National Records of Scotland

Beyond 2011

Matching the Census 2011 to the NHS Central Register using the Ord Wood method

Published on 9 January 2014

Contents

List of Tables

1. Background

This paper reports a project run in 2012/2013 to match the results of the 2011 Census in Scotland to the NHS Central Register (NHSCR) extract as it stood at the end of May 2011. The method used to undertake this matching involved extending the usual Fellegi-Sunter method of record linkage. This paper reports the stages of the matching procedure and the results obtained. The technical details of this method are presented in the paper 'The Ord Wood Project: A method of calculating match probabilities from record linkage output'.

2. The linkage of the Census to the NHSCR extract

The aim of this work was to undertake a match between the NHSCR extract and the 2011 Census. This project was to help better understand the over- and under-coverage present in both data sources as well as further develop the matching techniques used within National Records of Scotland (NRS). There were 5,000,834 records in the Census file representing completed Census returns. The NHSCR extract is a subset of the NHSCR database which is used for statistics and research within NRS. The NHSCR extract used contained 5,672,245 records of people alive and registered with a Scottish Health Board on Census Day.

2.1 The deterministic stage

The first stage of the matching was deterministic. In step 1a, a match was defined as complete agreement on first name, last name, date of birth, gender and post code. A total of 2,922,199 records pairs showing this level of agreement were identified. The probability of each of these being a true match was calculated. The probabilities were so close to one that, when subtracted from one and multiplied by 2.9 million to give an estimate of the number of false positives accepted, the result, to the nearest integer, was zero. All these record pairs were accepted as being matches and were removed from their respective files.

In steps 1b and 1c, a match was defined as complete agreement on first name, last name, date of birth and gender; and that this combination of first name, last name, date of birth and gender should be unique in that it did not exist elsewhere in either file. In step 1b, at least one of the postcodes was missing so that no comparison could be made on this field, while in step 1c the postcodes were allowed to be different[1]. A total of 777,204 record pairs were identified in step 1b and a further 357,453 in step 1c. Of these, it was estimated that 29 out of the 777,204 and 208 out of the 357,453 were false matches. Again these records were removed from their respective files.

2.2 The probabilistic stage

In step 2, the remaining records (944,732 Census and 1,633,735 NHSCR) were then probabilistically matched using the Link Plus package. Even after the removal of the "safe" deterministic links, the run took 36 hours.

Footnote

[1] Appendix A to this paper gives an analysis of health board allocations for the record pairs in stage 1b and 1c and of the post code differences for the record pairs in stage 1c.

A total of 944,079 record pairs were output by Link Plus. A cut-off score was set at a sufficiently high level to justify accepting record pairs with higher weights as matches without further confirmation. The number of pairs satisfying this criterion was 604,403. The number of false links was estimated as 108.

2.3 The tailored calculations

In order to identify further probable matches before undertaking a clerical review it is necessary to make use of further information. This involved undertaking calculations tailored to data not used in steps 1 or 2.

The first of these (step 3a) concerned surnames. The surname used in these stages was that listed in the NHSCR extract at the end of May 2011. However the NAMES file which is part of the extract contains the name history, where one exists, of the persons listed in the PERSON file. It was decided to give credit for agreement on surname if (i) the Census and NHSCR surnames were different and (ii) a surname could be found in that person’s name history which agreed exactly with the Census surname. The most likely reason for a change in surname is marriage or divorce. This found a total of 21,587 record pairs where the match probability was now greater than 99.5% and where neither the Census record nor the NHSCR record was currently identified as a match. The estimated number of false links was 24. These links were then added to the bank and removed from the residual files.

The second tailored calculation (step 3b) was that the residual files after stage 2 were run through a second matching software package called Rec Link and the output scrutinised to find any record pairs were (i) the match probability was greater than 99.5% and (ii) neither the Census record nor the NHSCR record was currently identified as a match. A further 6,986 pairs were identified in this way. The estimated number of false matches was 7. These matches were then added to the bank and removed from the residual files.

