2 : Contemporary international sources

Daryl Lloyd

There are obviously a large number of sources of surname data but in this session the intention is to focus on sources with comprehensive coverage of particular countries. The ideal such source would have as close as possible to 100% coverage of the nation's population, have representative geographical coverage and have no significant bias as between different ethnic and socio economic groups. There are a number of possible sources which can claim in varying degrees to be comprehensive, some of which are available across the world, others of which are specific to individual countries. These can be predominantly split down into:

1. 'Official' registers, such as the Electoral Register
2. Telephone directories
3. Ownership registers

Each source of data collection is less than 100% complete, with various levels of uncertainty. Not everyone completes the forms sent to them by Electoral Registrars, for instance, and it is not uncommon to opt out of national telephone directories for privacy reasons.

Furthermore, when making comparisons between countries at a regional level, difficulties can occur as a result of different methods or levels of georeferencing. There are relatively few, if any, geographical schemes which are fully compatible between countries, and even within a single country various sources of data may utilise a wide range of incompatible base units.

Anglophone Countries

Great Britain (not Northern Ireland)

Our main contemporary source from the UK comes from is the 1997 / 1998 version of the Electoral Register supplied by Experian. This covers the whole of the UK other than Northern Ireland. This, therefore, immediately has a number of limitations in comparison to the use of the Census, as Kevin Schürer discusses in his paper. First, only

individuals over 17 at the time of compilation will be recorded. In addition to this, not everyone who is entitled to be vote will be registered. According to Dance (1999) up to 14 per cent of the possible population will be missing from the register, thereby decreasing the total number of names that will appear. This percentage is likely to increase over time and we are therefore fortunate to have access to the last electoral register compiled in the UK prior to the introduction of arrangements whereby electors could 'opt out' of the supply of their entries to third parties. However, ONS figures quote the GB population in 2001 as 45,643,126 (over 16) and the total number of people registered to vote on our 1997/1998 electoral register is 43,340,557, only 5 per cent difference.

For GB98 we have introduced a minimum cut-off of 100 occurrences, meaning that only the surnames that occur at least 101 times in the register have been retained in our database for analysis. This has led to a total of 25,730 unique names (after standardising Mc, Mac and O' type names), with a total number of 37,278,477 occurrences. As will be discussed in the papers by Ken Tucker, and Mike Batty and Paul Longley, the frequencies of occurrence by name are very uneven. Therefore the largest 100 surnames (such as Smith, Jones and Williams) account for 24 per cent of the total population, whilst the top 10 per cent largest names contain in excess of 72 per cent of the total occurrences.

The ratio of UK minimum occurrences to total population has been used to derive cut-off values for some of the other (non UK) datasets. To make the 1998 analyses comparable to those based on earlier GB data sources (i.e. the 1881 Census of Population), this ratio is based only those names which have been classified as having British origins (see Richard Webber's paper on Taxonomies). This reduced the 1998 dataset to a population of 35,654,408 (23,186 unique names). When using this in conjunction with each other country's base population (or, more strictly speaking, the total number of surname occurrences) this produces a new cut-off. The final cut-offs used for all datasets currently used are show in Table 1.

Country	Population	Calculated cut-off
GB 1998	37,278,477	101 (non-calculated)
GB 1881	28,225,211	79.9
USA	80,921,677	229
Canada	9,148,211	25.9
Australia	7,784,676	100 (non-calculated)
New Zealand	934,686	2.6

Table 1: Minimum cut-offs for databases

We hold two levels of georeferencing for GB98, both of which are based of postcode geography. Our priority level is the *postcode area* (that is the first letter/s of the postcode – e.g. **WC**1E 6BT or **N**22 6RB). There are a total of 120 of these in GB, each with a different size of population, ranking from 1,440,771 (**B** – Birmingham) down to 17,820 (**ZE** – Shetland Islands) – a complete list of all relevant postcode area populations can be found in Table 2.

The second geographic level at which data has been held is *postcode sector* (the whole postcode aside from the final two letters – e.g. **WC1E 6**BT or **N22 6**RB). There are circa 9,000 of these in GB, and we have extracted a full set of names and addresses for two of the country's 659 parliamentary constituencies - Falmouth and Camborne in Cornwall and Copeland in Cumbria. In these areas we have been able to look at finer-scale surname patterns and distributions by virtue of holding information at the level of the full postcode (eg N6 4AN) (see session 20).

