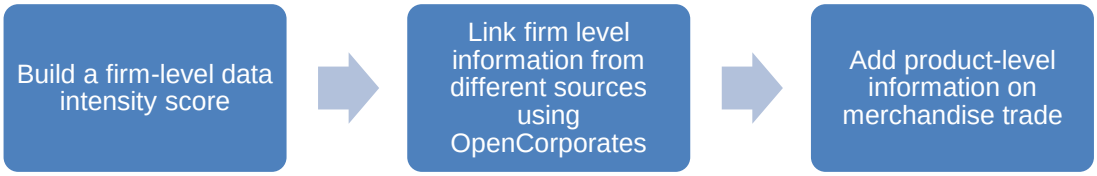
4. Methodology and empirical strategy

24. The approach combines Natural Language Processing techniques (NLP) to construct an indicator of data-intensive firms and integrates this indicator into a firm-level panel dataset spanning information on firm location, employees, productivity, financial assets, ownership and export behaviour. The main data sources are text data retrieved from online job advertisements provided by Lightcast in the United Kingdom, the firm-level database Orbis and HMRC's UK trader dataset on merchandise trade. The data contain information on skills and task requirements in each job advertisement, firm assets and ownership structures as well as product-level trade information (for an extensive discussion the data sources see Section 3).

25. The approach is broken down into three steps:

- Step 1: Derive a firm-level data intensity score using online job advertisement and the methodology detailed in Schmidt, Pilgrim and Mourougane (2023[7]) using NLP.
- Step 2: Match Lightcast firm-level data to Orbis data using the firm ID provided by OpenCorporates.
- Step 3: Use information from the UK trader database to provide insights into the trade behaviour of specific firms identified with the previous matching for the period 2016-2020.

Figure 1. Overview of the approach



Source: Authors' illustrations.

Step 1: Deploying a natural language processing algorithm to calculate data intensity at firm-level

26. This paper uses NLP techniques to extract relevant information from online job advertisements and to derive the share of jobs involved in data production, referred to as "data intensity", at firm level. It builds upon the approach introduced in Schmidt, Pilgrim and Mourougane (2023[7]), who applied the approach to derive data intensity shares at occupation, industry and economy level.

27. The definition of data production jobs follows the framework set out by Corrado et al. (2022[6]) and Statistics Canada (2019[8]) and identifies the relevant skills/tasks used in the specific position. The methodology aims at distinguishing between value creation at three stages of data production and defines three types of data assets, forming a "modern data stack":

- **Data entry:** These tasks relate to work with raw records that have been stored but not yet cleaned, formatted or transformed for analysis (e.g. data scraped from the web). Raw records also refer to data collected from experiments, statistical surveys or administrative records.
- **Databases:** Jobs in this category work on transformed raw data, records that have been cleaned, formatted and structured to be used in data analytics or visualisation.
- **Data analytics:** These tasks reflect jobs using advanced tools to analyse data (e.g. machine-learning algorithms).

28. The approach relies on an open-source NLP pipeline provided by the 'spacy' python library (spaCy, 2022[25]). The pipeline combines different NLP models to efficiently perform advanced text processing operations in an iterative fashion. The output of the pipeline can be used for tasks such as text classification or analysing phrase frequencies.

29. The main steps in the NLP pipeline are as follows:

1. Extract from the job advertisement the skills/tasks that are related to the production of data and transform this information into mathematical objects using natural language processing.
2. Based on a set of rules, use the mathematical objects to classify the job as data-intensive or not, linking it to data entry, database or data analytics activities. The criteria for the three types of data-related roles are independent from each other.
3. Aggregate the data-intensive jobs to get estimates by firm.

30. The resulting data intensity shares range from 0-100. These indicators can then be included in a firm-level dataset to analyse characteristics of data intensive firms (for a detailed explanation of the methodology see Annex A).

Step 2: Matching Lightcast and Orbis data

31. A key contribution of this paper is to match Lightcast and Orbis data using a specific validation algorithm. After cleaning and processing the data, the panel dataset includes only a subset of the two datasets that were merged, namely a total of 12 000 companies for the period 2015-21. In addition, it is necessary to validate the company names contained in the Lightcast data set with OpenCorporates. This has two distinct advantages: the company is then validated to be active in the United Kingdom (as no such a validation is performed in Lightcast); secondly it allows the construction of the unique company identifier which facilitates the matching based on a company ID (instead of a match based on company names).

32. Lightcast data was processed to remove all rows where company names and job-level information were missing (about 50% of the available job postings across years). Furthermore, all job advertisements posted by recruitment codes (referring to SIC Code 78300, 78200, 78109, 78101) were dropped to remove a bias in the firm-level dataset towards recruitment agencies. Additional validation checks were designed to ensure that firm names were correctly spelled. Finally, firms had to post at least ten job vacancies per year to be included in the analysis and ensure that the data intensity share by firm is reliable across all years. Earlier research shows that Lightcast data present some similarity with ONET but offer more granularity (Schmidt, Pilgrim and Mourougane, 2023[4]).

33. For Orbis data, the challenge of cleaning the data lies in creating consistent firm-level information with regards to financial and ownership variables. The cleaning process followed in large aspects processes described in Kalemli-Ozcan et al. (2015[26]; 2023[27]), Gal (2013[28]), Calvino et al. (2022[16]) and Andrews, Criscuolo and Gal (2019[29]). The only aspect that could be omitted relates to merging of Orbis data across vintages, as the data used in this paper are already harmonised over time. Overall, it is important to mention that the Lightcast sample is not representative of the universe of UK firms and specifically the small firms selected into the sample are more likely to be data-intensive than the omitted ones, due to the range of firms included in Lightcast.

34. With the objective to ensure consistent information of financial assets, export revenue and ownership information the dataset is cleaned based on an established set of rules. Companies that lack joint information on total assets and operating revenues, sales and employment are dropped. A company is dropped in a respective year if total assets, sales, fixed assets or employment are negative, or if employment exceeds 2 million. Furthermore, outliers which display implausible performance indicators are removed following (Kalemli-Ozcan et al., 2023[27]). This is the case, for instance, when sales are large than 99.9 percentile of the distribution. Finally, for a given company ID and year, missing strings which are unlikely to change over time (e.g. country, company name, city, region, postal code and legal form etc.) are replaced with values for the respective company for other years.

35. A key contribution of the methodology lies in the design of an algorithm that allows the validation of company names based on the OpenCorporates OpenRefine Reconciliation API (version 0.4.8). OpenCorporates is an open-source register that makes legal-entity data accessible. The algorithm allows to cross-check whether a company is actually registered in a specific jurisdiction. The query takes a list of company names and connects to the OpenCorporate database, identifies the suitable company name, and extracts a score that reflects the certainty of such a match. The information provided enables the researchers to construct the company ID of the respective entity. The latter enables a match based on the ID, instead of the company name, thus limiting data loss incurred when the matching is done through name matching (Dernis et al., 2015[30]). Table 3 reports the share of Lightcast firms retained at the different steps.

Table 3. Share of companies retained compared to the Lightcast dataset

Per cent

Year	Step 1: Validation via Open Corporates	Step 2: Matching to Orbis
2015	66.8	24.8
2016	56.7	49.5
2017	55.2	44.7
2018	75.4	20.9
2019	59.5	52.2
2020	68.3	39.4
2021	80.5	11.1
Total	73.3	22.7

Source: Authors' calculation based on Lightcast data.

Step 3: Extending firm panel data to merchandise product level

36. In a third step, the firm-level data set spanning information from both Lightcast and Orbis is extended to include product-level data on merchandise trade from the UK trader dataset. The company links are established using the same algorithm as above, first extracting the company IDs based on a list of company names from the UK trader data set. The firm-panel data is then linked to the UK trader dataset based on the company IDs. This allows for insights into the type of goods exported and imported by UK companies in the period 2016-2020 (down to HS six-digit level).

Representativeness of the panel dataset

37. The constructed, experimental panel dataset provides a first basis to investigate data-intensive firms based on the companies coming from the Lightcast dataset. It contains data on 12 000 firms and includes core variables related to the geographical location and financial assets of the firm, its productivity, capital assets, size, ownership status and export revenue (see Table 4). The analysis spans over 2015-21 at firm-level and 2016-20 at product-level.

38. The dataset use is bound by certain limitations. Due to the nature of online job advertisements, the data is biased towards firms that hire mainly online. Moreover, as Lightcast scraps data from online portals, only about half of the advertisements have information on company names and firm characteristics, which limits the representativeness with regards to occupations and industries. While the preprocessing ensures that biases related to recruitment agencies posting more jobs than other firms are taken into account, Lightcast does not ensure that the firm sample is consistent over time, which does not allow to follow specific companies over time. Only online job advertisement data for 2015-21 was deemed of sufficient quality to examine firm performance over time.

39. With regards to key indicators, data intensity is generally skewed across companies, with very few companies having high levels of data intensity, and a large majority showing low values. In terms of company size, the representativeness of small and medium enterprises is relatively good, yet the majority of companies in the sample employ more than 250 employees. This factor also explains the descriptive statistics on gross output, capital, productivity, MNE status and export revenue.

40. For the product-level extension, additional caveats have to be taken into account. The data source provides micro-data only for merchandise trade for 2016-2020.

Table 4. Descriptive statistics of key firm indicators in the panel dataset

Variable	Mean	Standard deviation	Minimum	Median	Maximum
Data Intensity (per cent)	0.05	0.12	0	0	1
Data entry	0.01	0.07	0	0	0.88
Database	0.01	0.05	0	0	0.98
Data analytics	0.02	0.03	0	0	0.93
Gross output (million GPB)	269 649 557	3 158 575 281	33	25 001 066	261 195 500
Capital (million GPB)	104 659 505	1 645 453 751	0	3 836 111	118 131 820
Employment	1006	5649	1	215	231 223
MFP	10.94	1.01	2.64	10.84	19.04
Export revenue (million GBP)	283 784 893	2280987558	57	8 339 828	39 613 198

Source: Authors' compilation.

