Uncovering hidden innovators: Insights through linked data

Linking administrative, survey, and alternative big datasets to identify alternative populations of innovators[1]

Dr Enrico Vanino

University of Sheffield

January 2026

Highlights

- It is possible to link administrative, survey, and alternative big datasets at the firm-level to identify alternative populations of innovative businesses in the UK.
- Depending on the definition of innovation applied, there could be between 35,000 and 140,000 hidden innovators in the UK, firms that are not normally considered as innovative by standardised definitions applied to official statistics.
- More than 75% of innovators are only captured by administrative data sources, such as HMRC and UKRI data, a huge proportion of businesses that would be hidden from the traditional measures.
- Accounting for hidden innovators increases the share of total employment in innovative firms, by around 20% particularly in high-tech sectors and in economically prosperous regions.
- Young businesses, those formed within the last 10 years, are more likely to be hidden from official statistics. Similarly small firms are more likely than large firms to be hidden innovators, as well as those in emerging industries.

Uncovering Hidden Innovators

Official measures of innovation traditionally used by statistical authorities, including research and development (R&D) investment, are linked to strict definitional requirements. These measures allow international comparability, alongside changes in time, but don't represent the full picture of innovation as captured with a broader definition.

[1] Acknowledgments: This project was supported by the British Academy Innovation Fellowship funding scheme. We thank the following organisations for their collaboration and for sharing their data: ONS, HMRC, UKRI, Innovate UK, IPO, the Welsh Government, Invest NI, Scottish Enterprise, INTERFACE, HIENT, and DataCity. We are grateful for the feedback received on previous versions of this work from GEOINNO 2024, DSIT, ONS, AAG 2024, ESCOE 2024, and the EC-JRC. Special thanks to Laura Requena and Beck Keane of the ONS Survey and Economic Indicators Division, and to the Innovation Research Caucus, for their support.

Incorporating wider data and definitions of innovation, alongside official measures, can enable a more complete understanding of the private-sector innovative landscape to support effective, evidence-informed policymaking.

Private sector innovation is often concentrated within large firms, typically operating in high-technology industries and situated in dense urban centres. Consequently, most policy attention and support for business research, development and innovation has focused on these types of enterprises.

These insights are largely based on traditional measures of innovation, normally focussed on a science-based model of formal R&D, typically derived from surveys, such as the ONS Business Enterprise Research and Development Survey (BERD), which may not capture all types innovation due to its structure and purpose. As economies evolve and shift, it is important to consider wider sources and data alongside these measures, as many commercial innovations do not draw on the latest scientific or technological breakthroughs – the official definition of innovation. Instead, they often take the form of “hidden” innovation, which doesn’t fit into the definitions captured in the official measures.

To address these issues, we linked the Business Enterprise Research and Development (BERD) survey data with 8 other firm-level longitudinal datasets, covering a broad range of R&D and innovation (RDI) activities between 2015 and 2022, building a novel linked database allowing a comprehensive assessment and analysis of innovation in the UK. This included a number of administrative data sources from public bodies, and an alternative data source using machine-learning analysis of company website text.

This integrated database enabled the identification of “hidden innovators”, firms not recognised as innovative according to the definitions used in the BERD survey, and the analysis of potential innovative activities for the entire business population. The database aggregated at the region (ITL2) and industry (SIC 20 macro-sectors) for the period 2015-2022 is available for all users and researchers on the IRC project website, together with additional details on the data sources used provided in the Data section of this report.

Identifying Hidden Innovators

We initially classify firms as innovators if they appear in any given year in at least one of the sources used, the broadest definition of innovation that could be considered. Figure 1 shows the proportion of innovators identified exclusively in each source, as well as those captured by multiple datasets.

The Business Enterprise Research and Development (BERD) survey identifies between 2% and 8% of all innovators under this broader definition, with coverage improving in 2022 following the ONS’s development of BERD to improve sample coverage. Over half of these firms are also identified in administrative datasets.

Administrative data consistently capture the largest share, over 80% of innovators in every year. By broadening the definition for this experimental analysis, we can see 75% in each year

are only present in the administrative data, demonstrating the additional value this source has in providing more granular information for the identification of firms involved in broader innovation activity.