Australia

The Australian data has been provided by Pacific Micromarketing, a marketing information business based in Melbourne. As with the GB98 database, these names have been digitised from the Electoral Register, the result of which is that the number of occurrences is considerably lower than the actual population. There is a much greater difference between possible voters and the number of people actually registered in Australia as there is in GB98. In 2001 in Australia, the population over 15 stood at 14,856,774, whilst the total number of registered electors was only 7,784,876. However, one of the likely reasons for this difference is that the data as provided already contained a cut-off of 100, as with GB98. This means that there could be a very high proportion of the population with a surname with fewer occurrences than this which are lost because of the artificially introduced threshold.

This data is also georeferenced at two levels. Comparable (in terms of the number of registered voters) with the larger of British postcode areas are the Australian States, of which there are eight. Below this there are Statistical Sub-Divisions (SSD), of which we have a total of 197 individual units (though there should be a total of 207 resulting from the 2001 Census Australian Bureau of Statistics, 2001: 13). As with GB98 the range of this is quite large, with Daly (SSD 71020 – NT) only containing 6 entries up to 436,771 in Brisbane City (SSD 30505 – QL).

New Zealand

There are two sources of data from New Zealand. The more comprehensive of the two is based on the telephone directory, containing 934,686 entries, but this will only have one entry per household, and no indication to the number of people with the same or different surnames in the household.

The second source contains even fewer occurrences (608,747) and originates from the National Land Title Database, which is the register of owner-occupies for the whole of New Zealand. There are some advantages of this database, even though it contains fewer occurrences. It is more official than the telephone directory, and has few opportunities for the individuals to opt out. Furthermore it contains, where relevant, the second surname of an owner-occupier household. This occurs in 218,997 cases this resulting in a combination of both the first and second surnames which produces a total of 827,768 unique entries. However, it is fair to say that this sample is biased towards households in which the owner-occupiers are both married and jointly own the property.

The level of aggregation on which both sets of data are supplied is based on Census geography, and the data has been summarised at each level of this geography. The base level is the Meshblock (38,336) which is aggregated up to the level of the Region. There are a total of 18 Regions in NZ, but only 16 in the database as two regions make up the Areas outside of Territorial Authorities and the offshore Chatham Islands District for which no records have been supplied.

In the case of the larger telephone directory, there are 77,693 unique names made up from 934,686 occurrences. However, only 851 unique names appear at least 50 times and 44 occur 500 times or more. With such a relatively poor coverage (the true population of New Zealand is just over 4 million) it would be expected that many of the smaller GB names would be missing from the database altogether.

US

Currently we only hold national totals for surnames, covering 145,242 unique names and a total number of occurrences of almost 81 million.

Canada

Currently we only hold national totals for surnames, covering 33,355 unique names and a total number of occurrences of 9.15 million

Non-Anglophone Countries

Sweden

Our Swedish data is sourced from the telephone directory and contains 9,281 unique names. The base georeferencing level is County, of which there are 25 throughout the whole country, and the total number of occurrences is 3,067,858 (out of true population of about 8,900,000). This number seems roughly correct, as publicly available figures show that there are roughly 3,800,000 households, but 5,790,000 conventional / land telephones (many of which would be for business use) (European Union, 2000). However, it has been pointed out by one Swede that the total number of telephone lines for his name in the database is smaller for the whole country than it should be for just his County.

Netherlands

The Dutch database is drawn from their telephone directory, but with an extra level of difficulty. As well as containing the surname in which a telephone line is registered, it also contains the names of businesses which are in the telephone directory, and this is what has lead there to being 1,436,352 unique entries and a total of 8,036,624 occurrences. EU estimates suggest that in 1999 there were only 6,800,000 households in the Netherlands. Furthermore, there we do not have any sub-national breakdowns.

Other sources

For contemporary Britain there are a number of other available sources, such as the NHS Central Register (more details of which can be found on Philip Dance's website - http://homepages.newnet.co.uk/dance/webpjd/intro/nhscrweb.htm), though one important thing to note about this potentially very useful database is that it covers England and Wales only.