5. Key features of data-intensive firms

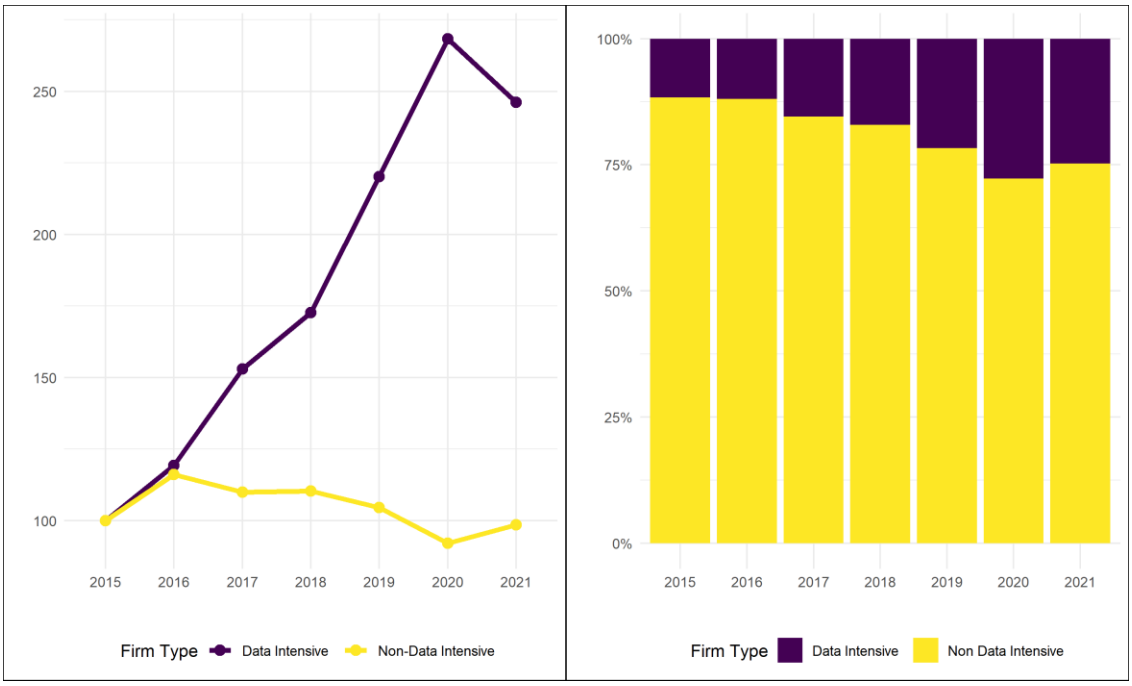
The number and share of data-intensive firms has increased since 2015

41. The number and share of data-intensive firms increased sharply from 2015 to 2021, with a peak in 2020 (Figure 2). In 2015, there were roughly 2 500 firms in the sample of 12 000 firms with no demand for data-related skills compared to about 250 firms with demand for data-intensive skills. This is consistent with the observed increase in the number of firms in digital industries over the past decade (Calvino et al., 2023[31]).The rise is particularly remarkable in a context of subdued economic growth in the United Kingdom during most of the period following the 2016 Referendum and the decision to leave the European Union. The economy slowed markedly after 2016, particularly as a result of lower investment growth (OECD, 2020[32]).

42. The number of data-intensive firms declined somewhat in 2021. The share of data-intensive firms reached 23% in 2021 (Figure 2).This could be interpreted as an adjustment to an overshooting in 2020, when the demand for data skills surged to adapt to the move to remote working and measures to limit human-to-human interactions to slow the spread of the COVID-19 virus. When the situation normalised, the demand for digital skills was less acute and resumed its pre-crisis trend. Firms may also have limited hiring and investment spendings in 2021, in the context of heightened uncertainties related to the duration of the pandemic and the timing of the withdrawal of emergency support (Bloom et al., 2018[33]).

Figure 2. Number and share of data intensive firms over time, United Kingdom

Growth of firms (2015=100) (left) and share of firms in per cent (right)

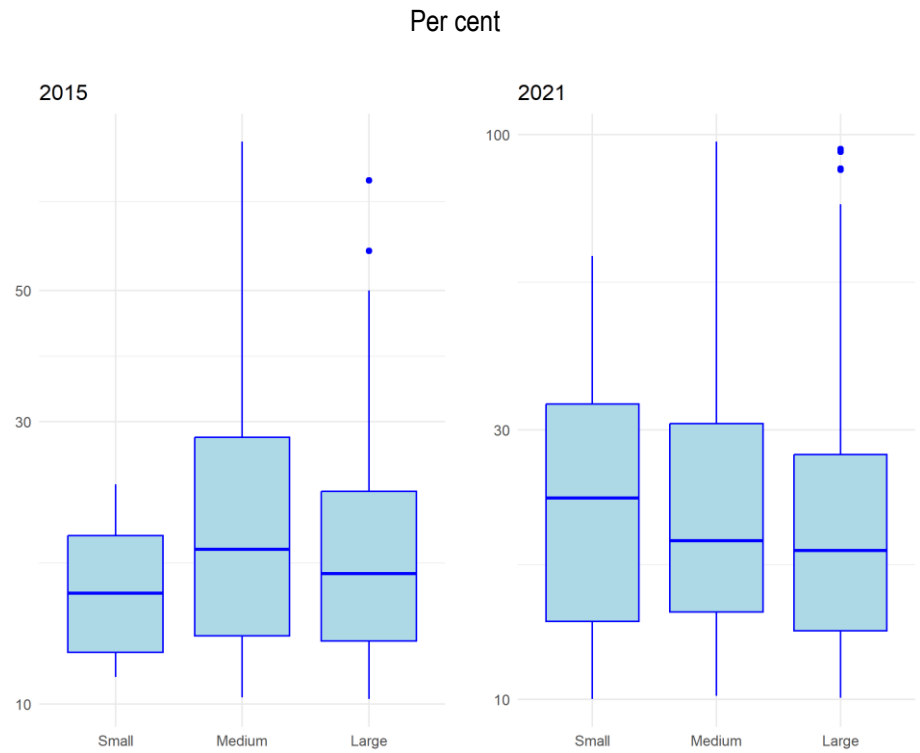


Note: Only firms with a data intensity > 10% and more than ten vacancies per year were selected. The sample includes a total of 12 000 firms over the period 2015-2021.
Source: Authors' calculations based on Lightcast and Orbis data.

The group of data-intensive firms is heterogenous but concentrated in specific industries and regions

43. Data-intensive firms can be found in all size groups which are found to be highly heterogeneous across industries and regions (Figure 3). Data intensity rates, measured as the share of job postings demanding data-related skills in the total of online postings, have increased across all size classes from 2015 to 2021. On average, the companies displaying the highest level of data intensity were medium sized in 2015 but were small sized in 2021. This reflects the nature of the data intensity indicator, which is capturing a flow rather than a stock, i.e. the new hiring related to data production rather than all staff in the company currently working in data production. In other terms, this suggests that small firms have, on average, accelerated their hiring rate of data-related skills more than medium-sized and large firms. In addition, in 2015, the small firms operated in computer software and finance industries, while in 2021 small, data-intensive firms belong to the banking, insurance, ITC and biotechnology industries.

Figure 3. Data intensity by company size, United Kingdom



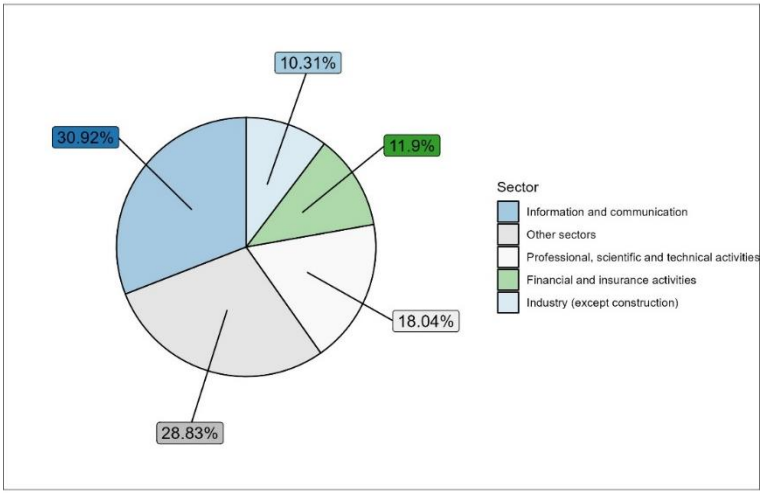
Note: The line in the box depicts the average data intensity in the firm size class. The small dots at the top of the scale are outliers that go beyond 1.5 times the interquartile range. Only firms with a data intensity > 10% and more than ten vacancies per year were selected. The y axis is log scaled. Small companies: < 9 employees, medium companies: 10-249 employees, large companies > 250 employees. The sample includes a total of 12 000 firms over the period 2015-2021.
Source: Authors' calculations based on Lightcast and Orbis data.

44. Most data-intensive firms (71%) operated in three industries, the information and communication sector, the professional scientific and technical activities and the finance and insurance activities (Figure 4). Those account for only 32% of the total firms in the UK economy. By contrast, very few data-intensive firms could be found in the water supply, sewerage and waste management sector, construction and accommodation and food services. As mentioned before, those results should be interpreted as industries which hired staff involved in data production, not industries which necessarily employ a lot of staff in this field.

