British Social Attitudes 1983 to 2014

2: Resolution of type conflicts

 (Draft only: 25 April 2016)

Problem variables 2000 to 2008

Files were combined in reverse year order

	2008	2007	2006	2005	2004	2003	2002	2001soc	2001	2000
SIntDate	edate10	edate10		a50	a50				a8	a50
StTIM	time8	time8		a50	a50	a50				
EndTIM	a10	time8			a8	a8				
Intdate		edate10		a50	a50	a50	a50		a8	f9
int_num		f4	f4	a4	f4					
IntNum		f4				a4	a4		a4	
LACode					a50	a50	a50		a4	a50
councode						a50	a50		a2	a50
postcode							a50		a4	a50

Conflicts in red:

	2008	2007	2006	2005	2004	2003	2002	2001soc	2001	2000
SIntDate	edate10	edate10		f8	a50				a8	a50
StTIM	time8	time8		time8	a50	a50				
EndTIM	time8	time8			a8	a8				
Intdate		edate10		f8	a50	a50	a50		a8	f9
int_num		f4	f4	a4	f4					
IntNum		f4				a4	a4		a4	
LACode					a50	a50	a50		a4	a50
councode						a50	a50		a2	a50
postcode							a50		a4	a50

Rather than check through each wave, I found it quicker to run **ADD FILES** and then deal with any inconsistent types identified.

I prefer to work with **abbreviated syntax** in **lower case**. In the SPSS syntax below I have colour coded the abbreviations to tally with that of the full commands, keywords and operators.

Creating a new variable **[date]** from various formats in different waves.

1995
Source variable: **dateint** (f9)

NB: This variable is called **Intdate** in all other waves.

```
compute day = trunc (dateint/1000000).
formats day (f2.0).
desc var day.
list var day /cases 5.

compute month = trunc (dateint/10000) - (day*100).
formats month (f2.0).
desc var  month.
list var  month /cases 5.

missing values day month (99).
freq day month.

compute date =  DATE.DMY(day,month,year).
list date /cases 5.
var lab day 'Day of interview'
   /month 'Month of interview'
   /year 'Year of interview'
   /date 'Date of interview (ddmmyyyy)'.
disp dic /var day month year date.

list dateint /cases 5.

*1995.
```

Variable	Label	Level	Width	Format
day	Day of interview	Scale	10	F2
month	Month of interview	Scale	10	F2
year	Year of interview	Scale	10	F4
date	Date of interview (ddmmyyy)y	Unknown	10	Edate10

2005, 2006, 2008

Source variable: qdate (f4)

```
desc var qdate.
list qdate /cases 5.

missing values qdate (-1,99).
compute day = trunc (qdate/100).
list day /cases 3.
compute month = mod  (qdate/ 100).

formats day month (f2.0).
list month /cases 3.
compute date =  DATE.DMY(day,month,year).

var lab day 'Day of interview'
   /month 'Month of interview'
   /year 'Year of interview'
   /date 'Date of interview (ddmmyyyy)'.

disp dic /var day month year date.
```

2007

Source variables: **[intdate] [sintdate]** but perhaps a new variable **[date]**?

Variable	Position	Label	Level	Width	Format
IntDate	10	Interviewer: Check Date of Interview and alter if not correct Q36	Scale	10	EDATE10
SIntDate	11	Computer IntDate Q37	Scale	10	EDATE10
StTIM	12	Start time Q38	Scale	10	TIME8

2002 to 2008 complete

2001 has **qdate qdated qdatem** Done

1999 – 2005 done.

1998

```
add files
        file *
        /file 'dataset13'.
```

Incompatible types

```
list  sintdate intdate sttim strttime postcode /cases 5.
```

sintdate: 20051998 **intdate**: 20051998 **sttim**: 18:56:15 **strttime** 1856 **postcode**: N18

```
disp dictionary
  / var sintdate intdate sttim strttime postcode.
```

Original

Variable	Position	Label	Level	Width	Format
sintdate	19	Computer IntDate DDMMYYYY Q27	Nominal	10	A75
intdate	18	Date interview completed DDMMYYYY Q26	Nominal	10	A75
sttim	20	Start time HH:MM:SS Q29	Nominal	10	A75
strttime	21	Start time HHMM Q29	Scale	8	F4
postcode	1106	Post code sector Census1	Nominal	8	A75

alter type sintdate intdate (f10.0) sttim (f8.0) postcode (a50).

