

# Data Explained

---

## Annual Survey of Hours and Earnings Linked to Census 2011 and O\*NET Data

Working towards an environmentally sustainable and equitable future?  
New evidence on green jobs from linked administrative data in the UK

**Author:** Damian Whittard

**Date:** May 2025

---

This Data Explained summarises experiences and learning from working with the Annual Survey of Hours and Earnings and Census 2011 linked dataset, which was linked to O\*NET (Occupational Information Network) datasets to study green jobs in England and Wales. This publication is intended to help guide future researchers using this data and to provide feedback into future dataset development and documentation.

The data used in this research project comes from the Office for National Statistics (ONS) (ASHE linked to Census 2011) and O\*NET's database of worker attributes in the United States (US), created and maintained by the Department of Labour. The ONS data was accessed through ONS's Secure Research Service (SRS). The data was not originally collected for research and as such, there are gaps and inconsistencies in its recording. The O\*NET data is freely available via O\*NET Online and was ingested into the SRS to link to the ASHE and Census data. The O\*NET dataset is based on the US labour market and therefore has limitations when applied to analysing the labour market of England and Wales. The main challenges are detailed in the following article.

### Project details

Given the urgency to transition to net-zero, there is a need for a robust evidence-base to support an environmentally sustainable and equitable changeover, but intelligence on green jobs and their impact on different groups is lacking. The initial study by Whittard et al., (2024) focussed on exploring the characteristics of employment in green occupations and the potential impact these roles can have on pay.

The analysis reveals that males are much more likely to be employed in green occupations, as are salaried workers, fulltime employees, those working in smaller business, and foreign owned companies. Those that do work in green occupations receive a pay premium of around 4% compared to their counterparts working in other occupations.

The greening of the economy offers the potential for a more inclusive and just transition. That said, of particular concern to policy makers should be the dual inequality that green occupational employment appears to engender. Not only are female and ethnic groups underrepresented in green employment, when they are employed in green jobs, they are paid less than their counterparts.

### Initial research questions

- [1] What are the characteristics of the people that work in green jobs?
- [2] What are the characteristics of green jobs and the firms that employ them?
- [3] Is there a pay advantage for working in a green job?
- [4] If so, how does this impact on different groups?

### Research methodology

The three main data sources used in the study are ASHE, Census 2011, and O\*NET. Initially the Wage and Employment Dynamics Project was funded by ADR UK, who would work with ONS to link, document and quality assure the ASHE to Census 2011 dataset – for more information on its characteristics see Forth et. al, (2022). By doing so, this enabled a rich set of personal characteristics from Census to be added to the accurate components of pay and employment information coming from the ASHE dataset.

Green occupations were identified in this dataset through an additional link to the O\*NET dataset. Although the O\*NET dataset is based on the US labour market, it can be mapped to the UK via an occupation code using international occupation classification. Mapping for this project used a direct crosswalk between the ONET-SOC (eight-digit) and UK SOC systems (four-digit), developed by the 'LMI for All' (2025). By linking the datasets, it is possible to identify green jobs as both a binary variable - where 0 indicates non-green and 1 indicates green - as well as a weighted continuous variable reflecting greenness of job.

Our study used regression analysis to explore the characteristics of those working in a green job and estimated the pay premium for working in a green job. Cross sectional analysis was applied to

the latest data (2018) to exploit the individuals' characteristics. A fixed effect model was used for the full panel (2011-2018), addressing issues of endogeneity by accounting for all time-invariant characteristics of the individuals (e.g. ability). This approach allowed for a focus on the net effect of the variables that change over time.

## Key variables

### Dataset: ASHE-Census

Variable	Categories	Code
Basic hourly wage	Continuous variable calculated by the ratio of basic weekly earning to the total number of basic weekly paid hours.	bpay/bhr
Female	Categorical variable (Male=1, Female=2)	sex
Age Age-squared	Continuous variable with squared value.	age age_squared
Ethnicity	Categorical variable: White = 1; Mixed/multiple ethnic groups = 2; Asian/Asian British = 3; Black/African/ Caribbean/ Black British = 4; Other ethnic group = 5.	aggethpuk11
Education	Categorical variable. Self-reported level of highest qualification: No qualification = 0; Up to A-level = 1; Apprenticeship = 2; Other/vocational qualification = 3; Degree or above = 4).	hlqpkuk11
Marital status	Dummy variable (Single = 0, Married or in a registered same-sex civil partnership = 0).	marstat
Born outside UK	Categorical variable (1 = born within the UK, 2 = born outside the UK).	aggcobbuk113
Health – Fair to very bad	Dummy variable for self-reported health (Very bad, bad or fair health = 0; Good or very good health = 1).	health
Dependent child	Categorical variable to indicate whether the individual is responsible for a dependent child: No dependent child in family = 0; Pre-school age (0-4) = 1; Primary school (5-11) = 2; Senior school (12-18) = 3.	dpcefamuk11
Public transport user	Categorical variable whether individual is a public transport user (Public transport user = 1, 2 Not a public transport user = 2).	ptranspuk11
Disability	Categorical variable for those who report a disability that interferes with their day-to-day activity (Disability – no limits = 1; Disability limited activity = 2).	disability
Basic paid hours	Basic weekly paid hours worked.	bhr