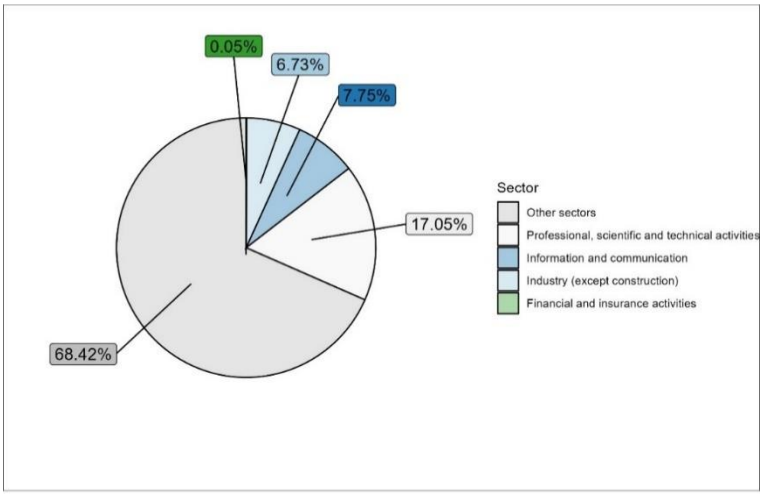
Figure 4. Share of firms across industries in the United Kingdom

Per cent, 2021

A – Data intensive firms by industry



B – Total firms in the UK economy by industry

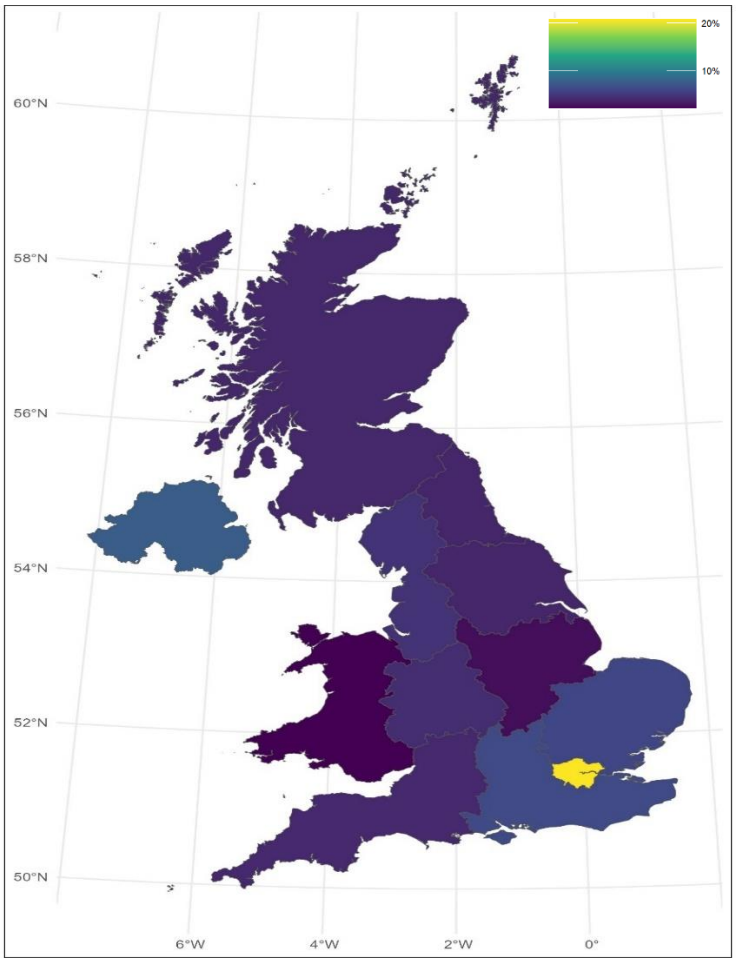


Note: In Panel A, only firms with a data intensity > 10% and more than ten vacancies posted per year were selected. The sample includes a total of 12 000 firms over the period 2015-2020, and 3 303 firms for 2021. In Panel B, the number of firms in the Finance and Insurance industries is for 2019. ONS data on business activity suggests this share has been broadly stable since 2019.
Source: Authors' calculation based on Lightcast data, OECD Structural Business Statistics ISIC4.

45. Most data-intensive firms were located in the Greater London region and in London specifically in 2021 (Figure 5). Many were also found in Northern Ireland, which houses a number of data or tech hubs in particular in Dublin, but also in the South East and East of England. By contrast, very few data-intensive firms were registered in the East Midlands and Wales. Implied regional disparities in terms of firm data-intensity, overlap to a large extent to regional disparities in terms of firm productivity, as reported in Gal and Egeland (2018[34]).

Figure 5. Regional concentration of data intensive jobs, United Kingdom

Per cent, 2021

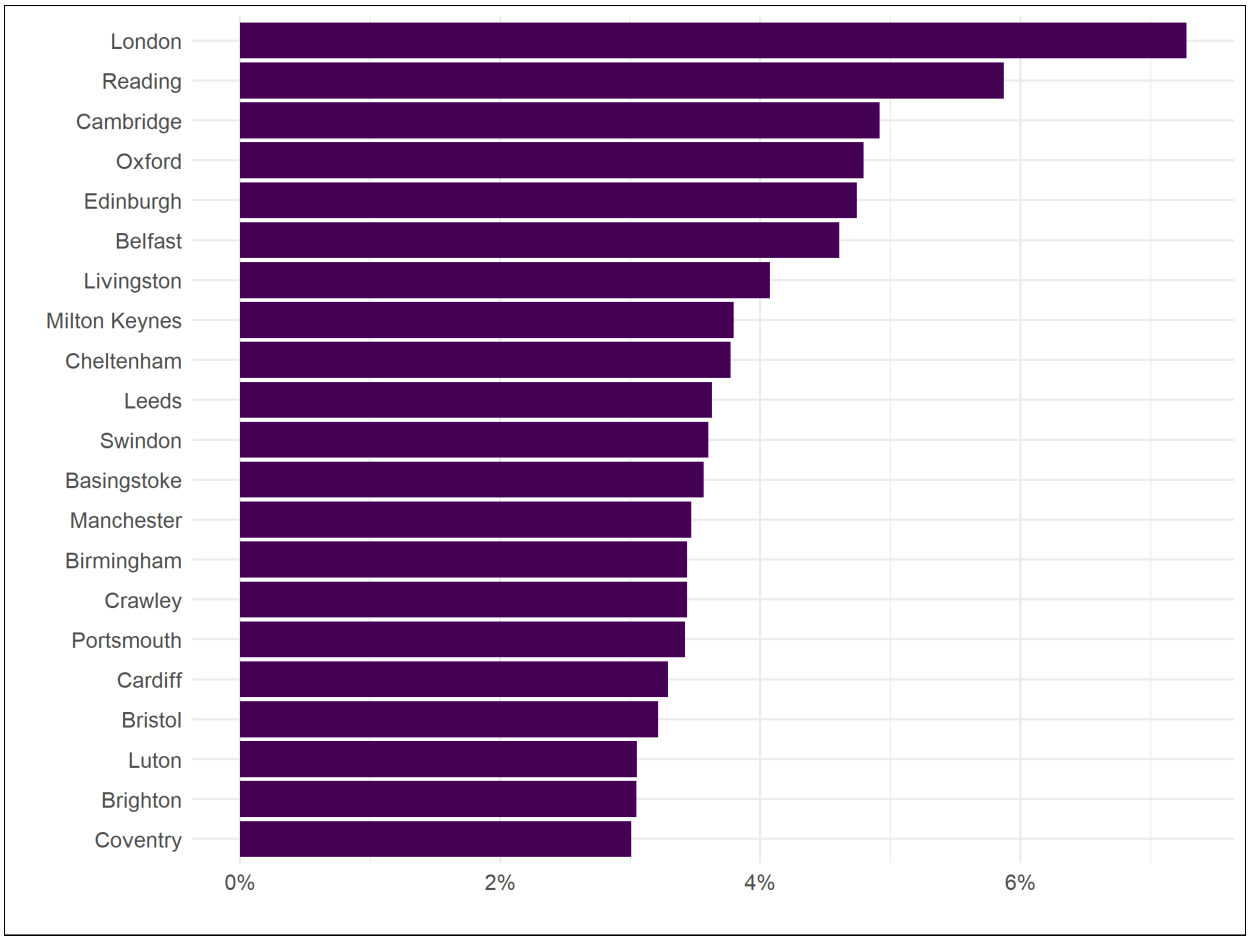


Source: Authors' calculation based on Lightcast data.

46. At a more granular level, data-intensive firms were concentrated in London and urban centres linked to research and industry (Figure 6). Large urban centres such as London, Edinburgh, Belfast, Birmingham, Cardiff and Manchester are among the top 20 cities with demand for data skills. Small urban centres with important links to bigger cities equally show a strong demand for data skills. Examples include Reading (5.9%), Basingstoke (3.6%), and Crawley (3.4%), close to London. Naturally, university cities such as Cambridge (4.9%) and Oxford (4.8%) display a high need for data-related jobs, as well as cities close to such university towns such as Cheltenham or Milton Keynes (3.8% respectively).

Figure 6. UK cities with the highest demand for data skills

Data intensity, per cent, 2021



Note: Data Intensity takes values between 0-100. Only cities with more than 50 000 inhabitants are selected.
Source: Authors' calculation based on Lightcast data.