Alternative sources, based on web-scraped data, identify on average 10% of innovators that aren't present in the survey or administrative data. This highlights the value of utilising alternative sources to identify "hidden innovators", alongside official measures to understand innovation more broadly within the business population.

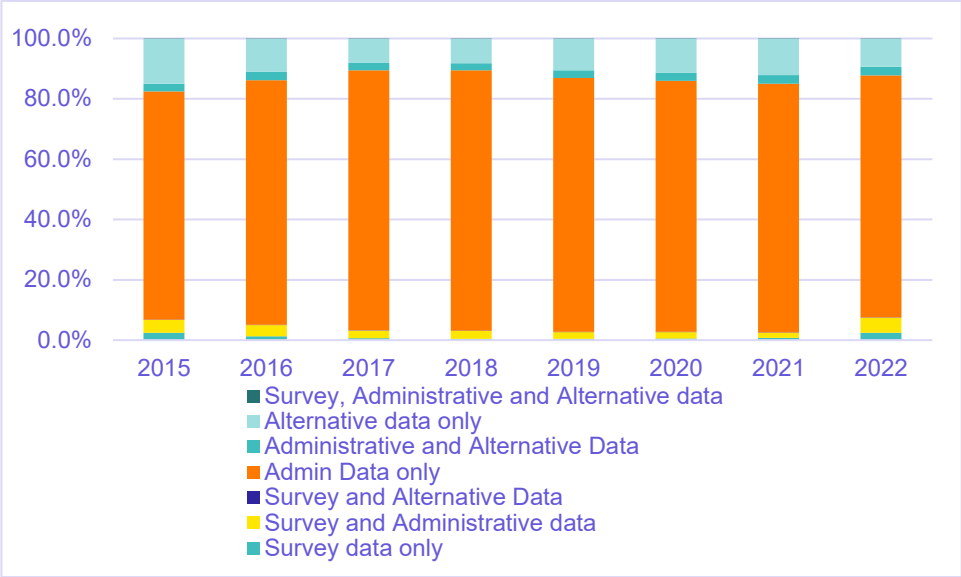

Figure 1: Share of innovative firms identified across data sources

Notes: Share of innovative firms in the UK identified using survey, administrative or alternative dataset over the period 2015-2022.

Alternative definitions of Innovators

By including businesses in any of the data sources considered as innovative, we identify over 140,000 hidden innovators. Over the same period, the Business Enterprise Research and Development (BERD) survey recorded only around 14,000 innovators, meaning that administrative and alternative data increase the identified population almost tenfold. Figure 2 shows the number of hidden innovators identified by merging the various datasets, firms that have never been recorded as innovators in the ONS BERD survey, but which appear in at least one administrative or alternative source.

This baseline definition of innovation may be overstating the extent of business innovations though, as administrative and alternative sources may classify firms as innovative even when they are not actively engaged in research, development or innovative activities. This risk is particularly acute for the HMRC R&D tax credit dataset, which accounts for nearly 90% of

innovators identified in administrative data, but has been acknowledged to be subject to error and fraud.[2]

To better understand these issues, we develop a number of alternative and gradually narrower definitions of innovators:

- All HMRC – the broadest baseline definition where if a business appears in any single data set at any point in time it is defined as an innovator. This provides an estimate of 141,000 hidden innovators
- HMRC+ – the narrowest definition, where a business is considered innovative only if it appears in the HMRC tax credit data and at least in another administrative dataset, for instance in public support schemes, or in the registration of patents and designs. This definition suggests that 39,000 firms are hidden innovators.[3]
- HMRC+Persistence – an alternative narrow definition, where a business is considered an innovator if it appears persistently in the HMRC tax credit data for a duration of at least 3 years or in any of the other administrative data sources. This definition estimates that there are 87,000 hidden innovators.