The third tailored calculation (step 3c) was that the record pairs unbanked after stage 3b were submitted to a recalculation of the log likelihood ratio using further information from names and postcodes. For first and last names, these changes were made in order to make use of patterns of similarity information between record pairs which had been observed manually but which would not be picked up by the standard Fellegi-Sunter method. For postcodes, the change was made to tailor the calculation of the linkage weight to the structure of postcodes, which is not done in the standard text comparison algorithm in Link Plus. This package was developed in the USA and supports zip codes but not post codes. A further 35,772 pairs were identified in this way. The estimated number of false matches was 15. These matches were then added to the bank and removed from the residual files.

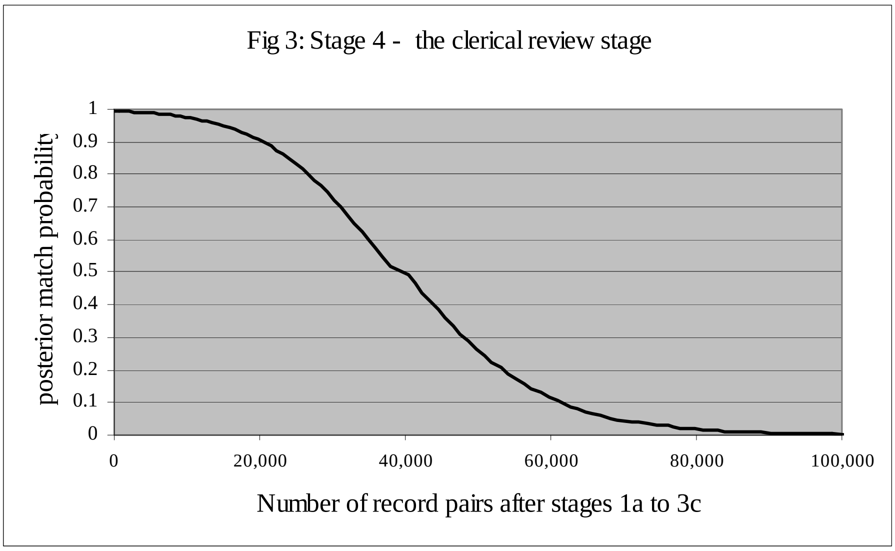
Table 1 below summarises the results of the deterministic and probabilistic stages and the tailored calculations. It can be seen that over four-fifths of the census records have been linked without the use of computer intensive specialist software and over 93% without clerical review. The loosening of the match criterion in steps 1b, 1c and 2 brings an associated cost but at fewer than 400 errors out of almost 4.69 million, it is not excessive

Table 1: Summary of stages 1, 2 and 3

Step	Number of matches accepted	Matches as % of Census file	Cumulative links accepted	Cumulative links as % of Census file	Estimated number of errors	Cumulative estimated errors	Cumulative errors as % of all links
1a	2,922,199	58.43%	2,922,199	58.43%	0	0	0.000%
1b	777,204	15.54%	3,699,403	73.98%	29	29	0.001%
1c	357,453	7.15%	4,056,856	81.12%	208	237	0.006%
2	604,403	12.09%	4,661,259	93.21%	108	345	0.008%
3a	21,587	0.43%	4,682,846	93.94%	24	369	0.008%
3b	6,986	0.14%	4,689,832	93.78%	7	376	0.008%
3c	35,772	0.72%	4,725,604	94.50%	15	391	0.008%

2.4 The clerical review stage

This left 946,293 NHSCR extract records and 275,228 census records available to go to clerical review. Almost all of the census records had been matched to a NHSCR record at stage 2 but with match probabilities varying from 99.5% to effectively zero. These record pairs were sorted by match probability and the first 100,000 of them are plotted in figure 3.



Summing the probabilities we estimate that the expected number of further matches to be found is 40,897[2] (with a standard deviation of 102) if we were to conduct a clerical review to find the remaining matches.