Zooming into high data-intensive firms and multinationals

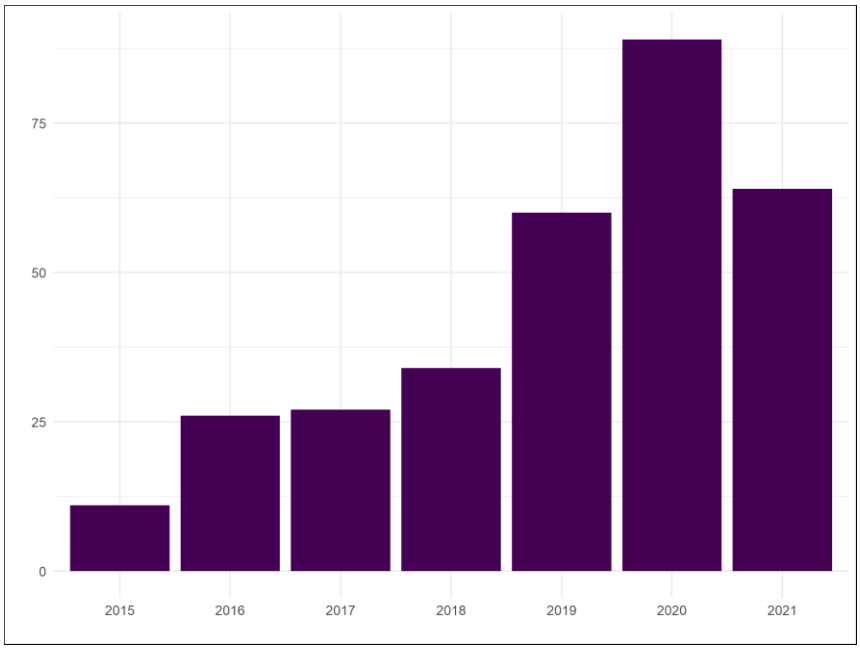
47. Given the high heterogeneity of data-intensive firms, the focus of this section is on two groups: first, the Top 20 data-intensive firms, whose data intensity is above 50%, and second multinational firms, defined as companies with at least two subsidiaries in other countries and data intensity above 10%. This could be useful to target policy directed to these groups or encourage firms to increase their data-intensity, given the link with productivity. It should be noted, however, that the composition of the Top 20 highly data-intensive firms can vary significantly from one year to the other, in line with a rapid evolving landscape of digital skill demand industries (Sostero and Tolan, 2022[35]). The findings for this group should thus be interpreted with care.

Highly data-intensive firms

48. Over time, the number of highly data-intensive firms follows the same pattern as the total number of data-intensive firms, with a rapid increase after 2017, likely following the push of artificial intelligence technologies into the market, a peak in 2020 and a subsequent fall in 2021 (Figure 7).

Figure 7. Number of highly data-intensive firms in the United Kingdom

Number of firms, 2015-2021



Note: Highly data-intensive companies are firms with a data intensity > 50% and those which posted more than ten vacancies per year were selected. The sample includes a total of 12 000 firms over the period 2015-2021.
Source: Authors' calculations based on Lightcast data.

49. Another common feature of most Top 20 data-intensive companies is that they were mostly hiring jobs with data analytics skills in 2021. Interestingly, database skills were less demanded. Reasons for that could be that firms want to move into advanced data analytics to gain more insights from their data, which would drive the demand for data scientists and analysts up. This does not mean that they do not engage in database-related activities, it may well be that the demand for those activities is already satisfied.

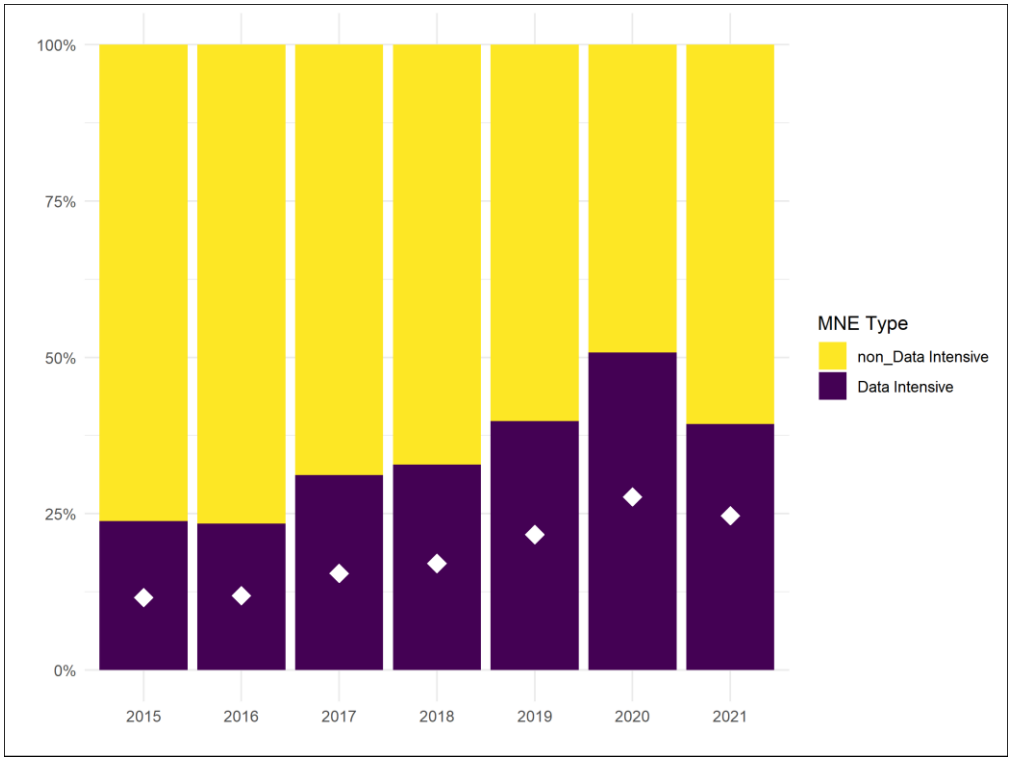
The number of data-intensive MNEs grew up until 2020

50. The number of data-intensive UK MNEs follow closely the pattern observed for both all data-intensive forms and highly data-intensive firms (Figure 8). Their number rose markedly from 2015 to 2020, before falling back somewhat in 2021. The main difference regards the share they represent compared to their non data-intensive counterparts. In 2020, in particular, data-intensive MNEs accounted for half of all MNEs covered in the analysis. The corresponding share was only 25% for all (MNE and non MNE) firms.

51. The demand for data skills amongst MNEs is highly heterogeneous. While many MNEs seek employees with data-related skill sets, a set of MNEs appears to heavily hire in the data entry domain. Others are mostly focused on advertising jobs in the database domain.

Figure 8. Share of data vs. non data-intensive MNEs, United Kingdom

Per cent, 2015 - 2021



Note: Data-intensive firms are those with a data intensity > 10% and more than ten vacancies per year. The white diamond corresponds to the share of data-intensive firms in the overall sample. MNEs are defined as companies with at least two subsidiaries in other countries. The sample includes a total of 12 000 firms over the period 2015-2021.
Source: Authors' calculations based on Lightcast and Orbis data.

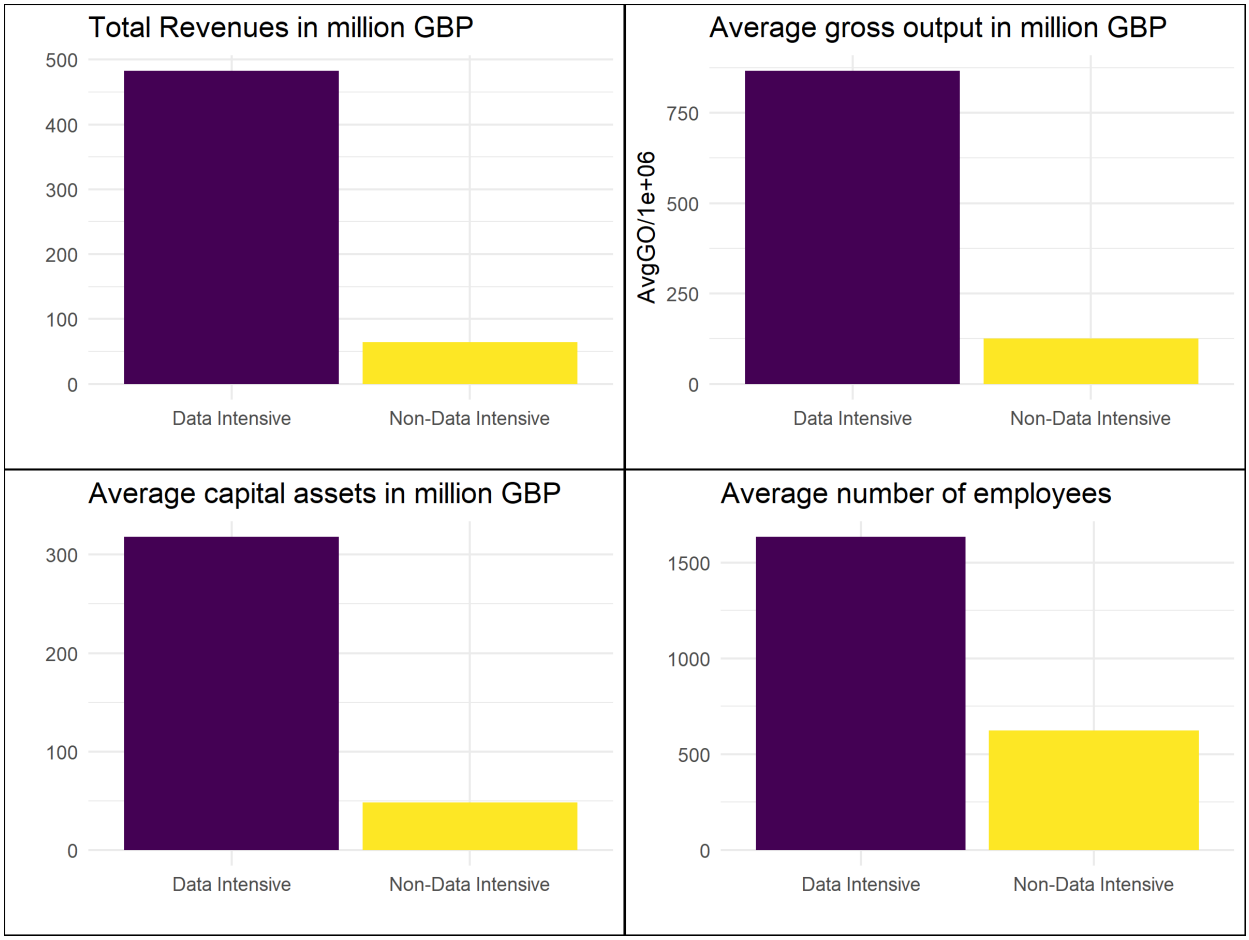
6. Performance of data-intensive firms

Data-intensive firms tend to employ more staff and be more capitalised than other firms