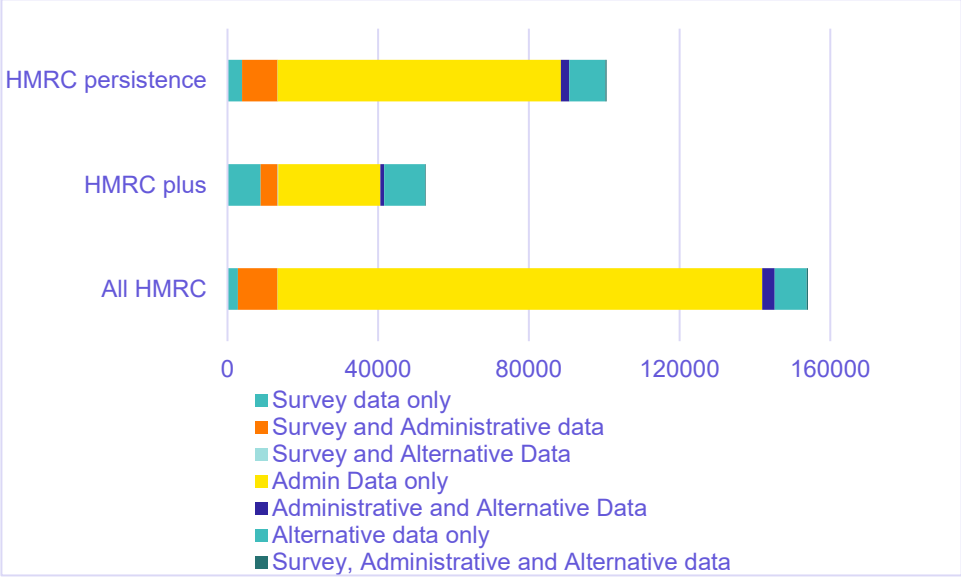

Figure 2: Total number of innovative firms identified across data sources

Notes: Total number of innovative firms in the UK identified using survey, administrative or alternative dataset between 2015-2022.

[2] See for instance the Approach to Research and Development tax reliefs 2023 to 2024 report.
[3] It should be noted however, that a company may validly claim for R&D tax-credit relief whilst not appearing in other sources, for example if the R&D is privately funded, or has not reached the point of patent applications.

Given the wide range of estimates of hidden innovators, we present the remainder of the analysis using both the broadest (All HMRC) and the narrowest (HMRC+) definitions. This approach provides upper and lower bound estimates of the significance of hidden innovators.

Characteristics of Hidden Innovators

To understand the limited overlap between survey, administrative, and alternative data sources, we investigated patterns that might explain why many innovative firms are absent from the Business Enterprise Research and Development (BERD) survey. By running a regression analysis, we estimated the likelihood of a firm being a hidden innovator, considering firm size, age, industry, and location. Figure 3 presents the results.

Small firms (fewer than 50 employees) are significantly more likely to be hidden innovators. The same holds for young firms (less than 10 years old), suggesting that start-ups, often among the most dynamic and tech-intensive businesses, are systematically underrepresented. These are both factors we now understand to be true in BERD before its redevelopment in 2022.

