Wage & Employment Dynamics

THE WED PROJECT

ASHE – 2011 CENSUS DATA LINKAGE

User Guide Drop 1 of the ASHE – 2011 Census Dataset

Abstract

In order to expand the number of personal characteristics that are observed for employees in the Annual Survey of Hours and Earnings (ASHE) dataset, the Office for National Statistics (ONS) has linked the personal details of employees observed in the 2011 ASHE to those of individuals observed in the 2011 Censuses for England and Wales. The linking process includes two phases. After Phase 1, robust links have been made for around 62% of eligible records. Phase 2 is expected to increase this percentage. Where a link has been established, it has been possible for ASHE to incorporate employee-level data from the 2011 Census on a range of personal characteristics, including the employee's educational qualifications, country of birth, ethnicity, religion, and disability status. It has also been possible to incorporate some limited information about their household circumstances. These Census data items are available for the matched individual in any year that they appear in ASHE. This *User Guide* (Version 1.1) provides researchers with an overview of the linkage process and linkage outcomes from Phase 1. It also describes the weights that have been constructed to account for linkage biases, and it explains how potential users can gain access to the linked data.

John Forth, Van Phan, Felix Ritchie, Damian Whittard, Lucy Stokes, Alex Bryson and Carl Singleton

Version 1.1: April 2022

Contents

This work was produced using statistical data from ONS. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

1. Introduction

The Annual Survey of Earnings and Hours (ASHE) is an annual survey based on a 1% sample of employee jobs, drawn from the PAYE register, and conducted by the Office for National Statistics (ONS). The survey collects information on employees' earnings, paid hours and occupation, along with some employer characteristics. However, the research dataset contains relatively few personal characteristics for the employee (restricted to gender, age and residential location).

In order to expand the number of personal characteristics that are observed for employees in ASHE, ONS has sought to link the personal details of employees observed in the 2011 ASHE to those of individuals observed in the 2011 Censuses for England and Wales (CEW11).[1] The linking process is proceeding in two phases, described as 'Phase 1' and 'Phase 2'. The datasets generated in each phase are labelled 'Drop 1' and 'Drop 2' respectively.

In Phase 1 of the linking process, robust links have been made for 62% of the employee job records in the 2011 ASHE belonging to employees resident in England and Wales. In many cases where a robust link has not been made, this is because missing data on one or more of the core linkage variables (name, gender, age or residential postcode) has prohibited the reliable identification of the ASHE employee in the Census dataset. Phase 2 of the linking process will bring in additional linkage variables in order to achieve robust matches for some of these cases. It is anticipated that the linkage rate may increase by up to 10 percentage points after Phase 2, although the outcomes from this additional stage of linking are necessarily uncertain at this stage.

In cases where a robust link has been made under Phase 1, it has been possible for ASHE (the "recipient dataset") to incorporate employee-level data from CEW11 (the "donor dataset") on a range of personal characteristics, including the employee's educational qualifications, country of birth, ethnicity, religion, and disability status. It has also been possible to incorporate some limited information about their household circumstances. These Census data items are available for the matched individual in any year that they appear in ASHE from 1999-2018.

This current version of the *ASHE-CEW11 User Guide* provides researchers with an overview of the linkage process and linkage outcomes from Phase 1. It also describes the weights that have been constructed to account for linkage biases, and it explains how potential users can gain access to Drop 1 of the ASHE-CEW11 dataset.[2]

The linked ASHE-CEW11 dataset is being made available to researchers via the ONS Secure Research Service and provides opportunities for new research on issues such as the returns to education and wage inequality. Guidance on accessing the dataset is provided in Section 12.

In a separate exercise, the Northern Ireland Statistics and Research Agency (NISRA) has linked the ASHE data for Northern Ireland to the 2011 Census for Northern Ireland. These data are made available by NISRA (see Section 13).

The WED team has submitted a bid to link ASHE to the 2021 Census for England and Wales but this has yet to be approved. The WED team are also exploring opportunities to link ASHE to the 2011

[1] Record linking is also known as record matching. We use the terms linking and matching interchangeably in this User Guide.

[2] The Guide does not provide a general discussion of data linkage; those who are unfamiliar with the main approaches are referred to Sections 1-5 of Mayer and Stockdale (2021).

and/or 2021 Census for Scotland, although the Censuses for Scotland are governed under different arrangements to those for England and Wales. There are no plans to link ASHE to the 2001 Census, although this is technically feasible (see Jenkins, 2008).

2. Motivation for linking ASHE to Census 2011

ASHE has many features of value to researchers: it provides rich information on earnings and working hours, collected from payroll records for a large and nationally representative sample of employee jobs. However, one limitation of ASHE is that the dataset contains few employee characteristics. These are limited to gender (Male/Female), age (in years) and residential location (Census Output Area). The limited number of employee characteristics arises because the survey is completed by employers, rather than by employees themselves.

The CEW11 microdata provides one source of information on the personal characteristics of employees. It has three specific attributes that make it suitable as a potential donor of such information for ASHE sample members. First, it is linkable to ASHE: the source datasets for ASHE and CEW11 held by ONS share a set of common fields that can be used to link individual records with a high degree of confidence (at least, in cases where these fields are fully populated). Second, the reference date is close to that of ASHE: the 2011 ASHE survey has a reference date of 13th April 2011, whilst CEW11 has a reference date of 27th March 2011. Third, it is comprehensive: by definition, one would expect almost all employees working in England and Wales on 13th April 2011 to be present in the CEW11 dataset.[3]

Figure 1 provides an overview of the information provided in ASHE and CEW11. It can be seen that CEW11 provides information on a variety of additional personal characteristics of relevance to the analysis of earnings and hours. Adding these characteristics to the ASHE dataset allows for a range of new research opportunities, including:

- Using data on educational attainment to examine the wage returns to human capital
- Identifying wage gaps across a range of characteristics protected under the Equality Act 2010 (e.g. ethnicity, disability, religion, sexual orientation)
- Exploring wage progression among migrants to the UK
- Examining the relevance of partner status in explaining wage variance and job mobility

[3] The exceptions would be those not resident in England and Wales on 27th March 2011 and any non-respondents to the Census.

Figure 1: Summary of topics covered in ASHE and the 2011 Census for England and Wales

ASHE 2011
Census 2011
ASHE-Census linkage

Personal characteristics: • Age • Gender • Residential location Job characteristics: • Earnings • Paid hours • Occupation Employer characteristics: • Size (employees) • Industry • Legal status • Workplace location	Personal characteristics: • Age • Gender • Residential location • Country of birth • English language proficiency • Ethnicity • Religion • Marital or civil-partnership status • Health and disability • Educational and vocational qualifications • Household circumstances Job characteristics: • Usual hours • Occupation Employer characteristics: • Industry • Workplace location	Personal characteristics: • Age • Gender • Residential location • Country of birth • English language proficiency • Ethnicity • Religion • Marital or civil-partnership status • Health and disability • Educational and vocational qualifications • Household circumstances Job characteristics: • Paid hours • Usual hours • Occupation Employer characteristics: • Size (employees) • Industry • Legal status • Workplace location

3. Overview of linkage outcomes and limitations

Outcomes from Phase 1

The ASHE and CEW11 datasets do not share a common, unique identifier at the level of the individual (e.g. National Insurance number), and so the linking process sought to link individuals by seeking a robust match on a set of common fields.

In Phase 1, the link was created by identifying all individuals observed in ASHE 2010, 2011 or 2012 and then seeking to locate these individuals in CEW11 by matching on a combination of name, gender, age and residential postcode. The linking process called upon ASHE records from 2010, 2011 and 2012 to allow for the possibility that a linkable individual may be missing from ASHE 2011 due to a temporary period of non-employment or employer non-response in that year.

To provide an overview of the linking outcomes from Phase 1, we focus on ASHE 2011, where the full WED dataset contains information from 178,247 unique individuals and 182,970 employee jobs (see Table 1). Robust links to CEW11 have been made for a total of 96,619 individuals.[4] Around 20,000 of the individuals in the full ASHE dataset are not recorded to be resident in England or Wales (ASHE variable: HGOR); these are very unlikely to get matched to CEW11 for obvious reasons, although matches are established for a minority.[5] Focusing then on those 159,096 individuals who are recorded in ASHE as being resident in England and Wales (1<=HGOR<=10), robust links have been made for 97,901 of these individuals: a linking rate of 61.5% for Phase 1. The 159,096 individuals resident in England and Wales hold a total of 163,185 employee jobs in the ASHE 2011 dataset. CEW11 data has been linked to 61.7% of these jobs.

