Longitudinal Linkage of Cross-Sectional NCDS Data Files Using SPSS

What are we doing when we merge data from two sweeps of the NCDS (i.e. data from different points in time)?

We are adding new information (i.e. new variables) for the same set of individuals. Each variable refers to an attribute at that specific point in time, not necessarily a lasting attribute. All variable names in NCDS are unique, so there is no danger of mixing up two variables from different NCDS surveys, even if they are asking the same question (e.g. what is your job now?).

The variable **serial** is used as the link variable to ensure the correct case is matched at one time-point to the same cohort member at a different point in their life.

As an example, suppose we want to merge the data from NCDS4 with NCDS5:

	serial	n622_4	n4region	resps4	n4113	n4118
1	010002M	2	2	1	729	1
2	010003P	1	1	1	740	1
3	010004R	1	2	1	717	1
4	010006V	1	2	1	608	1
5	010007X	2	10	1	209	1
6	010008Z	1	2	1	32	1
7	010009B	1	2	1	179	1
8	010010L	1	2	1	652	1
9	010013S	2	2	1	9	1
10	010014U	2	2	1	652	1
11	010015W	1	2	1	229	1
12	010017A	1	8	1	559	1
13	010019E	2	2	1	87	1
14	010020P	1	2	1	166	1
15	010021R	2	2	1	503	1
16	010022T	2	2	1	648	1
17	010023V	2	2	1	509	1

	serial	person	n622_5	n5region	n5gor	resps5
1	010001K	1	1	8	2	1
2	010003P	1	1	8	2	1
3	010004R	1	1	9	10	1
4	010006V	1	1	8	2	1
5	010007X	1	2	10	11	1
6	010010L	1	1	8	2	1
7	010011N	1	1	8	2	1
8	010012Q	1	1	8	2	1
9	010013S	1	2	8	2	1
10	010014U	1	2	8	2	1
11	010016Y	1	2	2	3	1
12	010017A	1	1	2	3	1
13	010019E	1	2	8	2	1
14	010022T	1	2	4	6	1
15	010023V	1	2	8	2	1
16	010026B	1	2	8	2	1
17	010027D	1	1	8	2	1

When downloading from the UK Data Archive, data files are normally made available in SPSS portable format (e.g. ncds4.por), so they can be read by other software. But for the sake of the examples in this document, it's assumed that each file has been converted to an SPSS system file (e.g. ncds4.sav).

Each cohort member is referred to as a *case*, and the data for each case occupy one line of the displayed matrix. So we can see 17 cases in each of the diagrams above.
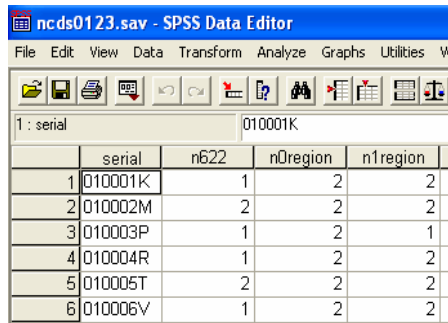
The variable names are here listed along the top line. For NCDS4 they are: serial, n622_4, n4region, resps4, etc.

For NCDS5 (cohort member data) they are: serial, person, n622_5, n5region, n5gor, etc.

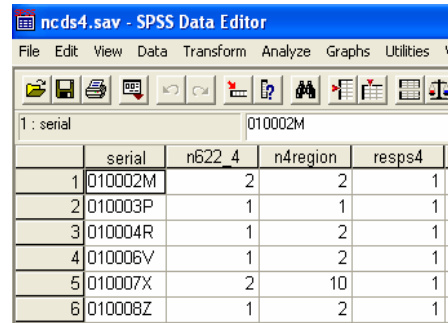Note that all the variable names are different, except for the link variable **serial**.

In merging the two files, we are concatenating the two sets of variables. This works perfectly for cases like 010003P, 010004R, 010006V, which are present at both time points.

But what about cases like 010002M, 010008Z and 010009B, which are present at NCDS4 but not NCDS5? Or cases like 010001K, 010011N and 010012Q which are present at NCDS5 but not at NCDS4?