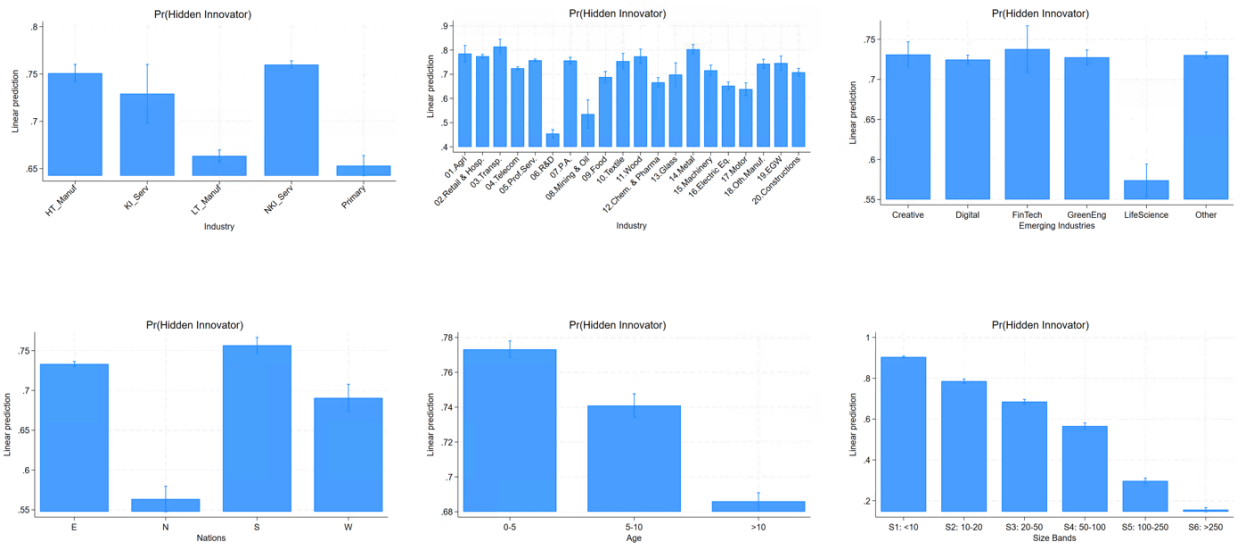

Figure 3: Factors affecting the likelihood of being a hidden innovator

Notes: Results of a regression analysis explaining the probability of being a hidden innovator based on several UK firms' characteristics over the period 2015-2022.

Location also matters. Firms in the devolved nations, especially Northern Ireland, are more likely to be identified as innovators in official statistics, whereas those in England are more likely to be hidden.

Industry is a particularly strong predictor. Firms in service sectors, both knowledge-intensive and non-knowledge-intensive, are much more likely to be hidden innovators. By contrast, those in primary industries and low-tech manufacturing are less likely to be hidden, possibly due to lower overall innovation levels. At a more detailed level, firms in transport, public services (including education and health), and construction are among the most frequently missed, followed by those in wood and metal manufacturing and agriculture. Emerging industries, such as start-ups in service sectors are especially likely to be hidden, notably in the creative industries and the fintech sector. In contrast, firms in life sciences are far less likely to be hidden innovators, indicating that official survey data performs relatively well in capturing innovation in high-tech sectors.

We conduct a similar analysis to examine the characteristics of firms identified as innovators solely through the alternative DataCity dataset. The results indicate that alternative big data sources are particularly effective at detecting innovation among micro-enterprises (fewer than 10 employees) and very young firms (less than five years old). This is especially evident in England, and more so in urban areas. From an industry perspective, alternative data sources are most likely to identify innovators in non-knowledge-intensive service sectors, with a disproportionate concentration in emerging digital and fintech industries.

Potential scale of hidden innovators RDI activities

Although when we broaden our definition, we identify a large number of hidden innovators, their importance could be negligible if these firms account for only a small share of employment or RDI investment, as survey data typically captures the largest firms and most significant innovators.

To address this, Figure 4 compares the share of employment in innovative firms identified by survey data with that identified using administrative sources.

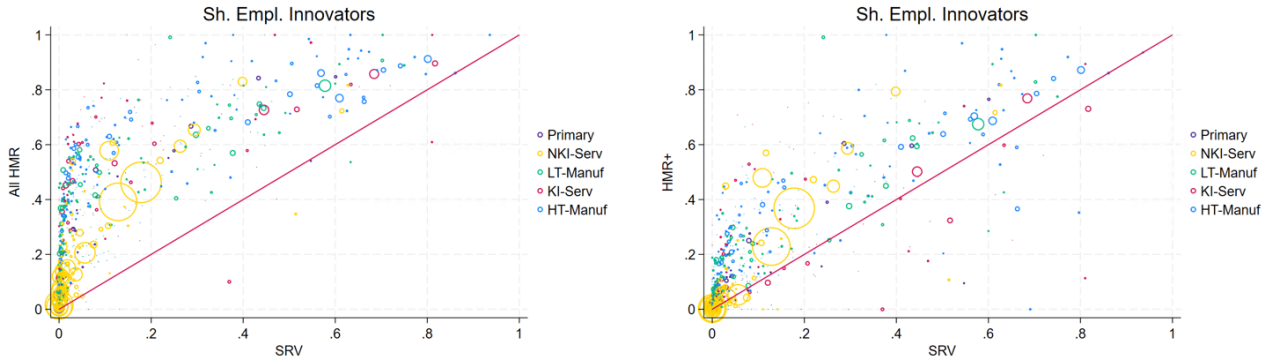

Figure 4: Share of industry employment in innovative firms identified using survey or administrative data sources.

