First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.2991

Big Data Matching Using the Identity Correlation Approach

McCormack, Kevin and Smyth, Mary
Methodology Division, Central Statistics Office, Cork, Ireland

Abstract

The Identity Correlation Approach (ICA) is a statistical technique developed for matching big data where a unique identifier does not exist. This technique was developed to match the Irish Census 2011 dataset to Central Government Administrative Datasets in order to attach a unique identifier to each individual person in the Census dataset (McCormack & Smyth, 2015[1]). The unique identifier attached is the PPS No. (Personal Public Service No.[2]). By attaching the PPS No. to the Census dataset, each individual can be linked to datasets held centrally by Public Sector Organisations. This expands the range of variables for statistical analysis at individual level. Statistical techniques developed here were undertaken for a major European Structure of Earnings Survey (SES) compiled by the CSO using administrative data only, and thus eliminating the need for an expensive business survey to be conducted (NES, 2007[3,4,5]). A description of how the Identity Correlation Approach was developed is given in this paper. Data matching results and conclusions are presented here in relation to the Structure of Earnings Survey (SES)[6] results for 2011.

Keywords: *Identity Correlation Approach, Big Data matching, Unique identifier*

1. Identity Correlation Approach

The Identity Correlation Approach (ICA) was developed as part of a big data matching Project known as the SESADP (Structure of Earnings Survey Administrative Data Project) carried out by the Central Statistics office, Ireland (McCormack & Smyth, 2015[1]). The aim of the SESADP was to produce data to meet the EU SES 2014 Regulation from administrative data sources, from 2011 on an annual basis going forward. This eliminated the need for an expensive business survey to be conducted each year (NES, 2007[3,4,5]).

The ICA involves combining a number of individual variables for each person until a unique identifier is arrived at (McCormack,K 2015[7]). An example of this is combining the individual characteristics of each person in the Irish Census Dataset. Beginning with the variable for *date of birth*, then combine it with the variable *gender*, then adding variable for *county,& marital status, etc.* until a unique identifier is arrived at for each person. This is illustrated in Figure 1 below.

2.1 Theoretical Application

There was 1.6m employees in Ireland in 2011. Of these, an average of 65,000 persons were born in the same year (years 1946 to 1995), as illustrated in Figure 1.

Figure 1: Identity Correlation Approach: Simple Model - Combining Variables

Operation	Variable	No. of Records
	Approx. No. of births each year	65,000
Divide by:	No. days in the year	365
	No. Persons with same DoB	178
Divide by:	Gender	2
	No. Persons with same DoB and gender	89
Divide by:	No. Counties	26
	No. Persons with same DoB, Gender, County	3
Divide by:	Marital Status (married & other)	2
	No. Persons with same DoB, Gender, County, marital status	1

The 65,000 persons born in the same year are divided by 365 days in the year to give approximately 178 persons with the same date of birth. The 178 persons with the same date of birth (DoB) can be divided by 2 for gender, to give 89 persons with the same DoB and gender. Dividing by the no. of counties a person lives in (89 divided by 26) results in 3 persons with the same DoB, gender & county. The 3 persons can be further subdivided by marital status resulting in 1 person with the same DoB, gender, county and marital status. Other variables used to further breakdown the data are industrial sector (NACE[5] code), no. of dependent kids, etc. A unique combination of variables for each person allows a person to be uniquely identified. This method is termed the Identity Correlation Approach (ICA).

2.2 Complexity and Variables Added

Complexity added to the simple model for the Identity Correlation Approach is shown in Figure 2. Populations are not evenly distributed, thus we allow for the fact that up to a third of the working population may be based in Co. Dublin. As a result, duplicates increase 10 fold for the No. of persons with the same DoB, gender & county. Similarly, the employee population is not evenly distributed in the various NACE sectors (industrial sector). Other variables added include no. of dependent children, which allows further breakdowns. In this way the Identity Correlation Approach arrives at a unique identity for each individual by combining a number of personal characteristics for the person.