Table 1: Linkage outcomes from Phase 1 at individual-level and job-level, ASHE 2011

Phase 1 outcomes	**All records**	
	Number	**%**
Individuals:		
All individuals	178,247	100.0
Of which: Linked to CEW11	98,617	55.3
Individuals resident in England or Wales	159,096	100.0
Of which: Linked to CEW11	97,901	61.5
Employee jobs:		
All employee jobs	182,970	100.0
Of which: Linked to CEW11	101,506	55.5
Jobs held by employees resident in England or Wales	163,185	100.0
Of which: Linked to CEW11	100,698	61.7

Notes: (i) Individuals are identified via PIDEN; (ii) see footnote 4 for the definition of a linked record.

As a result of Phase 1 of the linking process, Drop 1 of the linked dataset includes (for employees resident in England and Wales) around 10,500 job records belonging to members of a non-white ethnic group (CEW11 variable: ETHPUK11), around 4,800 records belonging employees with a long-term health problem or disability (CEW11 variable: DISABILITY), around 11,700 records belonging to employees who were born outside the UK (CEW11 variable: LRESPUK11) and around 36,700 records belonging to employees with a degree-level qualification (CEW11 variable: HLQPUK11).

Individuals may be linked across years in the ASHE dataset via a unique personal identifier (PIDEN), based on their National Insurance number. Hence, any information that has been linked to ASHE from CEW11 in the linkage year can be copied over to other years of ASHE in which that same individual is present (currently 1999-2018) – see Figure 3 in Section **Error! Reference source not**

[4] A good-quality link is one with a match score equal or greater to 0.82. The match score is discussed in Sections 5 and 8.

[5] Around 17,000 are resident in Scotland; a link is only likely to be made for these cases if the individual's postcode is partial or in error and a good match is found on other variables. A further 3,000 are unobserved on residential location, reducing the likelihood of a match.

found. for details.[6] This allows some research questions (e.g. the size of the ethnic wage gap) to be explored across multiple years.

Improvements intended for Phase 2

In some cases, a robust match could not be made in Phase 1 because missing data on one or more of the four linkage variables prohibited the reliable identification of the ASHE employee in the Census dataset. Such linkage failures were most commonly due to the employee's name being missing on a subset of ASHE records. Some 23% of ASHE records did not include information on the employee's name; this suggests that the linking rate among individuals with complete data may be as high as 80%. Phase 2 of the linkage process will bring information on the employee's occupation, industry sector and workplace postcode into the matching algorithm in order to identify robust matches for at least some of these cases where a link has not yet been made in Phase 1.

Limitations to ASHE-Census linkage

There are a number of limitations to the ASHE-Census linkage.

First, linkage has only been possible for a subset of ASHE records in Phase 1, and cannot be expected to be universal in Phase 2. This means that linkage biases need to be investigated and, where present, accounted for. Nevertheless, our analysis (discussed later) indicates that the linkage biases in Phase 1 are relatively minor.

Second, Census information is not observed for those individuals who may appear in ASHE in earlier or later periods, but who were not observed in ASHE 2010, 2011 or 2012. One example would be a person who retired from paid employment in 2009, or a person who entered the labour market for the first time in 2013. The implication is that the availability of Census information depletes as one moves further away from the linkage year. For instance, less than half (45%) of the employee jobs in ASHE 2016 held by employees resident in England and Wales have linked data from CEW11.

Third, as the employee characteristics that are inherited from CEW11 are necessarily measured at a single point in time, the ASHE-Census dataset is best suited to the analysis of characteristics that are fixed, such as ethnicity or migrant status. Opportunities to analyse characteristics which may with time, such as disability status, are more limited, since these characteristics are increasingly likely to be measured with error as one moves further away from 2011.

Further details on each of these points is provided in later sections, which discuss the linking process and outcomes in more detail.

4. Data linking issues

Any effort to link records from two separate datasets encounters challenges in ensuring that the correct records are linked to one another. There are two types of record linkage errors:

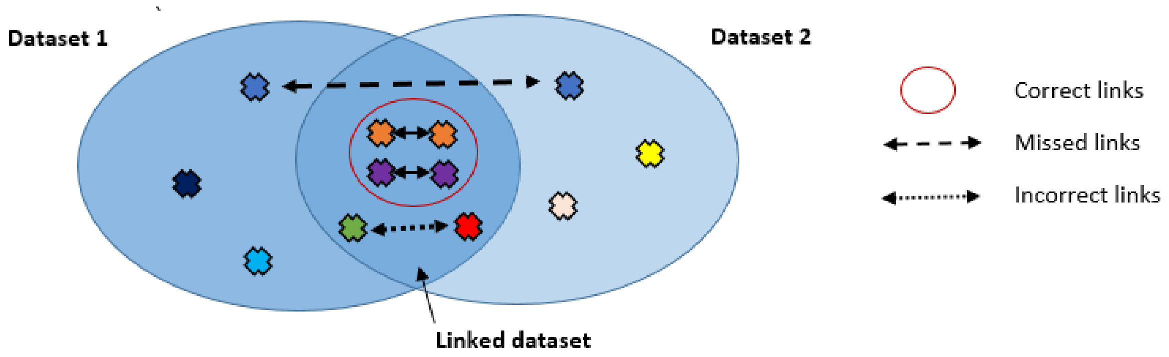
(i) Missed links (false negatives), where records that refer to the same individual are not linked;
(ii) Incorrect links (false positives), where records belonging to different individuals are erroneously linked together.

These errors are illustrated in Figure 1. Such errors can occur due to changes in persons' details (e.g. an address change, or name change after marriage) which may cause the details held in one dataset

[6] The ASHE WED dataset extends from 1997-2018 but PIDENs are not present in 1997 or 1998.

to differ from those held in the other. Errors can also occur when data fields are missing or contain mistakes (e.g. if a person's first name has not been recorded, or if the digits in a person's date of birth or postcode have been accidentally transposed).

Figure 1: Types of linkage error



Note: Adapted from Gammon (2016)

These linkage errors can lead to two potential problems for analysts:

1. *Erroneous inclusion or exclusion from the linked dataset:* The implication is that linked cases may not be representative of the full dataset, potentially leading to biased estimates. Exclusion may also reduce the number of records available for analysis, limiting the statistical power of the linked dataset.
2. *Misclassification or measurement error:* The implication is that incorrect inferences may be made about the characteristics of persons in the linked dataset, again potentially leading to biased estimates.[7]

As Doidge et al (2020) indicate, it is impossible with large datasets to measure linkage accuracy comprehensively. And in cases such as the ASHE-Census linked dataset, where the identifiers used in linking are withheld from the public dataset due to privacy issues, it is impossible for analysts to interrogate linkage accuracy directly. However, three types of information can be useful in informing analysts about potential linkage errors and their implications for analyses:

1. Good-quality documentation of the linkage process
2. Information on any quality-assessment processes that have already been undertaken
3. Information on any features of the linked dataset that enable further quality-assessment to be undertaken by analysts themselves.

Much of the remainder of this *User Guide* focuses on the provision of such information. Section 5 provides a description of the linkage process and quality-assurance work carried out by ONS. Section 6 provides information on linkage outcomes. Section 7 outlines work we have undertaken to address selection bias into the linked dataset through the creation of weights. Sections 7 and 9 explain how

[7] There is also the potential for missing data, if the donor dataset does not contain complete information on the donor characteristics of interest; this can be seen as a particular form of (1). Doidge et al (2020) also discuss the potential for split or merged records, but we do not consider these here, as they are not particularly germane to ASHE-Census linkage.

to access the linked data and outline the content of the linked dataset. Section 12 provides links to further information on ASHE, CEW11 and data linking methodology.

The information provided adheres, where possible, to the GUILD guidelines (Gilbert et al, 2017), which provide guidance on the information that should be made available about data linkage processes to improve the transparency of those processes.

5. The linkage process in detail

The process of linking ASHE to CEW11 was undertaken by ONS Data Engineering Team. The WED team have provided assistance in quality assuring and documenting the linked dataset.

Previous experience within ONS indicated that the optimal linkage methodology would comprise a combination of deterministic and probabilistic linking. Manual linking (also known as clerical linking) was rejected due to the time and expense required to manually link such a large number of records, although clerical checks have been conducted as part of the quality assurance process (see below).

Deterministic linking (also known as deterministic matching) links records from two datasets by looking for exact or partial agreement on across a combination of fields. Records are judged to be true links (or matches) when all or some identifiers are identical across the two records. A threshold is set for the degree of partial agreement that is acceptable.

Probabilistic linking (also known as fuzzy matching) links records by estimating the conditional probabilities of observing agreement on each matching variable. These conditional probabilities are then used to calculate the likelihood that two given records refer to the same entity. Record pairs with probabilities above a certain threshold are considered to be matches, while pairs with probabilities below the threshold are considered to be non-matches.

The linkage process contains five broad stages.

