Dataset Name: Farm Structure Survey
Dataset ID: 2122-A009

Linkage Summary Report
Stage 1: Preprocessing

Number of Input Records:	33,797	
Number of Input Farms (associated with a person):	33,637	
Data completeness (at Farm level)		
valid age band	31,338	93.2%
filled forename	9,696	28.8%
forename initial only	20,685	61.5%
any forename (including initial)	30,381	90.3%
filled surname	29,156	86.7%
valid sex	31,338	93.2%
valid postcode	33,543	99.7%
filled UPRN	23,642	70.3%
valid complete details: age band, forename (initial or full forename), surname, sex and postcode	27,441	81.6%

Further pre-processing:
Soundex codes of NYSIIS (following ISD Scotland algorithm) of Surname added to reformatted file

ADR storage and demographic keys	
Distinct storage keys	33,797
Distinct demographic keys	33,797
Farm ID with one storage and one demographic key	33,477
Farm ID with two storage and two demographic keys	160
Farm ID with any other key combination	0

records with personal information corresponding to single farm were carried forwards for probabilistic linkage to the spine using BigMatch 27,984

Stage 2: BigMatch Linkage against the Indexing Spine

BigMatch is a linkage software program developed and used in-house by the Statistical Research Division, U.S. Bureau of Census. It has been designed to undertake timely matching of very large files (e.g. linking the US census, 300 million x 300 million).

The program is strictly a linkage engine and implements traditional probabilistic record linkage methodology.

The Bigmatch program is designed to extract plausible matches from a large file using several blocking criteria without having to sort the file before each blocking run.

Further details at https://www.census.gov/srd/papers/pdf/rrc2007-01.pdf
A low initial score threshold (5.0) was selected as it was expected scores would be lower than usual as more field values are missing.

Block number	Block description
1	Exact matches on UPRN, age band, sex, full forename and full surname
2	Exact matches on UPRN, age band, sex, first name initial and full surname
3	Exact matches on UPRN
4	Exact matches on full postcode, age band, sex, full forename and full surname
5	Exact matches on full postcode, age band, sex, first name initial and full surname
6	Exact matches on full postcode, age band, sex, first name initial and Soundex surname
7	Exact matches on full postcode, age band, sex, middle name initial and full surname
8	Exact matches on full postcode, sex, full forename and full surname
9	Exact matches on full postcode, age band, full forename and full surname
10	Exact matches on full postcode, age band, sex and full surname
11	Exact matches on full postcode, age band, sex and full forename
12	Exact matches on 1st 4 char postcode, age band, sex, full forename and full surname
13	Exact matches on 1st 4 char postcode, age band, sex, first name initial and full surname
14	Exact matches on 1st 4 char postcode, age band, full forename and full surname
15	Exact matches on 1st 4 char postcode, sex, full forename and full surname
16	Exact matches on 1st 4 char postcode, age band, sex, full forename and Soundex surname
17	Exact matches on 1st 2 char postcode, age band, sex, full forename and full surname

Dataset Name: Farm Structure Survey
Dataset ID: 2122-A009

National Records of Scotland

Stage 3: Deduplication of match pairs

Number of pairs above threshold score output from all blocks per batch:

Batch Number	ExtID in batch	Number of pairs	Unique ExtID/SpineID combinations above threshold(s)	Unique ExtID above threshold(s)	Unique SpineID above threshold(s)	Unique ExtID/SpineID combinations at best match score
1 (whole dataset run)	20,002	62,806	62,806	20,002	50,840	22,460
TOTAL	**20,002**	**62,806**	**62,806**	**20,002**	**50,840**	**22,460**

Stage 3: Deduplication

Identify where there are duplicate ExtID across multiple SpineID

Number of ExtID/SpineID combinations at best match score (per ExtID)	**22,460**	
Number of ExtID matched to single SpineID at best match score	18,711	
Number of spineID matched to a single ExtID at best match score	20,995	

An automated process is carried out in order to ensure that each ExtID can appear a maximum of only once in the final linked dataset.