Figure 2: Identity Correlation Approach: Complexity & Variables Added

Operation	**Variable**	**No. of Records**
	Approx. No. of births each year	65,000
Divide by:	No. days in the year	365
	No. Persons with same DoB	178
Divide by:	Gender	2
	No. Persons with same DoB and gender	89
Divide by:	No.Counties (allowing for approx. one third employees living in Dublin)	3
	No. Persons with same DoB, Gender, County	30
Divide by:	NACE industrial code (15) - allow for one fifth employees in same NACE Sector	5
	No. Persons with same DoB, Gender, County, NACE	6
Divide by:	Marital Status (married & other)	2
	No. Persons with same DoB, Gender, County, NACE, marital status	3
Divide by:	No. of dependent kids (3 groups)	3
	No. Persons with same DoB, Gender, County, NACE, marital status, no. dependent kids	1

2. Practical Application

2.1 Census 2011 data

The identity Correlation approach was applied to the Irish Census Data 2011 as described above. This allowed for a Unique Identifier (UI) to be applied to each individual by combining their personal characteristics (i.e. DoB, gender, county residence, etc.). The unique identifier is called the matching variable (matchvar) which is used to link each person's record to other datasets (see Fig. 3).

Figure 3: Applying Identity Correlation Approach to Create Unique Identifier (Matchvar)

Date of				Marital		Matchvar
Birth	Gender	County	NACE	Status	No.Kids	(all variables)
15031949	M	CORK	42	M	0	15031949MCORK42M 0
11021945	F	LIMERICK	31	S	1	11021945FLIMERICK31S1
21111954	M	DUBLIN	25	D	2	21111954MDUBLIN25D2
19051964	M	CARLOW	55	O	2	19051964MCARLOW55O2
22091966	M	GALWAY	82	M	3	22091966MGALWAY82M3
24031971	F	CAVAN	84	M	0	24031971FCAVAN84M0

2.2 Public Sector Administrative Datasets (ADS)

A single master Administrative Dataset (ADS) was created by linking a number of Public Sector Administrative Datasets (Revenue Commissioners Tax data, Social Security Administrative Datasets and the CSO's Administrative Datasets (e.g. Central Business Register (CBR), Earnings and Labour Force Survey)). These datasets were combined using the PPS No. for each individual and the CBR Enterprise No. for Establishment Surveys. The identity Correlation approach was applied to the master Administrative Dataset (ADS) also, allowing for a Unique Identifier (UI) to be applied to each individual by combining their personal characteristics (i.e. DoB, gender, county residence, etc.). This Unique Identifier known as the match variable (matchvar) was then used to link to the UI (matchvar) in Census.

2.3 Linking Census to ADS

Variables common to both the Census dataset and the master Administrative Data Source (ADS) were identified (e.g. DoB, gender, etc.). These common variables were joined to each other to create a Unique Identifier on each dataset using the Identity Correlation Approach. By linking the two datasets using the Unique Identifier, a PPS No. could be applied to each individual person in the Census. This is shown in Figure 3. Once the PPS No. was assigned to the Census dataset, it enabled Census data to be linked to any Public Sector Administrative Dataset.

3. Dataset Matching

3.1 Census dataset

A total of 1.6 million employee records were extracted from the 4.6 million Census Records. Approximately 200,000 records had a unique Business No. identifier attached (CBR No.). Another 500,000 records had a CBR No. attached using the Employer's Business name on the Census. The first matching variable (Matchvar1) created for Census used the following variables combined: CBR No., Dob, gender, county, NACE 2, marital status, No. of kids. A second matching variable was created (Matchvar2) excluding NACE2 (see Figure 4). Up to ten matching variables (Matchvar1 – Matchvar10) were created. Each matching variable is similar to the previous one with one change to the composition variables for each subsequent matching variable created. Figure 4 illustrates the construction of each subsequent matching variable.