3. Discussion

If the figures given in the previous section are accurate then they should be consistent with a visual inspection of the table of record pairs available for clerical review. As depicted in figure 3, the matched pairs were sorted in descending order of the posterior match probability. Inspection of this sorted table indicated that the 40,897th record pair came in a block of record pairs with a weight of 13.20 and a match probability of 49.2%. Visual examination of this block revealed that a

Footnote

[2] This number is only about 37,000 less than the figure based on the output after stage 2, indicating that the 64,300 pairs found in stage 3 accounted for only 37,000 of the estimated links. This is in turn evidence that the extra evidence used in stage 3 was indeed of value in identifying a few extra high quality links.

common agreement pattern within it was similarity (though not always agreement) on first name and last name; agreement on gender and on two of day, month or year of birth; disagreement on the third date element; and a missing postcode. A clerical review would probably accept many of these pairs where the reviewer could see how the disagreement had occurred but could also see that none of the rules applied by the computer would pick it up. In these cases, the match probability would be an underestimate and hence the actual number of matches in figure 3 is probably greater than 40,897.

A consistency test was undertaken between the method proposed here and a clerical review. It involved two reviewers from Alternative Sources Branch who took part in a trial review exercise without being told in advance what was the aim of the trial. The reviewers looked at the same 240 pairs but worked separately so that they did not know each other's decisions. The decision in each case was to accept or reject the pair as a true match. The 240 record pairs were selected so that for 120 of them, the match probability was between 0.90 and 0.95 while for the other 120 it was between 0.80 and 0.90. The probabilities were approximately evenly distributed over these intervals, so the method predicts that the expected numbers of true matches in the two sets should be 111 and 102 with standard deviations of 2.9 and 3.9 respectively. One reviewer accepted 108 and 100 pairs in the two subsets (3 below and 2 below expectation) while the other accepted 113 and 105 (2 above and 3 above). All four of these are well within error tolerance of two standard deviations, suggesting that the method calculates probabilities which are consistent with the behaviour of real clerical reviewers.

4. Conclusion

The application in section 2 has suggested that, while there is scope for fine-tuning the method for particular applications, it is basically sound and offers the scope for estimating with reasonable accuracy what would be the outcome of a clerical review if one were to be undertaken. In addition to estimating the total number of matches, the method can also be used to estimate how many links would be found in subgroups defined by gender, age, geography or any other variable or combination of variables. If only ‘golden records’ are counted (with the associated undercoverage being accepted), then for example if the method indicates that a given record pair is a match with a probability of 0.68 and has gender ‘female’ on both files, then 0.68 can be added to the subtotal of ‘female’ matches. The same action is taken if the gender is ‘female’ in one file but missing in the other. If the fields disagree however then 0.34 is added to both ‘male’ and ‘female’ match subtotals. Age and geography can be handled in the same way.

Appendix A

Health Boards

For the record pairs in steps 1b and 1c it is possible to assess the extent of geographical agreement by comparing the health board posting from the NHS Central Register (NHSCR) extract with the health board derived from the postcode on the census record. This was done separately for the two steps and the results are given in tables A1 and A2. In each case the NHSCR posting is down the side and the census posting along the top.

For step 1b, it can be seen that the majority of the cases (99.2%) fall on the major diagonal, indicating agreement at least at the geographical area of the health board. Where cases fall outside the major diagonal, the largest cell totals (a further 0.3%) are in the cells for G (Glasgow) with L (Lanarkshire), C (Clyde) and A (Ayrshire and Arran); and S (Lothian) with B (Borders) or V (Forth Valley), all of which are geographically contiguous areas.

For step 1c, the correlation between the two axes is noticeably smaller, reflecting the fact that a degree of difference in geographical affiliation is guaranteed at this step. However it is still the case that the majority of the cases (87.8%) fall on the major diagonal. Again, where cases fall outside the major diagonal, the largest cell totals (a further 6.0%) are in the cells for G (Glasgow) with L (Lanarkshire), C (Clyde), A (Ayrshire and Arran) and V (Forth Valley); and S (Lothian) with F (Fife), B (Borders) or V (Forth Valley), again areas in the same part of Scotland. Taken together, the analysis of health board postings suggests that record pairs in steps 1b and 1c are consistent with the hypothesis that they refer to the same people who move from one part of Scotland to a contiguous part.

