UK Government

UCL

RAF122/2223 Estimating clean energy R&D spending using tax credit data

Authors

Will McDowall and Hamid Nejadghorban, UCL Institute for Sustainable Resources.

Views expressed in this report are those of the researcher and not necessarily those of the UK Government.

HMRC Datalab disclaimer: This work contains statistical data from HMRC which is Crown Copyright. The research datasets used may not exactly reproduce HMRC aggregates. The use of HMRC statistical data in this work does not imply the endorsement of HMRC in relation to the interpretation or analysis of the information.



© Crown copyright 2026

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit nationalarchives.gov.uk/doc/open-government-licence/version/3.

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned.

Contents

Introduction 4
Background 4
Methods and data 5
Overview 5
Patenting firms 5
Matching clean energy patents to firms 5
Using company patent portfolios to estimate activity share 7
Non-patenting firms 9
TheDataCity 9
Machine learning approach 10
Results 14
Patenting firms 14
Estimates from non-patenting firms 15
Machine learning estimates of clean energy share of R&D in non-patenting firms 15
Estimated R&D spending by non-patenting firms 17
Combined estimates: patenting and non-patenting firms 18
Headline conclusions from our work 19
Trends in R&D spending and in the number of firms 19
UK clean energy R&D in context 20
Limitations 22
Relationship between patenting and R&D 22
Lag structure between R&D and patent applications 22
Missing or misleading geographic information in patent records 23
Challenges in matching patents to firms, and firms to tax records 23
Machine learning approach 24
Time horizon limitations related to non-patenting firms 24
Conclusions and next steps 25
Conclusions 25
Next steps 25

Introduction

This document provides a report of a project carried out by UCL for DESNZ. The project aimed to provide an initial estimate of UK private sector R&D expenditure related to climate mitigation, particularly in energy technologies, to inform net zero policy making. To meet this aim, we have used on R&D tax credits accessed through the HMRC Datalab.

This report provides an overview of the background, the approach taken, and it presents the results of the analysis data, along with comments on the limitations and appropriate next steps.

Background

DESNZ plays a critical role in the UK's ecosystem for energy R&D. With limited budgets, policy choices must seek to make the most effective use of policy support for energy innovation. Currently DESNZ must make such choices without a full picture of the UK's energy innovation system because of a key knowledge gap: the scale and distribution of private sector spending on clean energy research and development (R&D) is not known. Without such knowledge, policymakers are missing important information about how effectively policies are re-directing innovation away from unabated fossil fuels and towards clean energy.

The lack of knowledge of private-sector clean energy R&D spending is a more general problem globally. It has been repeatedly highlighted by analysts studying energy innovation processes, including in the recent Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), which argued that

> *"This inability to precisely quantify private investments in energy R&D leads to a patchy understanding of the energy innovation system" (Blanco et al. 2022).*

Similarly, in a review of the global energy R&D landscape Zhang et al. (2021) describe data on private energy R&D spending as *"spotty or non-existent".* R&D is a fundamentally important input into the energy innovation process, and a lack of data on R&D holds back analysis of energy innovation.

A handful of previous studies have attempted to estimate private-sector R&D spending on clean energy for the EU (Wiesenthal et al (2012); Breyer et al. (2013) and Pasimeni et al (2019)). Our work represents an advance on those earlier studies in several respects: First, our method is the first to both encompass a broad scope (rather than a small number of specific firms as in Wiesenthal et al. 2012) and extend beyond the population of patenting firms, thus addressing a key limitation of earlier studies, since not all technological innovations are patented, and R&D spending does not always generate successful outcomes that later appear as patents. Furthermore, previous methods have relied almost entirely on publicly listed firms that report their R&D spending, and have extrapolated results from those large corporate

actors to small and medium-sized firms. Some previous studies have also attempted to understand energy R&D spending by looking at the results of the UK's R&D survey by SIC code.[1]

Methods and data

Overview

Our approach is based on exploiting the HMRC's records of firm-level spending on R&D. When firms seek tax relief through the R&D tax credit system, they report their qualifying R&D expenditure to HMRC. This data provides a good basis for estimating R&D spending on clean energy activities, as take-up of R&D tax incentives is high, particularly amongst the most R&D-active firms.

During the period analysed, there were two schemes government used to provide incentives for R&D through the tax system: the Research and Development Expenditure Credit (RDEC), and the more generous R&D tax relief for small and medium-sized enterprises.[2] In both cases, when businesses make a claim, they were required to state their total eligible R&D expenditure, and this data is available to researchers through the HMRC Datalab.

In order to use R&D tax credit data to estimate clean energy R&D, two key steps are required. First, identification of firms that are carrying out clean energy R&D. Second, identification of the share of their total R&D activity that is related to clean energy, and the shares of activity associated with specific energy technologies. We have used two types of data to identify firms and their shares of activity: patents, and the net zero list of firms produced by TheDataCity. Our approach to each is described below.

Patenting firms

Matching clean energy patents to firms

When firms file applications at patent offices, patent examiners assign classification codes that describe the technology areas to which the invention belongs. In recent years, patent offices have also assigned patent codes specific to technologies related to climate change mitigation. These are the Y02 patent codes.

To identify firms involved in clean energy R&D in the UK using patent data, we combined several datasets.

These were:

[1] Office for National Statistics (2025) 'Business enterprise research and development, UK'.
[2] This has now changed: the RDEC and SME schemes have merged, and there is a new Enhanced R&D Intensive Support (ERIS) scheme.

1. PATSTAT, the global patent database managed by the European Patent Office. PATSTAT is the main source of patent information used by innovation researchers globally.
2. Companies House, which provides data on UK-registered companies.
3. WIPO patentscope, the database of the World Intellectual Property Organization
4. Orbis IP, a commercial database in which patents have been matched to firms.

Patent documents list both an applicant (typically a company or university, but sometimes an individual) and one or more inventors (individuals). Information on the address of the applicant and inventors is available in the patent document. We exploited this information to identify R&D activity that takes place in the UK, using PATSTAT to identify all Y02 patents that have either an applicant or an inventor with a UK address.

In order to match this to firms, we conducted a fuzzy string matching procedure to link these patents to UK firms listed in Companies House. This generated a total of 3,977 companies.