- Stage 1: Data preparation
- Stage 2: Deterministic matching
- Stage 3: Probabilistic matching
- Stage 4: Clerical review
- Stage 5: Additional matching on industry or occupation

In practice, the process was iterative, with the results of initial clerical reviews informing final decisions about the methods to be employed. However, the process is presented chronologically here for simplicity.

Stages 1-4 have been included in Phase 1 of the linking. Stage 5 will be added in Phase 2.

Stage 1: Data preparation

The linkage process was conducted using a non-anonymised version of the ASHE microdata. Apart from the presence of personal identifiers, this dataset was otherwise equivalent to the ASHE microdata that forms the input to the WED ASHE dataset. Records for which the residential location of the employee was recorded to be in Scotland were removed prior to linking.

The resulting subset of ASHE records were linked to a non-anonymised version of the CEW11 microdata held by ONS; this dataset was equivalent to the 100% CEW11 person-level data file held in the Secure Research Service, except that it contained a wider range of variables, from which 317 were selected to be part of the dataset for linking.[8]

These non-anonymised versions of the ASHE and CEW11 datasets had the following common fields:

- Employee name (forename initial, middle name initial, surname)
- Sex (male, female)
- Date of birth (day, month, year)
- Age (years)
- Home postcode (postcode area, postcode district and inward code)
- Work postcode (postcode area, postcode district and inward code)
- Occupation (SOC(2010) Unit Group – four digits)[9]
- Industry sector (SIC(2007) Class or Sub-class – five digits)

These data fields were first standardised to have the same format in each dataset, removing any punctuation and white space, exploding double-barrelled names and so on. A set of 'matching variables' were then defined on each dataset to hold these fields at varying levels of completeness. For instance, one variable would hold the full home postcode, whilst another would contain only the postcode area. The full set of matching variables are shown in Appendix Table A1, along with rates of missing data on these items in ASHE.

A set of 'core matching variables' comprising the employee's name, age, date of birth, home postcode and sex were then selected to be used in Stages 2-4 of the linking process; work postcode, occupation and industry will not feature until Stage 5 (Phase 2).[10] A set of 46 match keys were devised to identify different combinations of these core matching variables. Examples of the match keys are listed in columns 1 and 2 of Table 1 below; a full list is provided in Appendix Table A2.

Table 1: Examples of match keys generated for ASHE-Census linkage

Match key	Description	Score	Uniqueness
1	FORENAME_I(full) + SURNAME(full) + DOB(full) + HOME_POSTCODE(full) + SEX(full)	1.0	100.0
5	FORENAME_I(full) + SURNAME(full) + DOB(6) + HOME_POSTCODE(full) + SEX(full)	0.94	99.9
20	FORENAME_I(full) + SURNAME(soundex) + DOB(6) + HOME_POSTCODE(full) + SEX(full)	0.69	99.2
43	FORENAME_I(full) + SURNAME(full) + DOB(full) + HOME_POSTCODE(0) + SEX(full)	0.75	96.4

Each match key was given a score, shown in column 3 of Table 1. This is a numerical value to indicate the extent to which the match key draws on complete information. To derive this value, each component of the match key is given a weight (or partial score) which reflects its observed

[8] The ONS Census team do not allow the full set of variables to be released. The selection of 317 variables was made by the WED team, choosing those variables most likely to be useful in the analysis of earnings and hours.

[9] SOC(2010) is not present on ASHE 2010 and so SOC(2000) codes were used in that case, but converted to SOC(2010) codes using ONS cross-walk tables.

[10] These variables did not feature in Stages 2-4 was because occupation and industry were felt to be more prone to coding errors than the core matching variables, and work postcode prone to respondent error.

usefulness in matching. The component weights are shown in Appendix Table A3 and are determined after a trial and error process to identify those component items which best help to discriminate between true and false matches (component items which help most are given higher weights). The weights for each component of the match key are summed and then divided by the maximum total weight (12) to give the score shown in Table 1.

Each match key was also given a uniqueness value, which indicates the uniqueness of the specific combination of items within the data to be linked. A high uniqueness value indicates that a match key discriminates well between different records in the data to be linked.

With these elements in place, the match keys were then ranked according to the score and uniqueness and numbered accordingly from 1 to 46 (see Appendix Table A2). The match keys would be applied in this order in Stage 2 (beginning with match key #1 and ending with match key #46).

The final stage of data preparation involved removing duplicate records and then generating the match keys. In ASHE, duplicates were identified as cases for which there was more than one record with identical information for: PIDEN, FORENAME_I(full), SURNAME(full), DOB(full), HOME_POSTCODE(full) and SEX(full). After duplication, each record in the ASHE dataset for linking comprised a unique combination of PIDEN and these core matching variables.[11]

Stage 2: Deterministic matching

In the deterministic matching stage, each ASHE record was taken, one at a time, and the linkage algorithm worked progressively through the match keys, starting with Match Key 1 and moving progressively to Match Key 46. If a match was found to a single Census record at any one of these stages, the matching process stopped and the matched records were kept as a 'matched pair'. The matching process then proceeded to the next ASHE record.

If no match was found on any of the 46 match keys, the ASHE was placed into the residual pot and progressed to Stage 3. If a match was found to two or more Census records, the ASHE record was also placed into the residual pot and progressed to Stage 3.

Stage 3: Probabilistic matching

In the probabilistic matching stage, a match was sought between any of the residual ASHE records passed from Stage 2 and any Census records that remained unmatched from Stage 2. In this stage, the rules for matching were relaxed somewhat, so as to allow for typographic errors in the content of the matching variables, such as the transposition of two or more characters in a surname or the transposition of digits in a postcode.

As a specific example, a typographic error may lead to the surname "Duck" being entered as "Djck" in one or other of the datasets. This would result in a missed link under Stage 2, since an ASHE record with surname "Duck" would not match to a Census record with surname "Djck" under any part of the deterministic matching, despite 75% of the letters being identical. A probability of 0.75 can then be attached to the match between this pair of surnames, however, and this probability can be multiplied by the component weight for surname (shown in Table A3 to be a weight of 3). The resulting value (2.25 in this example) enters the computation of the overall match score for this pair

[11] It is possible for the same PIDEN to generate multiple rows of data, if the underlying ASHE records have different values for the employee's name, date of birth etc recorded against different jobs in ASHE 2010, 2011 or 2012. However, the linkage process will identify the record which links best to the Census, and then match the same Census donor information to all instances of PIDEN in ASHE. In other words, all jobs held by the same individual (PIDEN) in ASHE in any year inherit the same Census information.

of records, which necessarily also takes account of the extent of the match across the other core matching variables. As in the deterministic matching phase, this overall match score indicates the likelihood that the pair of records relates to the same individual.

A probabilistically-derived match score is obtained for all potential matches between the residual ASHE and CEW11 records entering Stage 3, and the pair of records with the highest match score is retained. Records matched via this probabilistic process are assigned Match Key 0.

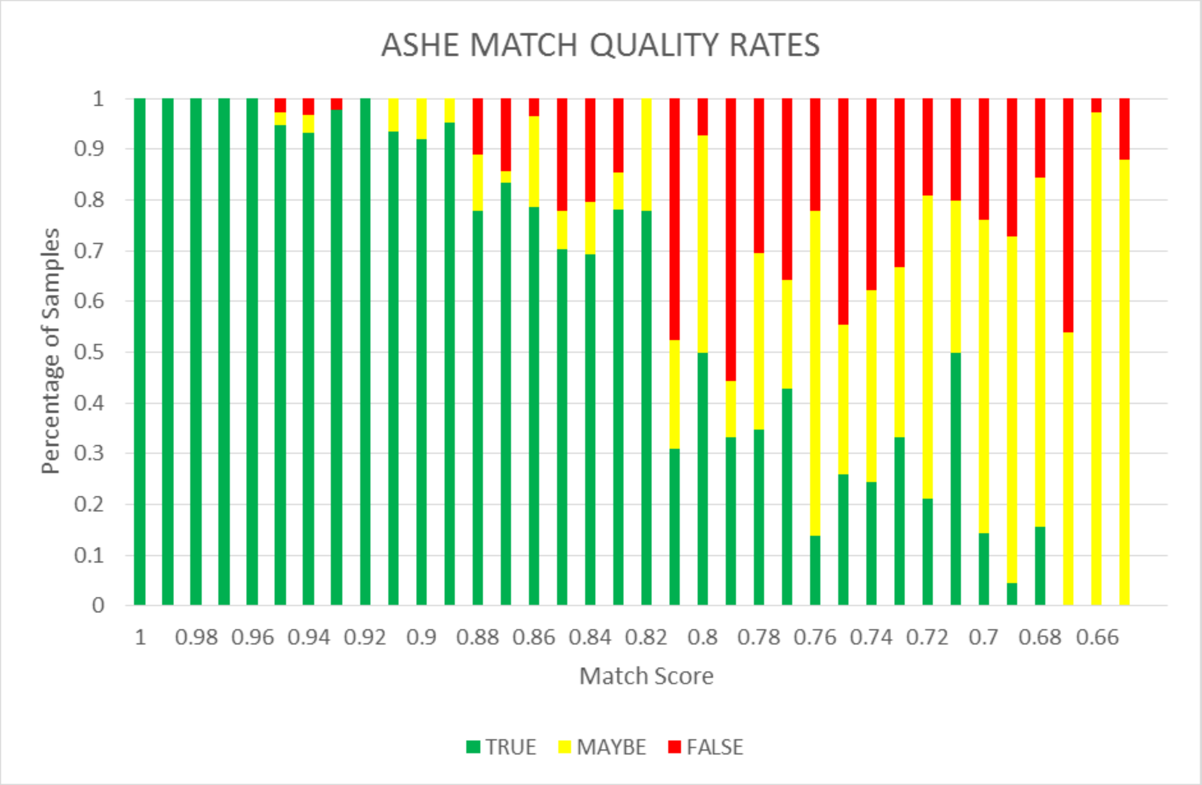
The process recovered only a small number of additional matches: just 328 (0.3%) of the 101,506 employee jobs in ASHE 2011 that achieved a robust match to CEW11 were matched via probabilistic linking.

