Office for National Statistics

Regional UK business research and development, methods

Exploring models and data linkage to develop estimates of business expenditure on research and development at International Territorial Level (ITL) 1.

Contact:
Jim Hawkins
Subnational.Development@ons.gov.uk
+44 1329 444824

Release date:
17 April 2023

Next release:
To be announced

Table of contents

1 . Overview

The Office for National Statistics (ONS) is undertaking an ambitious transformation of its business statistics, including improvements to the Business Enterprise Research and Development (BERD) Survey. These improvements aim to enhance the coverage of small to medium-sized businesses that perform research and development (R&D) but have not previously been identified as R&D performers. These businesses were therefore not accounted for in BERD estimates.

We explore methods for attributing R&D expenditure of enterprises to UK regions. This work investigates how expenditure on R&D, by enterprises who did not disclose the location of where they perform R&D, can be attributed to UK International Territorial Level (ITL) 1 geographies.

The exploratory methods we look at could potentially be used to attribute business R&D expenditure to lower levels of geography. This supports the goals outlined in the UK Statistics Authority's Statistics for the public good five-year strategy and the Government Statistical Service's Subnational data strategy, which outline the ONS's ongoing aim to improve the granularity of UK statistics.

When carrying out this work, we found that outputs of a machine learning model developed to attribute business R&D expenditure to ITL1 regions were very similar to outputs produced by more simplistic methods, such as apportionment by employee counts. We believe there is scope to build upon this learning by further investigating the impact of research intensity on the regional distribution of business expenditure on R&D.

2 . Apportionment methods

In subnational statistics, regional apportionment refers to using a proxy metric to attribute a statistic to regions or other subnational geographies. This is generally applied when specific data pertaining to the regional distribution of a statistic are not available.

The Business Enterprise Research and Development (BERD) Survey incorporates data collected through several sources. Businesses identified as having the largest expenditure on research and development (R&D) are given a long-form survey, which allows businesses to provide details of the operating locations where R&D is performed. Other businesses are either provided with a short-form survey or their expenditure on R&D is imputed from other sources.

Neither short-form nor imputed data includes a breakdown of the operating locations where R&D is performed. Data on businesses who perform R&D in Northern Ireland are provided through a separate survey conducted by the Northern Ireland Statistics and Research Agency (NISRA). More information on the current methods used by the BERD Survey can found in our Business Enterprise Research and Development Survey QMI.

To attribute expenditure by businesses included in the short-form and imputed BERD data to UK countries and regions, three methods were used; the differences between the outputs of each method were then compared. The first method was to develop a machine-learning model that would classify businesses as either being headquarter-focused R&D performers, where R&D is likely to be performed at a centralised location, or non-headquarter focused R&D performers, where R&D expenditure is distributed across a business's operating locations.

Businesses categorised as headquarter-focused would have their R&D expenditure attributed to a single headquarters location, while non-headquarter focused R&D performers would have their R&D expenditure apportioned across all operating locations based on employee counts. More details and an explanation of the model are provided in Section 3: Approaches to modelling business R&D expenditure.

The second method used R&D intensity to classify businesses by industry as belonging to either "high-intensity R&D" or "low-intensity R&D" industries. We use the term "R&D intensity" as defined by the Organisation for Economic Cooperation and Development (OECD) in their insight paper on R&D intensity at industry level: how does UK compare with top performing OECD countries? Businesses in industries classified as having low-intensity R&D expenditure were attributed to a headquarters location, while businesses in industries classified as having high-intensity R&D expenditure were apportioned across operating locations by employee counts.

R&D intensity is defined as a business's expenditure on R&D divided by its turnover. Examples of high R&D intensity industries include the pharmaceutical and tech industries, whereas retail is considered a low R&D intensity industry.

The third method used for comparison is simple apportionment, where all expenditure on R&D by businesses included in the short-form and imputed BERD data is apportioned across all operating locations by employee counts.

In the current BERD methodology, all R&D expenditure by businesses featured in the short-form or imputed data is attributed to the headquarter locations of the businesses.

Our analysis shows that apportionment of business expenditure on R&D using more complex modelling techniques provides no statistically significant different regional expenditure estimates at International Territorial Level (ITL) 1 when compared with simpler apportionment methods based on employee count at business site locations. However, for some ITL1 regions, apportionment using modelling techniques does provide a statistically significant difference in regional expenditure when compared with apportionment using R&D intensity.