52. Data-intensive firms appear to be, on average, larger than non data-intensive firms along a number of dimensions. Their average number of employees is around 1 500, compared to only 700 for their non data-intensive counterparts (Figure 9). This is in line with Nguyen and Paczos (2020[36]) who find that data-intensive firms employ more staff than their non data-intensive counterparts. Data-intensive firms also appear to be endowed with about six times more capital assets than non data-intensive firms on average and to generate seven times higher revenues and higher gross output (Figure 9).

Figure 9. Comparison between data and non data-intensive firms, United Kingdom

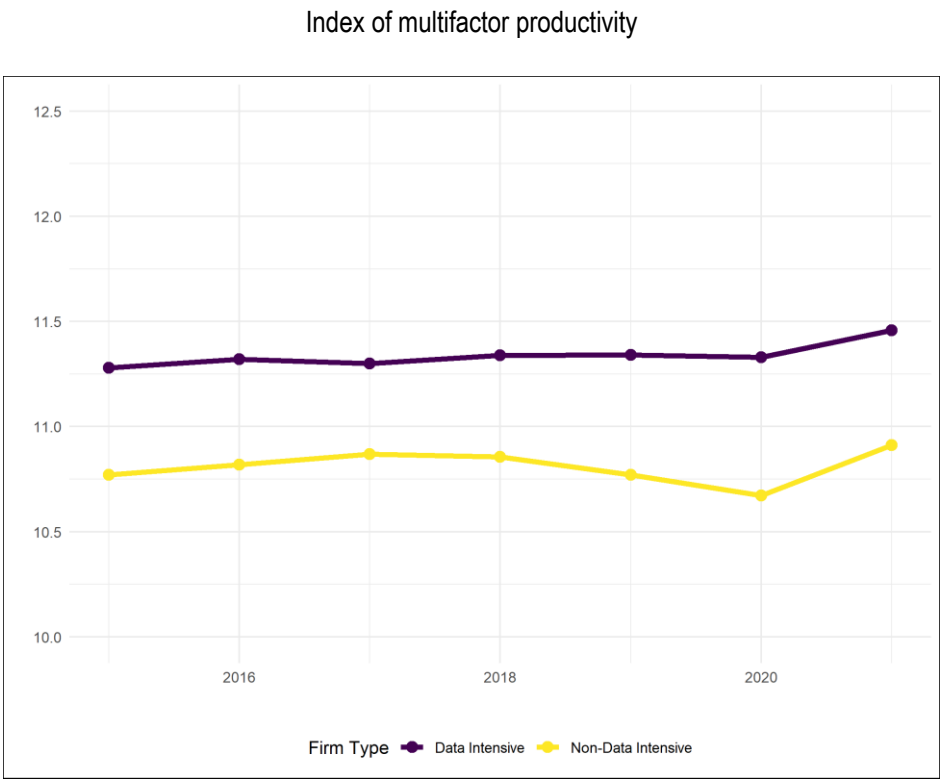
2021



Note: Only firms with a data intensity > 10% and more than ten vacancies per year were selected. The sample includes a total of 12 000 firms over the period 2015-2021.
Source: Authors' calculations based on Lightcast and Orbis data.

53. Reflecting partly a higher engagement in trade (see below), data-intensive firms also appear to be more productive than non data-intensive firms. Average multifactor productivity among data-intensive firms is found to be half a percentage point higher than those of the non data-intensive counterparts (Figure 10). Those findings do not account for confounding factors such as industry composition. This is in line with existing research (Sorbe et al., 2019[17]; Gal et al., 2019[11]; Brynjolfsson and McElheran, 2016[37]). The gap between the two groups of firms appears to have been broadly constant since 2015, with only a marginal increase at the start of the COVID-19 crisis.

Figure 10. Average multifactor productivity levels by firm type, United Kingdom



Note: Only firms with a data intensity > 10% and more than ten vacancies per year were selected. Multifactor productivity is computed as a residual following Woolridge (2009). The sample includes a total of 12 000 firms over the period 2015-21. Data are not corrected for industry fixed effects.
Source: Authors' calculations based on Lightcast and Orbis data.

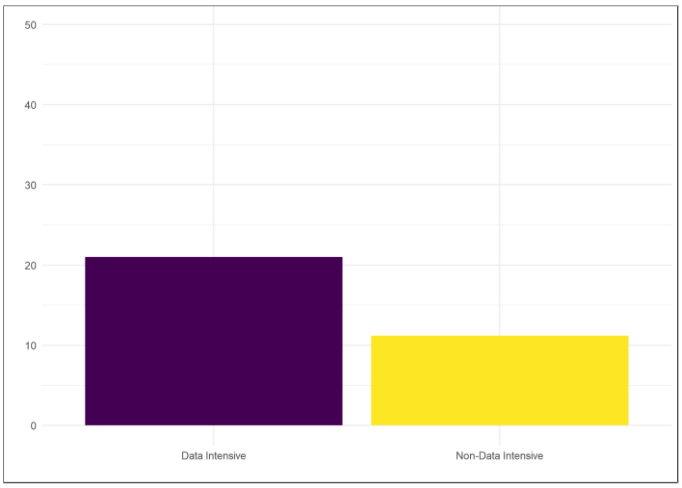
Data-intensive firms trade internationally more than other firms

54. International trade in the United Kingdom increased steadily in value terms from 2015 to 2019. It was hit by the COVID-19 crisis in 2020 and recovered somewhat in 2021. During the whole period, the United Kingdom was a net importer of goods and a net exporter of services and firms were deeply integrated into global value chains (GVCs). Although the size of the services sector has been growing since the 1990s and accounts for 80% of the UK economy both in terms of output and of employment, the focus of the subsequent sections is on goods which still represented almost half of total export values and 70% of imports (ONS, 2023[38]). Trade in goods contribute to the UK's engagement in GVCs. According to the latest OECD TiVA estimates, the UK manufacturing industry embodied more foreign value added in its exports (25%) than the services industry (11%) in 2020. Trade in goods also accounts for a large share of the current account deficit. When data allow, some indications will be provided on trade in services.

55. Data-intensive firms differ from other firms in their engagement in international trade (intensive margin). The share of exporters is around 20% in data-intensive firms, compared to only 12% for the non data-intensive counterparts (Figure 11). Looking at the industry breakdown, the differences between the two categories of firms are particularly marked in manufacturing, but also in a number of services industries such as transportation and storage as well as information and communication (Figure 12). There is little difference in finance and insurance activities, professional scientific and technical activities or industries where exports are limited such as education and human health.

Figure 11. Engagement in exporting of data and non data-intensive firms, United Kingdom

Share of exporting firms average across 2015-2021, per cent



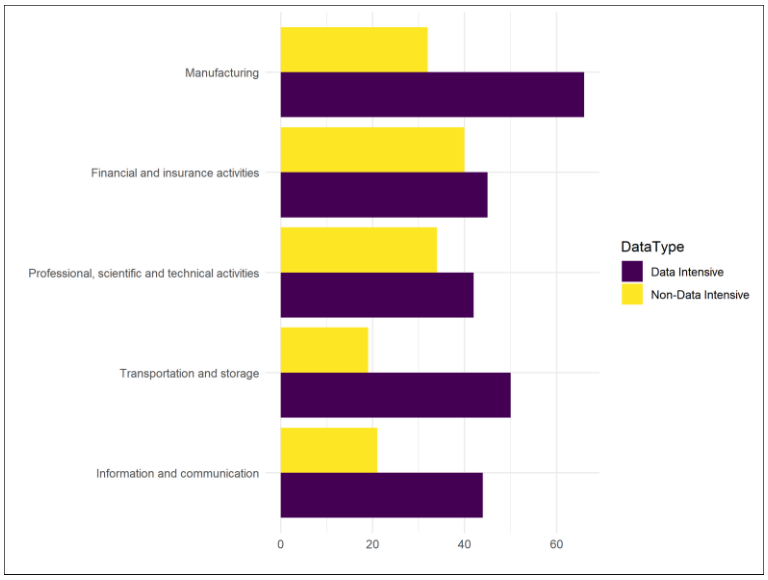
Note: Firms are selected based on a data Intensity > 10%, and more than ten vacancies posted per year. The sample includes a total of 12 000 firms over the period 2015-2021.
Source: Authors' calculations based on Lightcast and Orbis data.

56. Looking at the product level and focusing on trade in goods, there is no difference in the number of products traded, a metric for product diversification, between data and non data-intensive firms in either exports or imports (Figure 13). While a marginal decline in product diversification can be observed during the period 2015-2020, there is almost no change for non data-intensive firms.