Figure 4: Matching Variables

Date_of Birth	Gender	County	NACE	Ent No.	Marital Status	No.Kids	Match Var 1	Match Var 2	Match Var 3
15031949	M	CORK	42	EN12345678	M	0	15031949MCORK42EN12345678M0	15031949MCORK42EN12345678M	15031949MCORK42EN12345678
11021945	F	LIMERICK	31	EN52345679	S	1	11021945FLIMERICK31EN523456791	11021945FLIMERICK31EN52345679S	11021945FLIMERICK31EN52345679
21111954	M	DUBLIN	25	EN52795680	O	2	21111954MDUBLIN25EN527956802	21111954MDUBLIN25EN52795680O	21111954MDUBLIN25EN52795680
19051964	M	CARLOW	55	EN32795681	D	2	19051964MCARLOW55EN327956812	19051964MCARLOW55EN32795681D	19051964MCARLOW55EN32795681
22091966	M	GALWAY	82	EN22795682	M	3	22091966MGALWAY82EN227956823	22091966MGALWAY82EN22795682M	22091966MGALWAY82EN22795682
24031971	F	CAVAN	84	EN52795683	M	0	24031971FCAVAN84EN527956830	24031971FCAVAN84EN52795683M	24031971FCAVAN84EN52795683
28021977	F	DUBLIN	71	EN84355684	S	1	28021977FDUBLIN71EN843556841	28021977FDUBLIN71EN84355684S	28021977FDUBLIN71EN84355684
30061990	F	KERRY	35	EN73795687	M	1	30061990FKERRY35EN737956871	30061990FKERRY35EN73795687M	30061990FKERRY35EN73795687

3.2 ADS (Public Sector Administrative Datasets)

The records in the master Public Sector Administrative Dataset (ADS) also contained all the above variables used in Census to create the matching variables (Matchvar1 – Matchvar10). Therefore the matching variables created in the ADS were used to match to the same variable in the Census.

3.3 Incremental Matching Process

To match both datasets (Census to ADS) the matching variables (Matchvar1 – Matchvar10) were used.

First Matching (using *Matchvar1*)

In practice, duplicates will occur when the Unique Identifier (*Matchvar1*) is created. This can be due to errors in coding. Therefore in the dataset linking process, each dataset (Census and the ADS) is edited to extract records where the Matchvar1 is unique. If there are duplicate occurrences of *Matchvar1*, then these records are deleted from the matching process. The two datasets (Census and the ADS) were matched using *Matchvar1*. This yielded a match of 307,300 Census records to the ADS dataset.

Second Matching (using *Matchvar2*)

In the second stage of matching, the remaining unmatched records in each dataset (Census and the ADS) are edited to extract records where the variable *Matchvar2* is unique. Only unique occurrences of Matchvar2 are used to match both datasets. This yielded a match of 75,400 Census records to the ADS dataset.

Third and consecutive matches (Matchvar3 to Matchvar11)

The matching process continued in this incremental process using Matchvar3 up to Matchvar10. This yielded a total of 797,100 Census records to the ADS dataset.