3 . Approaches to modelling business R&D expenditure

Our aim is to produce a machine learning model that accurately classifies businesses as either headquarter-focused research and development (R&D) performers or non-headquarter focused R&D performers. These classifications can then be used to decide whether in-house performed R&D expenditure by businesses is attributed to a single headquarter location or apportioned by employment counts across all operating locations.

The location where R&D is performed is provided by businesses who complete the Business Enterprise Research and Development (BERD) long-form survey, and these data are used to train the model. This assumes that businesses who complete the long-form BERD survey are representative of businesses included in the BERD short-form and imputed data. This is discussed in more detail in Section 6: Limitations and areas for further research.

A further aim of our approach to modelling business R&D expenditure is to compare model outputs with outputs using other more simplistic apportionment methods.

Pre-processing

Model outputs

To create a binary classification problem, an output label was created consisting of two classes: headquartered and apportioned. These output labels were created on the long-form data that the model was trained on.

A headquartered business performs its R&D at one main site, while an apportioned business is assumed to perform its R&D across all of its sites.

To prepare the data for modelling, the following pre-processing steps were taken:

1. The total intramural expenditure of the business at each R&D performing location was used. Intramural R&D expenditure refers to R&D performed by a business "in house".

2. For each business operation location where R&D is performed as reported in the BERD long-form survey responses, the proportion of the whole enterprise's R&D performed at that location was calculated.

3. The operating location with the highest intramural R&D proportion was assumed to be the R&D headquarters of the business.

4. The Pearson's correlation coefficient between the enterprise intramural R&D expenditure and the total number of local R&D-performing operating locations was determined to be negative 0.56. This was used to weight the R&D headquarters of a business with the number of R&D-performing operating locations of that business. This equation shows how the output was produced:

$$\begin{aligned}&\text{Weighted Intramural Expenditure}\\&\quad\text{at R\&D headquarters}\\&\qquad=\\&\text{Percentage of R\&D performed}\\&\quad\text{at R\&D headquarters}\\&\qquad\times\\&\frac{0.56}{\text{Number of R\&D performing local units}}\end{aligned}$$

5. The corresponding outputs were then binned into two bins separated by the median. The "apportioned" class was assigned to the values that fell below the median. The "'headquartered"' class was assigned to the values that fell above the median.

Through this approach, a binary classification problem was formulated. As the binning was performed around the median, there was no issue of class imbalance. However, this binning was not based on a pre-existing definition of what constituted a business that performs its R&D at its headquarters or across all sites.

Additionally, the non-long form businesses provided no information on the locations of their R&D-performing headquarters and other operating locations, nor could this information be extracted from the UK Government's Inter-Departmental Business Register (IDBR), a repository of business data. This means that the model assumes that the BERD long-form data is representative of businesses included in the short-form and imputed data. This is discussed in more detail in Section 6: Limitations and areas for further research."

Inputs

The inputs were created through a combination of feature (or variable) selection and data wrangling. Features were initially selected based on the correlations found in exploratory data analysis and feature importance analysis. Domain knowledge was also used to select features for the model. For example, we can reason that businesses that operate across many regions of the country are more likely to apportion their R&D expenditure, as well as businesses that have more employees and operating locations.

The total employee numbers of each enterprise were taken from the IDBR database and merged into the BERD long-form data. The other inputs used were the industry, the total number of operating locations, and the total number of International Territorial Level (ITL) 1 regions that each enterprise operated in. As all the inputs needed to be numerical, the industry variable had to be processed via one-hot encoding. The models used for this analysis only accept numerical inputs, and one-hot encoding is a technique by which each industry category was assigned a unique numerical value.

Additionally, the total number of ITL1 regions in which each enterprise operated was extracted from the IDBR as this contained all of the postcodes of each business's operating locations. These postcodes were converted into ITL1 regions using a postcode look-up tool. The total number of operating locations for each business was also extracted from the IDBR data.

Standard Industrial Classification (SIC) Code Assignment

SIC codes are 5-digit codes used to classify industries. Each digit in the code represents a further layer of granularity, meaning that removing digits creates a broader level of industry classification. This is useful when industry classifications are too specific and industry sample sizes are too small for the intended purpose.