Our matching procedure involves an initial comparison of company names, followed by address-based secondary matching to resolve ambiguities. Initially, we converted all company names to lowercase to maintain uniformity and systematically removed common corporate terms such as “corporation”, “inc.”, “LLC”, and “limited”, among others, using regular expressions. Additionally, all special characters and punctuation were removed, leaving only alphanumeric characters to ensure consistent comparisons.

Subsequently, the cleaned datasets underwent fuzzy matching using Python’s “RapidFuzz” library. We employed the “Token Sort Ratio” algorithm, a robust method that accommodates variations in word order. The Token Sort Ratio algorithm involves:

1. Tokenisation: Splitting each string into separate words (tokens).
2. Sorting: Alphabetically sorting tokens to ensure matching irrespective of the word order.
3. Similarity Measurement: Comparing the sorted token strings using a method similar to the Levenshtein distance metric,[3] calculating the minimum number of single-character edits (insertions, deletions, substitutions) required to change one string into the other.

This metric produces a similarity score ranging from 0 to 100, indicating how closely the strings match. Only matches with scores equal to or exceeding 95 were accepted.

The primary advantage of this approach is its robustness to variations in word order, effectively matching company names regardless of token sequence. It is also flexible and efficiently handles variations in token placement, enhancing match accuracy. However, it may underperform with abbreviations or inconsistent naming conventions; we mitigated this risk by standardising company names and addresses.

[3] Levenshtein, V.I., 1966, February. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707-710).

Each company name from the PATSTAT dataset was matched against all entries from Companies House. Parallel processing was utilised by dividing the PATSTAT dataset into manageable chunks of 1,000 records each, significantly enhancing processing efficiency. In cases where multiple company names surpassed the similarity threshold, addresses provided a secondary resolution method. Standardising address formats (e.g., abbreviations, uniform postal code formats) was critical, similar to the procedure used for company names. Subsequently, we employed fuzzy matching (similar methodologies as name matching) to compare addresses and selected the record with the highest similarity score. The resulting matched dataset includes all necessary PATSTAT information along with the CRN from the Companies House database.

We supplemented the PATSTAT data with information from WIPO, the World Intellectual Property Organisation's database, which facilitates searches for inventor country addresses in PCT patents. Additionally, we searched for Y02 patents in the Orbis IP database, a commercial source that matches patents to firms. Combining the data from WIPO and Orbis IP resulted in 1,234 firms, of which 803 were unique additions to the dataset.

The resulting dataset contains 4,780 companies active in the UK that have filed Y02 patents.

Using company patent portfolios to estimate activity share

Many of the companies that have filed Y02 patents associated with a UK address are large companies with a diverse range of activities. Large energy, engineering, and manufacturing companies are often involved in both climate mitigation-related R&D, and other R&D projects. For firms with a patent filing record that includes both clean and non-clean energy patents, we have used the relative share of clean energy patent applications to estimate the share of R&D expenditure that is focused on clean energy (following Popp and Newell 2012). This provides an estimate of clean energy R&D spending among patenting firms, with a clear breakdown of spending by energy technology area (based on patent codes, which enable identification of clear technology categories). The mapping of patent codes to categories is based on that provided by the International Energy Agency[4]. The technology categories—and associated patent codes—are shown in Table 1.

Table 1. Technology categories and associated CPC patent codes

Technology category	Patent codes
All climate mitigation	Y02
Climate mitigation in buildings (including building heating and cooling technologies)	Y02B

[4] International Energy Agency (2022) 'Methodological notes for patent counts'.

Technology category	Patent codes
All mitigation technologies related to energy generation, transmission and distribution	Y02E
Wind power	Y02E10/70
Solar power	Y02E10/50
Nuclear power	Y02E10/30
Carbon capture and storage	Y02C
Climate mitigation in transport	Y02T
Industrial energy efficiency	Y02P10/00', 'Y02P20/00', 'Y02P30/00', 'Y02P40/00', 'Y02P70/00', 'Y02P80/00', 'Y02P90/00'
Smart grids and storage	'Y02E60/10', 'Y02E60/13', 'Y02E60/14', 'Y02E60/16', 'Y02E70/30', 'Y02E40/00', 'Y02E60/60', 'Y04S'
Hydrogen and fuel cells	'Y02E60/30', 'Y02E60/32', 'Y02E60/34', 'Y02E60/36', 'Y02E60/50', 'Y02T90/40'

For all the 4,780 firms that have filed a Y02 patent associated with a UK address, we collected all patents filed since 2010. We then used information from patent codes to infer the relative share of R&D activities on different technology areas (following the approach by Pasimeni et al. 2019[5]). For instance, for calculating the Y02 proportion:

$$Proportion\ of\ Y02\ Patents = \frac{The\ number\ of\ Y02\ patents}{Total\ number\ of\ patents}$$

We calculated this proportion for each firm and all years between 2010-2024, and for the categories of technology shown in Table 1. In using patent portfolios to estimate the share of R&D that has been focused on clean energy over time, we made several assumptions.

[5] Francesco Pasimeni, Alessandro Fiorini and Aliki Georgakaki (2019) 'Assessing private research and development spending in Europe for climate change mitigation technologies via patent data' World Patent Information, volume 59, article 101927

First, and perhaps most important, we have assumed that the relationship between R&D spending and patents is the same across technology areas within each firm. The propensity to patent per unit R&D reflects a) 'efficiency' of R&D (probability of producing a patentable invention per unit R&D) and b) strategic choices around IP protection (i.e. probability of filing a patent having developed a patentable invention)[6]. In this approach, we have assumed that these are properties of the firm. This is a limitation: there is evidence to show that patent propensity can differ across sectors[7], and it is likely that patent propensity may differ across different areas of clean energy technology.

Second, R&D projects take time, and those that are successful may not immediately generate patentable outcomes. We assumed that the typical lag between R&D spending and patent outcomes is two years. This lag is based on the literature, but it is important to note that in reality there will be a wide variation across different firms and projects.

Finally, many R&D-active firms do not file patents in all years. In order to use patent data to estimate the share of R&D by technology, we used a moving average approach. We multiplied the moving average of the patent ratio by the eligible R&D spending for each firm in each year. For example, if a hypothetical Firm A has, on average, 25% of its patents in Y02 technologies between 2018 and 2020 (years t to t+2), we applied this ratio to the firm's eligible R&D spending in 2017 (t-1) to estimate its R&D spending on Y02 technologies for that year.