Stage 4: Clerical review

Clerical checks were conducted on samples of matched records, with the aim of classifying matched pairs according to whether they were correct links or incorrect links (‘false positives’). The objective was to identify a threshold for the match score above which one could be reasonably certain that a pair of records had been correctly linked.

Only those records with a match score greater than, or equal to, 0.65 were subject to clerical review as initial review showed that very few, if any, robust links could be made below this level. The initial cut-off was chosen so as to be confidently below any likely match quality threshold. This subset of matched records with match scores of at least 0.65 were stratified into around 50 groups based on their match score and a random sample of 18 records was selected within each group. Each record was manually inspected and classified as TRUE (clearly a match), FALSE (clearly not a match) or MAYBE (matches well on some items, but others cause suspicion, e.g. different location). In some cases, the clerical outcome was cross-checked by another ONS staff member, to ensure consistent decision making was being undertaken. The results were then analysed to determine the share of TRUE, FALSE and MAYBE matches within each group. These results are presented in Figure 2.

Figure 2: Outcomes of clerical review



Note: Based on a sample of 895 matched records

From this analysis, ONS determined that a match score of 0.82 represented a valid means of discriminating between cases that were likely to represent correct (true) links and those that were likely to represent incorrect (false) links. Within the sample of 895 records:

- 410 true matches had a match score of 0.82 or above
- 47 maybe/false matches had a match score of 0.82 or above
- 92 true matches had a match score below 0.82.

In other words, 90% of links with match scores at or above the threshold of 0.82 were true (also known as the ‘precision rate’). And 82% of all true links were found at or above this threshold (known as the ‘recall rate’).

The matched records generated by the linkage process have been filtered using this match score threshold, such that only those links with a match score of 0.82 or above have been included in the final ASHE-Census dataset.

Stage 5: Matching on industry or occupation

The process of clerical review described above (Stage 4) identified some record-pairs with match scores below the threshold of 0.82 in which the consideration of additional data fields led reviewers to suspect that the pair was in fact a true match. Many of these were cases in which the ASHE record was missing some part of the employee’s name. Three additional fields have therefore been introduced into the linking process in Stage 5, namely: occupation; industry sector; and work postcode. These additional fields are being used in an attempt to establish good-quality matches among the residual cases that were either not matched in Stages 1-4 or which have been matched with a score below 0.82.

Note: Stage 5 is taking place in Phase 2 of the linking process. Any cases linked via this method will appear in Drop 2 of the ASHE-CEDW11 dataset.

6. Linkage rates

The overall linkage rate among the 163,185 jobs in ASHE 2011 held by employees resident in England and Wales is 62% (Table 1). However, this linkage rate varies by employee, job and employer characteristics.

Table 2 presents the marginal effects of the probit model for ASHE 2011 in which a dummy variable (0,1) identifying Census linkage is regressed on a set of employee, job and employer characteristics. Column (1) shows the marginal effects from an unweighted model, thereby indicating the relative linkage rates for different characteristics in the raw dataset. Column (2) shows the marginal effects from the same model after applying the standard ASHE weight (CALWGHT); this model is employed in Section 7 when deriving weights to address linkage biases.

Column 1 shows that linkage rates are substantially lower for older workers than for younger workers or those in middle-age. Linkage rates rise with tenure, and tends to rise with pay level, although there is some attenuation in the highest deciles. Linkage rates are fairly consistent across regions, with the exception of London, where the predicted linkage rate is around 9 percentage points lower than in most other regions; this is a common outcome in data linkage projects. The sample biases created by these differential linkage rates are addressed in Section 7.

Table 2: Marginal effects from probit model to predict Census linkage in 2011 for employees resident in England and Wales

	(1)		(2)	
VARIABLES	Unweighted	S.E	Weighted	S.E
Male	0.053***	(0.003)	0.056***	(0.003)
Age group (reference group: 16-19)				
+ 20-24	-0.137***	(0.008)	-0.133***	(0.008)
+ 25-29	-0.177***	(0.008)	-0.174***	(0.008)
+ 30-34	-0.112***	(0.008)	-0.111***	(0.008)
+ 35-39	-0.052***	(0.008)	-0.052***	(0.008)
+ 40-44	-0.022***	(0.008)	-0.023***	(0.008)
+ 45-49	-0.037***	(0.008)	-0.038***	(0.008)
+ 50-54	-0.137***	(0.008)	-0.139***	(0.008)
+ 55-59	-0.350***	(0.008)	-0.353***	(0.008)
+ 60-64	-0.547***	(0.009)	-0.549***	(0.009)
+ 65 plus	-0.506***	(0.011)	-0.508***	(0.011)
Tenure (reference group: less than 1 year)				
+ 1 year but less than 2 years	0.025***	(0.005)	0.025***	(0.005)
+ 2 years but less than 5 years	0.055***	(0.004)	0.055***	(0.004)
+ 5 years but less than 10 years	0.079***	(0.004)	0.080***	(0.004)
+ 10 years but less than 20 years	0.097***	(0.005)	0.098***	(0.005)
+ 20 years or more	0.036***	(0.005)	0.036***	(0.006)
+ Missing / invalid tenure	0.055***	(0.007)	0.055***	(0.008)
Region (reference group: North East)				
+ North West	-0.010*	(0.006)	-0.009	(0.006)

+ Yorkshire	-0.009	(0.007)	-0.007	(0.007)
+ East Midlands	0.003	(0.007)	0.005	(0.007)
+ West Midlands	-0.007	(0.007)	-0.006	(0.007)
+ South West	-0.009	(0.007)	-0.007	(0.007)
+ East of England	-0.008	(0.006)	-0.005	(0.007)
+ London	-0.093***	(0.006)	-0.090***	(0.006)
+ South East	-0.016***	(0.006)	-0.014***	(0.006)
+ Wales	-0.007	(0.007)	-0.007	(0.007)
Private sector	-0.013***	(0.004)	-0.012***	(0.004)
Pay decile (reference group: 1st decile)				
+ 2nd decile	0.032***	(0.005)	0.029***	(0.005)
+ 3rd decile	0.040***	(0.005)	0.039***	(0.005)
+ 4th decile	0.042***	(0.005)	0.039***	(0.006)
+ 5th decile	0.047***	(0.006)	0.045***	(0.006)
+ 6th decile	0.059***	(0.006)	0.056***	(0.006)
+ 7th decile	0.062***	(0.006)	0.058***	(0.006)
+ 8th decile	0.065***	(0.006)	0.062***	(0.006)
+ 9th decile	0.054***	(0.006)	0.052***	(0.007)
+ 10th decile	0.033***	(0.007)	0.032***	(0.007)
Basic paid hours (reference group: 1-15 hours)				
+ 16-29 hours	-0.012***	(0.004)	-0.011**	(0.004)
+ 30-47 hours	-0.010***	(0.004)	-0.009***	(0.004)
+ 48 hours or more	-0.030***	(0.009)	-0.029***	(0.009)
Observations	*163,185*		*163,185*	
Pseudo-R^2	*0.085*		*0.084*	

Note: ASHE 2011 records for employees resident in England or Wales. Marginal effect generated from probit model. Column (1) reports an unweighted model. Column (2) reports the same model weighted by the standard ASHE weight (*weight*). Other control variables included in the models are: occupation (SOC10, 1 digit), industry (SIC07, 1 digit).