Because of the wide range of SIC codes present, we processed SIC codes from 5-digit codes to 2-digit codes. We then grouped these codes into industry classifications. We found that four industry classifications made up 93% of all industry classifications in the long-form data, and were therefore the only industry classifications with sufficient representation in the sample. As a result, only businesses with these industry classifications were selected for the analysis.

The selected industry groups with their corresponding 2-digit SIC codes were:

- manufacturing: 10 to 33
- wholesale: 45 to 47
- information and Communication: 58 to 63
- professional, scientific, and technical: 69 to 75

Model Development

Methodology

Development of a machine learning model can be split into three tasks: training, validation and testing. The data for each model were split into a train-test split of 80% and 20%, respectively. Additionally, in order to take potential uneven class distribution into consideration, the output labels were stratified during the train-test split.

The training set was the data where different models were tested to optimize the model. Within this training dataset, the mean accuracy of each model was measured via 10-fold cross validation. Training accuracy is the accuracy of the model on the training outputs. The model is run on this dataset once.

10-fold cross validation involves splitting the dataset into 10 portions, with a different portion in each iteration being used as a validation dataset. The average of the model accuracy on this dataset over 10 iterations is taken to produce the validation accuracy.

The test dataset is a holdout dataset, which tests the generalisability of the model on unseen data. To avoid confirmation bias, the test dataset was chosen at random using the Python scikit-learn machine learning library. The test accuracy is the accuracy when the model that has originally been produced using the train dataset has been applied to the test dataset.

The criteria used to select candidate models included:

- training and validation accuracies
- minimal discrepancies between train, validation and test accuracies

Models with higher training accuracies than test accuracies are overfitting, and models with lower train accuracies than test accuracies are underfitting. Therefore, the models with the highest generalisability had minimal discrepancies between the train, validation and test accuracies.

Candidate models

Various algorithms can be tested for binary classification problems. The candidate models were logistic regression, support vector machine, random forest, K-nearest neighbours, and a stack of models. A stack of multiple models was also tested, with the outputs of all models going through a final logistic regression model. After the training and validation phase, support vector machine and logistic regression models were the main candidates that were fine-tuned based on the criteria listed above.

Training

Each model was initiated on default parameters. This provided an early indication of the models that needed to be fine-tuned and further developed. This is how the optimal model was determined. All models were tested on a variety of hyperparameters using a grid search cross-validation approach, which produced the hyperparameters that best fit the model.

Results

To select the optimal model, the metrics we used included:

- accuracy; the number of correct predictions as a fraction of the total number of predictions
- recall; the proportion of known class value (for example, "headquartered") being predicted as "headquartered"
- precision; the proportion of predicted class value (for example, "headquartered") actually "headquartered"
- F1-score; a balance between the precision and recall

K-nearest neighbours performed poorly at the initial training stage, so we did not use this model. The logistic regression and random forest models outperformed the support vector machine in validation accuracy. Small sample sizes increase the risk of overfitting in random forest classification models, therefore random forest classification was not deemed a suitable candidate. While the stacked model also produced relatively high validation accuracies, the complexity of the model made its use less appealing, since other simpler models also have high validation accuracies. To ensure generalisability, we used the logistic regression model to produce the final estimates.

During the model fine-tuning process, we kept the maximum iterations of the model as high as possible (1,000) to enable the model to fit the data more optimally. On the test dataset, the metrics in Table 1 show the results of the logistic regression model. The test accuracy closely corresponded with the validation accuracy of 0.86.

Table 1: Performance of a logistic regression model on the test dataset
The precision, recall, and F1 scores are all 0.86 or above, further confirming the robustness of the model

	Precision	Recall	F1-score
Apportioned	0.90	0.87	0.89
Headquartered	0.86	0.89	0.88
Accuracy			0.88

Source: Business Enterprise Research and Development (BERD) long-form survey, 2020 from the Office for National Statistics

4 . Different apportionment methods explained

This section describes the steps carried out for each of the three methods of apportioning business research and development (R&D) expenditure to UK countries and regions.

Simple apportionment

- For each enterprise, apportion intramural expenditure by employee counts at each site.