Post codes

For the record pairs in step 1c, where both post codes existed but were not identical, the following comparative study was undertaken. The distance between two postcodes was measured using a similarity scale based on the components of which a postcode is composed. If the post code regions were different (e.g. “FK” and “TD” then the similarity was zero. If the post code regions were the same but had different outward parts (e.g. “FK3” and “FK6”) then the similarity was one. If the outward parts were the same but the post code sectors were different (e.g. “FK3_7” and “FK3_9”) then the similarity was two. If the post code sectors were the same but the post codes were different (e.g. “FK3_6LS” and “FK3_6RD” then the similarity was three. The similarity measure could not take its maximum value of four (where the postcodes were identical) since these cases had gone into step 1a.

The distribution of the similarity measure amongst the 357,453 step 1c record pairs which could be compared is given in table A3. Of the 13.2% where the post code region was different, the two regions were geographically contiguous in a large majority of the cases. Overall, the table gives an impression of most moves being within a city or comparable geographical area. If anything, the number of small moves within the same post code sector is larger than would be expected but some of these can be accounted for as coding or transcription errors rather than real moves.

Table A1: Cross tabulation of NHSCR health board posting (vertical axis) with Census health board posting (horizontal axis) for step 1b: values less than six have been deleted

	A	B	C	F	G	H	L	N	R	S	T	V	W	Y	Z
A	62598	6	38	13	97	14	19	25	.	38	10	12	.	10	.
B	8	15818	.	.	15	10	11	13	.	121	9	6	.	.	.
C	89	.	50318	17	267	18	58	45	.	35	8	21	.	6	.
F	6	11	12	79053	29	13	13	27	.	106	117	22	.	.	.
G	123	10	360	80	105297	40	491	49	.	74	35	54	14	38	.
H	23	11	16	16	34	27646	7	96	.	33	15	6	23	.	.
L	50	9	42	22	311	9	83888	7	.	98	10	43	.	11	.
N	23	7	13	28	25	76	13	76196	.	49	57	21	.	8	.
R	.	6	.	.	.	.	.	.	1176	.	.	.	.	.	.
S	45	121	39	111	64	45	94	42	.	93685	52	91	7	12	.
T	15	10	9	82	28	17	22	59	.	47	56234	51	.	8	.
V	12	6	43	29	71	12	38	18	.	90	23	43159	.	12	.
W	.	.	.	.	14	20	.	12	.	.	.	.	3192	.	.
Y	18	.	8	6	32	.	9	.	.	20	9	35	.	16090	.
Z	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1661

Table A2: **Cross tabulation of NHSCR health board posting (vertical axis) with Census health board posting (horizontal axis) for step 1c: values less than six have been deleted**

	A	B	C	F	G	H	L	N	R	S	T	V	W	Y	Z
A	18491	16	380	42	1282	46	200	92	.	172	53	50	.	76	.
B	6	6132	24	33	63	27	46	32	.	606	29	28	.	23	.
C	549	13	25096	50	2885	84	307	131	.	210	60	141	22	29	.
F	33	29	58	18846	220	79	77	177	.	859	643	185	9	21	.
G	977	65	3035	239	57659	334	4786	439	35	935	314	602	93	173	21
H	31	14	68	41	253	14394	34	370	17	206	102	34	49	14	9
L	209	36	258	77	3535	50	31259	110	.	508	88	477	8	37	.
N	63	43	104	149	343	398	93	32847	36	569	431	99	30	29	26
R	.	.	.	.	9	15	.	25	1485	14	.	.	.	.	.
S	169	554	206	971	838	300	502	547	30	55449	423	807	25	132	42
T	73	40	101	653	414	130	115	455	12	599	22450	234	14	32	11
V	54	27	152	203	706	70	275	135	7	632	220	16129	10	39	10
W	.	.	12	.	75	59	11	20	.	21	6	7	2165	.	.
Y	90	15	26	12	158	17	41	34	.	120	20	28	.	7103	.
Z	.	.	.	.	12	8	.	28	.	11	.	.	.	.	1490

Table A3: Distribution of similarity measure for post code pairs in step 1c

Value	Frequency	Percentage	Meaning
0	47,225	13.2%	Different regions (eg EH3 6NA and FK11 2RL)
1	140,367	39.3%	Same region but different outward half (eg EH3 6NA and EH8 4RQ)
2	54,478	15.2%	Same outward half but different sector (eg EH3 6NA and EH3 4LH)
3	115,383	32.3%	Same sector but different postcode (eg EH3 6NA and EH3 6KF)