Non-patenting firms

TheDataCity

Not all firms that conduct R&D file patents[8], and an estimate based solely on patenting firms would underestimate total clean energy R&D. This is offset somewhat by the fact that large firms account for more than half of spending on R&D in the UK[9], and large firms are much more likely to have patents than small firms.

In order to capture non-patenting firms, we have used TheDataCity's list of firms involved in the net zero supply chain. This enables us to identify additional firms that may be conducting R&D related to climate change mitigation, even where they are not filing related patents.

TheDataCity uses a machine learning platform, trained on the text of company websites, to identify companies involved in activities that are poorly represented by existing industrial classification codes. TheDataCity has worked with various stakeholders to develop a list of

[6] Anthony Arundel, and Isabelle Kabla (1998)
'What percentage of innovations are patented? Empirical estimates for European firms'
Research Policy, volume 27, issue 2, pages 127–141
[7] Anthony Arundel, and Isabelle Kabla (1998)
'What percentage of innovations are patented? Empirical estimates for European firms'
Research Policy, volume 27, issue 2, pages 127–141
[8] Bronwyn H. Hall, Christian Helmers, Mark Rogers and Vania Sena (2013) 'The importance (or not) of patents to UK firms' NBER Working Paper, number 19089 National Bureau of Economic Research
[9] Based on HMRC tax credit data

companies involved in net zero. We have brought the DataCity's list of net zero firms into the HMRC datalab to match with R&D tax records.

Some of these firms are specialists in net zero technologies. Many, however, are firms that pursue clean energy technologies alongside other activities. In the absence of patent data, it is difficult to infer what share of activity of such firms is related to net zero. In this project, we have run a machine learning approach trained on the subset of firms that are both in TheDataCity list, and which have filed patents. We have sought to identify whether this approach can support estimates of the clean energy activity shares of non-patenting firms.

We have then matched this list of firms, along with our estimates of the share of activity that is low-carbon, with HMRC's R&D tax credit file. This allows us to estimate the share of R&D spending that is related to climate mitigation, and specific technology categories.

Machine learning approach

The objective of this part is to develop a robust machine learning framework to predict the share of R&D spending allocated to clean technologies for non-patenting firms, a measure that cannot be directly observed since it is inherently defined for patenting firms as the ratio of clean technology patents to total patents. We use the observed ratio from patenting firms as a proxy to train our predictive model, which is then applied to the test set of approximately 20,000 non-patenting firms in DataCity database. Given the high dimensionality and complexity of the dataset, including numerous categorical variables, missing values, and panel-specific characteristics, we employ Extreme Gradient Boosting (XGBoost) algorithm (Chen and Guestrin, 2016)[10], an advanced gradient boosting algorithm known for its superior performance, computational efficiency, and flexibility in handling missing data. The methodology incorporates extensive data preprocessing steps, such as generating binary indicators, one-hot encoding of categorical variables, and decomposing multi-valued fields, as well as a hyperparameter tuning strategy to optimise model performance. This report details the theoretical foundations of the XGBoost algorithm, including its objective function and regularisation techniques to ensure the insights are accessible to both technical and non-technical audiences.

XGBoost is a widely used and efficient implementation of gradient boosting, an ensemble learning technique that combines the predictive power of many weak learners (typically regression trees) into a single strong model. In regression problems, XGBoost approximates the mapping from input features to a continuous target variable by summing the contributions of individual trees. The model's prediction for an instance x_i is given by:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \qquad eq(1)$$

[10] Tianqi Chen and Carlos Guestrin (2016), 'XGBoost: A scalable tree boosting system', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, pp. 785–794

Here, $\hat{y}_i$ is the predicted continuous output (i.e., the share of R&D spending on clean technology), and x_i denotes the input features (include variables such as "Sector Keywords", "SIC code", and "location"). Each $f_k(.)$ is an independent regression tree[11] characterised by its structure q (which defines the tree's splits) and leaf weights w.

XGBoost extends the basic gradient boosting framework by integrating:

1. Systems Optimisation:
    - Techniques like tree pruning, parallel processing, and hardware-level optimisations make XGBoost highly efficient.
2. Algorithmic Enhancements:
    - Regularisation (via $L1$ and $L2$ penalties) to avoid overfitting,
    - Sparsity awareness to efficiently handle missing values, and
    - A weighted quantile sketch algorithm to determine optimal split points.

Together, these enhancements allow XGBoost to deliver high predictive accuracy while using fewer computing resources.

The objective function that XGBoost minimises is:

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \qquad eq(2)$$

where:

- $l(y_i, \hat{y}_i)$ is a differentiable convex loss function that measures the difference between the predicted ($\hat{y}_i$) and the observed (y_i) outcome value. In regression, a common choice is the squared error loss:

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \qquad eq(3)$$

- $\Omega(f_k)$ penalises the complexity of the model that helps to smooth the final learnt weights to avoid over-fitting. $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$, where T is the number of leaves in the tree f, and w are the leaf weights, and γ and λ are hyperparameters that control the regularisation strength of the objective function.

Exploiting the additive nature of Equation (1), the objective at the t-th iteration is:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \qquad eq(4)$$

[11] Regression tree is a type of decision tree that each leaf represents a numeric value. In each leaf, it uses average value of observations that meet the leaf condition and could move to the next condition or shows the average value for the rest of observations

To optimise this function, XGBoost uses a second-order Taylor expansion of the loss function:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)}\right) g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) + \Omega(f_t) \qquad eq(5)$$

Where, g_i is the first and h_i is the second derivatives (gradient and hessian) of the loss function with respect to $\hat{y}_i^{(t-1)}$. Therefore, the optimum value for $\mathcal{L}^{(t)}$ is:

$$\tilde{\mathcal{L}}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\lambda + \sum_{i \in I_j} h_i} + \gamma T \qquad eq(6)$$

where, $I_j = \{i | q(x_i) = j\}$ is the instance set of leaf j. For regression tasks, the loss function $l(.)$ is typically chosen to be the squared error.