Altered Types

Computer IntDate DDMMYYYY Q27	EDATE10	F10.0
Date interview completed DDMMYYYY Q26	EDATE10	F10.0
Check date of interview: DDMMYYYY Q802	F8	F10.0
Date completedDDMMYYYYA2.55bB2.79bC2.55b	F8	F10.0
Start time HH:MM:SS Q29	TIME8	F8.0
Post code sector Census1	A50	A50

Altered Types

Computer Interview Date DD:MM:YYYY :Q38	EDATE10	F10.0
Interviewer: Check Date of Interview and alter if not correct Q36	EDATE10	F10.0
Start time HH:MM:SS :Q39	TIME8	F8.0
Postcode Sector <spoint>	A50	A50

display dictionary
/variables sintdate intdate sttim strttime postcode.

Variable	Position	Label	Level	Width	Format
SIntDate	11	Computer Interview Date DD:MM:YYYY :Q38	Nominal	10	F10
IntDate	1051	Interviewer: Check Date of Interview and alter if not correct Q36	Scale	10	F10
StTIM	12	Start time HH:MM:SS :Q39	Nominal	10	F8
strttime	4384	Start time of interview HHMM Q32	Scale	8	F4
postcode	4004	Postcode Sector <spoint>	Nominal	50	A50

formats sintdate intdate (edate10) sttim (time8).
display dictionary
/variables sintdate intdate sttim strttime postcode.

Variable	Position	Label	Level	Width	Format
SIntDate	11	Computer Interview Date DD:MM:YYYY :Q38	Nominal	10	EDATE10
IntDate	1051	Interviewer: Check Date of Interview and alter if not correct Q36	Scale	10	EDATE10
StTIM	12	Start time HH:MM:SS :Q39	Nominal	10	TIME8
strttime	4384	Start time of interview HHMM Q32	Scale	8	F4
postcode	4004	Postcode Sector <spoint>	Nominal	50	A50

Compute new variable **[date]** to comply with other waves:

compute date = intdate.
formats sintdate intdate lintdate date (edate10) sttim (time8) strttime (time6).
variable labels year 'Year of interview'
/date 'Date of interview (ddmmyyyy)'.
display dictionary
/variables date sintdate intdate sttim strttime postcode year.

Variable Information

Variable	Position	Label	Level	Width	Format
date	1121	Date of interview (ddmmyyyy)	Unknown	10	EDATE10
sintdate	19	Computer IntDate DDMMYYYY Q27	Nominal	10	EDATE10
intdate	18	Date interview completed DDMMYYYY Q26	Nominal	10	EDATE10
sttim	20	Start time HH:MM:SS Q29	Nominal	10	TIME8
strttime	21	Start time HHMM Q29	Scale	8	TIME6
postcode	1106	Post code sector Census1	Nominal	8	A50
year	1120	Year of interview	Scale	10	F4

Variables in the working file

freq year.

year Year of interview

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1997	**5420**	7.1	7.1	7.1
	1998	**6292**	8.2	8.2	15.3
	1999	3143	4.1	4.1	19.4
	2000	3426	4.5	4.5	23.9
	2001	3287	4.3	4.3	28.2
	2002	**13740**	17.9	17.9	46.1
	2003	**17728**	23.1	23.1	69.2
	2004	**6398**	8.4	8.4	77.6
	2005	4268	5.6	5.6	83.2
	2006	4290	5.6	5.6	88.8
	2007	4124	5.4	5.4	94.1
	2008	4486	5.9	5.9	100.0
	Total	76602	100.0	100.0	

Oops! Duplicate cases (my fault when using **ADD CASES** on wrong file)

