

2011 Census – 2011 ASHE Linkage Report

PART A : PROJECT DETAILS							
DATE OF TTP MATCHING REPORT:	18/05/2021						
REPORT CREATED BY:	NISRA's Census Office Third Party Team						
PROJECT REFERENCE NUMBER:	TD_001						
TITLE OF RESEARCH PROPOSAL:	ASHE Census Linkage						
PART B : EXECUTIVE SUMMARY							
<p>METHODOLOGY-</p> <p>This project involves 1 matching process</p> <ul style="list-style-type: none"> Match (i): Census 2011 data to Annual Survey of Hours and Earnings (ASHE) 2011 data <p>SUMMARY OF RESULTS-</p> <p>The resulting matching has high coverage and accuracy:</p> <table border="1"> <thead> <tr> <th></th> <th>Unadjusted Match Rate</th> <th>Accuracy</th> </tr> </thead> <tbody> <tr> <td>ASHE 2011 to Census 2011</td> <td>88.7%</td> <td>100%</td> </tr> </tbody> </table> <p>RESULTING DATA:</p> <p>The research dataset contains a fully linked data set of 5,528 ASHE individuals or 5,770 ASHE records as a small number of individuals have more than one job.</p>			Unadjusted Match Rate	Accuracy	ASHE 2011 to Census 2011	88.7%	100%
	Unadjusted Match Rate	Accuracy					
ASHE 2011 to Census 2011	88.7%	100%					
PART C : BACKGROUND							
<p>Role of the Trusted Third Party (TTP)</p> <p>Statisticians in NISRA's Census Office act as a Trusted Third Party (TTP) for the ADR NI projects. This role involves linking together datasets using limited personal information (e.g. name, date of birth and postcode) provided by data suppliers for the purpose of a specific ADR NI project. TTP use these links to provide the Research Support Unit (RSU) team in NISRA with an anonymous index of IDs which link between the datasets involved. The role of RSU is to bring together the attribute data using this index into a research dataset. TTP never have access to the attribute data. Detailed in this paper is the work TTP have carried out to link 2011 Census data to 2011 ASHE linked and produce the linkage keys for the RSU.</p>							

PART D : DATA SOURCES	
DATASET 1	
DATA SUPPLIER:	Economic and Labour Market Statistics Branch
DESCRIPTION:	Annual Survey of Hours and Earnings (ASHE) - a UK wide business survey that provides estimates on hourly, weekly and annual earnings for employee jobs. The ASHE is a 1% sample taken from the HM Revenue and Customs Pay as You Earn System.
NUMBER OF RECORDS:	5,770 records for 2011 5,528 individuals for 2011
PRE-MATCHING CHECKS	
<ol style="list-style-type: none"> 1. The ASHE has a very high response rate and there are no imputed records in the data. 2. Individuals selected for ASHE 2011 were eligible for matching. This data was supplemented with 2012-2018 ASHE data for those individuals selected for 2011 only, in order to maximise the response rate. 3. Information available to the TTP for linkage purposes: <ul style="list-style-type: none"> • Name information which includes surname, forename initial and mid-name initial, • Home and work postcode, • Sex and • Date of birth. 	
DATASET 2	
DATA SUPPLIER:	Census Office
DESCRIPTION:	Extract from the 2011 Census (entire enumerated population aged 16 and over including those both economically active and economically inactive)
NUMBER OF RECORDS:	1,365,124
PRE-MATCHING CHECKS	
<ol style="list-style-type: none"> 1. The data did not include those individuals who were imputed due to no response which accounted for 5% of the total population. <i>Match rates have not been adjusted for imputation as the data only included enumerated individuals.</i> 2. All 1,365,124 records contained enough data to enable them to be available for linkage. 3. Information available to the TTP for linkage purposes : <ul style="list-style-type: none"> • Name information which includes forename, mid-name and surname, • Address information at the time of the Census, • Sex and • Date of birth. 	
PART E : MATCHING METHODOLOGY	

TTP use a 'rule based' or deterministic linkage approach using Match-Keys. Match-keys are created by putting together pieces of information to create unique keys that can be used for automated linking e.g.

Dataset1

ID	Forename	Surname	Sex	DOB	Postcode	Match-key 1: Forename, Surname, Sex, DOB, Postcode
1	JOHN	SMITH	M	01/01/1990	BT6XXX	JOHNSMITHM01011990BT6XXX

Dataset 2

ID	Forename	Surname	Sex	DOB	Postcode	Match-key 1: Forename, Surname, Sex, DOB, Postcode
2	JOHN	SMITH	M	01/01/1990	BT6XXX	JOHNSMITHM01011990BT6XXX

We would choose to accept a link between record 1 in dataset 1 and record 2 in dataset 2 because match-key 1 is identical for both records.

The variation in recording demographic information across datasets can occur in a number of different forms. A single match-key alone cannot resolve all of the differences that occur between data sources, hence the need for multiple Match-Keys. A series of Match-Keys have been developed, each of which is designed to resolve particular inconsistencies between match pairs.

An example of a non-exact match-key is one constructed from the first two characters of an individual's forename and surname (Bi-grams), combined with their date of birth and postcode district. The strongest level of linking is exact linking, which links pairs of records that are identical on all linking fields.

Typically, the Match-Keys are processed in a stepwise manner starting with the most exact match-key and working down to the last match-key. More information can be found at:

[DATA MATCHING USING NORTHERN IRELAND ADMINISTRATIVE DATA: A WORKED EXAMPLE](#)

PART F : DONOR IMPUTATION

In order to produce a fully linked dataset, donor imputation work was carried out to link the remaining unmatched ASHE individuals to Census individuals with similar characteristics. This work followed the same deterministic linkage approach using Match-Keys as was used for the initial linkage. Match-Keys were created by putting together pieces of information to create unique keys that can be used for automated linking. Additional variables were provided for this work. The variables used to create the Match-Keys were:

- Occupation Code (coded using the 2010 Standard Occupational Classification System),
- Business Activity Code (coded using the UK Standard Industrial Classification of Economic Activities 2007 (SIC 2007)),
- Gender,
- Age,
- Home Postcode and
- Work Postcode.

An iterative approach was used beginning with matching on what were deemed to be tighter Match-Keys and loosening the Match-Keys each time only accepting matches where there hasn't already been a match through the tighter Match-Keys. Some example match-keys used for this process can be seen below:



Match-Key	Variables
1	3 digit SIC07, 4 digit SOC10, Gender, Age, Postcode Sector
2	2 digit SIC07, 4 digit SOC10, Gender, Age, Postcode Sector
3	3 digit SIC07, 4 digit SOC10, Gender, Age, Postcode District
4	2 digit SIC07, 4 digit SOC10, Gender, Age, Postcode District

This process resulted in individuals in the ASHE dataset being linked to a pool of potential donor records in the 2011 Census dataset. This pool of individuals had similar characteristics to those in the ASHE dataset.

Once the unmatched ASHE records had been linked to a pool of potential donors within the Census data a random record was selected to be the match for that ASHE individual.