Step 1: Where bestblock for extid/spineid pair is lowest number (most stringent) bestblock for that extid	22,245	
Step 2: Where maximum score for extid/spineid pair is the maximum score for that spineid	21,582	
Step 3: Restrict to where only one extid per spineid	21,415	
Step 4: Restrict to where only one spineid per extid	18,442	
Following clerical review apply thresholds to select better quality matches	14,213	
Final number of external records with best matches to the Spine	**14,213**	
Matches found by deterministic linkage	**1,425**	
Combined matches	15,638	
Combined matches (after removing any competing spineID)	15,589	46.3%
Final number of external records with best matches to health data (CHI number)	**15,560**	**46.0%**

Dataset Name: Farm Structure Survey
Dataset ID: 2122-A009

Stage 4: Linkage Quality

Deterministic exact matches with no competing matches are assumed to have no false positives

The blocking criteria employed in this linkage and the block-specific linkage thresholds were determined iteratively over a number of BigMatch runs by clerically reviewing a limited sample of best match weight pairs per blocking strategy.
After the final BigMatch run and post-run processing , best match pairs were sampled using a stratified random approach. Best m atch pairs were stratified by the blocking criteria and the integer part of the probabilistic linkage score. Pairs were sorted within each strata by the linkage weight,and a random sample of up to 20 pairs were selected within each block and integer weight.
The final thresholds used in this linkage were set as follows:
Block 3,6,8 (score 8.0, 10.0 and 9.0), block 11 (score 12.0) and block 16 (score 11). Blocks 9, 10 & 15 were dropped on the b asis of too many competing matches/difficult to assess match status.

Summary Estimate of Precision from Pairs Sampling - by Blocking Strategy :-

	Description	N	%	Number Sampled	Estimated Precision	Estimated Precision inc low precision strata
Block 2	Exact matches on UPRN, age band, sex, first name initial and full surname	678	4.8%	20	95.0%	100.0%
Block 3	Exact matches on UPRN	117	0.8%	35	89.1%	46.8%
Block 4	Exact matches on full postcode, age band, sex, full forename and full surname	2250	15.8%	20	100.0%	100.0%
Block 5	Exact matches on full postcode, age band, sex, first name initial and full surname	7233	50.9%	20	100.0%	100.0%
Block 6	Exact matches on full postcode, age band, sex, first name initial and Soundex surname	261	1.8%	71	97.4%	94.0%
Block 7	Exact matches on full postcode, age band, sex, middle name initial and full surname	4	0.0%	4	100.0%	100.0%
Block 8	Exact matches on full postcode, sex, full forename and full surname	74	0.5%	20	100.0%	34.7%
Block 9	Exact matches on full postcode, age band, full forename and full surname	4	0.0%	4	100.0%	80.0%
Block 12	Exact matches on 1st 4 char postcode, age band, sex, full forename and full surname	454	3.2%	58	87.8%	41.9%
Block 13	Exact matches on 1st 4 char postcode, age band, sex, first name initial and full surname	2439	17.2%	60	96.4%	99.0%
Block 14	Exact matches on 1st 4 char postcode, age band, full forename and full surname	29	0.2%	29	100.0%	100.0%
Block 16	Exact matches on 1st 4 char postcode, age band, sex, full forename and Soundex surname	79	0.6%	26	87.2%	100.0%
Block 17	Exact matches on 1st 2 char postcode, age band, sex, full forename and full surname	432	3.0%	42	84.2%	13.4%
Overall		**14,213**	**100.0%**	**429**	**98.0%**	**81.7%**

Precision Est		
		98.0%
	95% Lo	**96.9%**
	95% Hi	**99.2%**

Deterministic matches were found on the basis of exact match on forename initial, full surname, age, sex and full postcode. Those cases matching a single spineid (where not matched by probabilistic linkage) were selected. Deterministic matches are assumed to be correct.

