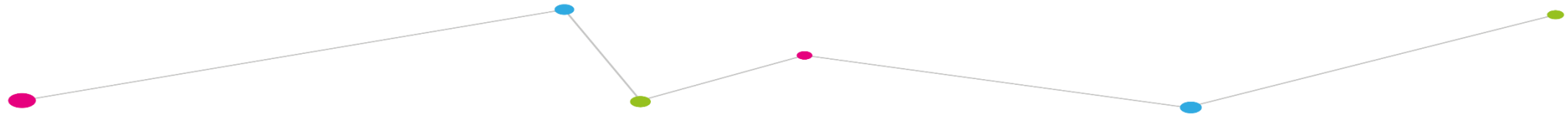
Creating sibling indicators in the NPD

Talk to NPD User Group Sept 2016

Tom Holt, FFT (tom.holt@fft.org.uk)
Becky Allen, Education Datalab (rebecca.allen@fft.org.uk)



The challenge of identifying siblings in the NPD

1. What do we mean by a sibling?
2. Can we develop routines to accurately and efficiently clean-up the address information and family names?

DfE commissioned FFT to investigate the practicalities involved in including some form of sibling indicator across the NPD

Existing research using siblings in the NPD

- Siblings identified by: Cheti Nicoletti and Birgitta Rabe, ISER, University of Essex
- Siblings defined as pupils in state schools aged 4-16 and living together at the same address in January 2007 (first year of full address details in NPD)
- Their main analysis sample is those taking KS4 exams in 2007, 2008, 2009 or 2010
- Leaves sample of sibling pairs (excluding twins) that includes about 10% of pupils

Their research questions

- Inequality in Pupils' Educational Attainment: How Much Do Family, Sibling Type and Neighbourhood Matter? https://www.iser.essex.ac.uk/files/iser_working_papers/2010-26.pdf
- School inputs and skills: Complementarity and self-productivity https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2013-28.pdf
- Sibling spillover effects in school achievement https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2014-40.pdf

Identifying Siblings

A sibling can be defined as:

- **SG(A)**: A pupil living at the same house as another pupil

or

- **SG(A+S)**: A pupil living at the same house and having the same surname as another pupil

Approach (Summary)

- Collating all addresses as expressed in census tables
- Grouping and matching into core address table
- Deriving sibling groups based on the definitions SG(A) and SG(A+S)

Datasets Included:

- Spring Census.
- Pupil Referral Unit (PRU) Census
- Alternative Provision (AP) Census
- Early Years Census (EYC)

Approach (Collating Addresses)

School Census accepts 2 address formats:

- BS7666 (preferred)
- Line Address Format

Standardising addresses from both formats into:

- Building Part No.
- Building Part Name
- Building Number
- Building Name
- Street Name
- Postcode

Approach (Grouping / Matching Addresses)

- Exact Match: Grouping addresses considered to be identical based on standardised fields
- Matching via Postal Address File (PAF)
  - Finding ‘best’ Unique Property Reference Number (UPRN) for each address - exact or fuzzy matching (e.g. postcode errors, mis-spellings)
  - Grouping addresses which are not identical but link to same UPRN

- Master table of Address IDs generated based on the final grouped addresses.

Approach (Deriving Sibling Groups)

- Each row in the census tables is linked to the relevant Address ID

Sibling Group A (Address Only)

- The Address ID is also used as the SG(A) ID

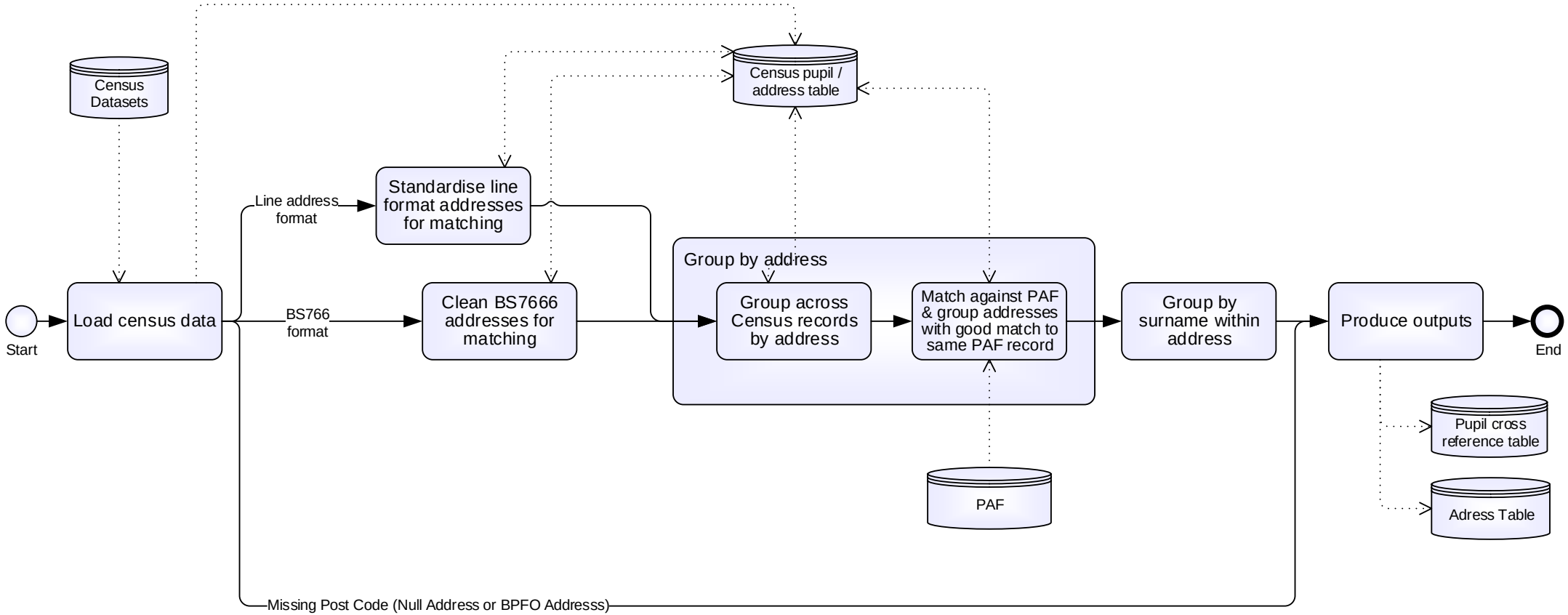
Sibling Group A+S (Address & Surname)