57. More importantly, differences are visible in the categories of products traded by data-intensive and non data-intensive firms. A large number of firms specialised in trading machinery and equipment, electrical machinery and optical, photographic medical instruments are data-intensive (Figure 14).

Figure 12. Export intensity by industry, United Kingdom

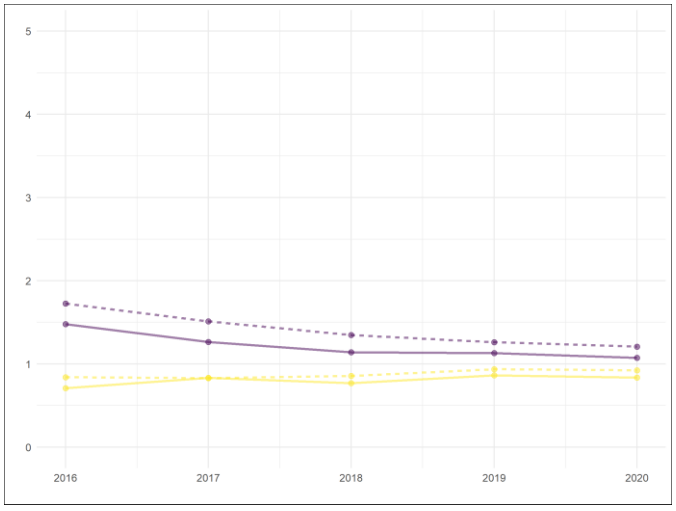
Export-to-output ratio, per cent, 2021



Note: Firms are selected based on a data Intensity > 10%, and more than ten vacancies posted per year. The y axis shows the export revenue divided by the total gross output across data/non data-intensive firms respectively. The sample includes a total of 12 000 firms over the period 2015-21.
Source: Authors' calculations based on Lightcast and Orbis data.

Figure 13. Average number of products traded, United Kingdom

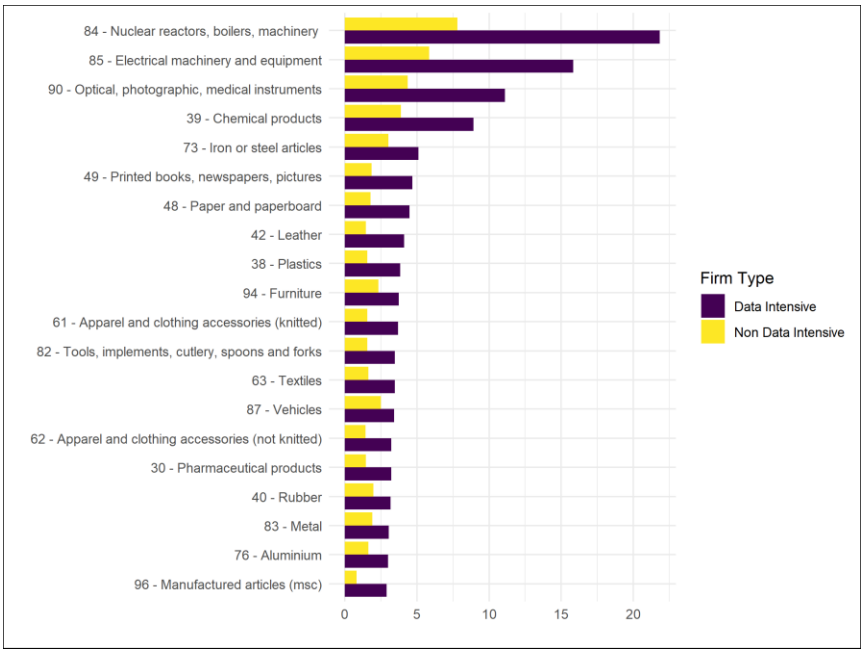
Average number of product types, 2016-2020



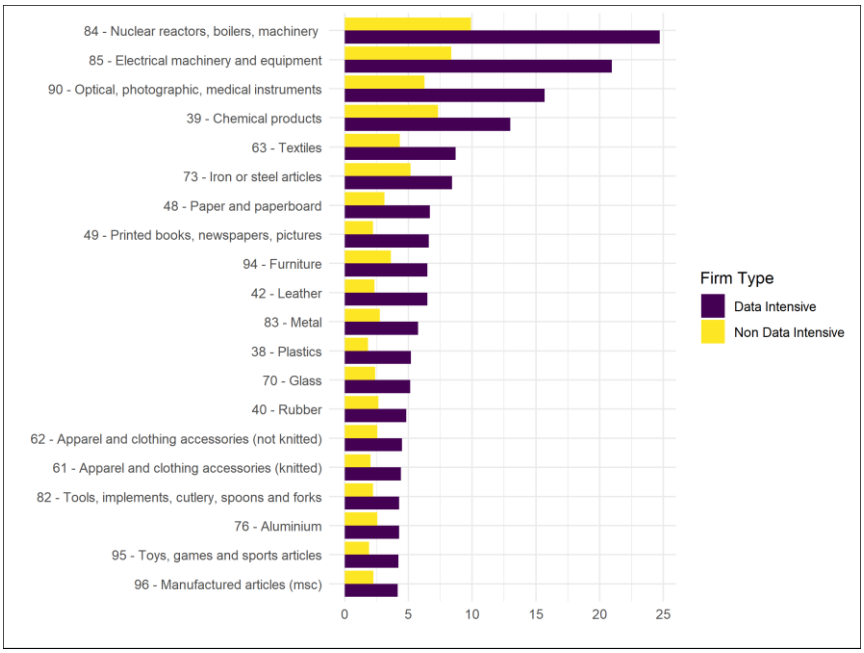
Note: The solid purple-line refers to data-intensive exports, the dotted purple line refers to data-intensive imports. The yellow lines refer to non data intensive firms (solid line for exports and dotted line for imports). The average of unique products traded is calculated by dividing the number of unique products grouped by trade flow and company type divided by the total number of companies active (data-intensive and non data-intensive respectively) in that year. The sample includes a total of 12 000 firms over the period 2015-21. 2021 had to be excluded due to a series break in methodology for this year.
Source: Authors' calculations based on Lightcast and UK trader dataset.

Figure 14. Share of data-intensive firms trading by product categories, United Kingdom

Panel A: Share of companies by exported products, per cent, 2020



Panel B: Share of imported products by company shares, per cent, 2020

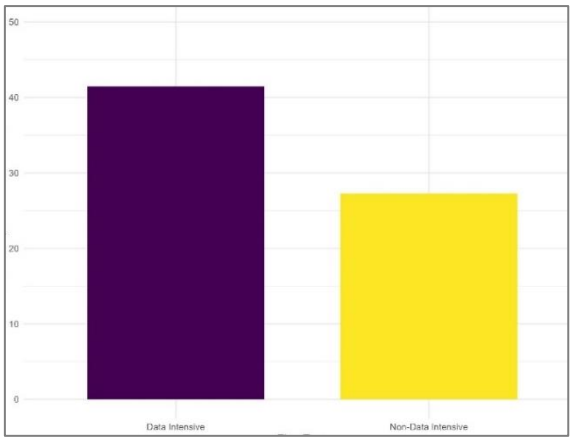


Note: The product categories are based on two-digit HS codes, classification system H5. The products are sorted by the share of data-intensive companies (descending order). The sample includes a total of 12 000 firms over the period 2015-2021.
Source: Authors' calculations based on Lightcast and UK Traders data.

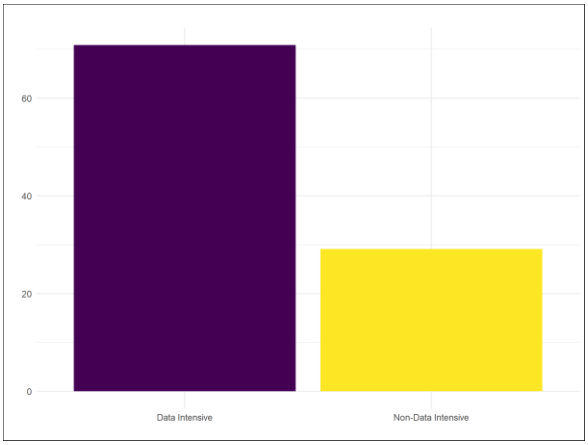
58. Data-intensive firms are also found to export more than non data-intensive firms (extensive margins). The aggregate share of exports in output of data-intensive firms exceeded 40%, as opposed to 27% for non data-intensive firms in 2021. Data-intensive firms received a higher part of their total revenues from exports (Figure 15).

Figure 15. Data-intensive firms received a higher part of their total revenues from exports

Panel A. Share of exports in total revenue, per cent, 2021



Panel B. Export-to-output ratio, per cent, 2021



Note: Firms are selected based on a data Intensity > 10%, and more than ten vacancies posted per year. The sample includes a total of 12 000 firms over the period 2015-2021.
Source: Authors' calculations based on Lightcast and Orbis data.

7. Conclusion

59. This paper combines several data sources, online job advertisement and two firm-level databases to shed lights on the nature and behaviours of data-intensive firms in the United Kingdom over the past decade. In addition to providing new insights on these firms, helping a better targeting of policy, a byproduct of the analysis is to offer a new methodology to match Lightcast data with Orbis, which is used extensively in economic analysis. This allows to provide a richer description of data-intensive firms, by adding information on their performances along a number of dimensions.

60. The main insights of the empirical analysis are as follows. The number of data-intensive firms, defined as those that are hiring data-related skills, increased sharply in the United Kingdom from 2015 to 2021, with a peak in 2020 and a small fall back in 2021. A similar pattern can be observed for highly data-intensive companies and data-intensive MNEs. Data-intensive firms are concentrated in a limited number of industries and regions. While data-intensive firms can be found in all size groups, the companies displaying on average the highest level of data intensity were medium sized in 2015 but are now small sized. Data-intensive firms tend to employ more staff and be more capitalised than non data-intensive firms. They are on average more productive, generate more revenues and engage more in foreign markets.