Dataset Name: Farm Structure Survey
Dataset ID: 2122-A009

Stage 5: Linkage Rates by Data and Demography

Personal data from a business survey (of farms) has been matched to individual personal data (home postcode)
We might expect lower match rates because of this data source incompatibility
Probabilistic and Deterministic matches have been combined
Analysis is at farm ID level (rather than data row level) as some farm ID's had more than one data row

data completeness (minimum of populated forename initial, surname, sex, age band and postcode)

complete_data	Match			
	No	Yes	Total	% Match
No	6,080	116	6,196	1.9%
Yes	11,968	15,473	27,441	56.4%
Total	18,048	15,589	33,637	46.3%

Completeness of the available personal details is a major factor in linkage success

have full forename	Match			
	No	Yes	Total	% Match
No	12,730	11,211	23,941	46.8%
Yes	5,318	4,378	9,696	45.2%
Total	18,048	15,589	33,637	46.3%

Surprisingly having full forename details doesn't affect the match rate much (we might expect full forename to be better than forename initial).
However, missing values for other variables could interfere with matching

have full forename	Match			
(with all other details complete)	No	Yes	Total	% Match
No	8,762	11,143	19,905	56.0%
Yes	3,206	4,330	7,536	57.5%
Total	11,968	15,473	27,441	56.4%

Having complete data increases the match rate. Full forename details give a small extra benefit compared to forename initial

Age

Age (years)	Match			
	No	Yes	Total	% Match
16-24	76	42	118	35.6%
25-34	490	326	816	40.0%
35-44	2,268	1,685	3,953	42.6%
45-54	4,083	3,592	7,675	46.8%
55-64	4,245	4,073	8,318	49.0%
65-99	4,606	5,852	10,458	56.0%
missing	2,280	19	2,299	0.8%
Total	18,047	15,589	33,637	46.3%

The match rate appears to increase for older people.
One possible reason for this is that young people might be more likely to be working on somebody else's farm (not their home)

Age (controlling for all other details being complete)

	Match			
Age (years)	No	Yes	Total	% Match
16-24	50	42	92	45.7%
25-34	369	325	694	46.8%
35-44	1,706	1,674	3,380	49.5%
45-54	3,066	3,571	6,637	53.8%
55-64	3,223	4,056	7,279	55.7%
65-99	3,554	5,805	9,359	62.0%
Total	11,968	15,473	27,441	56.4%

Still see an association so age may be a factor in likelihood of linking data
Could there be an association between data completeness and age?

Age	complete_data			
Age (years)	No	Yes	Total	% with complete data
16-24	26	92	118	78.0%
25-34	122	694	816	85.0%
35-44	573	3,380	3,953	85.5%
45-54	1,038	6,637	7,675	86.5%
55-64	1,039	7,279	8,318	87.5%
65-99	1,099	9,359	10,458	89.5%
missing	2,299	0	2,299	0.0%
Total	6,195	27,441	33,637	81.6%

Older people do appear to be slightly more likely to provide complete details

Dataset Name: Farm Structure Survey
Dataset ID: 2122-A009

Stage 5: Linkage Rates by Data and Demography - continued

Sex

	Match			
Sex	No	Yes	Total	% Match
Female	3,093	2,236	5,329	42.0%
Male	12,675	13,334	26,009	51.3%
missing	2,280	19	2,299	0.8%
Total	18,047	15,589	33,637	46.3%

There is a slightly lower match rate for females. This could be due to the marriage surname not being updated on the population spine.

Now controlling for all personal details being complete

	Match			
Sex	No	Yes	Total	% Match
f	2,431	2,219	4,650	47.7%
m	9,537	13,254	22,791	58.2%
Total	11,968	15,473	27,441	56.4%

The same pattern of slightly lower match rate for females persists. With complete data, match rates are higher for both males and females.

UPRN

	Match			
Have UPRN	No	Yes	Total	% Match
No	6,438	3,557	9,995	35.6%
Yes	11,610	12,032	23,642	50.9%
Total	18,048	15,589	33,637	46.3%

There is a lower match rate when UPRN is missing. Does UPRN improve matching or is it a reflection of data completeness?

	Have complete data			
Have UPRN	No	Yes	Total	% with complete data
No	3,838	3,503	7,341	47.7%
Yes	8,198	12,020	20,218	59.5%
Total	12,036	15,523	27,559	56.3%

If UPRN is provided then it is more likely that complete personal details are provided.

The match rate is improved by having UPRN. This is a more specific match on property (resident at).
As 70% of FSS records have uprn, we might expect an even higher match rate using UPRN. The lower observed match rate may reflect low data quality.