SPSS to the rescue

Data > Identify Duplicate Cases

(from PASTE)

```
SORT CASES BY Serial(A).
MATCH FILES
  /FILE=*
  /BY Serial
  /FIRST=PrimaryFirst
  /LAST=PrimaryLast.
DO IF (PrimaryFirst).
COMPUTE  MatchSequence=1-PrimaryLast.
ELSE.
COMPUTE  MatchSequence=MatchSequence+1.
END IF.
LEAVE  MatchSequence.
FORMATS  MatchSequence (f7).
COMPUTE  InDupGrp=MatchSequence>0.
SORT CASES InDupGrp(D).
MATCH FILES
  /FILE=*
  /DROP=PrimaryFirst InDupGrp MatchSequence.
VARIABLE LABELS  PrimaryLast 'Indicator of each last matching case as Primary'.
VALUE LABELS  PrimaryLast 0 'Duplicate Case' 1 'Primary Case'.
VARIABLE LEVEL  PrimaryLast (ORDINAL).
FREQUENCIES VARIABLES=PrimaryLast.
EXECUTE.


crosstabs year by primarylast.
```

year Year of interview * PrimaryLast Indicator of each last matching case as Primary Crosstabulation

Count

		PrimaryLast Indicator of each last matching case as Primary		
		0 Duplicate Case	1 Primary Case	Total
year Year of interview	1997	**4065**	1355	5420
	1998	3146	3146	6292
	1999	0	3143	3143
	2000	0	3426	3426
	2001	0	3287	3287
	2002	**10305**	3435	13740
	2003	**13296**	4432	17728
	2004	**3199**	3199	6398
	2005	0	4268	4268
	2006	0	4290	4290
	2007	0	4124	4124
	2008	0	4486	4486
Total		34011	42591	76602

Primary case numbers match original, so:

```
select if primarylast= 1.
frequencies year.
```

year Year of interview

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1997	1355	3.2	3.2	3.2
	1998	3146	7.4	7.4	10.6
	1999	3143	7.4	7.4	17.9
	2000	3426	8.0	8.0	26.0
	2001	3287	7.7	7.7	33.7
	2002	3435	8.1	8.1	41.8
	2003	4432	10.4	10.4	52.2
	2004	3199	7.5	7.5	59.7
	2005	4268	10.0	10.0	69.7
	2006	4290	10.1	10.1	79.8
	2007	4124	9.7	9.7	89.5
	2008	4486	10.5	10.5	100.0
	Total	42591	100.0	100.0	

Phew! The one time it's OK to use **select if** without a previous **temporary.**

File > **Save As**

bsa2008-1997.sav

Proceed to add 1996 then 1997. Rather than spend time combing through the source files, it's quicker to run **ADD FILES** as this immediately throws up the inconsistent variable types.

```
*1996au.
freq rsex.
```

rsex Sex of respondent Q28

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	1576	43.5	43.5	43.5
	2 Female	2044	56.5	56.5	100.0
	Total	3620	100.0	100.0	

```
*1996bu.
freq rsex.
```

rsex Sex of respondent Q28

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	17	40.5	40.5	40.5
	2 Female	25	59.5	59.5	100.0
	Total	42	100.0	100.0	

```
*1996au.
add files file *
   /file 'dataset16'.
freq rsex.
```

rsex Sex of respondent **Q28**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	1593	43.5	43.5	43.5
	2 Female	2069	56.5	56.5	100.0
	Total	3662	100.0	100.0	

File > **Save As**

bsa1996.sav
(dataset 15)

```
*bsa 1997-2008.
add files file *
   /file 'dataset15'.

disp dic /var  ward newdc nwname.
```

Variable Information

Variable	Position	Label	Level	Width	Format	Missing
ward	3844	Ward code Q22	Nominal	9	A10	"ZZZZ"
newdc	5491	Local authority district	Nominal	6	A30	
nwname	5493	New ward name <Scotland>	Nominal	27	A30	

Variables in the working file

```
*bsa 1996.
list ward newdc nwname /cases 5.
```

Variable Information

Variable	Position	Label	Level	Width	Format	Missing
ward	1046	Local authority ward code Q	Nominal	5	A75	
newdc	1050	1996+ local auth.district [Scotland] Q	Nominal	5	F2	-1
nwname	1052	1996+ LA ward name [Scotland] Q	Nominal	27	A75	