Apportionment using Organisation for Economic Co-operation and Development (OECD)-defined R&D intensity

- For each enterprise, classify by high R&D intensity and low R&D intensity.
- For high-intensity businesses, apportion intramural expenditure across all operating locations by employee counts.
- For low-intensity businesses, assign all intramural expenditure to headquarters.

Machine learning classification model

- For each enterprise, classify by headquartered and apportioned.
- For apportioned businesses, apportion intramural expenditure across all operating locations by employee number.
- For headquartered businesses, assign all intramural expenditure to headquarters.

R&D intensity categorisation

The OECD classified businesses into five intensity categories (PDF, 1.92MB):

- high
- medium-high
- medium
- medium-low
- low intensities

For the purposes of this analysis, these were reduced to two categories:

- high intensity – consisting of high and medium-high intensities
- low intensity – consisting of medium, medium-low and low intensities

This grouping was informed by the fact that high and medium-high-intensity businesses contributed 66.7% of business R&D expenditure in Great Britain from 2015 to 2019, with the remainder made up by the lower-intensity businesses.

5 . Comparison of apportionment methods

Analysis of businesses with unknown site expenditure

The businesses in the Business enterprise research and development (BERD) dataset can be grouped into two categories.

Category 1

Where the intramural research and development (R&D) expenditure can be assigned to known International Territorial Level (ITL1) regions:

- long form businesses
- single site businesses
- non-long form multi-site businesses where all operating locations are situated within the same ITL1 region

Category 2

Where the intramural R&D expenditure cannot be assigned to known ITL1 regions: non-long form multi-site businesses where operating locations are not situated within the same ITL1 region.

The in-house R&D expenditure for the unknown portion makes up 10.8% of the total in-house expenditure for all businesses in the dataset. Because of the similarity of the initial results on the whole dataset, a more granular analysis of only the unknown portion was performed. This was intended to understand if there were any discrepancies between the three methods in predicting the apportionment for the unknown businesses.