61. Although the paper offers new insights further work is needed to provide a more complete picture. In particular providing more information on trade in services, which accounts for an increasing share of trade in OECD countries, and in the United Kingdom in particular would be particularly useful. Another useful area would be to expand the analysis to other countries, to investigate whether UK data intensive firms differ significantly from their counterparts in other countries, in particular in the UK's main trading partners.

References

Ahmad, N. and P. Schreyer (2016), "Measuring GDP in a Digitalised Economy", *OECD Statistics Working Papers*, No. 2016/7, OECD Publishing, Paris, https://doi.org/10.1787/5jlwqd81d09r-en. [5]

Anderton, R., V. Botelho and P. Reimers (2023), "Digitalisation and productivity: gamechanger or sideshow?", No. 06, Science Po. [21]

Andrews, D., C. Criscuolo and P. Gal (2019), "The Best versus the Rest: Divergence across Firms during the Global Productivity Slowdown", *CEP Discussion Paper No 1645*, https://cep.lse.ac.uk/pubs/download/dp1645.pdf. [29]

Andrews, D., C. Criscuolo and P. Gal (2016), "The Best versus the Rest: The Global Productivity Slowdown, Divergence across Firms and the Role of Public Policy", *OECD Productivity Working Papers*, No. 5, OECD Publishing, Paris, https://doi.org/10.1787/63629cc9-en. [1]

Bajgar, M. et al. (2020), "Coverage and representativeness of Orbis data", *OECD Science, Technology and Industry Working Papers*, No. 2020/06, OECD Publishing, Paris, https://doi.org/10.1787/c7bdaa03-en. [47]

Berlingieri, G., P. Blanchenay and C. Criscuolo (2017), "The great divergence(s)", *OECD Science, Technology and Industry Policy Papers*, No. 39, OECD Publishing, Paris, https://doi.org/10.1787/953f3853-en. [9]

Berlingieri, G. et al. (2020), "Laggard firms, technology diffusion and its structural and policy determinants", *OECD Science, Technology and Industry Policy Papers*, No. 86, OECD Publishing, Paris, https://doi.org/10.1787/281bd7a9-en. [10]

Bernard, A. et al. (2018), "Global Firms", *Journal of Economic Literature*, Vol. 56/2, pp. 565-619, https://doi.org/10.1257/jel.20160792. [2]

Bloom, N. et al. (2018), "Really uncertain business cycles", *Econometrica*, Vol. 86/3, pp. 1031-1065, https://www.jstor.org/stable/44955229. [33]

Bloom, N., R. Sadun and J. Reenen (2012), "Americans Do IT Better: US Multinationals and the Productivity Miracle", *American Economic Review*, Vol. 102/1, pp. 167-201, https://doi.org/10.1257/aer.102.1.167. [18]

Bricongne, J., S. Delpleuch and M. Lopez Forero (2021), *Productivity slowdown, tax havens and MNEs Intangibles: where is measured value creation?*, https://publications.banque-france.fr/sites/default/files/medias/documents/wp835.pdf. [14]

Brynjolfsson, E. and K. McElheran (2016), *Data in Action: Data-Driven Decision Making in U.S. Manufacturing*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2722502. [37]

Calderón, J. and D. Rassier (2022), "Valuing the U.S. Data Economy Using Machine Learning and Online Job Postings", *U.S. Bureau of Economic Analysis | NBER Working Paper*, https://conference.nber.org/conf_papers/f159271.pdf. [45]

Calvino, F. et al. (2023), "What technologies are at the core of AI?: An exploration based on patent data", *OECD Artificial Intelligence Papers*, No. 6, OECD Publishing, Paris, https://doi.org/10.1787/32406765-en. [31]

Calvino, F. et al. (2022), "Identifying and characterising AI adopters: A novel approach based on big data", *OECD Science, Technology and Industry Working Papers*, No. 2022/06, OECD Publishing, Paris, https://doi.org/10.1787/154981d7-en. [16]

Cameraat, E. and M. Squicciarini (2021), "BurningGlass Technologies data use in policy-relevant analysis: An occupation-level assessment", *OECD Science, Technology and Industry Working Papers*, OECD Publishing, Paris, https://doi.org/10.1787/cd75c3e7-en. [23]

Carrasco, S. and R. Rosillo (2021), *Word Embeddings, Cosine Similarity and Deep Learning for Identification of Professions & Occupations in Health-related Social Media*, https://aclanthology.org/2021.smm4h-1.12.pdf. [41]

Cette, G., S. Corde and R. Lecat (2017), "Stagnation of productivity in France: a legacy of the crisis or a structural slowdown", *Economie et Statistique / Economics and Statistics*, Vol. 494-495-496, pp. 11-36, https://doi.org/10.24187/ecostat.2017.494t.1916. [12]

Corrado, C. et al. (2021), "New evidence on intangibles, diffusion and productivity", *OECD Science, Technology and Industry Working Papers*, No. 2021/10, OECD Publishing, Paris, https://doi.org/10.1787/de0378f3-en. [20]

Corrado, C. et al. (2022), "The value of data in digital-based business models: Measurement and economic policy implications", *OECD Economics Department Working Papers*, No. 1723, OECD Publishing, Paris, https://doi.org/10.1787/d960a10c-en. [6]

Coyle, D. et al. (2022), "Are digital-using UK firms more productive?", No. 2022-06, ESCoE. [13]

Coyle, D. and D. Nguyen (2018), *Cloud Computing and National Accounting*, https://www.escoe.ac.uk/publications/cloud-computing-and-national-accounting/. [15]

Crocetti, G. (2015), "Textual Spatial Cosine Similarity", https://arxiv.org/ftp/arxiv/papers/1505/1505.03934.pdf. [43]

Dernis, H. et al. (2015), "World Corporate Top RandD Investors: Innovation and IP bundles", No. JRC94932, JRC and OECD. [30]

Gal, P. (2013), "Measuring Total Factor Productivity at the Firm Level using OECD-ORBIS", *OECD Economics Department Working Papers*, Vol. No. 1049, https://doi.org/10.1787/5k46dsb25ls6-en. [28]

Gal, P. and J. Egeland (2018), "Reducing regional disparities in productivity in the United Kingdom", *OECD Economics Department Working Papers*, No. 1456, OECD Publishing, Paris, https://doi.org/10.1787/54293958-en. [34]

Gal, P. et al. (2019), "Digitalisation and productivity: In search of the holy grail – Firm-level empirical evidence from EU countries", *OECD Economics Department Working Papers*, No. 1533, OECD Publishing, Paris, https://doi.org/10.1787/5080f4b6-en. [11]

Grundke, R. et al. (2018), "Which skills for the digital era?: Returns to skills analysis", *OECD Science, Technology and Industry Working Papers*, No. 2018/09, OECD Publishing, Paris, https://doi.org/10.1787/9a9479b5-en. [19]

HM Revenue and Customs (2023), *Trade data - Bulk datasets information pack*, https://www.uktradeinfo.com/trade-data/latest-bulk-datasets/bulk-datasets-information-pack-includes-post-eu-exit-changes/. [24]

Jurafsky, D. and J. Martin (2023), *Speech and Language Processing*, https://web.stanford.edu/~jurafsky/slp3/. [39]

Kalemli-Ozcan, S. et al. (2023), "How to Construct Nationally Representative Firm level Data from the Orbis Global Database: New Facts on SMEs and Aggregate Impications for Industry Concentration", *AEJ Macro Submission*, https://econweb.umd.edu/~kalemli/assets/workingpapers/Data_Paper_AEJ_Macro_Submission2023.pdf. [27]

Kalemli-Ozcan, S. et al. (2015), "How to construct nationally representative firm level data from the Orbis Global database: New facts on SMEs and aggregate implications for industry concentration", *National Bureau of Economic Research*, https://doi.org/10.3386/w21558. [26]

Manning, C. and H. Schutze (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, https://nlp.stanford.edu/fsnlp/. [44]

Nguyen, D. and M. Paczos (2020), "Measuring the economic value of data and cross-border data flows: A business perspective", *OECD Digital Economy Papers*, No. 297, OECD Publishing, Paris, https://doi.org/10.1787/6345995e-en. [36]

OECD (2022), "Measuring the value of data and data flows", *OECD Digital Economy Papers*, No. 345, OECD Publishing, Paris, https://doi.org/10.1787/923230a6-en. [3]

OECD (2020), *OECD Economic Surveys: United Kingdom 2020*, OECD Publishing, Paris, https://doi.org/10.1787/2f684241-en. [32]

ONS (2023), *UK Balance of Payments, The Pink Book 2023*, https://www.ons.gov.uk/economy/nationalaccounts/balanceofpayments/bulletins/unitedkingdombalanceofpaymentsthepinkbook/2023. [38]

Reddivari, S. and J. Wolbert (2022), "Calculating Requirements Similarity Using Word Embeddings", *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 438-439, https://doi.org/10.1109/COMPSAC54236.2022.00079. [42]

Schmidt, J., G. Pilgrim and A. Mourougane (2023), "What is the role of data in jobs in the United Kingdom, Canada, and the United States?: A natural language processing approach", *OECD Statistics Working Papers*, Vol. 2023/05, https://doi.org/10.1787/fa65d29e-en. [7]

Schmidt, J., G. Pilgrim and A. Mourougane (2023), "What is the role of data in jobs in the United Kingdom, Canada, and the United States?: A natural language processing approach", *OECD Statistics Working Papers*, No. 2023/05, OECD Publishing, Paris, https://doi.org/10.1787/fa65d29e-en. [4]