7. Weighting to remove linkage biases

The effect of the differential linkage rates shown in Section 6 is to skew the profile of the ASHE-Census sample away from the profile of the full ASHE sample to some extent. This is shown in Table 3, where column (1) shows the profile of the full ASHE sample in 2011, column (2) shows the profile of the linked ASHE-Census sample in 2011. Column (3) presents the square of the absolute difference between columns (1) and (2) across the categories of each variable as a measure of the bias in the profile of the ASHE-Census sample on each specific characteristic – termed the "squared error". The mean squared error (MSE) is then computed across the categories of each variable as a summary measure of the extent to which the profile of the ASHE-Census sample is biased across that variable as a whole. Larger values of MSE indicate a larger variation between the profile of the full ASHE sample and the ASHE-Census linked sample; in other words, a larger degree of sample bias. It is apparent that the degree of bias is relatively modest for most characteristics, but greatest in respect of employee gender and employee age.

Table 3: Profile of full ASHE sample compared with ASHE-Census linked sample in 2011 for employees resident in England and Wales

	(1)	(2)	(3)	(4)	(5)
	ASHE sample	**ASHE-Census sample**	**Squared error (1-2)^2**	**ASHE-Census sample**	**Squared error (1-4)^2**
	Standard weight	**Standard weight**		**ASHE-Census weight**	
	Col %	Col %		Col %	
Gender					
+ Female	49.32	47.46	3.46	49.45	0.02
+ Male	50.68	52.54	3.46	50.55	0.02
Mean squared error (MSE)			**3.46**		**0.02**
Age group					
+ 16-19	3.61	4.06	0.20	3.72	0.01
+ 20-24	9.13	8.72	0.17	9.29	0.03
+ 25-29	11.35	10.50	0.72	11.52	0.03
+ 30-34	11.46	12.17	0.50	11.70	0.06
+ 35-39	11.52	13.48	3.84	11.79	0.07
+ 40-44	13.00	15.83	8.01	13.36	0.13
+ 45-49	13.25	15.82	6.60	13.61	0.13
+ 50-54	10.99	11.36	0.14	11.34	0.12
+ 55-59	8.46	5.50	8.76	8.64	0.03
+ 60-64	5.27	1.80	12.04	3.52	3.06
+ 65 plus	1.96	0.76	1.44	1.50	0.21
Mean squared error (MSE)			**3.86**		**0.35**
Tenure					
+ less than 1 year	13.84	12.85	0.98	13.97	0.02
+ 1 year but less than 2 years	11.10	10.73	0.14	11.24	0.02
+ 2 years but less than 5 years	24.93	25.28	0.12	25.26	0.11
+ 5 years but less than 10 years	21.75	22.85	1.21	21.82	0.00
+ 10 years but less than 20 years	16.05	17.00	0.90	15.86	0.04
+ 20 years or more	9.22	8.31	0.83	8.73	0.24
+ Missing/invalid tenure	3.10	2.98	0.01	3.11	0.00
Mean squared error (MSE)			**0.60**		**0.06**
Pay decile					
+ 1st decile	9.84	8.86	0.96	9.82	0.00
+ 2nd decile	9.12	8.52	0.36	9.16	0.00
+ 3rd decile	9.28	8.88	0.16	9.37	0.01
+ 4th decile	9.31	9.06	0.06	9.33	0.00
+ 5th decile	9.37	9.30	0.00	9.36	0.00
+ 6th decile	9.69	9.90	0.04	9.73	0.00
+ 7th decile	10.11	10.54	0.18	10.14	0.00
+ 8th decile	10.64	11.30	0.44	10.67	0.00
+ 9th decile	10.94	11.60	0.44	10.91	0.00
+ 10th decile	11.70	12.04	0.12	11.53	0.03
Mean squared error (MSE)			**0.28**		**0.00**
Private/ Public					
+ Public	27.13	27.85	0.52	27.05	0.01
+ Private	72.87	72.15	0.52	72.95	0.01
TOTAL Mean squared error (MSE)			**0.52**		**0.01**
Basic hours worked					
+ 0-15 hours	12.93	11.95	0.96	12.79	0.02
+ 16-29 hours	17.13	16.25	0.77	16.94	0.04
+ 30-47 hours	68.03	70.03	4.00	68.33	0.09
+ 48 hours or more	1.91	1.77	0.02	1.94	0.00
Mean squared error (MSE)			**1.44**		**0.04**

Occupation					
+ Managers, directors and senior official	9.27	9.72	0.20	9.28	0.00
+ Science, research, engineering and technology professionals	19.98	20.61	0.40	19.71	0.07
+ Associate professional and technical occupations	13.60	14.28	0.46	13.68	0.01
+ Administrative and secretarial occupations	12.42	12.45	0.00	12.50	0.01
+ Skilled trades occupations	7.97	8.06	0.01	7.84	0.02
+ Caring, leisure and other service occupations	9.55	9.15	0.16	9.64	0.01
+ Sales and customer service occupations	9.01	9.01	0.00	9.16	0.02
+ Process, plant and machine operatives	6.00	5.69	0.10	5.96	0.00
+ Elementary occupations	12.20	11.04	1.35	12.23	0.00
Mean squared error (MSE)			**0.30**		**0.02**
Industry					
+ Agriculture, Mining, Manufacturing, Electricity, Water	11.56	12.1	0.29	11.45	0.01
+ Construction	3.57	3.49	0.01	3.56	0.00
+ Wholesale, retail, repair of vehicles	15.76	16.12	0.13	15.94	0.03
+ Transport, and storage	4.37	4.49	0.01	4.39	0.00
+ Accommodation, and food service	5.02	4.30	0.52	5.03	0.00
+ Information, and communication	3.86	4.05	0.04	3.86	0.00
+ Financial and insurance activities	4.45	4.65	0.04	4.51	0.00
+ Real estate activities	1.23	1.23	0.00	1.23	0.00
+ Professional, scientific, and technical activities	5.69	5.76	0.00	5.72	0.00
+ Admin and support services	5.90	5.31	0.35	5.89	0.00
+ Public admin and defence	5.25	5.44	0.04	5.26	0.00
+ Education	15.94	16.01	0.00	15.75	0.04
+ Health, and social work	13.63	13.42	0.04	13.64	0.00
+ Art, entertainment, and creation	1.76	1.78	0.00	1.79	0.00
+ Other service activities/ missing	2.01	1.87	0.02	1.98	0.00
Mean squared error (MSE)			**0.10**		**0.01**
Region					
+ North East	4.62	4.77	0.02	4.62	0.00
+ North West	12.32	12.51	0.04	12.35	0.00
+ Yorkshire	9.47	9.66	0.04	9.48	0.00
+ East Midlands	8.16	8.46	0.09	8.17	0.00
+ West Midlands	9.83	10.04	0.04	9.84	0.00
+ South West	9.87	9.92	0.00	9.85	0.00
+ East of London	10.91	11.12	0.04	10.94	0.00
+ London	13.53	12.04	2.22	13.51	0.00
+ South East	16.04	16.11	0.00	16.00	0.00
+ Wales	5.27	5.38	0.01	5.24	0.00
Mean squared error (MSE)			**0.25**		**0.00**
Observations	*163,185*	*100,698*		*100,698*	

Note: ASHE 2011 records for employees resident in England or Wales. The squared error is the square of the absolute difference between the full ASHE sample and the ASHE-Census sample. The standard ASHE weight (weight) is used in both column (1) and column (2).

To address the sample biases that occur as a result of the linkage process, we construct a weight which adjust the profile of the linked ASHE-Census to (approximately) match the profile of the full ASHE sample. The weight can be applied when producing estimates from the ASHE-Census sample to make those estimates representative of all jobs held by employees resident in England and Wales in the sampled year.