## Data Explained

Experience Experience-squared	Continuous variable and squared term. Time in job calculated by year of observation, minus employment start year plus one.	empstart_y
Part-time job	Dummy variable to indicate whether job is part-time (Part-time = 0; Full-time=1).	fulltime
Hourly paid	Dummy variable to indicate whether the individual was hourly paid (Not hourly paid = 0; Hourly paid = 1).	hourly_paid
Size of employer	Categorical variable for the employer size band (number of employees): 0-9; 10-49; 50-249; 250 and over.	emp_size_band
Collective agreement	Dummy variable to indicate whether the individual is subject to a collective bargaining agreement (No collective agreement = 0; Collective agreement = 1).	coll_agt
Foreign owned enterprise	Dummy variable to indicate whether the individual works for a foreign owned company (UK owned = 0; Foreign owned = 1).	for_own
Sector	Categorical variable of 11 industrial sectors (Primary = 0, Manufacturing = 1, Utilities = 2, Construction = 3, Sales = 4, Service = 5, Finance and Law = 6, Public = 7, Health = 8, Creative = 9, Other = 10).	sector
Region	Categorical variable. Government office region at workplace from NUTS1 (North East = 1; North West = 2; Yorkshire = 3; East Midlands = 4; West Midlands = 5; South West = 6; East = 7, London = 8; South East = 9; Wales = 10).	region

### Dataset: O\*NET/LMI Crosswalk

Variable	Categories	Code
Green Task and Occupation (Binary Variable)	<p>Dummy variable to identify UK occupations which match to one or more of the US green occupations, and occupations with green tasks as identified via the O*NET project.</p> <p>The O*NET database distinguishes three occupational categories that differ regarding the effect of the transition to a climate-neutral and sustainable economy on occupations. There are two categories which are considered directly green: Green new and emerging – this includes new occupations and those with unique green tasks; and Green enhanced skills). There is also one</p>	<p>max_green_occ max_green_task</p> <p>max_GN&amp;E max_GES max_GID</p>

	category which is assessed to be indirectly green as it supports the transition to a green economy, but without having any green tasks; Green in demand.	
Green Occupation continuous variable)	Continuous variable which weights occupation date to identify probability or greenness of occupation.	wght_mean_green_occ wght_mean_green_task wght_mean_GN&E wght_mean_GES wght_mean_GID

## How you dealt with data limitations

Below we describe the main data limitations which could have introduced bias and reduced the accuracy of our findings.

### ASHE linked to Census 2011

The initial linkage between ASHE and the Census created a high-quality employment and earnings dataset enriched with valuable personal characteristics. Unfortunately, this linkage did not include complete information on employees' family or household circumstances which would have boosted the research value. At time of writing the WED team are continuing to work with ONS on the development of this dataset to address this, with the aim of adding further variables to describe the economic activity of other members of employees' family or household.

Other challenges with this dataset include the fact that the availability of Census-linked data within the ASHE panel diminishes over time. For example, in ASHE 2011 nearly three-quarters of employee jobs held by workers residing in England and Wales had linked data from the 2011 Census, but this dropped to less than half in 2018. In addition, employee characteristics inherited from the Census are self-reported at a single point in time meaning that fluid characteristics (e.g. disability status, religion) are increasingly prone to measurement error the further they are from 2011.

Finally, due to data owner concerns with disclosure, available geographic markers were limited to local authority level and above. However, such concerns of disclosure could instead be addressed through a data-set specific output checking processes which would allow the inclusion of lower-level geographic markers.

### O\*NET

The O\*NET database enables classification of occupations based on the environmental relevance of their associated tasks. It uses a broad definition of green jobs across twelve sectors identified as significantly impacted by decarbonisation efforts. The value of the task and occupation-based approach is that it captures green employment across sectors. This therefore broadens the definition of green jobs beyond those just working in green industries.

By mapping this US occupation data to UK occupation data, the study completed an occupation based, bottom-up approach to identifying green jobs in the UK. A potential limiting assumption of this study was that it made the key assumption that tasks undertaken within occupations are the

same in the UK and US, and that those occupations considered green in the US are also considered green in the UK. The only alternative would be to link a UK based dataset which identifies green occupations but unfortunately, to the author's knowledge, such a dataset is currently not available.