The diagram below shows what happens in the merged file. We simply have missing values for all the variables corresponding to the sweep (NCDS4 or NCDS5) in which that cohort member wasn't present:

	serial	n622_4	n4region	resps4	n4113	n4118	person	n622_5	n5region	n5gor	resps5	resp5cmi	resp5wyt
1	010001K	.	.	.	.	.	1	1	8	2	1	1.00	1.00
2	010002M	2	2	1	729	1	.	.	.	.	.	.	.
3	010003P	1	1	1	740	1	1	1	8	2	1	1.00	1.00
4	010004R	1	2	1	717	1	1	1	9	10	1	1.00	1.00
5	010006V	1	2	1	608	1	1	1	8	2	1	1.00	1.00
6	010007X	2	10	1	209	1	1	2	10	11	1	1.00	1.00
7	010008Z	1	2	1	32	1	.	.	.	.	.	.	.
8	010009B	1	2	1	179	1	.	.	.	.	.	.	.
9	010010L	1	2	1	652	1	1	1	8	2	1	1.00	1.00
10	010011N	.	.	.	.	.	1	1	8	2	1	1.00	1.00
11	010012Q	.	.	.	.	.	1	1	8	2	1	1.00	1.00
12	010013S	2	2	1	9	1	1	2	8	2	1	1.00	1.00
13	010014U	2	2	1	652	1	1	2	8	2	1	1.00	1.00
14	010015W	1	2	1	229	1	.	.	.	.	.	.	.
15	010016Y	.	.	.	.	.	1	2	2	3	1	1.00	1.00
16	010017A	1	8	1	559	1	1	1	2	3	1	1.00	1.00
17	010019E	2	2	1	87	1	1	2	8	2	1	1.00	1.00
18	010020P	1	2	1	166	1	.	.	.	.	.	.	.

How do we go about merging files, using analysis software like SPSS?

To merge two data files in SPSS (e.g. the combined childhood dataset ncds0123.sav and the 23-year-old dataset ncds4.sav), two approaches are possible. If you are familiar with SPSS syntax, this is the quickest and most elegant way. If not, you can use the drop-down windows menus. In section (a) below we give detailed instructions for using syntax, and in section (b) we explain the menus approach.

	serial	n622	n0region	n1region
1	010001K	1	2	2
2	010002M	2	2	2
3	010003P	1	2	1
4	010004R	1	2	2
5	010005T	2	2	2
6	010006V	1	2	2

	serial	n622_4	n4region	resps4
1	010002M	2	2	1
2	010003P	1	1	1
3	010004R	1	2	1
4	010006V	1	2	1
5	010007X	2	10	1
6	010008Z	1	2	1

(a) Syntax approach

```
** open each cross-sectional file, sort by serial and save.
get file='c:\ncds4.sav'.
sort cases by serial.
save outfile='c:\ncds4.sav'.

get file='c:\ncds0123.sav'.
sort cases by serial.
save outfile='c:\ncds0123.sav'.

** append ncds4.sav to the 'active' file ncds0123.sav.
match files file=*
  /file='c:\ncds4.sav'
  /by serial.
save outfile='c:\ncds01234.sav'.
```

You can continue in this way to merge more files to the 'active file'. For instance, having first sorted ncds5cmi.sav by serial, you could proceed by executing the following syntax in place of the above *save outfile='c:\ncds01234.sav'* command:

```
match files file=*
  /file='c:\ncds5cmi.sav'
  /by serial.
save outfile='c:\ncds012345cmi.sav'.
```

(b) Drop-down menus approach

First we need to sort the two data files. It makes sense for this to be done in reverse order, so that we end up with ncds0123.sav as the currently 'active' file at the end of the sorting process.