First, we use the probit model shown in column (2) of Table 2 to derive the predicted probability for jobs in ASHE being linked with the 2011 Census: *p(linked)*. We then create the ASHE-Census weight (ACEW11_WT) as:

$$acew11\_wt = weight * \frac{1}{p(linked)}$$

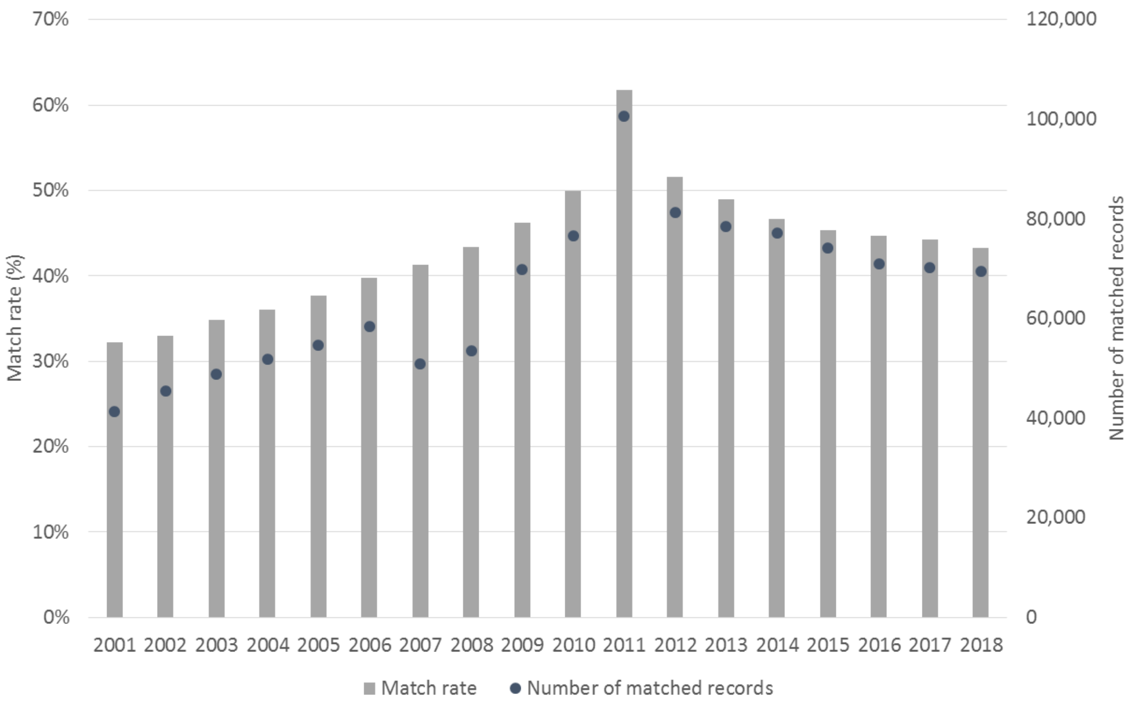
This new weight is trimmed in order to remove the influence of outliers, following the approach recommended in Valliant and Dever (2018: 157), that is capping the maximum and minimum weight values as equal to the median value of the weight, plus or minus three times the value of the interquartile range. The final weight is scaled to have the same total as the original ASHE weight.

Columns (4) and (5) of Table 3 show the performance of ACEW11_WT in correcting the profile of the ASHE-Census sample. The most substantial biases are reduced and the mean squared error is no more than 0.35 for any characteristic shown in the table. The Kish design effect of the new weight is also close to that of the original ASHE weight (1.09 compared with 1.04). The new weight therefore performs well in terms of reducing bias without the penalty of larger standard errors.

Weighting for linkage bias in other years

As noted earlier, the linked Census data is spread out across the different years of the ASHE dataset via the unique employee identifier (PIDEN). An employee who appears in 2007, 2011 and 2015 (for example) will then have the same Census data appended to their job records in each year. The share of records with linked Census data declines as one moves further away from 2011 (see Figure 3) and the linkage bias may not be identical in each year, so we have derived ACEW11_WT separately for each year of the ASHE panel. The approach is identical to that described above, except that the coefficients of the probit model used to generate *p(linked)* are allowed to vary by year, and the trimming is also carried out within-year. This approach is flexible in allowing the nature of the linkage bias to differ across years, although it assumes that the dimensions across the bias differs are the same.

Figure 3: Share of ASHE records linked to CEW11, by year, for employees resident in England and Wales



Note: ASHE records for employees resident in England and Wales

8. Linkage quality

In this section, we provide an overview of linkage quality in the linked ASHE-Census dataset. We begin by looking at the distribution of match scores. Then we examine the degree of correspondence on variables that ASHE and Census have in common.

Table 4 shows the distribution of match scores for all linked records in 2011. Around 80% of linked cases have a match score of 1.0. Almost all of these records were linked via Match Key 1. Around 90% of linked cases have a match score of at least 0.90.

Table 4: Distribution of match scores for linked cases, 2011

Match score	Per cent (%)
1.0	78.8
0.95-0.99	3.9
0.90-0.94	7.8
0.85-0.89	6.5
0.82-0.84	3.1
All	**100.0**

Note: ASHE-Census linked records for employees resident in England and Wales

Table 5 reports on the degree of correspondence between ASHE and Census on variables that are common to both datasets. The degree of correspondence is necessarily high for the core matching variables (gender, age, home region). However, some discrepancies are apparent; this may arise when the case has been linked by virtue of the strength of the match on other data items. In cases where the two datasets do not agree on a particular item, either dataset could be in error; there is no *a priori* reason to prefer one source over the other and analysts may wish to check whether their results are sensitive to the choice of source dataset for such items.

Table 5: Degree of correspondence between ASHE and Census

Variable	Consistency
Gender: Male/Female	99.69%
Age: Years	94.19%
Home region: Government Office Region	99.32%
Occupation: SOC(2010) Major Group)	67.88%
Industry: SIC(2007) Section	73.22%

Note: ASHE-Census linked records for employees resident in England and Wales

9. Features of the ASHE-Census dataset

The ASHE-Census dataset contains four sets of variables:

ASHE variables: The dataset contains all ASHE variables from the WED version of the ASHE microdata made available in the SRS, with one exception: the ASHE-Census dataset contains none of the ASHE variables relating to residential geography below the level of Local Authority District (LAD: ASHE variable HLA). The removal of ASHE residential geography below LAD-level from the published ASHE-CEW11 dataset was a condition of the data owners of CEW11 in order for matching to go ahead, and was intended to maintain the confidentiality of Census respondents. However, workplace geography is available at Output Area level, and Census variables down to Output Area level are present for linked cases. Naming conventions for ASHE variables follow the WED version of the ASHE data. Variables and values are labelled.

Census variables: The dataset contains 317 variables taken from the CEW11 individual-level data file held in the Secure Research Service. All CEW11 variables on the ASHE-Census dataset have the name suffix "_census". Variables are labelled but values remain unlabelled; code frames are provided in the SRS Core Documentation Library for Census 2011. At the time of writing (April 2022), the WED Team are seeking to add further variables to describe the economic activity of other members of the employee's family or household; this is work in progress.

Linkage variables: The Drop 1 dataset contains 12 variables that document aspects of the linkage process (see Table 6). All have the name suffix "_link"; variables and values are labelled. These variables are provided so that researchers can examine the sensitivity of their results to the inclusion or exclusion of cases with differing levels of match quality (e.g. selecting only those cases with a match score of 1.0).

Table 6: Linkage variables included on the dataset

Variable name	Type	Description
matchkey_link	Numeric	The number of the match key used to link the case
desc_link	String	Description of the match key used to link the case
score_link	Numeric	Overall match score
forename_i_score_link	Numeric	Component score for FORENAME_I
surname_score_link	Numeric	Component score for SURNAME

sex_score_link	Numeric	Component score for SEX
dob_4_score_link	Numeric	Component score for DOB_4
dob_6_score_link	Numeric	Component score for DOB_6
dob_end_score_link	Numeric	Component score for DOB_END
home_postcode_2_score_link	Numeric	Component score for HOME_POSTCODE_2
home_postcode_4_score_link	Numeric	Component score for HOME_POSTCODE_4
home_postcode_end_score_link	Numeric	Component score for HOME_POSTCODE_END

Weighting variables: The weight variable described in Section 7 is named ACEW11_WT. This weight variable is observed for all cases belonging to employees resident in England and Wales (1<=HGOR<=10) and has been derived in each year of the data from 2001 to 2018. The weight seeks to ensures that estimates based on those ASHE job records that could be linked to CEW11 are representative of all jobs held by employees resident in England and Wales.

10. Illustrative analyses

This section provides a brief illustration of the analytical capabilities of the ASHE-Census linked dataset.

Table 7 presents mean gross hourly wages across a range of selected employee characteristics in 2011 for employees resident in England and Wales. All estimates are produced from the ASHE-Census linked dataset using the ASHE-Census weight described in Section 7. Whilst gender is observed in the standard ASHE dataset, the remaining characteristics are observed from the Census. The table shows clear heterogeneity in mean gross hourly wages by ethnicity, with employees in the Indian and Chinese ethnic groups earning more than those in the White ethnic group, whilst employees in other non-White groups earn considerably less. The table also shows a clear wage premium for those with a degree-level qualification, and wage penalties for those with a disability and those born outside the UK.