There were also practical challenges with the linking the O\*NET database. For example, there is a mismatch in terms of granularity between the US occupations and the UK occupations. O\*NET records occupations at the eight-digit level, whereas the occupational classification system used in ASHE is the UK Standard Occupational Classification (UK SOC) 2010 recorded at the more aggregated four-digit level.

When linking the data, a direct crosswalk was applied which was developed by the 'LMI for All' data portal and funded by the UK Department for Education (2025). The initial mapping resulted in multiple matches between US and UK occupations, generating a binary variable - 0 indicating a non-green and 1 indicates green. As UK occupation codes are more aggregated compared to the detailed O\*NET occupations, this meant that whole occupational categories in the UK may be classified as green even if only a single US sub-category is considered green.

To address this issue, we created cautious weighted continuous measure of green jobs calculated using weights accounting for the US employment share of matched occupations to UK occupation, and weights accounting for estimates using the UK employment share of those occupations (which avoids double counting of US O\*NET occupations matched to multiple UK occupations).

An example of this approach is follows:

- A UK occupation (OCC1) has three US occupations attached to it, one of which is green. The US green occupation accounts for 20% of the employment from the three US occupations matched to the UK occupation.
  - OCC1 initial weighted =  $1 * 0.2$
- The US green occupation is mapped to two UK occupations (OCC1 and OCC2). The employment share between OCC1 and OCC2 is 40% and 60% respectively.
  - Therefore, the final estimate for OCC1 =  $0.2 * 0.4 = 0.08$

A further limitation of applying the O\*NET data is that it assumes that an occupation's green status remains constant across firms, regardless of the firm's environmental impact. In future studies, to address this issue, researchers could consider combining information on occupations with employers (e.g. net-zero), so they can identify occupations which are green, and which are located in firms that are contributing to the challenge of environmental sustainability.

While recognising the limitations, an advantage of the study, is that green jobs are sometimes overstated in size due to broad definitions or misclassification of roles that only tangentially relate to sustainability. This overstatement can skew analyses and lead to inflated estimates of the green economy's size. O\*NET data helps to address this, as it enables identification of both directly and indirectly green jobs across sectors. This allows the researcher to refine the definition of green jobs, focusing on core green roles rather than including peripheral positions.

In addition, in order to mitigate against a number of these limitations, multiple regression approaches were used to test the robustness of findings. To take account of binary and non-binary measures of green jobs, both ordinary least squares (OLS) and logistic regression were applied. Panel models were also used to account for different relationships and data structures. For example, fixed effects models were used to account for time invariant unobserved heterogeneity (differences that we don't see in the data that don't change over time – e.g. an individual's innate abilities).

## Additional data which would help to further develop the research

To improve the longitudinal characteristics of the dataset and improve the geographic coverage, the O\*NET dataset could also be linked to an ASHE data linked to the Census 2021 for England and Wales (the WED project are currently working on this linkage), and the 2011 and 2022 Census for Scotland.

Linking the data to firm level information (e.g. Business Structure Database) could improve the quality of the dataset by allowing green jobs to be identified by combining occupation with firm activity.

## References

1. Forth, J., Phan, V., Ritchie, F., Whittard, D., Stokes, L., Bryson, A. and Singleton (2022) SHE – Census 2011 Data Linkage: User Guide for Drop 2 of the ASHE-Census 2011 Dataset. Version 2.1. Available at: <https://www.wagedynamics.com/wp-content/uploads/2023/01/ASHE-CEW11-User-Guide-Version-2.1-Drop-2.pdf>
2. LMIFORALL (2025). *Labour Market Information for All*. Available at: <https://www.lmiforall.org.uk/>
3. Whittard, D., Bradley, P., Phan, V., and Ritchie, F. (2024). *Working Towards an Environmentally Sustainable and Equitable Future? New Evidence on Green Jobs from Linked Administrative Data in the UK*. ESCoE Discussion Paper No. 2024-13. Available at: <https://escoe-website.s3.amazonaws.com/wp-content/uploads/2024/10/17165556/ESCoE-DP-2024-13.pdf>

## Disclaimer

This work was produced using administrative data accessed through ONS's SRS. The use of the data in this work does not imply the endorsement of SRS or data owners in relation to the interpretation or analysis.

This work uses research datasets which may not exactly reproduce National Statistics aggregates.

## Acknowledgements

This work is supported by ADR UK (Administrative Data Research UK). ADR UK is a partnership transforming the way researchers access the UK's wealth of public sector data, to enable better informed policy decisions that improve people's lives. ADR UK is an Economic and Social Research Council (ESRC) investment (part of UK Research and Innovation). [Grant number: ES/Y001184/1]

## Contact

Name: Damian Whittard

Email: [damian2.whittard@uwe.ac.uk](mailto:damian2.whittard@uwe.ac.uk)

---