Variables in the working file

```
formats ward (a10) newdc nwname (a30).
```

Weird but saved as **bsa2008-1996.sav**

1995 went straight in:

year Year of interview

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1995	3633	7.3	7.3	7.3
	1996	3662	7.3	7.3	14.6
	1997	1355	2.7	2.7	17.3
	1998	3146	6.3	6.3	23.6
	1999	3143	6.3	6.3	29.9
	2000	3426	6.9	6.9	36.8
	2001	3287	6.6	6.6	43.4
	2002	3435	6.9	6.9	50.3
	2003	4432	8.9	8.9	59.2
	2004	3199	6.4	6.4	65.6
	2005	4268	8.6	8.6	74.1
	2006	4290	8.6	8.6	82.7
	2007	4124	8.3	8.3	91.0
	2008	4486	9.0	9.0	100.0
	Total	49886	100.0	100.0	

```
alter type  endtim  (f10.0).
formats endtim (time8).

add files file *
   /file 'dataset1'.
freq year.
```

year Year of interview

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1995	3633	5.2	5.2	5.2
	1996	3662	5.3	5.3	10.5
	1997	1355	2.0	2.0	12.5
	1998	3146	4.5	4.5	17.0
	1999	3143	4.5	4.5	21.6
	2000	3426	4.9	4.9	26.5
	2001	3287	4.7	4.7	31.3
	2002	3435	5.0	5.0	36.2
	2003	4432	6.4	6.4	42.6
	2004	3199	4.6	4.6	47.2
	2005	4268	6.2	6.2	53.4
	2006	4290	6.2	6.2	59.6
	2007	4124	6.0	6.0	65.5
	2008	4486	6.5	6.5	72.0
	2009	3421	4.9	4.9	76.9
	2010	3297	4.8	4.8	81.7
	2011	3311	4.8	4.8	86.5
	2012	3248	4.7	4.7	91.2
	2013	3244	4.7	4.7	95.8
	2014	2878	4.2	4.2	100.0
	Total	69285	100.0	100.0	

Last bit: add 1983-1994 to 1995 -2014

```
*1995-2014.
list ward sector censusdc ccname /cases 10.
freq  ward sector censusdc ccname.
disp dic /var
        ward sector censusdc ccname.
```

Variable Information

Variable	Position	Label	Level	Width	Format	Missing Values
ward	5431	Ward code Q22	Nominal	9	A10	"ZZZZ"
sector	6969	Postcode sector Q11	Nominal	50	A10	
censusdc	7529	local authority/district ONS code Q	Nominal	8	A25	
ccname	8160	county name/region name <A25>	Nominal	27	A10	

Variables in the working file

ward a6 sector a6 values censusdc a2 ccname a15

```
*1983-1994.
list ward sector censusdc ccname /cases 10.
freq  ward sector censusdc ccname.
disp dic /var
        ward sector censusdc ccname.
```

Variable Information

Variable	Position	Label	Level	Width	Format	Missing
ward	2043	Wards in GB versions A+B	Ordinal	6	F4	
sector	3109	ENTER POSTCODE SECTOR	Nominal	8	A75	
censusdc	3895	R's district council correct census dv	Ordinal	8	F3	
ccname	3545	R know the name of regional council?Q346	Ordinal	6	F1	-2

Variables in the working file

Therein lies a problem. Not only different formats, but different variables?

1995 - ward a6 sector a6 censusdc a2 ccname a15 || missing ward ‘zzzz’

1983 - ward f3 sector a6 censusdc f3 ccname f1 || missing sector ‘xxxxxx’

1983

```
rename var
        (ccname=knowname) (ward=numward)(censusdc = numcensusdc) (ccname=knowname)
```

From 1995-2014:

```
add files file *
   /file 'dataset18'.
```

Check:

```
freq year.
```

year Year of interview

		Frequency
Valid	1983	1761
	1984	1675
	1985	1804
	1986	3100
	1987	2847
	1989	3029
	1990	2797
	1991	2918
	1993	2945
	1994	3469
	1995	3633
	1996	3662
	1997	1355
	1998	3146
	1999	3143
	2000	3426
	2001	3287
	2002	3435
	2003	4432
	2004	3199
	2005	4268
	2006	4290
	2007	4124
	2008	4486
	2009	3421
	2010	3297
	2011	3311
	2012	3248
	2013	3244
	2014	2878
	Total	95630

Tallies with original files so:

File > Save As:

File name: **bsa1983-2014jfh1.sav**
Size: 1.04 gb
Number of cases: 95,630
Number of variables: 10,764

Still got to check missing values and measurement levels. Since the files were combined in reverse year order, this may not be as taxing. Can't just copy the **Labels** column as the variables are not in the same sequence in each file. I need to work through waves to find the cleanest labels, missing values and measurement levels, then use **APPLY DICTIONARY** /from <file>, but it will still take a very long time.