The performance of XGBoost can be enhanced by tuning its hyperparameters, which fall into three main categories:[12]

1. General Parameters:
    - booster: Determines the type of model to use (e.g., “gbtree” for tree-based models).
    - nthread: Sets the number of parallel threads for training.
2. Booster Parameters:
    - eta (learning rate): Shrinks the contribution of each tree to prevent overfitting.
    - gamma: Minimum loss reduction required for a split (see Equation 6).
    - max_depth: Maximum depth of the trees, controlling model complexity.
    - lambda ($L2$ regularisation) and alpha ($L1$ regularisation): Regularisation parameters that penalise complex models.
3. Learning Task Parameters:
    - objective: Specifies the loss function to be minimised (e.g., “reg:squarederror” for regression).
    - eval_metric: Defines the metric for model evaluation (e.g., RMSE, MAE, or R^2).

In regression problems, feature importance can be evaluated using several measures:

1. Weight: It shows how frequently each feature is used in splitting the data across all trees. Features with higher weight are involved in more splits, suggesting they are repeatedly helpful in partitioning the dataset.

[12] We used the “xgboost” package in Python, and the hyperparameters are referred to by their names as defined in the package.

2. Cover: It measures the average number of observations affected by splits on a given feature. A higher cover indicates that the feature's splits influence a larger portion of the dataset.
3. Gain: It reflects the average improvement in the loss function when a feature is used for splitting. Features with higher gain lead to larger reductions in prediction error and thus have a strong impact on model accuracy.
4. Permutation Importance: Measures the increase in prediction error when the feature's values are randomly permuted.

For regression tasks, model performance is typically assessed using:

1. Mean Squared Error (MSE): MSE is the average of the squared differences between the predicted values ($\hat{y}_i$) and the actual values (y_i). Squaring the errors penalises larger deviations more heavily, making this metric particularly sensitive to outliers. Lower MSE values indicate better model performance.
2. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE. By taking the square root, the error is expressed in the same units as the target variable, which makes the interpretation more intuitive. A lower RMSE suggests that the typical prediction error is small.
3. Mean Absolute Error (MAE): MAE calculates the average absolute difference between the predictions and the actual values. It treats all errors equally without penalising larger errors more than smaller ones, offering a straightforward measure of average error magnitude. A lower MAE indicates that, on average, predictions are closer to the true values.
4. Coefficient of Determination (R^2): R^2 quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. It is defined as one minus the ratio of the residual sum of squares to the total sum of squares. An R^2 value closer to 1 implies that the model explains a larger fraction of the variance, meaning better performance.

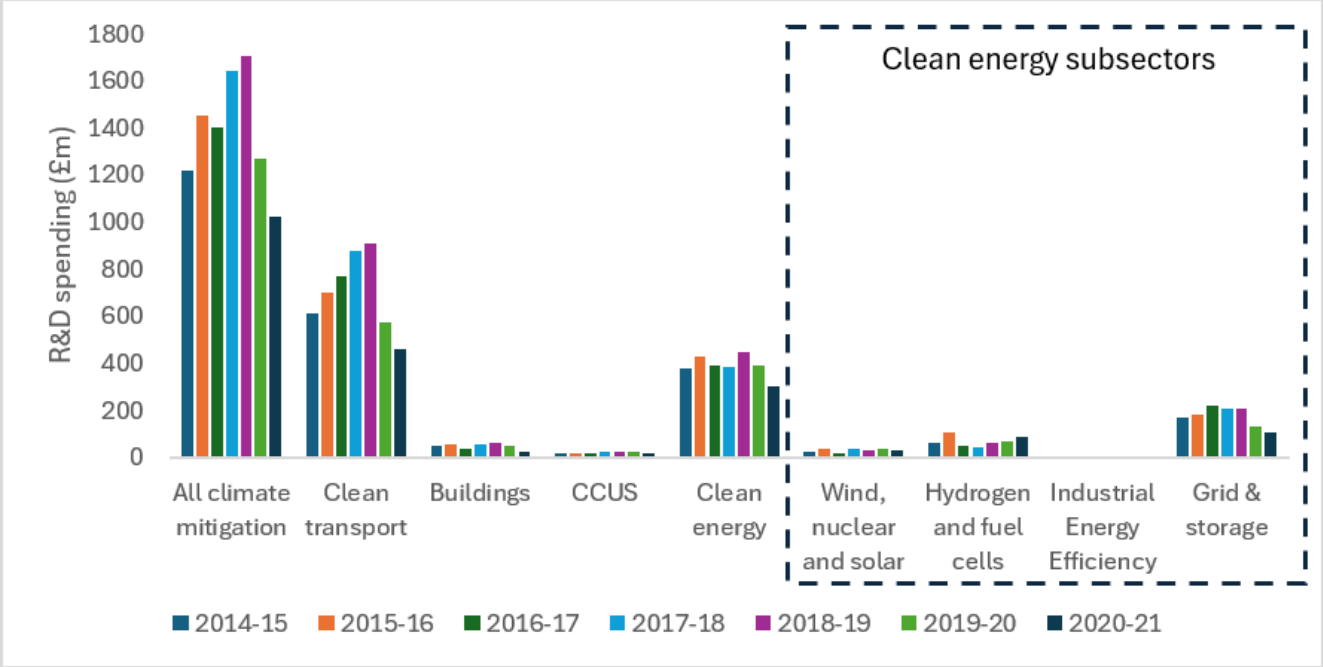
These metrics quantify how well the model's continuous predictions match the observed data.

Results

Patenting firms

The estimates of R&D among patenting firms is presented in Figure 1 and Table 2. The estimates suggest that spending on climate mitigation is dominated by energy and transport. Apparent declines in spending between 2018-2019 and 2020-2021 are largely driven by apparent reductions in R&D spending related to transportation.

Figure 1. R&D spending by technology area by patenting firms.



Note: 'clean energy' includes wind, nuclear, solar, and some of hydrogen and fuel cells, grids & storage, and industrial energy efficiency. Source: Calculations based on HMRC administrative datasets

Table 2. Estimated Private R&D spending in patenting firms (nominal £m)

	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21
All climate mitigation	1221.8	1455.7	1401.7	1645.9	1707.5	1267.3	1025.4
Clean transport	614.3	698.4	768.2	875.6	907.3	573	461.2
Buildings	47.8	54.1	35.1	54.9	59	47.2	23.3

	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21
CCUS	17.1	13.9	16	20.9	24.5	24.1	19.1
Clean energy	378.5	425.6	389.3	383.5	449.4	392.1	298.4
Wind, nuclear and solar	23.9	34.7	16.7	36.5	31.4	32.6	28.2
Hydrogen and fuel cells	63.2	106	48.6	44.9	62.9	69.7	85.1
Industrial Energy Efficiency	2.9	0.6	0.4	0.5	0.3	1.1	1.6
Grid & storage	166	178.9	221.2	206	207.4	133.3	104.2

Source: Calculations based on HMRC administrative datasets

Estimates from non-patenting firms

Machine learning estimates of clean energy share of R&D in non-patenting firms

This analysis reveals a promising yet moderate performance of our predictive model. Utilising a panel dataset comprising 718 matched unique firms (3,604 observations) between patenting and DataCity firms with 1,780 features, we trained an XGBoost regression model employing 5-fold cross-validation and 24 combinations of hyperparameters, resulting in 120 fitted models. The model performance for each technology is presented in Table 3.

