

FEATURE

Jamie Jenkins
Office for National Statistics

Linking the Annual Survey of Hours and Earnings to the Census: a feasibility study

SUMMARY

This article describes a project to link the Annual Survey of Hours and Earnings (ASHE – formerly the New Earnings Survey) and the 2001 Census. The investigation looks at the feasibility of linking the two data sets using a sample of the census data. Linking the data would enhance the ASHE data set by adding the personal characteristics of individuals. The results show that there is the potential to link the two data sets although further work would be needed using the whole census data set to ensure the matched data was not biased.

Linking survey data to the census provides opportunities to enhance the value of the original data source by adding supplementary variables for analysis and research. It can provide information that is not possible to obtain through the survey or reduce the burden on respondents by substituting survey questions with information from the census. Other potential benefits are the opportunities to cross-validate common variables between the survey and the census.

This article considers the potential to link the Annual Survey of Hours and Earnings (ASHE), formerly known as the New Earnings Survey (NES), to the census. ASHE is used for most UK micro- and macroeconomic analysis of earnings within the labour market and is a 1 per cent (0.8 per cent in 2007) sample of all employees in the UK. The survey reference period is such that it includes a pay period for a date in April of each year. While ASHE has a wealth of information on the earnings and hours of employees, it contains very little on their personal characteristics.

The purpose of linking the two data sets is to supplement the ASHE data set with variables on the personal characteristics of individuals. This would enhance the knowledge of the distribution of earnings and the numbers of people paid below the National Minimum Wage. It can also be used to generate a pay inequality model for areas such as gender, ethnicity and disability. Matching also allows for a comparison of common variables such as the number of hours worked per week

and industry/occupation of work. Any information published would be subject to the same confidentiality rules as the main census. A linked data set would also be anonymised to protect the identity of individuals.

This article explains the data sources used for the data linkage project. The next section explains the method used to link the data sets. The subsequent section will describe the success in linking the data. The final section considers any potential bias in using a linked data set and the path for future work.

Data sources NES and ASHE

NES was an annual survey, run every April from 1970 to 2003, of the earnings of employees in Great Britain. It was based on a sample of 1 per cent of employees registered for the pay-as-you-earn (PAYE) tax collection system run by HM Revenue & Customs, formerly Inland Revenue. The information collected related to gross earnings before tax, national insurance or other deductions and was completed from employers' pay records.

ASHE was introduced in 2004 to replace NES. ASHE includes improvements to the coverage of employees not originally in the NES sample, imputation for item non-response and the weighting of earnings estimates to overcome unit non-response. The questionnaire for ASHE was improved in 2005 and included improvements to the collection of data relating to allowances and incentive pay. The Department of

Enterprise, Trade and Investment conducts a similar but separate survey in respect of employees in Northern Ireland to allow for UK estimates. This was also the case for NES.

For the rest of this article, the term ASHE will be used to mean NES or ASHE interchangeably.

Census

Since 1801, every ten years, the nation has set aside one day for the census – a count of all people and households. It is the most complete source of information about the population that is available. It is the only source of information which provides a detailed picture of the entire population, and is unique because it covers everyone at the same time and asks the same core questions everywhere. For the 2001 Census, information is available for personal characteristics such as qualifications, ethnicity, religion, marital status, economic activity, employment status, socio-economic class, country of birth and health status.

Linking methodology

It should be possible to link information for all individuals in the ASHE sample in 2001 to their census record, as the census also took place in April of that year, and they have some common variables. In order to assess the feasibility of linking the two sources, a subset of the census data has been made available. This subset is split into:

- Cornwall in the Government Office Region of the South West
- Bexley in London, and
- Bedfordshire in the East

Each individual is uniquely identified in ASHE through their National Insurance

number. The census does not collect this information, so direct linkage through this unique identifier is not possible. As there should be close to complete overlap between the data sets, it is possible to use a combination of common variables for matching. Both ASHE and the census contain personal details such as:

- date of birth
- first name
- middle name
- surname
- gender

Data in ASHE are those which are contained on the sample file, while in the census they are provided by the respondent who fills out the form. It is possible to use this information to link the two data sources although there are some issues. One is that matching may not be unique, as there may be instances where two or more individuals have the same personal details. Another is that personal information is recorded incorrectly, differently or insufficiently in either or both data sources.