- Records with same Address ID and surname grouped into SG(A+S) ID
- Records are also grouped if a previous, alternative or fuzzy surname match is found

Approach (Sibling Group Information)

- Quality of address matching within sibling group
- Ignore flag set for duplicate NPD Pupil IDs – records are excluded from outputs
- Counts of valid records per sibling group, and birth order ranks within each group are calculated

Final Process Adopted



Quality Assurance - Matching

- Quality of matched records (2013 school census)

Category		Number of matches	%age of matches
	Matches made between Census records	3,201,240	100.0%
A	Near exact matches where the address was also found on PAF	3,062,432	95.7%
B	Matches made via PAF with a strong match to PAF	81,465	2.5%
C	Near exact matches where the address was not found on PAF	52,803	1.6%
D	Less certain match via PAF matching	4,540	0.1%

Table 3.A- Breakdown of matches by level of confidence

- Quality of Unmatched records (2013 school census)

Category		Number of records	%age of records included in matching
	Unmatched census records	2,688,776	32.6%
E	With match to PAF	2,552,773	30.9%
F	Without match to PAF	136,003	1.6%

Table 3.B - Breakdown unmatched records

Breakdown by Size of Sibling Group

(2013 school census)

Number of siblings in group	Methodology SG(A)						Methodology SG(A+S)
	%age of people	Breakdown by numbers of surname clusters in group					%age of people
		1	2	3	4	5 or more	
1	33.04%	100.00%					39.84%
2	41.97%	89.45%	10.54%				40.96%
3	17.19%	76.35%	13.94%	9.71%			14.13%
4	5.46%	64.18%	13.05%	13.74%	9.02%		3.77%
5	1.58%	54.40%	11.93%	13.09%	11.68%	8.90%	0.93%
6	0.50%	49.80%	9.69%	11.05%	11.35%	18.09%	0.27%
7	0.16%	0.08%	0.17%	0.46%	2.00%	13.07%	0.08%
8	0.05%	0.02%	0.05%	0.13%	0.57%	5.87%	0.02%
9	0.02%	0.01%	0.01%	0.04%	0.18%	2.61%	0.01%
10-15	0.01%	5.96%	1.72%	2.26%	3.43%	74.89%	0.00%
16+	0.02%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%
Total	100.00%	88.42%	7.79%	2.70%	0.76%	0.33%	100.00%

Outputs

Pupil cross reference tables

Tables for the two types of sibling group, with NPD Pupil ID from census, and sibling group identifiers

SG(A) Grouping Table	Data Year: 2016
NPD Id	
SG(A) group Identifier	
Number of pupils in group	
Pupil Birth Order	
Pupil DOB	
Address identifier	
Address matching quality indicator	

SG(A+S) Grouping Table	Data Year: 2016
NPD Id	
SG(A+S) group Identifier	
Number of pupils in group	
Pupil Birth Order	
Pupil DOB	
Address identifier	
Address matching quality indicator	

Outputs (2)

Address Table

Table of addresses as expressed in Census tables (not year-specific)

Address identifier	
SAON (Secondary Addressable Object Name)	BS7666 only
PAON (Primary Addressable Object Name)	BS7666 only
STREET	BS7666 only
ADDRESSLINE1	Line Address Format only
ADDRESSLINE2	Line Address Format only
ADDRESSLINE3	Line Address Format only
POSTCODE	
UPRN	

Timetables

- Late October: Spring Census 07/08 and Spring Census 12/13
- November: Spring Census 15/16 (requires updated PAF file)
- Future: possibly ILR and HESA data

What now?

- How good are these sibling indicators?
- Would you use them in your research?
  - To understand the impact siblings have on each other
  - To account for similar family backgrounds but different educational experiences

Thanks for listening!

Tom Holt, FFT (tom.holt@fft.org.uk)
Becky Allen, Education Datalab (rebecca.allen@fft.org.uk)