Notes: Share of innovative firms' employment in the UK across BERD cells (industry-nation-size category) identified using survey or administrative dataset over the period 2015-2022.

Administrative data significantly surpasses survey data, capturing a greater share of employment across Business Enterprise Research and Development (BERD) cells (defined by industry, nation, and firm-size category). This is evident in cells above the 45-degree line, mostly those with lower employment (smaller dots in the diagrams), and particularly in high-tech manufacturing and knowledge-intensive service sectors. Notable examples include medium-sized firms in devolved nations, especially in chemicals and pharmaceuticals, electrical equipment, and R&D sectors.

We can repeat the same analysis looking specifically at emerging industries, finding that administrative data identifies far larger shares of employment in innovative companies in particular in the digital sector.

A similar pattern exists when you look at the regional distribution, as shown in Figure 5. Administrative data identifies a substantially larger share of employment in innovative firms across all regions, with West Yorkshire exceeding 40% of total employment in innovative businesses.

When comparing administrative and survey data, the largest differences are found in regions with sizeable business populations (represented by the larger dots in the diagrams). Notable examples include the central south of England, particularly Oxfordshire, Hertfordshire, Bedfordshire, and Buckinghamshire, and the central east, including Derbyshire, Nottinghamshire, and South Yorkshire. In these areas, administrative data captures between 12% and 14% more employment in innovative firms than the survey data.

These results suggest that survey data represents similar levels of innovation in rural or less populated areas as administrative data, but has lower levels when looking at economically active regions with larger business bases

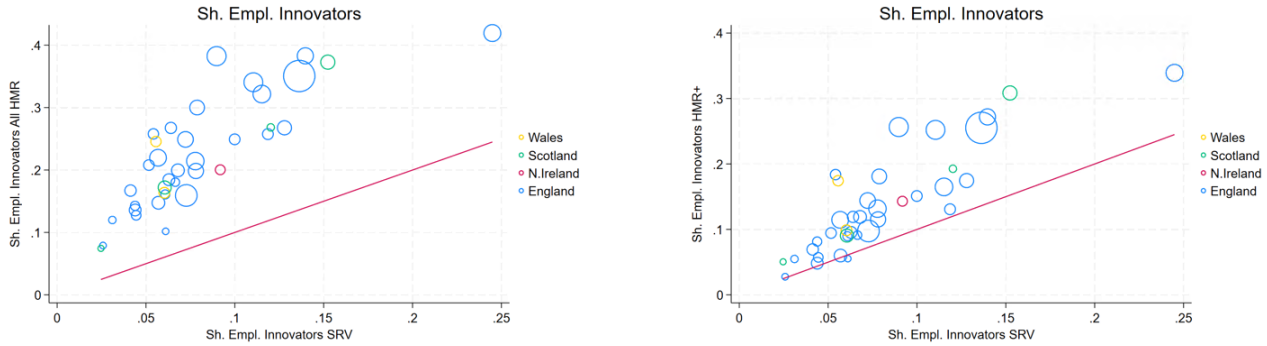

Figure 5: Share of regional employment in innovative firms identified using survey or administrative data sources.

Notes: Share of innovative firms' employment in the UK across ITL2 regions identified using survey or administrative dataset over the period 2015-2022.

We can also assess how the use of administrative data might affect estimates of business R&D expenditure in the UK, applying our different definitions of innovators and comparing the results with survey-based estimates. Figure 6 presents this analysis, showing total business R&D expenditure by data source and innovator definition.

We estimate R&D expenditure by identifying firms appearing in both BERD and HMRC records. We then calculate the ratio between their HMRC tax credit claims and BERD reported expenditure, for all firms within each BERD cell and then apply the cell average to all other HMRC-only observations. This provides an inferred estimate of R&D investment based on tax credit claims under each innovator definition.