Information on individual names is stored differently between the two data sources. In ASHE, the first name, middle name and surname are separate although for first name and middle name, information is only available for the first initial, limiting its use. With a large number of individuals not having a middle name and some have more than one, this variable is not used for matching. In the census, the first name, middle name and surname are stored collectively, so manipulation is needed to separate the variables. For some individuals on the census, there are instances where there are some missing letters on the surname; for example a

person called 'BARNARD' is stored as 'BARNAR'. There are also some instances where, through scanning error, a surname named 'COLES' is stored as 'COLBS'. To overcome these problems, only the first three letters of the surname are used for matching and possible false matches are considered later.

Exact matching is used to link ASHE to the census using the personal characteristics described. The matching procedure uses five iterations such that the first iteration uses the strictest matching criteria, while relaxing the criteria for subsequent iterations. Once matched, an individual is then excluded from subsequent iterations. The matching criteria can be seen in **Table 1**.

The census represents individuals where they live, while in 2001 the ASHE survey only collected information on where they worked (since 2002 information is also available on where they live). While the majority of people live and work in a large region, such as the UK, when disaggregating to smaller regions, it is more common for an individual to work in one area and live in another. **Table 2** shows the number of people who work and live in each of three regions as a percentage of the number of people who work in the region, using information from the 2007 ASHE survey.

The high percentage of people working and living in Cornwall is a consequence of the landscape of the region, with only a few local authorities bordering with another region. At the other extreme, Bexley is part of Greater London, which is in close proximity to a number of other regions and also benefits from good public transport modes allowing for easier commuting to and from work.

Outcome of linking

Before undertaking the record linkage, duplicates in the ASHE sample file (where an individual has more than one job) were removed and each of the three regions were extracted using postcode information provided by the employer on the survey. This resulted in 950 records in Cornwall, 442 records in Bexley and 1,008 records in Bedfordshire. The three regions Cornwall, Bexley and Bedfordshire are matched in turn and the results are now discussed.

Cornwall

There are around 490,000 census records to link to the 950 ASHE records and the linkage results for each of the iterations are shown in **Table 3**. It should be remembered

Table 1

Matching criteria for each iteration

	Combination of variables
Iteration 1	Date of birth (DMY) + gender + first name + surname
Iteration 2	Date of birth (MY) + gender + first name + surname
Iteration 3	Date of birth (DY) + gender + first name + surname
Iteration 4	Date of birth (DM) + gender + first name + surname
Iteration 5	Date of birth (DMY) + gender + surname

Table 2

Percentage of individuals who work in a region and live in the same region

Region	% who work and live in same region
Bexley	48.9
Cornwall	95.0
Bedfordshire	69.3

Source: ASHE 2007

Table 3
Results of matching exercise for Cornwall

	Number to match	Matched
Iteration 1	950	722
Iteration 2	228	25
Iteration 3	203	12
Iteration 4	191	25
Iteration 5	166	23
Unmatched	143	

Table 4
Results of matching exercise for Bexley

	Number to match	Matched
Iteration 1	442	165
Iteration 2	277	8
Iteration 3	269	5
Iteration 4	264	13
Iteration 5	251	5
Unmatched	246	

Table 5
Results of matching exercise for Bedfordshire

	Number to match	Matched
Iteration 1	1,008	501
Iteration 2	507	28
Iteration 3	479	7
Iteration 4	472	51
Iteration 5	421	7
Unmatched	414	

Table 6
Number of potential false matches for the three regions combined

	Number of matches	Possible false matches	% of false matches
Iteration 1	1,388	4	0.3
Iteration 2	61	28	45.9
Iteration 3	24	12	50.0
Iteration 4	89	57	64.0
Iteration 5	35	5	14.3

that a linkage will not be made if an individual works in Cornwall but lives outside the region. There may also be 'false' matches where individuals are matched when they are different people. As the matching criteria are relaxed, this is more likely, and false matches are considered later. For the first iteration, where the matching criteria are most strict, 722 of the 950 records were matched, representing a linkage rate of 76 per cent. When relaxing the matching criteria and running through all the iterations, 807 (85 per cent) individuals are matched.