Table 3. The performance of the best fitted model for each technology

Measure /Technology	*Y02*	*Y02B*	*Y02C*	*Y02E*	*Y02WSN*	*Y02HFC*	*Y02IEE*	*Y02G&S*
R^2	62.49%	23.55%	86.06%	60.62%	74.50%	80.20%	72.66%	60.74%

Note: this table presents the performance of the best fitted models which used to predict the patent ratio for non-patenting firms in order to estimate the share of R&D spending in clean technologies.

While the DataCity database originally contains 85 variables, after data cleaning and feature engineering, we generated a total of 1,780 features. The target variables are patent ratios for selected technologies, with a primary focus on the main outcome variable, referred to as the Y02 patent ratio. This ratio represents the proportion of a firm's R&D effort directed toward clean technologies, as inferred from the share of Y02 patents among all patents held by patenting firms.

To gain insight into which features most strongly influence the model's predictions, we examined feature importance from three different perspectives: Weight, Gain, and Cover. Looking across these importance scores, we found out that certain financial indicators (e.g., turnover, assets, intangible assets) and variables related to corporate activities, Real-Time Industrial Classifications (RTICs)[13], structure, and location appear consistently among the top predictors. Their prominence suggests these firm characteristics are pivotal in explaining differences in Y02 patent ratios.

While the current XGBoost model demonstrates moderate performance, there remains substantial potential to enhance predictive accuracy through further refinements in future work. Two primary avenues for improvement include advanced feature engineering and more comprehensive hyperparameter tuning.

Regarding hyperparameter tuning, exploring a broader and more detailed range of parameters, such as learning rate, tree depth, regularisation terms, and subsampling ratios, could result in a better-optimised model. Additionally, employing methods like Bayesian Optimisation or Randomised Search Cross-Validation may offer more efficient exploration of high-dimensional hyperparameter spaces compared to traditional grid searches, though these approaches still demand considerable computational resources.

In terms of advanced feature engineering, leveraging the panel structure of the dataset (characterised by repeated observations over time) by creating lagged variables or trend indicators may better capture temporal dynamics, further enhancing predictive capability.

However, it is important to note that extensive hyperparameter tuning can significantly extend model training durations, from minutes to potentially days per run, depending on the complexity of the search space and available computational resources. Furthermore, given the extensive hyperparameter space and the high dimensionality of the dataset (comprising 1,780 features), thorough tuning substantially increases computational demands.

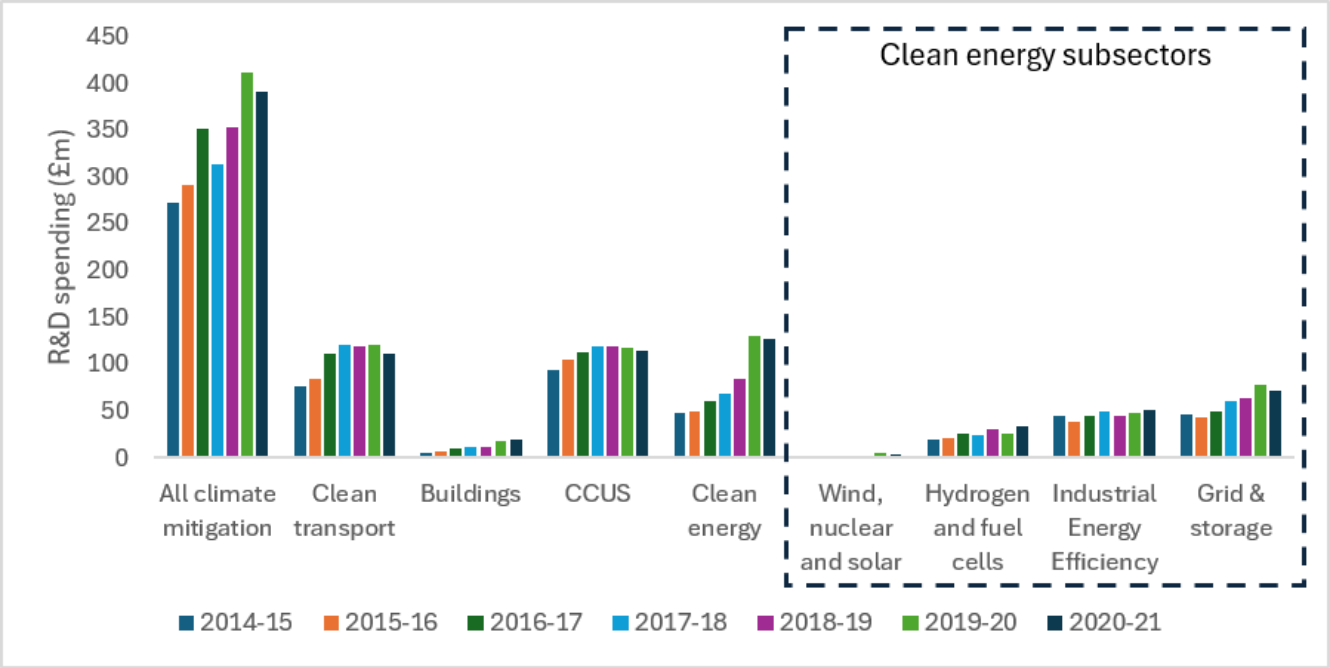
Consequently, while advanced feature engineering and detailed hyperparameter optimisation have the potential to significantly boost model performance, these enhancements require substantial resource investment. This includes additional financial expenditures, extended time for experimentation and model validation, and specialised expertise to develop and implement domain-specific features effectively.

[13] RTICs are The Data City's proprietary form of industrial classification, based upon web-scraped data and built using machine learning.