Appendix – Base population sizes for GB98

Postcode area code	Postcode area name	Population
AB	Aberdeen	378,871
AL	St. Albans	186,846
B	Birmingham	1,440,771
BA	Bath	332,604
BB	Blackburn	370,762
BD	Bradford	424,784
BH	Bournemouth	443,538
BL	Bolton	297,356
BN	Brighton	641,865
BR	Bromley	240,479
BS	Bristol	737,262
CA	Carlisle	254,884
CB	Cambridge	329,601
CF	Cardiff	790,946
CH	Chester	531,678
CM	Chelmsford	495,804
CO	Colchester	327,123
CR	Croydon	306,312
CT	Canterbury	379,266
CV	Coventry	627,219
CW	Crewe	241,306
DA	Dartford	322,719
DD	Dundee	219,083
DE	Derby	565,314
DG	Dumfries	119,802
DH	Durham	253,410
DL	Darlington	289,284
DN	Doncaster	575,568
DT	Dorchester	168,283
DY	Dudley	327,275
E	London E	611,021
EC	London EC	24,805
EH	Edinburgh	679,274
EN	Enfield	260,876
EX	Exeter	430,818
FK	Falkirk	211,714
FY	Blackpool	239,525
G	Glasgow	979,402
GL	Gloucester	470,383
GU	Guildford	568,849
HA	Harrow	352,286
HD	Huddersfield	202,319
HG	Harrogate	113,097
HP	Hemel Hempstead	375,958
HR	Hereford	134,807
HS	Harris	22,040
HU	Hull	355,137
HX	Halifax	121,514

IG	Ilford	226,538
IP	Ipswich	452,327
IV	Inverness	162,467
KA	Kilmarnock	303,651
KT	Kingston-upon-Thames	425,055
KW	Kirkwall	40,889
KY	Kirkcaldy	284,943
L	Liverpool	693,443
LA	Lancaster	273,062
LD	Llandrindod Wells	40,640
LE	Leicester	743,257
LL	Llandudno	423,418
LN	Lincoln	215,936
LS	Leeds	615,659
LU	Luton	243,225
M	Manchester	873,448
ME	Medway	444,169
MK	Milton Keynes	367,286
ML	Motherwell	303,810
N	London N	607,067
NE	Newcastle upon Tyne	941,595
NG	Nottingham	898,113
NN	Northampton	478,558
NP	Newport	384,119
NR	Norwich	570,079
NW	London NW	411,635
OL	Oldham	357,075
OX	Oxford	488,104
PA	Paisley	269,173
PE	Peterborough	657,833
PH	Perth	127,476
PL	Plymouth	434,443
PO	Portsmouth	647,434
PR	Preston	413,660
RG	Reading	597,901
RH	Redhill	399,659
RM	Romford	375,602
S	Sheffield	1,080,103
SA	Swansea	576,487
SE	London SE	689,059
SG	Stevenage	307,355
SK	Stockport	495,151
SL	Slough	278,481
SM	Sutton	166,284
SN	Swindon	337,165
SO	Southampton	521,825
SP	Salisbury	180,018
SR	Sunderland	208,636
SS	Southend-on-Sea	413,126
ST	Stoke-on-Trent	520,973
SW	London SW	705,446

SY	Shrewsbury	262,212
TA	Taunton	250,590
TD	Galashiels	88,707
TF	Telford	157,859
TN	Tunbridge Wells	523,101
TQ	Torquay	231,372
TR	Truro	229,427
TS	Cleveland	483,428
TW	Twickenham	372,632
UB	Southall	272,737
W	London W	448,373
WA	Warrington	482,023
WC	London WC	31,077
WD	Watford	200,977
WF	Wakefield	396,482
WN	Wigan	248,415
WR	Worcester	232,093
WS	Walsall	342,823
WV	Wolverhampton	306,965
YO	York	438,855
ZE	Lerwick	17,820

Table 2: GB 1998 postcode area populations

Australian Bureau of Statistics, 2001; Statistical Geography Volume 1: Australian Standard Geographical Classification, Australian Bureau of Statistics, Canberra, ACT.

Dance, P.J., 1999; Modern British Surnames, Available: http://homepages.newnet.co.uk/dance/webpjd/index.htm [Accessed: 4th April 2004]

European Union, 2000 (November); Basic facts and indicators: Sweden, Available: http://www.eu-esis.org/Basic/SEbasic00.htm [Accessed: 19th April, 2004]