Bexley

There are around 206,000 census records to link to the 442 ASHE records and the linkage results for each of the iterations are shown in **Table 4**. For the strictest criteria, 165 of the 442 records (37 per cent) are matched, increasing to 44 per cent when combining all matches for each of the five iterations.

Bedfordshire

There are around 360,000 census records to link to the 1,008 ASHE records, with the results of the linking shown in **Table 5**. For the strictest criteria, 501 of the 1,008 records (50 per cent) are matched, increasing to 59 per cent when combining all matches for each of the five iterations.

False matches

As there are not a large number of records to consider in the matching process, it is possible to look at the iterations in turn to identify those that could be false. For the first iteration, there will only be a false match if an individual has the first three letters of the surname the same but the remaining letters are different.

Combining the three regions and looking at the first iteration, this appears to have happened in only four of the 1,388 matches (see **Table 6**). There are a small number of further matches where the surnames are not exactly the same but these could be

explained by scanning errors. For iteration 2, the day is removed from the date of birth for matching. Of the extra 61 matches, closer inspection suggests that 28 (46 per cent) of these are false matches, with the remaining 33 looking plausible with the day being slightly different, possibly through scanning error. The false matches are identified as both digits in the day of birth are different and the full surname is not similar.

For iterations 3 and 4, around 50 per cent and 65 per cent, respectively, of matches appear to be false. For iteration 5, where the first name is not used in the matching, in some of the extra matches, the letter of the first name on the census corresponds to the letter of the individual's middle name in ASHE. This suggests that names may be transposed in ASHE for some individuals. Some of the other matches can be explained where there is a difference in the first name letter through scanning error, with the remaining 14 per cent of the matches looking false. The results show that when not all of the date of birth is used, the number of false matches increases, suggesting this is an important variable and all of its detail should be used.

Characteristics of those matched versus unmatched

Where individuals are not matched, this may introduce bias and so analysis using the matched data must be treated with caution if those respondents who are matched differ in characteristics from those who are unmatched.

The following section will look at the characteristics of those matched and unmatched for gender, occupation group, employment status (full- or part-time), age group and gross weekly earnings.

Gender

For each of the three regions combined, there were 1,388 of the 2,400 records matched and 1,012 unmatched. Of the total, 49.6 per cent are male and 50.4 per cent are female. However, of those who were matched, 45.5 per cent are male with the remaining 54.5 per cent female. This compares with 55.2 per cent male and 44.8 per cent female for those unmatched. This shows that males are more likely to be part of the unmatched group than their female counterparts. When considering the three regions separately, the pattern is the same but with Bedfordshire having the largest difference and Cornwall the least.

Occupation

Figure 1 shows the percentage in each occupational major group for the unmatched and matched groups. There are some differences between the major groups. There is a significantly higher percentage of records within the managers and administrators, professional occupations and clerical and secretarial occupations for the unmatched group when compared with those that are matched. For personal and protective service occupations and other occupations, there is a significantly higher percentage of records in the matched group than in the unmatched group. Other occupations include occupations such as postal workers and cleaners.

Combining the three regions does hide variations between them. For example, in Bexley there is a significantly higher percentage within the first four occupational major groups for the

unmatched group than in the matched group. These occupations would be better paid and people may be more willing to travel further, consistent with them working and living in different regions and being unmatched. It should be noted that the census relies on descriptions about the occupation from each individual; in ASHE, the descriptions come from the employer and so these can vary.

Employment status

When considering employment status, for the matched group, 66.5 per cent are full-time with the remaining 33.5 per cent part-time. This compares with 74.5 per cent full-time and 25.5 per cent part-time for the unmatched group.