Estimated R&D spending by non-patenting firms

This section presents the estimates of spending on clean energy R&D by non-patenting firms (in Figure 2 and Table 4). This has involved matching firms with R&D tax records, and then using the machine-learning approach set out above to predict each firm's share of R&D that is focused on various clean energy categories.

Figure 2. Estimated R&D spending by non-patenting firms



Source: Calculations based on HMRC administrative datasets

Table 4. Estimates of R&D spending in nominal £m related to clean energy and climate mitigation in non-patenting firms, by area of technology. Source: Calculations based on HMRC administrative datasets

	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21
All climate mitigation	272.1	291.4	351.2	313	353.4	410.7	389.8
Clean transport	76.1	84.4	110.7	119.7	118.9	119.7	110.8
Buildings	5.5	7.2	10	11.8	11.9	17.1	19.9

	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21
CCUS	94	103.9	113.1	118.4	118	116.5	114.2
Clean energy	48.1	48.5	59.9	67.8	83.5	130.4	127
Wind, nuclear and solar	0.8	0.8	1.3	1.6	2.3	5.2	2.8
Hydrogen and fuel cells	18.7	21	25.1	23.7	30.9	24.8	34.1
Industrial Energy Efficiency	43.7	37.5	45	49.6	45.2	46.9	50.3
Grid & storage	45.3	42.8	49.8	59.9	62.8	77.5	71.7

The share of R&D attributed to patenting and non-patenting firms differs substantially across technology areas. In particular, the share of R&D attributed to non-patenting firms by our analysis is much higher for CCUS-related technologies, and for industrial energy efficiency. These estimates are uncertain, and the machine learning is based on a small training set given the low numbers of firms patenting in these areas. They should be treated as initial indicative estimates.

It is also worth noting that TheDataCity bases its classification of firm as ‘net zero’ or not based on today’s website. It is possible that firms have entered or exited climate-related R&D over the observation period. This would mean that our analysis mis-classifies firms. We are unable to quantify this source of uncertainty in the estimates.

Combined estimates: patenting and non-patenting firms

Combining the patenting and non-patenting firms gives an estimate of total R&D related to clean energy in the UK (Table 5).

Table 5. Combined estimate of clean energy R&D spending by patenting and non-patenting firms. Source: Calculations based on HMRC administrative datasets

	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	2020-2021
All climate mitigation	1494	1747	1753	1959	2061	1678	1415
Clean transport	690	783	879	995	1026	693	572
Buildings	53	61	45	67	71	64	43
CCUS	111	118	129	139	143	141	133
Clean energy	427	474	449	451	533	523	425
Wind, nuclear and solar	25	35	18	38	34	38	31
Hydrogen and fuel cells	82	127	74	69	94	94	119
Industrial Energy Efficiency	47	38	45	50	46	48	52
Grid & storage	211	222	271	266	270	211	176

Headline conclusions from our work

Trends in R&D spending and in the number of firms

From 2014 to the end of the 2020-2021 tax year, our results suggest an annual average spending in the UK on climate mitigation technologies of at least around £1.4bn. The decline in our estimated spending from a peak in 2018-2019 to 2020-2021 could be driven by several factors,

1. Lags in the publication of patents (since we use R&D data based on the three-year moving average of patents over the years immediately prior to the R&D estimate). Patents are typically published around 18 months after application. This means that our estimates for the year 2020-2021 may be affected by patent publication lags. Any firm that only begins filing Y02 patents in 2023 risks being omitted from our

estimates for 2020-2021 because of patent publication lags. However, this effect is likely to be small.

2. There has been a global decline in rates of patenting of climate-change-related inventions since around 2010. Numerous factors have been suggested as possible causes, including the possibility that maturing of many major renewable energy technologies may have led to a focus on incremental (and non-patented) innovation[14]. Long-term trends in the relationship between R&D and patenting could affect our estimates, potentially biasing them downwards for firms that conduct climate-related R&D alongside other R&D projects.
3. A decline in global oil prices may have reduced private incentives for investment in energy efficiency technologies, particularly in the transport sector (which sees the largest decline in patents).

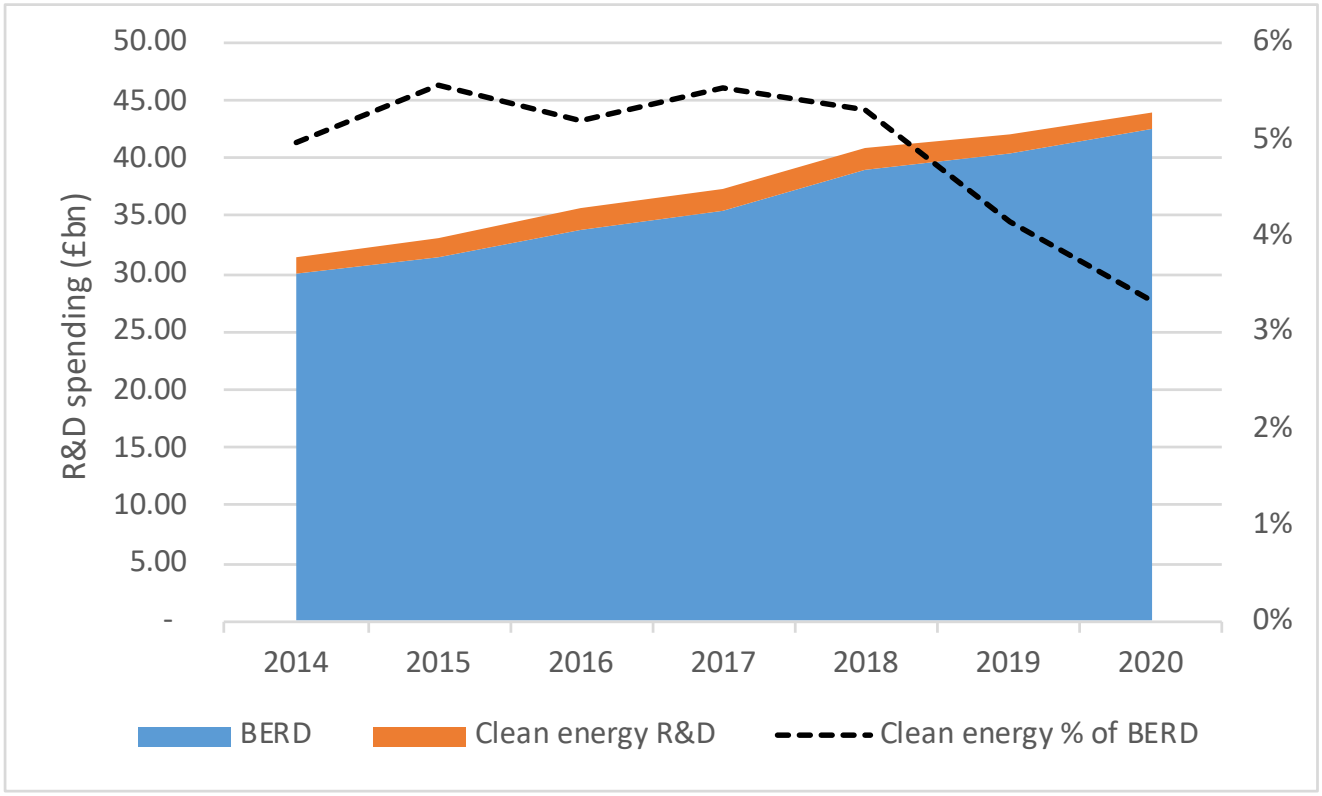
The data also show an increase during the observation period of the number of firms that are both filing Y02 patents and claiming tax credits. This may relate to a broadening of the UK's ecosystem for net zero-related innovation.