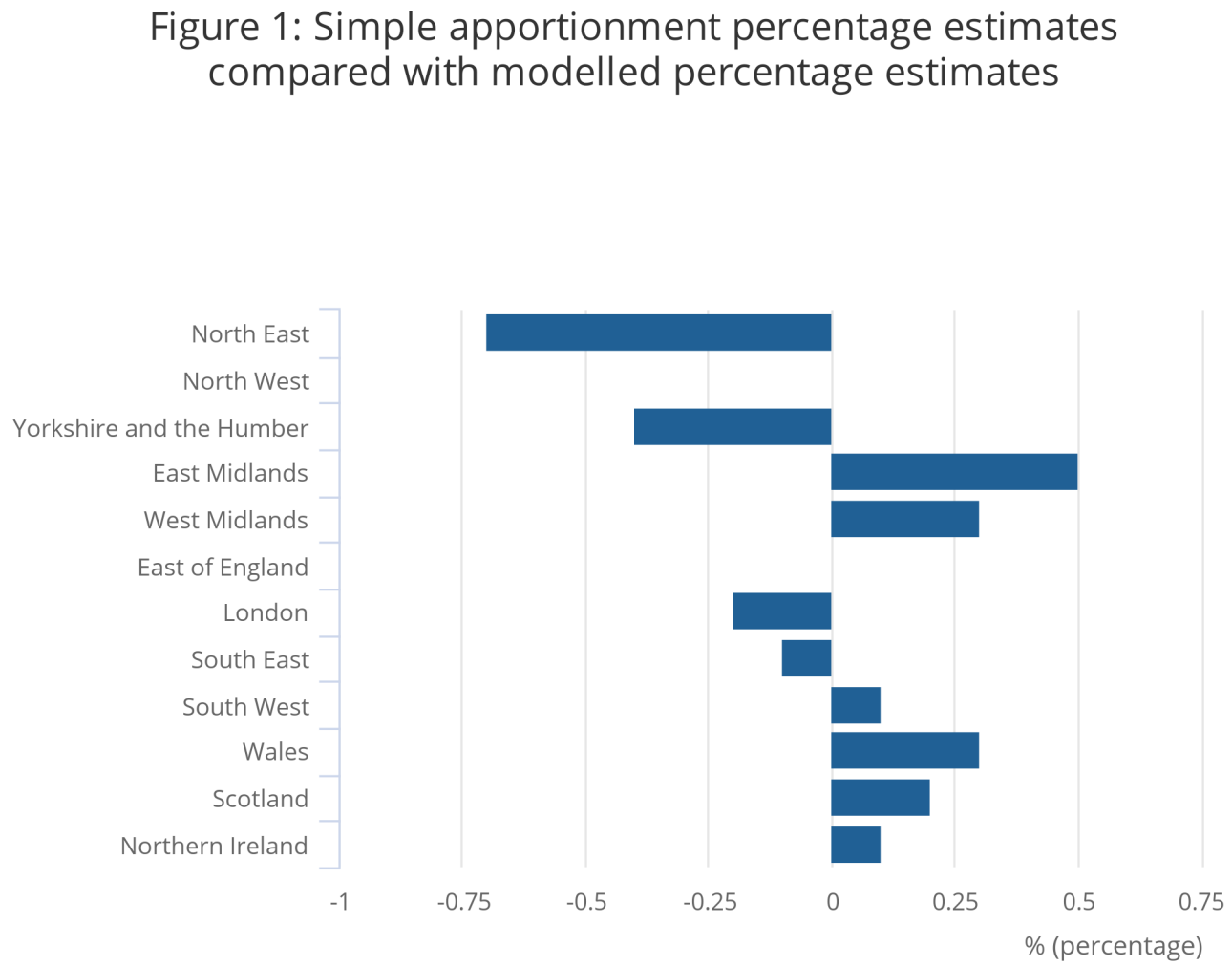
Analysis of previous years

The apportionment estimates for unknown portions of 2018 and 2019 were also produced using the three methods. The estimates were consistent across all years, providing further evidence for the robustness of methodology and results.

Headline results

On the 2020 BERD data, the three apportionment methods produced similar outputs. A close association can be seen between the model estimates and the simple apportionment estimates, as can be seen by the minimal percentage change. The R&D intensity estimates appear more distinguishable from the other two methods, with a greater percentage change overall. A two-sample t-test was found to have found no statistically significant differences between apportionment using modelling techniques and simple apportionment techniques. However, there was some statistical significance between estimates produced by modelling techniques and R&D intensity (see Table 2).

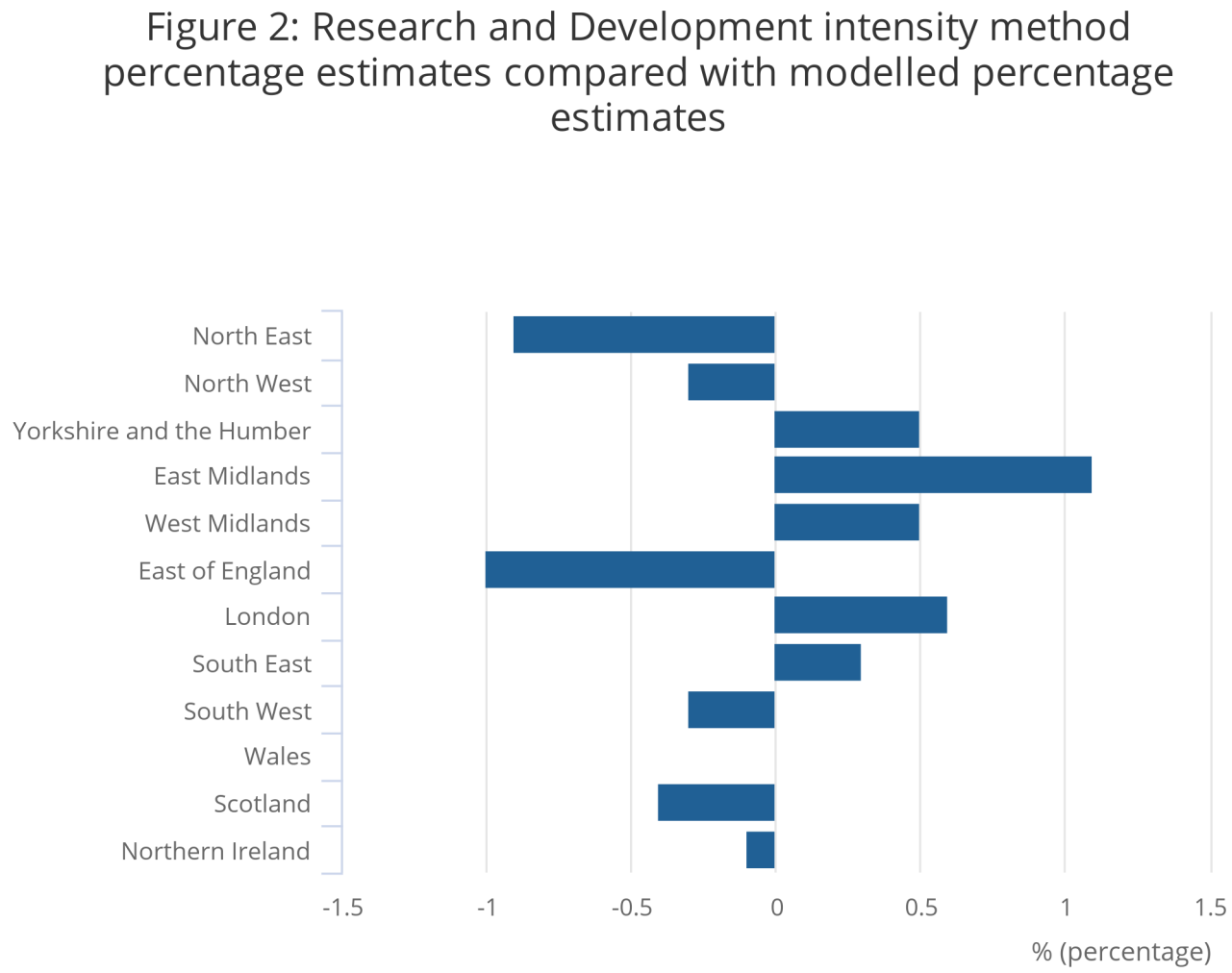
Figure 1: Simple apportionment percentage estimates compared with modelled percentage estimates