Our results show that using the broadest definition (All HMRC) produces much higher R&D expenditure estimates, on average estimating R&D expenditure of £20bn more than estimates of R&D expenditure from BERD, while the narrowest definition (HMRC+) yields an estimated total R&D investment which is much more in line with the BERD previsions. This demonstrates the potential of combining survey and administrative datasets to produce broader estimates of business innovation expenditure in the UK, reflecting the wider definitions of innovators.

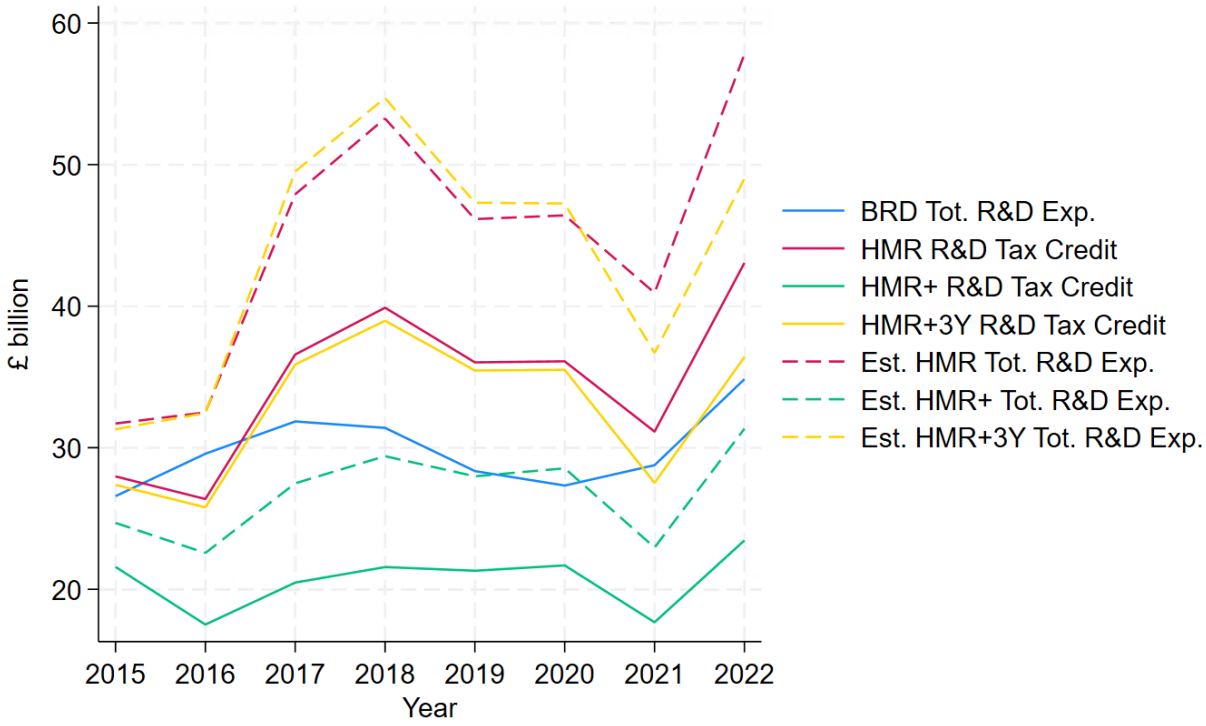

Figure 6: Total R&D expenditure according to different data sources and innovators definitions

Notes: Total or estimated R&D expenditure in the UK over the period 2015-2022 according to BERD and HMRC data sources.

Implications for statistics and policy making

By integrating administrative and alternative data sources with existing survey evidence we have been able to build a more complete, and accurate, picture of business innovation in the UK. The current reliance on survey-based measures, such as the ONS Business Enterprise Research and Development (BERD) survey, tends to capture only a narrow subset of innovative activity, in line with the definitional requirements. Although this is useful for international comparability and changes over time, it can overlook other emerging innovation by only considering these sources in isolation.