Age group

For the broad age groups there is a significantly higher percentage of those in

the 25 to 39 age group (see **Figure 2**) for those unmatched when compared with the matched group. The reciprocal occurs for the 40 to 54 age group, where there is a higher percentage in the matched group than in the unmatched group. This may be reflected in the fact that the 25 to 39 age group is more mobile and likely to work and live in a different region from those in the 40 to 54 age group.

Earnings

Earnings vary across the country and London consistently tops the regional list in terms of gross weekly earnings. This is reflected in the fact that high proportions of the labour force are employed in higher-paying industries and also receive allowances for working in the capital. In ASHE, the median is the most common measure used to display average earnings. This is the middle point of the population, with exactly the same number of people earning below this amount as above it. In some instances it can be more suitable to present the median rather than the mean, as the latter can be influenced by the relatively few extreme values in a pay distribution. Looking at the three regions in turn, in Cornwall, the median gross weekly pay of those that were matched was £296.93 (see **Table 7**), close to the £294.38 for those unmatched. For Bexley, the median gross weekly pay for those matched was £389.52 and for those unmatched £374.28. Looking at the mean gross weekly pay for those matched, it was £432.82, lower than the £468.57 for those unmatched. This suggests that there are higher earners in the unmatched group, consistent with the findings that the unmatched group is concentrated in more higher-paid occupations. In Bedfordshire, the median gross weekly pay for the matched group is £344.84, with the matched group being £391.56.

Conclusion

Overall, this article shows that there is the potential to link ASHE to the census. A number of iterations to link the two data sources have been used but, when relaxing the information on the date of birth, the chances of a false match increase significantly. Lacking information on where a person lives in the 2001 ASHE brings limitations to the matching process. The linkage rate was highest for Cornwall and lowest for Bexley, which is expected using the percentages of people who work and live in the same region in the most recent (2007) ASHE survey.

Figure 1
Percentage of cases matched and unmatched for Cornwall:
by occupational major group



Figure 2
Percentage of cases matched and unmatched: by age group

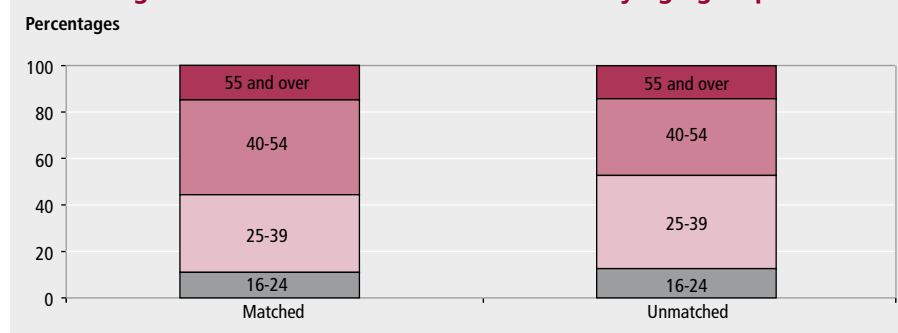


Table 7
Median and mean gross weekly earnings for each region for those records matched and unmatched

	Matched		Unmatched	
	Median	Mean	Median	Mean
Cornwall	296.93	333.23	294.38	348.42
Bexley	389.52	432.83	374.28	468.57
Bedfordshire	344.84	386.09	391.56	449.70

In order to increase the linkage rates, information from the whole census would be needed and larger regions, similar to travel to work areas, used to ensure someone who works in one area will be picked up living in another. A further option is to link to a census extract that classifies individuals to the region they work. However, this was not available for this exercise.

There are some significant differences between the unmatched and matched groups, in particular for Bexley and Bedfordshire, and so there is potential bias in using the current matched data. The resource to carry out full matching using

all the census data will have to be assessed against the benefits for improved research by adding variables on individuals' personal characteristics to ASHE. As ASHE contains the same cohort of individuals each year, if they are in employment, it is possible to carry forward their personal characteristics; but how relevant these are in future years, for example highest educational attainment, needs to be determined.

CONTACT

 elmr@ons.gsi.gov.uk