Source: Business Enterprise Research and Development (BERD) survey, 2020 from the Office for National Statistics

In Figure 1 there are minimal discrepancies from model estimates to simple apportionment methods for in-house R&D expenditure for businesses with unknown site expenditure in 2020. Only 2 of 12 regions have a percentage difference of 0.5% or more.

Figure 2: Research and Development intensity method percentage estimates compared with modelled percentage estimates



Source: Business Enterprise Research and Development (BERD) survey, 2020 from the Office for National Statistics

In Figure 2 there appears to be greater percentage change from the model outputs to outputs from the R&D intensity method. The East Midlands regions and East of England regions have the largest percentage differences of 1.1% and negative 1.0% respectively.

Significance testing

To measure whether the differences between the estimates produced by the different methods for each year were statistically significant, a two-sample t-test was performed. Outliers were removed and a Box Cox transformation was applied for data to fit a normal distribution -- an important assumption for a two-sample t-test. Each sample was also tested for equal variance. Other important assumptions such as the dependent variable being continuous, each sample having data obtained independently from each other, and data in each sample being chosen randomly, were all satisfied.

Hypothesis definition

Null hypothesis

There is no difference between mean intramural expenditure estimates by ITL1 region produced by the model and by the other methods.

Alternative hypothesis

There is a difference between mean intramural expenditure estimates by ITL1 region produced by the model and by the other methods.

Hypothesis testing was carried out and p values were observed to confirm whether the null hypothesis would be rejected.

Methodology

If there was no statistically significant difference between the estimates using the different methods, the simple method would be taken forward because of less resource cost. If there was a statistically significant difference, the model can be tested on future iterations of BERD.

The following steps were taken to test for statistical significance:

1. Compare mean from modelled estimate with means from other methods for each year, resulting in two comparisons per year: model and R&D intensity estimates, and model and simple apportionment estimates

2. Compare only the ITL1 regions that had the greatest differences in each method

3. Perform a two-sample t-test to check for statistical significance

Results

When comparing the ITL1 regions of greatest mean difference between model estimates and other estimates a significant difference is found between the R&D intensity and the model, but not when comparing the model with simple apportionment.