Figure 5: Incremental Matching Process

<table>
<tr><th>Unique Identifier (UI)</th><th>Variables constituting UI</th><th>No. of Records Linked</th></tr>
<tr><td>Matchvar1</td><td>(CBRno|DoB|sex|NACE2|PP|County|MS)</td><td>307,393</td></tr>
<tr><td>Matchvar2</td><td>(CBRno|DoB|sex|NACE2|PP|County)</td><td>75,410</td></tr>
<tr><td>Matchvar3</td><td>(CBRno|DoB|sex|NACE2|PP|MS)</td><td>29,854</td></tr>
<tr><td>Matchvar4</td><td>(CBRno|DoB|sex|NACE2|PP)</td><td>14,885</td></tr>
<tr><td>Matchvar5</td><td>(DoB|sex|NACE2|PP|County|MS)</td><td>47,963</td></tr>
<tr><td>Matchvar6</td><td>(DoB|sex|NACE2|PP|County)</td><td>16,898</td></tr>
<tr><td>Matchvar7</td><td>(DoB|sex|NACE2|PP|MS)</td><td>25,996</td></tr>
<tr><td>Matchvar8</td><td>(DoB|sex|NACE2|PP)</td><td>15,170</td></tr>
<tr><td>Matchvar9</td><td>(DoB|sex|NACE2|County|MS)</td><td>204,099</td></tr>
<tr><td>Matchvar10</td><td>(DoB|sex|NACE2|County)</td><td>59,102</td></tr>
<tr><td></td><td><b>Total no. of records linked</b></td><td><b>796,770</b></td></tr>
</table>

3.4 False Positives

False positives can occur in the matching process if a variable is incorrect on one of the datasets. For example is the county variable has not been updated on the Social Welfare dataset then the county will be different on the person’s record on Census. Similarly, if the NACE code is incorrect on either dataset, then it will not match a person to their correct record.

4. Mathematical Representation of Identity Correlation Approach (ICA)

Creating a Unique Identifier (UI) for each record using the Identity Correlation Approach (ICA) will result in a perfect match of records across two datasets, if there is a sufficient overlap of variables on both datasets. The probability of matching records across two datasets can be calculated by a formula known as the Matching Rate of Unique Identifier (MRUI). This is illustrated in Table 1.

Table 1: Matching Rate for Unique Identifier

$$\mathbf{N} * \frac{\mathbf{1}}{\mathbf{V1_{ui}}} * \frac{\mathbf{1}}{\mathbf{V2_{ui}}} * \frac{\mathbf{1}}{\mathbf{V3_{ui}}} * \frac{\mathbf{1}}{\mathbf{V4_{ui}}} * \ldots\ldots \frac{\mathbf{1}}{\mathbf{Vx_{ui}}} = \mathbf{MRUI}$$

Where:

N = Population , V = Variable, x = no. of variables

ui = Uniqueness Factor. ui = no. of classes where variable is distributed evenly across all classes

MRUI The Matching Rate for Unique Identifier (MRUI) is the ability to identify a unique record in a dataset, given the combination of variables used to deduce the record. Mathematically it is assumed that variables are discreet (non-dependent).

5. Results & Conclusion

Results for matching Census records to Administrative Datasets are given in Figure 6, classified by NACE industrial sector. The lower rate of matching in some sectors (e.g. Construction) can be attributed to records not being updated for certain variables. If the theoretical MRUI value indicates a perfect match, but this is not reflected in practice, then there are issues with coding or with records not being updated.

Figure 6: Employee Population Coverage 2011 - Census & Administrative Datasets Matched

Nace Economic Sector	**No. Employees Total**
B-E Industry	55
F Construction	26
G Wholesale and retail trade	44
H Transporation and Storage	51
I Accommodation and Food Services	31
J Information and communication	52
K-L Financial, insurance, etc.	61
M Professional, scientific & technical	39
N Administrative and support services	26
O Public administration & defence	69
P Education	63
Q Health & social work	46
R-S Arts, entertainment, other services	41
Total	47

5.1 Impact of ICA Data Matching

The ICA (Identity Correlation Approach) enabled 50% of the employee population in Ireland (750,000 of 1.5 million employees) to be matched to the Census 2011 dataset, as part of the SESADP. This enabled the Census dataset to provide variables for the SES (e.g. education level and occupation). Outputs from the SESADP produced the SES data for 2011 to 2014 and avoided having to do an expensive business Survey. IT and statistical infrastructure are now in place to produce the SES on an annual basis going forward, reducing costs from €1.6million annually to €0.1million. A Cost/Benefit Analysis of the SESADP is given in Figure 8.