Table 7: Average gross hourly wages across selected employee characteristics in 2011 for employees resident in England and Wales

	Observations	Mean (£/hour)	Median (£/hour)
Gender			
+ Female	50,799	12.10	9.55
+ Male	49,899	16.11	12.46
Ethnicity			
+ White	89,955	14.18	10.92
+ Indian	2,745	16.07	12.15
+ Pakistani	1,047	12.88	9.80
+ Bangladeshi	367	11.49	9.01
+ Chinese	431	16.78	14.23
+ Black African	1,260	12.34	10.00
+ Black Caribbean	1,172	13.34	11.50
Education			
+ No qual.	7,483	8.98	7.74
+ GCSEs	33,401	10.58	8.87
+ Apprenticeship	2,871	12.17	10.88
+ A-level	15,825	11.73	9.81
+ Degree	36,747	19.71	16.68
+ Other/Vocational	3,868	10.66	8.64
Disability			
+ Non-disable	95,410	14.25	11.00
+ Disabled	4,809	12.53	9.84
Non-UK born			
+ UK born	98,010	14.21	11.00
+ non-UK born	2,209	12.48	8.10

Note: ASHE-Census linked records for employees resident in England and Wales. Mean and median are weighted estimates generated with the ASHE-Census weight.

Table 8 goes on to present the results of ordinary least squares (OLS) regressions of log gross hourly wages in 2011 for employees resident in England and Wales. Column 1 presents the wage result in the full ASHE sample with the standard ASHE weight (CALWGHT), and Column 2 shows the result controlling for ASHE variables only for the ASHE-Census sample with the adjusted weight (ACEW11_WT). Census variables are added in Column 3 with the same weight.

The coefficients from the models presented in Columns 1 and 2 are broadly consistent, indicating that the ASHE-Census sample is relatively unbiased when compared with the full ASHE sample, after applying relevant weights. The largest discrepancies occur in respect of the wage premium for being male, and the wage premium associated with living in London, both of which are estimated to be 0.9 log points higher in the ASHE-Census sample.

Column 3 of the table adds further control variables taken from the Census. These comprise person characteristics (ethnicity, education, disability, non-UK born, English language, health status), and family characteristics (couple family, carer duty, number of children, age of the youngest child). The R-squared for the regression increases from 0.578 to 0.611 with the addition of these control variables. Coefficients are shown for ethnicity, education and disability status, all of which show the expected signs. The effects on ASHE variables include small reductions in the variation of wages with age, in the male wage premium and in the premium associated with living in London.

Table 8: Ordinary Least Squares (OLS) regression of log gross hourly wages in 2011 for employees resident in England and Wales

VARIABLES	(1) Full ASHE sample	(2) ASHE-Census sample	(3) ASHE-Census sample
Age	0.038***	0.042***	0.037***
	(0.001)	(0.001)	(0.001)
Age square (/100)	-0.041***	-0.046***	-0.038***
	(0.001)	(0.001)	(0.001)
Male	0.120***	0.129***	0.121***
	(0.002)	(0.003)	(0.003)
Tenure	0.014***	0.014***	0.014***
	(0.000)	(0.001)	(0.001)
Tenure square	-0.024***	-0.026***	-0.025***
	(0.001)	(0.002)	(0.002)
Part-time	-0.061***	-0.060***	-0.066***
	(0.002)	(0.003)	(0.003)
Private	-0.069***	-0.066***	-0.061***
	(0.003)	(0.004)	(0.004)
Home Region			
+ North West	0.024***	0.027***	0.029***
	(0.005)	(0.006)	(0.006)
+ Yorkshire	0.007	0.006	0.012*
	(0.005)	(0.006)	(0.006)
+ East Midlands	0.030***	0.028***	0.029***
	(0.005)	(0.007)	(0.007)
+ West Midlands	0.018***	0.014**	0.017***
	(0.005)	(0.006)	(0.006)
+ South West	0.035***	0.032***	0.027***
	(0.005)	(0.006)	(0.006)
+ East of London	0.088***	0.090***	0.096***
	(0.005)	(0.006)	(0.006)
+ London	0.210***	0.219***	0.240***
	(0.005)	(0.006)	(0.007)
+ South East	0.119***	0.121***	0.122***
	(0.005)	(0.006)	(0.006)
+ Wales	-0.007	-0.007	-0.010
	(0.006)	(0.007)	(0.007)
Ethnicity			
+ Indian			-0.036***
			(0.009)
+ Pakistani			-0.045***
			(0.013)
+ Bangladeshi			-0.102***
			(0.021)
+ Chinese			0.009
			(0.020)
+ Black African			-0.134***
			(0.011)
+ Black Caribbean			-0.079***
			(0.010)
Education			

+ GCSEs			0.073***
			(0.004)
+ Apprenticeship			0.091***
			(0.008)
+ A-level			0.142***
			(0.005)
+ Degree			0.318***
			(0.005)
+ Other/Vocational qual.			0.075***
			(0.007)
Disabled			-0.045***
			(0.006)
Occupation	Y	Y	Y
Industry	Y	Y	Y
Other individual and family characteristics (from Census)	N	N	Y
Observations	146,720	90,921	82,295
R-squared	0.577	0.578	0.611

Note: This table reports wage equation estimates for employees in England and Wales, 2011. Column (1) reports the result for the full ASHE sample with the standard ASHE weight while columns (2) and (3) report the result for the ASHE-Census sample with the adjusted ASHE-Census weight. The control variables for columns (1) and (2) are from ASHE only. Column (3) adds Census variables, including person characteristics (ethnicity, education, disability, non-UK born, English language, health status), and family characteristics (couple family, carer duty, number of children, age of the youngest child).
*** p<0.01, ** p<0.05, * p<0.1

11. Accessing the linked data

Drop 1 of the linked ASHE-Census dataset is not yet available to researchers. However, it is anticipated that it will be released in July 2022. The dataset will be made available via the ONS Secure Research Service.

Users wishing to access the data should make an application via the Secure Research Service in the usual way.

12. Further information

- A general introduction to methods of data linking and potential linkage errors is provided by Doidge et al (2020).

- Further information on ASHE is provided on the WED website: http://www.wagedynamics.com/data-documentation/data-variable-description/

- Further information on the 2011 Census is provided on the ONS Census micro-site: https://www.ons.gov.uk/census/2011census

13. ASHE-Census data for Northern Ireland

The Northern Ireland Statistics and Research Agency (NISRA) have recently undertaken a similar process of matching the 5,770 records from the Northern Ireland ASHE 2011 to the 2011 Census for Northern Ireland. The resulting dataset is made available for research as the "Earnings and Employees Study (EES) 2011".

The main linkage process used by NISRA was similar to that used by ONS for England and Wales, except that a greater amount of manual input was dedicated by NISRA to achieving matches for residual cases that were not matched via linkage algorithms. ONS and NISRA have collaborated to compare matching outcomes and these are reported to be similar, once the lower levels of missing name data in the Northern Ireland ASHE are taken into account. However, donor imputation features as an additional, final stage in NISRA's process, resulting in Census information ultimately being matched to all 5,770 records. Donor imputation was not used by ONS in the matching process for England and Wales.

Further information on the Earnings and Employees Study (EES) 2011 is provided here: https://www.nisra.gov.uk/support/research-support/administrative-data-research-northern-ireland-adr-ni-themed-datasets

14. Bibliography

Doidge J, Christen P and Harron K (2020) *Quality Assessment in Data Linkage*, London: Office for National Statistics.

Gammon S (2016) "Data Linkage at ONS", presented at workshop on *Perspectives on Data Linkage – Techniques, Challenges and Applications*, Isaac Newton Institute, Cambridge, 16th September.

Gilbert R, Lafferty R, Haggar-Johnson G, Harron K, Zhang L, Smith P, Dibben C and Goldstein H (2017) "GUILD: GUidance for Information about Linking Data sets", *Journal of Public Health*, 40, 1: 191-198.

Jenkins J (2008) "Linking the Annual Survey of Hours and Earnings to the Census: a feasibility study", *Economic and Labour Market Review*, 2, 2: 37-41.

Mayer A and Stockdale J (2021) "Developing standard tools for data linkage: February 2021", ONS Working Paper Series, London: Office for National Statistics.

15. Appendix

Table A1: Matching variables used in ASHE-Census linking

Component	Variable	Example	Missing in ASHE (%)
Age	AGE	91	0
Date of birth (yyyymmdd)	DOC	19281118	0
Sex	SEX	M or F	0
Surname	SURNAME	Duck	23.3
Forename initial	FORENAME_I	D	23.8
Home postcode	HOME_POSTCODE	SW1A1AA	0.4
Work postcode	WORK_POSTCODE	EC1Y8TZ	Awaiting
Occupation (SOC(2010) Unit Group)	OCC	5221	0
Industry sector (SIC(2007) Class or Sub-class)	SIC	52220	Awaiting

Notes: Some missingness rates have to be provided by ONS. Missingness rates for the equivalent data items CEW11 have not yet been provided.