Table 2: Results of two-sample t-test for statistical significance between mean estimates produced by different methods

	Model and Intensity			**Model and Simple**		
	t-statistic	**p value**	**null hypothesis**	**t-statistic**	**p value**	**null hypothesis**
2020	-11.5	1.60e-28	rejected	0.26	0.79	accepted
2019	-11.6	3.60e-29	rejected	0.28	0.77	accepted
2018	-15.7	1.10e-51	rejected	-0.37	0.71	accepted

Source: Business Enterprise Research and Development (BERD) survey, 2020 from the Office for National Statistics

Using the ITL1 regions of greatest difference between the model and each of the other apportionment methods, the similarity between the model estimates and the simple apportionment estimates became clear, as did the differences from the R&D intensity-defined estimates. Analysis of 2018 and 2019 further confirmed these trends. The null hypothesis is rejected with a p value of below 0.05, indicating statistically significant differences in mean values. The t-statistic measures the size of the difference relative to variation in the sample data. A larger t-statistic is further evidence to reject the null hypothesis.

In summary, the statistical testing shows how the differences in in-house R&D estimates using modelling techniques is not statistically significant, but some statistically significant differences are present when modelling techniques are compared with R&D intensity-based apportionment.

6 . Limitations and areas for further research

The methods set out in this article describe several methods for regional apportionment of business research and development (R&D) expenditure for businesses in the Business enterprise research and development (BERD) sample frame who have not disclosed the locations where R&D is performed because of filling in the short-form survey.

Comparison of the methods has shown that outputs of a logistic regression model that demonstrates high model accuracy provides outputs that are not statistically significant from outputs of simple regional apportionment by employee counts. However, there is a statistically significant difference between modelled outputs and outputs from apportionment that incorporate research intensity.

Although modelled outputs showed a high accuracy score (0.88), further testing against empirical data collected by the transformed BERD survey will increase confidence in the robustness of the proposed methods. This is because the model is trained using data from respondents to the BERD long-form survey. The long-form sample is built of businesses with the most significant expenditure on R&D, therefore they are predominantly large multinational companies and may not be representative of the smaller businesses who complete the BERD short-form survey or whose R&D expenditure estimates are imputed.

In the transformed BERD survey some businesses who were previously included in the imputed or short-form data will receive the long-form. Once data from the transformed BERD survey are available it will be possible to compare the outputs from the three methods of apportionment used in this study with actual reported regional distributions of R&D expenditure for businesses previously included in the short-form or imputed BERD data.

This study explored the feasibility of using methods of apportionment to produce regional estimates of business R&D expenditure for International Territorial Level (ITL1) geographies, however, there is potential for the described methods to be used to produce estimates at lower levels of geography. The Government Statistical Service (GSS) Subnational Data Strategy outlines the ambition to produce statistics that are more granular and at lower levels of geography to meet user needs. Further research and testing of these methods using data from the larger sample produced by the transformed BERD survey could enable the production of business R&D expenditure estimates at more granular levels of geography in line with the aims outlined in the strategy.

7 . Related links

Business enterprise research and development survey
Bulletin| Released 22 November 2022
Spending and numbers employed on research and development by businesses in the UK, including data on sources of funds and regional spread. Produced by the Office for National Statistics (ONS).

Business enterprise research and development survey QMI
Methodology| Last revised 22 November 2022
Quality and Methodology Information for UK business enterprise research and development statistics, detailing the strengths and limitations of the data, methods used, and data uses and users. Produced by the Office for National Statistics (ONS).

Statistics for the Public Good
Publication| Released July 2020
The five-year strategy 2020 to 2025 of the UK Statistics Authority (UKSA).

Government Statistical Service (GSS) Subnational data strategy
Publication
Strategy outlining the GSS ambitions for subnational data.

UK public-funded gross regional capital and non-capital expenditure on research and development: financial year ending 2021
Publication| Released 17 April 2023
Experimental UK public-funded gross capital and non-capital expenditure on research and development (R&D) by International Territorial Level 1 (ITL1) geographies during the financial year ending 2021.

Measuring UK public-funded gross regional capital and non-capital expenditure on research and development
Publication| Released 17 April 2023
Methods used to produce experimental UK public-funded gross capital and non-capital expenditure on research and development, International Territorial Level 1 geographies, financial year ending 2021.

8 . Cite this methodology

Office for National Statistics (ONS), released 17 April 2023, ONS website, methodology, Regional UK business research and development methods