By linking survey, administrative, and alternative datasets, it is possible to identify a large set of "hidden innovators", which when combined with official statistics give a more complete picture of business innovation in the UK. Linking these data help give more detail on innovation that is otherwise not available from single data sources. By considering a broader definition of innovation, we can see the number of firms we capture across the linked data to be up to 10 times higher than the BERD survey sample, adding additional insight and granularity into the broader understanding of innovation. Administrative data, particularly HMRC R&D tax credit records, play a central role in understanding the coverage of innovation in a broader sense, while alternative data sources contribute unique insights into micro-enterprises, start-ups, and emerging industries.

A key implication of this study is the importance of modernising and strengthening the efforts to measure business innovation. By systematically linking administrative data and alternative sources under a broader definition, we can develop more detailed indicators of innovation and emerging patterns. Such integration would provide a richer evidence base at low cost, enabling more precise analysis of innovation activity across industries, regions, and firm types.

By linking data beyond the traditional R&D intensive firms as captured in BERD, this could allow policymakers to design targeted intervention that reflect the full diversity of the UKs innovative businesses, and reach businesses not currently doing R&D but likely to in the future.

The approach demonstrated here offers a blueprint for other OECD countries seeking to improve innovation analysis and understanding. By fostering stronger collaboration and data-sharing between statistical offices, government departments, and alternative data providers, we have shown the public good value of this type of data linking exercises, not only in the field of business innovations, but also in other fields of research.

Data

This study links multiple longitudinal micro-level datasets to analyse UK businesses' Research and Development and innovation (RDI) activities over an eight-year period (2015–2022). Using unique firm identifiers from Companies House, we integrate information from

statistical surveys, administrative sources, intellectual property records, and alternative big data, ensuring comparability of measures by following OECD's Frascati and Oslo manuals.

The analysis begins with the Office for National Statistics' (ONS) Business Enterprise Research and Development (BERD) survey, which samples 5,000–20,000 firms annually. BERD provides traditional metrics of R&D investment and innovation output, including R&D expenditure and employment in R&D roles.

Administrative data is then incorporated from several key sources. First, His Majesty's Revenue and Customs (HMRC) records identify all businesses receiving R&D tax credits—introduced in the early 2000s to incentivise R&D investment through corporation tax relief or cash payments for qualifying expenditure.

Additional national-level support data is obtained from UK Research and Innovation (UKRI) bodies. This includes Research Council funding—particularly from the Engineering and Physical Sciences Research Council (EPSRC) and the Medical Research Council (MRC), which emphasise industry collaboration—and Innovate UK, which delivers the majority of direct innovation grants and loans to firms.

One significant form of support comes from the Catapult network, a set of technology and innovation centres offering physical R&D facilities across sectors such as high-value manufacturing, digital technologies, and satellite applications.

Regional-level data is also included, covering all R&D and innovation funding provided by devolved authorities. Examples include SMART Wales (SMW), Invest Northern Ireland (INI), Scottish Enterprise (SCOTENT), Highlands and Islands Enterprise (HIENT), and INTERFACE, a regional technology adoption agency.

Innovation output data is enriched using UK Intellectual Property Office (IPO) records on intellectual property rights. In addition to patent registrations—commonly used as innovation indicators—the dataset includes other forms of IP, such as domestic and international registered designs. These measures are particularly valuable for capturing innovation in sectors where patents are less prevalent.

The authors also integrate alternative big data from the DataCity platform, which generates real-time, firm-level innovation indicators for the UK business base. DataCity aggregates information scraped from company websites with public and proprietary sources, applying AI and machine learning techniques to classify firms' innovation status. Each firm receives an innovation score, with "innovators" identified only at high confidence levels (above 75%). While this is a valuable supplementary source, we could not independently test the underlying algorithm, and could not fully confirm its validity.

Finally, all these datasets are linked to the Business Structure Database (BSD), which contains baseline firm information for the UK business population. This includes number of employees, turnover, ownership structure, location, and industry classification.

The full data spreadsheet can be accessed **via this link**.

Now that you have read our report, we would love to know if our research has provided you with new insights, improved your processes, or inspired innovative solutions.

Please let us know how our research is making a difference by completing our short feedback form via this link.

You are also welcome to email us if you have any questions about this report or the work of the IRC generally: info@ircaucus.ac.uk

Thank you

The Innovation & Research Caucus