UK clean energy R&D in context

In this section, we put our estimates in context, by providing comparisons to other figures on relevant R&D spending in the UK economy. First, in Figure 3, we show our estimates of R&D related to climate mitigation in comparison to total UK business enterprise R&D (BERD), as estimated by the Office for National Statistics. The data show that R&D related to climate mitigation has been between 3-5% of the UK's private sector spending on R&D during the period for which data is available.

[14] Benedict Probst, Simon Touboul, Matthieu Glachant and Antoine Dechezleprêtre (2021), 'Global trends in the invention and diffusion of climate change mitigation technologies', Nature Energy, volume 6, issue 11, pages 1077–1086

Figure 3. Stacked areas show total UK Business Enterprise R&D (BERD) that is unrelated to clean energy, our clean energy R&D estimates; dashed line shows the share of BERD that is clean energy (%)



Source: Calculations based on HMRC administrative datasets for clean energy R&D; BERD data from ONS.

Our estimates show private spending on R&D related to climate change mitigation is well above the average over the same period for public spending on clean energy R&D, which was around £700m per year (according to data reported by government to the International Energy Agency). Publicly funded R&D increased markedly from an average of £550m/year during the earlier period for which we have reported data (2014-2017) to an average of around £900m/year during the latter period (2018-2021). The private sector spending figures presented in this report show no such dramatic increase (as is clear in Figure 1).

Limitations

In this project, we have made a number of assumptions, and encountered some challenges in the use of R&D tax credit data to estimate clean energy R&D spending. In this section, we highlight some of these challenges and assumptions. With additional work, it would be possible to significantly reduce the impact of these issues on the accuracy of the estimates. It is our view that the limitations described below likely result in a bias towards underestimating the true value of R&D spending.

Relationship between patenting and R&D

In order to identify the share of R&D activity that is related to specific technologies, we have assumed that R&D spend is linearly related to patent counts. Within the literature, previous analysis has generally found a strong correlation between R&D spending and patenting (Hall and Ziedonis 2001[15]; Bolivar-Ramos 2017[16]). However, various other factors are known to influence this relationship, including capital expenditures (Piergiovanni and Santarelli 2013[17]) and sectoral and technology differences (Wiesenthal et al. 2012[18]). In particular, our assumption that the relationship between R&D and patenting for a given firm is constant across technologies may be inaccurate. Further work would be valuable to measure explanatory factors behind variation in the patent-intensity of R&D spending, and such work would strengthen the robustness of the findings.

Lag structure between R&D and patent applications

The length of time between a company performing R&D and subsequently filing a patent that has been enabled by that R&D is variable. For the data reported here, we have assumed that—on average—patent applications arise from the R&D conducted two years prior to the application, a widely used lag in the literature. However, the question of lag structure is complex (Wang and Hagedoorn 2014)[19]. Further work could focus on firms that *only* patent clean energy technologies (i.e. 'pure' clean energy firms), which would enable us to examine

[15] Bronwyn H. Hall and Rosemarie Ham Ziedonis (2001), 'The patent paradox revisited: An empirical study of patenting in the U.S. semiconductor industry, 1979–1995', RAND Journal of Economics, volume 32, issue 1, pages 101–128

[16] María Teresa Bolívar-Ramos (2018), 'The relation between R&D spending and patents: The moderating effect of collaboration networks', Journal of Engineering and Technology Management, volume 46, pages 26–38

[17] Roberta Piergiovanni and Enrico Santarelli (2012), 'The more you spend, the more you get? The effects of R&D and capital expenditures on the patenting activities of biotechnology firms', *Scientometrics*, volume 92, issue 1, pages 107–131

[18] Tobias Wiesenthal, Guillaume Leduc, Karel Haegeman and Hans-Günther Schwarz (2012), 'Bottom-up estimation of industrial and public R&D investment by technology in support of policy-making: The case of selected low-carbon energy technologies', *Research Policy*, volume 41, issue 1, pages 116–131

[19] Ning Wang and John Hagedoorn (2014), 'The lag structure of the relationship between patenting and internal R&D revisited'v, Research Policy, volume 43, issue 8, pages 1275–1285

whether the lag differs by technology area, and whether a different lag structure would be more accurate for energy technologies.

Missing or misleading geographic information in patent records

It is possible that international companies may undertake R&D activities within the UK, but register their patents from the address of the global headquarters. In this case, using the address of the patent applicant may be inaccurate. Similarly, firms with a UK headquarters may file patents using their UK address, even when the research underpinning the invention was carried out overseas. We have sought to limit the impact of this effect by using information on the addresses of inventors, who are the people that have actually conducted the R&D.

However, not all patents provide sufficient information on the addresses of applicants and inventors. This limits our ability to identify the geographic location of R&D activities using patent data. This effect may be reducing the coverage of our dataset of clean energy patents, as it may mean that some clean energy patents with UK involvement have not had a UK address reported on the patent (e.g. collaborative patents involving a UK inventor in an international research team may not record the UK inventor's address).