5.2 Data Quality

An analysis of the data was undertaken to determine if the ICA matched the Census correctly to the other administrative data sources (ADS). There was a 90% correlation with the individual's *employer name* on Census with the *employer name* on the Business Register. The 10% with a different business name were eliminated as false positives. In Figure 9 a distribution of employees by age group and NACE in the Census dataset shows a very good comparison with the SESADP (similar results were obtained for education, occupation & gender comparisons).

Figure 8: Cost/Benefit Analysis of the SESADP

	Business Survey former NES	SESADP Project	SESADP (Annual)
Survey Type	Annual Survey	Data for 4 years	Annual basis - going forward
Reference period	Years 2002 to 2009	2011 to 2014	2015 onwards
Cost	€ 1.6 million p.a.	€0.4m	€0.1m p.a.
Timeliness	T+ 18	2 years to develop	T+ 10
Data edits	Data Edits	No edits - Revenue data	No edits - Revenue data
Sample size	65k	800k	800k+
Coverage of employee population	4%	50%	50%+
Burden	70,000 employees	None	None
Burden	5,000 enterprises	None	None
Staff Nos.	15 FTEs	4	2
Savings	-	€6.0m	€1.5m p.a.

Figure 9: Employees Nos. in SESADP compared to Census dataset by Nace Sector and Age Group 2011

NACE Economic Sector	% difference in employee nos.					
	Age Group in years					
	15-24	25-29	30-39	40-49	50-59	60 and over
	%	%	%	%	%	%
B-E Industry	-1	-1	2	1	0	-1
F Construction	-1	-1	3	0	-1	-1
G Wholesale & Retail Trade; Repair of Motor Vehicles and Motorcycles	0	-1	1	1	0	-1
H Transportation and Storage	-1	-1	0	2	1	-1
I Accomodation and Food Service Activities	-1	-1	2	1	0	-1
J Information and Communication	-1	-2	3	0	1	-1
K-L Financial, Insurance and Real Estate	-1	0	3	0	0	-1
M Professional, Scientific and Technical Activities	0	1	3	-1	-2	-1
N Administrative and Support Service Activities	0	1	4	0	-2	-2
O Public Administration and Defence; Compulsory Social Security	-1	-1	2	1	-1	-1
P Education	-2	-2	2	2	0	0
Q Human Health and Social Activities	0	-1	1	1	-1	-1
R-S Arts, entertainment, recreation and other service activities	0	1	1	1	-1	-2

References

McCormack,K (2015). Constructing structural earnings statistics from administrative datasets: Structure of earnings survey – Administrative data project. *The Statistics Newsletter – OECD***,** 62, 3-5[8].

McCormack,K. & Smyth,M. (2015). Constructing structural earnings statistics from administrative datasets. *New Techniques and Technologies for Statistics (NTTS) 2015. Collaboration in Research and Methodology for Official Statistics*[1].

McCormack,K. & Smyth,M. (2015). Specific Analysis of the Public/Private Sector Pay Differential for National Employment Survey 2009 & 2010 Data. *Research Paper. Central Statistics Office, Ireland*[3]

NACE is the "statistical classification of economic activities in the European Community" [9]

National Employment Survey 2007 (2009), *Central Statistics Office, Ireland*[6].

National Employment Survey 2008 and 2009 (2011), *Central Statistics Office, Ireland*[5].

National Employment Survey 2009 and 2010 Supplementary Analysis (2012), *Central Statistics Office, Ireland*[4].

The European Union Structure of Earnings Survey (SES) in accordance with Council Regulation n° 530/1999 is conducted in the 28 Member States of the European Union as well as candidate countries and countries of the European Free Trade Association (EFTA) [7].

The Personal Public Service Number (PPSN) is a unique reference number that allows individuals access social welfare benefits, public services and information in Ireland. State agencies that use PPSNs to identify individuals include the Department of Social Protection, the Revenue Commissioners and the Health Service Executive (HSE). [2]