Sorbe, S. et al. (2019), "Digital Dividend: Policies to Harness the Productivity Potential of Digital Technologies", *OECD Economic Policy Papers*, No. 26, OECD Publishing, Paris, https://doi.org/10.1787/273176bc-en. [17]

Sostero, M. and S. Tolan (2022), “Digital skills for all? From computer literacy to AI skills in online job advertisements”, *JRC Working Papers Series on Labour Education and Technology*, https://joint-research-centre.ec.europa.eu/publications/digital-skills-all-computer-literacy-ai-skills-online-job-advertisements_en. [35]

spaCy (2023), *spaCy Model Architectures*, https://spacy.io/api/architectures. [40]

spaCy (2022), *Language Processing Pipelines*, https://spacy.io/usage/processing-pipelines. [25]

Statistics Canada (2019), *The value of data in Canada: Experimental estimates*, https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00009-eng.htm. [8]

Tsvetkova, A. et al. (forthcoming), *Representativeness of Lightcast web-scrapped vacancy data. An assessment for largest English-speaking countries*. [46]

Tsvetkova, A. et al. (2024), “How well do online job postings match national sources in large English speaking countries?: Benchmarking Lightcast data against statistical sources across regions, sectors and occupations”*, OECD Local Economic and Employment Development (LEED) Papers*, No. 2024/01, OECD Publishing, Paris, https://doi.org/10.1787/c17cae09-en. [22]

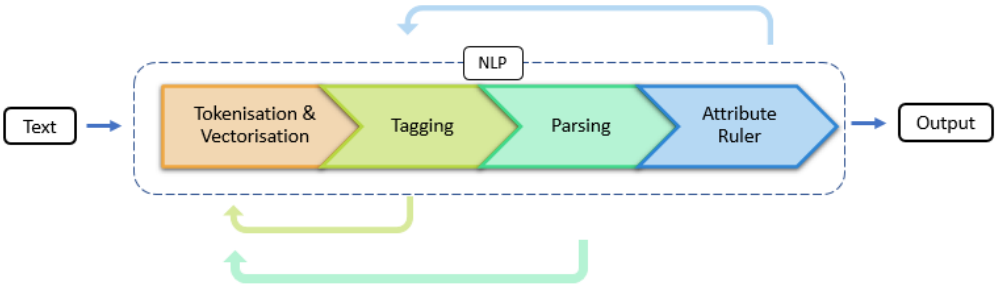
Annex A. Explaining the Natural Language Processing approach

This annex draws on Schmidt, Pilgrim and Mourougane (2023) and provides details on the NLP process used to derive data intensity of firms in the United Kingdom.

Deploying an NLP algorithm

This paper uses NLP techniques to extract relevant information from online job advertisements and construct a measure of data intensity at firm level. It relies on an open-source NLP pipeline provided by the 'spacy' python library (spaCy, 2022[25]). The pipeline combines different NLP models to efficiently perform advanced text processing operations in an iterative fashion (Figure B.1). The output of the pipeline can be used for tasks such as text classification or analysing phrase frequencies.

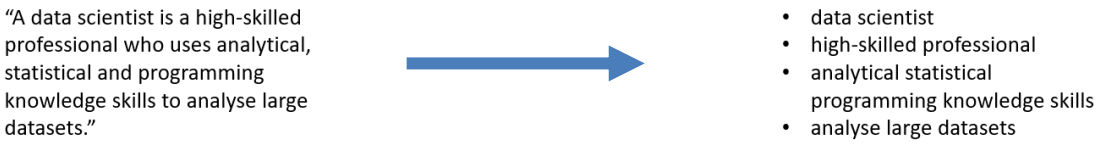
Figure A.1. A natural language processing pipeline



Source: Authors' illustration based on (spaCy, 2022[25]).

Tokenisation and vectorisation are the most relevant steps to the approach used in this paper. Tokenisation is the process of breaking down a text into smaller units, called tokens. These tokens can be words, phrases, or even individual characters, depending on the specific method used. In this paper, text from the job advertisements is processed and subsequently split into noun chunks (visible on the right in Figure B.2). This allows to harmonise the text for each job advertisement, as the chunks are consistent in their length and converted into lower case. It also automatically removes stop words (e.g. "the" or "and") and noise, such as special characters. This step is usually the first step in an NLP process and takes information from functions in the pipeline that identify the types of words (tagging), capture their grammatical structure (parsing) and customise rules for specific words in a sentence based on their characteristics (attribute ruler).

Figure A.2. Example of a tokenisation process



Source: Authors' illustration.

In a second step, called vectorisation, the pipeline transforms text into numerical vectors (embeddings) that can then be used to compare the similarity between noun chunks. The process of generating a vector for a text object is called vectorisation. Each word is assigned a unique vector that encodes its meaning and relationships with other words. The position and orientation of the vector in the multi-dimensional space capture the semantic similarities and differences between words. In this application, vectors are generated by a machine-learning model, and typically have 300 dimensions. The dimensions of the vector depend on the machine-learning model chosen that generates the vector (Jurafsky and Martin, 2023[39]).

Embeddings are generated such that two text tokens with similar linguistic usage will have vectors that are close to each other, i.e. the distance between them in vector space is small. However, the features of the embedding no longer have any meaning. In this paper, the spacy neural network model uses a machine-learning component "tok2vec" that learns how to produce vectors for each token to generate the embeddings (spaCy, 2023[40]).

The example below shows vectors for the noun chunks "data analysis", "data analytics" and "your information". The generated vectors for "data analysis" and "data analytics" are almost identical as the words represented are very close to each other compared to the noun chunk "your information".

```
Data analysis    = [1.5, -0.4, 7.2, 19.6, 3.1, ..., 20.2]
Data analytics   = [1.5, -0.4, 7.2, 19.5, 3.2, ..., 20.8]
your information = [7.5, -1.0, 7.2, 14.8, 2.8, ..., 19.0]
```

The resulting embeddings can be used to perform mathematical operations on the text, such as similarity measures. One advantage of the spacy pipeline is the availability of pre-trained models to generate embeddings that capture the meaning of the text more efficiently. The specific model has been pre-trained on large amounts of text data and can produce high-quality representations for a wide range of NLP tasks as shown in previous research (Carrasco and Rosillo, 2021[41]; Reddivari and Wolbert, 2022[42]). This saves time and computing power and allows for a quick deployment of spacy models on a variety of NLP tasks. The outcomes of the pre-trained model were checked on a validation set of over 10 000 manually classified job advertisements.

Classifying noun chunks into data entry, database and data analytics activities

After the processing by the NLP pipeline, the chunks are classified as data entry, database and data analytics activities based on three criteria, selected from a set of measures commonly used in NLP research and to ensure estimates are robust to small changes in thresholds (see below).

The cosine similarity, defined as the inner product of two vectors (x and y) divided by the product of their length (Equation 1), measures how similar the chunk is to the target word. When the similarity score is 1 then the two vectors are similar, or 0 then the two vectors are orthogonal. The threshold for a specific chunk to be data intensive is set at 50%, a threshold widely applied in the existing NLP literature indicating that the words are similar to each other (Crocetti, 2015[43]; Manning and Schutze, 1999[44]).

Equation 1: Cosine similarity measure

$$Cos\ (x, y) = \frac{x\ .\ y}{||x|| * ||y||}$$

x vector representation of noun chunk (x)

y vector representation of noun chunk (y)

The dispersion measure describes the frequency of a noun chunk in one occupation relative to the frequency of the same noun chunk in all job advertisements (Equation 2). The parameter is chosen based on a threshold x_c to adjust for biases in the sample of text data and avoid that the number of total noun chunks, which significantly varies across countries, drive the classification results of the algorithm. This ensures that the word is not widely used, such as common phrases e.g. 'your skillset' or 'your personal data', but instead likely to be a specific term only found in certain job advertisements, such as 'predictive modelling'.

Equation 2 Relative frequency measure

$$rel\_freq_n = \frac{Count\_by\_n\_occ/Count\_by\_occ}{Count\_by\_n/Count\_Total} > x_c$$

n noun chunk

occ occupation class

x_c data-specific threshold

The occurrence measure ensures that the noun chunks appear in one of three landmark occupations (namely data entry clerk, database administrator and data scientist). Those are chosen based on (Statistics Canada, 2019[8]; Calderón and Rassier, 2022[45]).

Finally, a noun chunk is classified as data intensive, if all three criteria of the classification rule are met. This is specified for data entry, database and data analytics respectively. The criteria for the three types of data-related roles are independent from each other, meaning noun chunks identified in one landmark occupation do not overlap with the others.

Classification rule

Noun chunk is data intensive IF:

Cosine similarity	> 0.5	AND
$\text{relative\_frequency}_{\text{noun chunk}}$	$> x_c$	AND
noun chunk	$\epsilon\ landmark$ (= data entry clerk, database administrator and data scientist)	

ELSE: 0

Deriving data intensity shares

The classified noun chunks at job level are used to construct an indicator of data intensity at firm-level.

In a first step the noun chunks are classified to identify whether a job is data intensive or not. If more than three noun chunks meet the classification criteria, the job advertisement is considered as a data-intensive job (1); otherwise, it is considered as non data-intensive (0). The breakdowns into data entry, database and data analytics activities are expressed as shares of total chunks classified and always sum to 1.

In a second step, the jobs are aggregated by firm using a weighted average (by the count of jobs advertised per firm) to calculate the share of data-intensive jobs in each grouping. Throughout all stages of aggregation, data integrity checks were implemented to ensure the consistency of classifications and the appropriate treatment of missing values in the data.