Table A2: Full list of match keys

No.	Description	Full Count	Null Count	Min. Score	Max. Score	Uniqueness (%)
1	FORENAME_I(full) + SURNAME(full) + DOB(full) + HOME_POSTCODE(full) + SEX(full)	5	0	1.00	1.00	100.0
2	FORENAME_I(full) + SURNAME(sorted) + DOB(full) + HOME_POSTCODE(full) + SEX(full)	4	0	0.75	0.75	100.0
3	FORENAME_I(full) + SURNAME(full) + DOB(full) + HOME_POSTCODE(full) + SEX(0)	4	1	0.92	0.92	100.0
4	FORENAME_I(full) + SURNAME(sorted) + DOB(full) + HOME_POSTCODE(full) + SEX(0)	3	1	0.67	0.67	100.0
5	FORENAME_I(full) + SURNAME(full) + DOB(6) + HOME_POSTCODE(full) + SEX(full)	4	0	0.94	0.94	99.9
6	FORENAME_I(full) + SURNAME(sorted) + DOB(6) + HOME_POSTCODE(full) + SEX(full)	3	0	0.69	0.69	99.9
7	FORENAME_I(full) + SURNAME(full) + DOB(full) + HOME_POSTCODE(4) + SEX(full)	4	0	0.95	0.95	99.9
8	FORENAME_I(full) + SURNAME(sorted) + DOB(full) + HOME_POSTCODE(4) + SEX(full)	3	0	0.70	0.70	99.9
9	FORENAME_I(full) + SURNAME(full) + DOB(6) + HOME_POSTCODE(full) + SEX(0)	3	1	0.85	0.85	99.9
10	FORENAME_I(full) + SURNAME(full) + DOB(full) + HOME_POSTCODE(4) + SEX(0)	3	1	0.87	0.87	99.9
11	FORENAME_I(full) + SURNAME(full) + DOB(6) + HOME_POSTCODE(4) + SEX(full)	3	0	0.89	0.89	99.8
12	FORENAME_I(full) + SURNAME(full) + DOB(4) + HOME_POSTCODE(full) + SEX(full)	4	0	0.88	0.88	99.7
13	FORENAME_I(full) + SURNAME(full) + DOB(full) + HOME_POSTCODE(2) + SEX(full)	4	0	0.85	0.85	99.7
14	FORENAME_I(full) + SURNAME(full) + DOB(6) + HOME_POSTCODE(4) + SEX(0)	2	1	0.80	0.80	99.7
15	FORENAME_I(full) + SURNAME(full) + DOB(full) + HOME_POSTCODE(2) + SEX(0)	3	1	0.77	0.77	99.6
16	FORENAME_I(full) + SURNAME(full) + DOB(4) + HOME_POSTCODE(full) + SEX(0)	3	1	0.79	0.79	99.5
17	FORENAME_I(full) + SURNAME(soundex) + DOB(full) + HOME_POSTCODE(full) + SEX(full)	4	0	0.75	0.75	99.3
18	FORENAME_I(full) + SURNAME(soundex) + DOB(full) + HOME_POSTCODE(full) + SEX(0)	3	1	0.67	0.67	99.3
19	FORENAME_I(full) + SURNAME(soundex) + DOB(full) + HOME_POSTCODE(4) + SEX(full)	3	0	0.70	0.70	99.2
20	FORENAME_I(full) + SURNAME(soundex) + DOB(6) + HOME_POSTCODE(full) + SEX(full)	3	0	0.69	0.69	99.2
21	FORENAME_I(0) + SURNAME(full) + DOB(full) + HOME_POSTCODE(full) + SEX(full)	4	1	0.83	0.83	99.1

22	FORENAME_I(full) + SURNAME(full) + DOB(4) + HOME_POSTCODE(4) + SEX(full)	3	0	0.83	0.83	98.9
23	FORENAME_I(0) + SURNAME(full) + DOB(full) + HOME_POSTCODE(full) + SEX(0)	3	2	0.75	0.75	98.9
24	FORENAME_I(0) + SURNAME(full) + DOB(full) + HOME_POSTCODE(4) + SEX(full)	3	1	0.78	0.78	98.9
25	FORENAME_I(0) + SURNAME(full) + DOB(6) + HOME_POSTCODE(full) + SEX(full)	3	1	0.77	0.77	98.8
26	FORENAME_I(0) + SURNAME(full) + DOB(full) + HOME_POSTCODE(4) + SEX(0)	2	2	0.70	0.70	98.7
27	FORENAME_I(full) + SURNAME(2) + DOB(full) + HOME_POSTCODE(full) + SEX(full)	4	0	0.75	1.00	98.4
28	FORENAME_I(full) + SURNAME(full) + DOB(6) + HOME_POSTCODE(2) + SEX(full)	3	0	0.79	0.79	98.4
29	FORENAME_I(0) + SURNAME(full) + DOB(6) + HOME_POSTCODE(full) + SEX(0)	2	2	0.69	0.69	98.4
30	FORENAME_I(full) + SURNAME(2) + DOB(full) + HOME_POSTCODE(full) + SEX(0)	3	1	0.67	0.92	98.4
31	FORENAME_I(full) + SURNAME(2) + DOB(full) + HOME_POSTCODE(4) + SEX(full)	3	0	0.70	0.95	98.3
32	FORENAME_I(full) + SURNAME(full) + DOB(4) + HOME_POSTCODE(4) + SEX(0)	2	1	0.74	0.74	98.3
33	FORENAME_I(0) + SURNAME(full) + DOB(4) + HOME_POSTCODE(full) + SEX(full)	3	1	0.71	0.71	98.2
34	FORENAME_I(full) + SURNAME(1) + DOB(full) + HOME_POSTCODE(full) + SEX(full)	4	0	0.75	1.00	98.2
35	FORENAME_I(0) + SURNAME(full) + DOB(full) + HOME_POSTCODE(2) + SEX(full)	3	1	0.68	0.68	98.2
36	FORENAME_I(full) + SURNAME(1) + DOB(full) + HOME_POSTCODE(full) + SEX(0)	3	1	0.67	0.92	98.1
37	FORENAME_I(full) + SURNAME(1) + DOB(full) + HOME_POSTCODE(4) + SEX(full)	3	0	0.70	0.95	98.0
38	FORENAME_I(full) + SURNAME(2) + DOB(6) + HOME_POSTCODE(full) + SEX(full)	3	0	0.69	0.94	98.0
39	FORENAME_I(0) + SURNAME(full) + DOB(6) + HOME_POSTCODE(4) + SEX(full)	2	1	0.72	0.72	98.0
40	FORENAME_I(full) + SURNAME(full) + DOB(6) + HOME_POSTCODE(2) + SEX(0)	2	1	0.70	0.70	97.6
41	FORENAME_I(full) + SURNAME(1) + DOB(6) + HOME_POSTCODE(full) + SEX(full)	3	0	0.69	0.94	97.3
42	FORENAME_I(full) + SURNAME(full) + DOB(0) + HOME_POSTCODE(full) + SEX(full)	4	1	0.75	0.75	96.7
43	FORENAME_I(full) + SURNAME(full) + DOB(full) + HOME_POSTCODE(0) + SEX(full)	4	1	0.75	0.75	96.4
44	FORENAME_I(full) + SURNAME(0) + DOB(full) + HOME_POSTCODE(full) + SEX(full)	4	1	0.75	1.00	96.0
45	FORENAME_I(full) + SURNAME(0) + DOB(full) + HOME_POSTCODE(full) + SEX(0)	3	2	0.67	0.92	95.8

46	FORENAME_I(full) + SURNAME(0) + DOB(6) + HOME_POSTCODE(full) + SEX(full)	3	1	0.69	0.94	95.1
0	Probabilistic link					

Notes:

(i) Value in brackets indicates the number of characters/digits used for the variable in compiling the match key (full=all; 6=first six; 4=first four etc; 0=none). ‘Sorted’ indicates that the characters were alphabetically sorted; this technique can be used to accommodate the accidental transposition of characters in hand-typed text. ‘Soundex’ indicates that a numerical value was used, representing the string as spoken; again this technique can be used to accommodate typographical errors.

(ii) Full count indicates the number of variables for which full information is used; Null count indicates the number for which no information is used

(iii) Scores are compiled by summing the components weights showed in Table A3 and dividing by 12 (the maximum total). Weights were determined by ONS on the basis of prior linkage work.

(iv) A high uniqueness value indicates that a match key discriminates well between different records in the data to be linked.

Table A3: Component weights used to determine the score for each match key

Match Key Component	**Weight**
FORENAME_I(full)	2
SURNAME(full)	3
DOB(4)	1.5
DOB(6)	2.25
DOB(full)	3
HOME_POSTCODE(2)	1.2
HOME_POSTCODE(4)	2.4
HOME_POSTCODE(full)	3
SEX(full)	1

Note: These component weights were derived in two stages. The initial set of weights were based on experience from previous data linking exercises. The ASHE-Census linking team then used information from clerical reviews of ASHE-Census linked records to iteratively arrive at a final set of weights which maximised the number of 'true matches' whilst accurately discriminating between good matches and poor ones.