Challenges in matching patents to firms, and firms to tax records

We have used fuzzy string matching to link patents to firms. Patent documents include the name of the company that has applied, and patent databases (including PATSTAT and Orbis IP) have worked to harmonise the names of applicants within their patent records. Alternative spellings, the inclusion or exclusion of 'PLC', 'Inc.', and other such variations are common. Within PATSTAT, we have used the 'harmonised name' field, benefiting from the efforts made by the European Patent Office to harmonise this information. However, the matching of these names to the names in Companies House is imperfect.

For this analysis, we have included matched firms only when we have a very high degree of confidence that the matched firms are correct. However, this likely means that we have failed to match some firms that are carrying out clean energy R&D.

We have also identified challenges with respect to matching between patenting firms and those carrying out R&D. There may be cases in which R&D is conducted by one member of a corporate group, but patents are filed by a different member. This would lead to an underestimate of private R&D spending in our dataset. Further work could identify all members of a corporate group, and explore the R&D and patenting activities of all members.

Finally, there were some weaknesses in the matching of our patent-company dataset to the HMRC's records of taxpayers. HMRC Datalab conducted the matching of our matched patent-company dataset to their database of taxpayers, based on a concordance between Company Registration Number (CRN) and HMRC's Unique Taxpayer Reference. Only 79% of UK firms

that we matched with Y02 patents could be matched with HMRC's Corporation Tax records, of which only some had submitted R&D tax credit claims. The result is that only 31% of the firms that we matched with Y02 patents are represented in the estimates.

Machine learning approach

The model is trained on patenting firms where the outcome (the share of clean technology patents) is observed, but it is applied to non-patenting firms where this ratio is not directly observable. This extrapolation assumes that the relationship between firm characteristics and clean R&D share in patenting firms holds for non-patenting firms, a potentially risky assumption if systematic differences exist between these groups. For instance, the outcome variable for patenting firms (the share of clean technology patents) might not perfectly translate to non-patenting firms. Differences in innovation processes could limit the predictive accuracy when the model is applied outside the training domain.

Additionally, with 1,780 features, the model faces a high-dimensional feature space. This increases the risk of overfitting, where the model captures noise rather than signal. However, XGBoost's built-in regularisation (via $L1$ and $L2$ penalties) helps to counteract overfitting. Further, thorough hyperparameter tuning (using grid or randomised search) can identify the optimal complexity.

Finally, the DataCity dataset includes missing values, which can complicate model training and interpretation. We leveraged the XGBoost's ability to natively handle missing values where appropriate, ensuring that the treatment of missing data is consistent and well-documented.

Time horizon limitations related to non-patenting firms

There are two limitations related to our estimates for non-patenting firms. First, firms are only observed by TheDataCity in 2024. We do not have any information on whether these firms have only recently entered net-zero activities, and we are unable to observe firms that have exited such activities. The results from patenting firms suggest that the number of firms engaged in clean energy R&D has increased over the observation period, which suggests that TheDataCity analysis may overstate clean energy R&D activities (by identifying firms as engaged in clean energy throughout the period, whereas they may only recently have diversified into this area).

Second, our clean energy R&D share predictions for non-patenting firms are time-invariant, i.e. the firm is assumed to have the same share of R&D spending related to clean energy over time. This necessary simplification creates additional uncertainty in the estimates.

Conclusions and next steps

Conclusions

The results presented here provide the first estimate of clean energy R&D spending for the UK based on administrative data. For the reasons discussed in section 5.1 above, the estimates remain uncertain. Relying as they do on patents and website text as an indicator of clean energy R&D activity, these estimates may miss important developments in the climate mitigation innovation ecosystem. For example, it is noticeable that R&D spending on nuclear energy is very small in these results, despite recent developments in both UK fission (notably Rolls Royce SMR) and fusion (e.g. Tokamak Energy). Some of these developments, such as the establishment of Rolls Royce SMR, would not be captured by our data, since the most recent findings are for 2020-2021.

Nevertheless, the approach using R&D tax credit data shows real promise at generating technology-level time series data on energy-related R&D. It thus provides an important addition to the UK's knowledge base on clean energy innovation. We plan to undertake further work to improve these estimates—funding permitting. The priority next steps are set out below.

Next steps

These preliminary estimates provide a first picture of private R&D spending related to net zero and clean energy. They are subject to several limitations, as set out previously. Here, we highlight the next steps that could be taken to improve the estimates, subject to funding being available.

The single most important step would be to link other datasets to our existing data, in order to better identify both a) which firms are involved in R&D related to climate change mitigation, and b) the share of those firms that is related to climate change mitigation. Datasets include:

1. Data on grant-funded projects including records of R&D grants from InnovateUK, DESNZ and European Horizon projects. Project descriptions for such projects provide both further evidence on the relative share of activities of firms. This would strengthen the estimates, as it would provide additional information (beyond patents) on the R&D activities of firms.
2. ONS data from the Low Carbon and Renewable Energy Economy survey. Respondent firms provide data on their share of turnover that is associated with specific low-carbon and renewable energy activities, which could be used to estimate their share of R&D activity in these areas. However, there are administrative barriers for such work, because of difficulties of data sharing between ONS and HMRC secure data systems. This currently prevents us from linking data from the ONS 'low carbon and renewable energy economy' survey. Linking these datasets would enable us to strengthen our estimates, both directly (by providing an additional source of information about

companies' share of activity in various clean energy activities) and indirectly, by providing insight into the relationship between R&D and patenting across different areas of technology.

There are various other steps that could be taken to improve the estimates provided here. These include:

Looking in more detail at 'pure clean energy' firms: i.e. those firms whose patent portfolios are heavily dominated (>95%) by patents related to climate change mitigation. Since we can assume that essentially all the R&D conducted by such firms is related to climate mitigation, we can remove variation over time in patenting rates, and instead explore variation in patent intensity across areas of technology. This may help to characterise uncertainty in the existing estimates.

Explore more fully whether/when R&D claims and patent applications are split across different members of a corporate group.

Finally, further work could broaden the array of technologies considered to include those related to fossil fuels. This would enable a better understanding of which energy technologies receive the largest volumes of public support.

This publication is available from: www.gov.uk/desnz

If you need a version of this document in a more accessible format, please email alt.formats@energysecurity.gov.uk. Please tell us what format you need. It will help us if you say what assistive technology you use.