So read in ncds4.sav (**File > Open > Data > Ncds4.sav**).
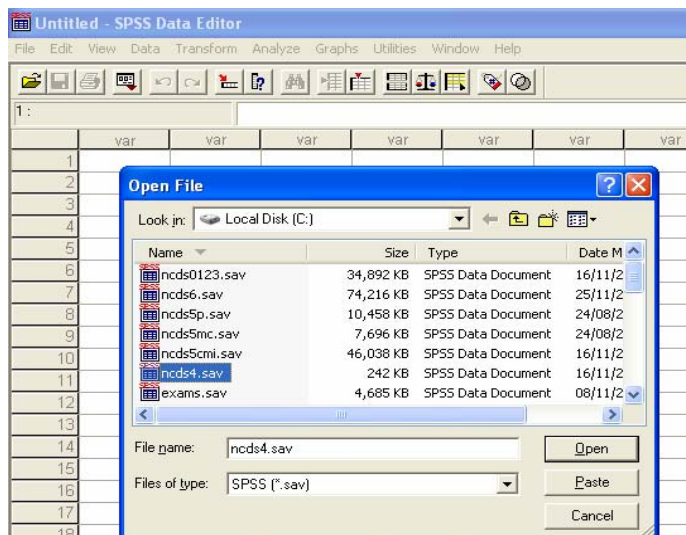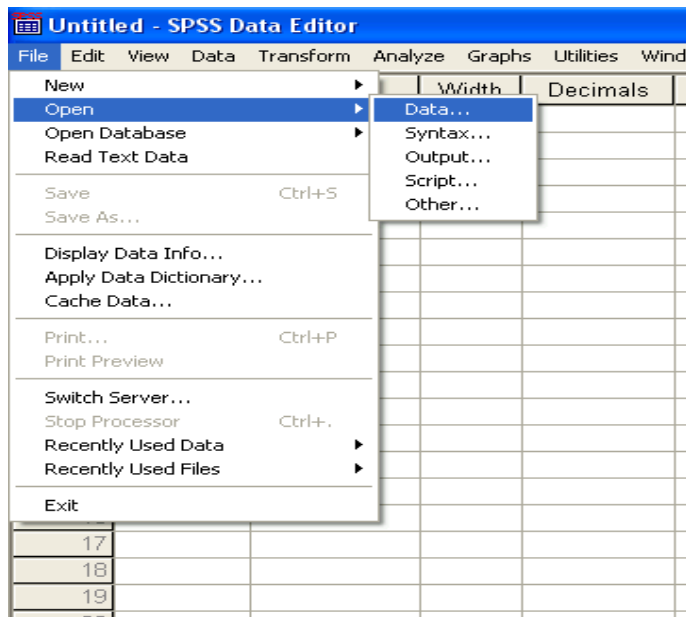
Then **Sort** the link variable (**Data > Sort Cases > by serial**)
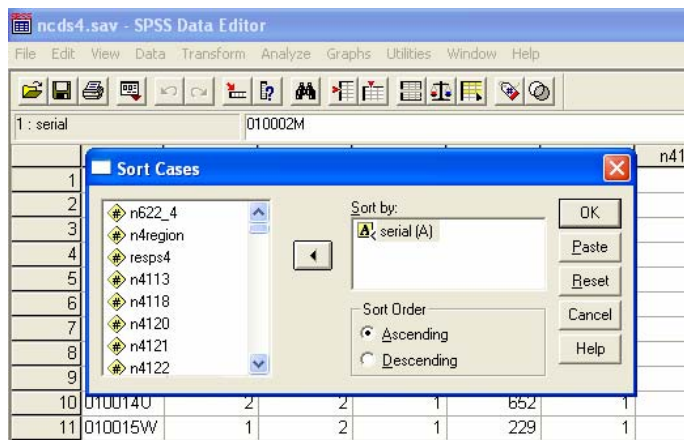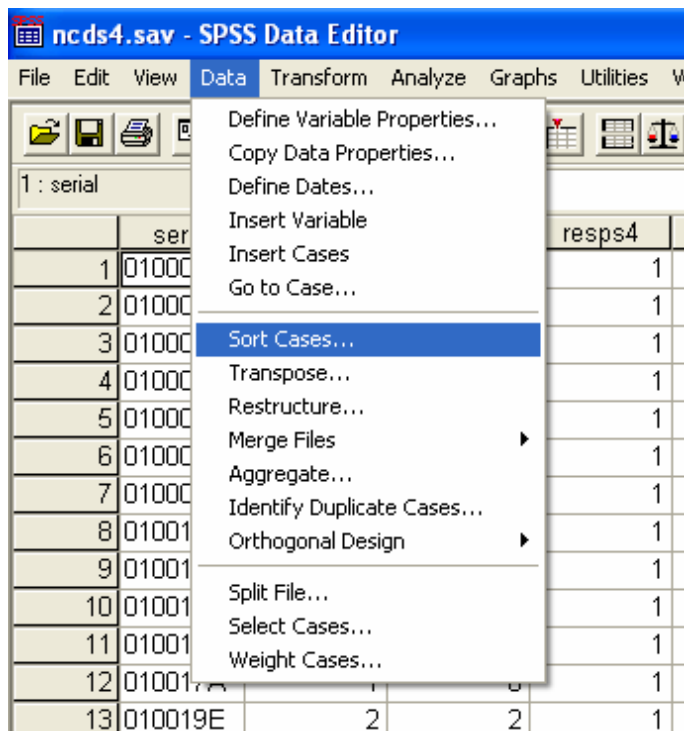
Now save the sorted file Ncds4.sav **(File > Save)**

Repeat the above procedure for the file Ncds0123.sav. At the end of this process, the sorted Ncds0123.sav will be the currently 'active' file.

Now we are ready to merge the files (**Data** > **Merge Files** > **Add Variables**)
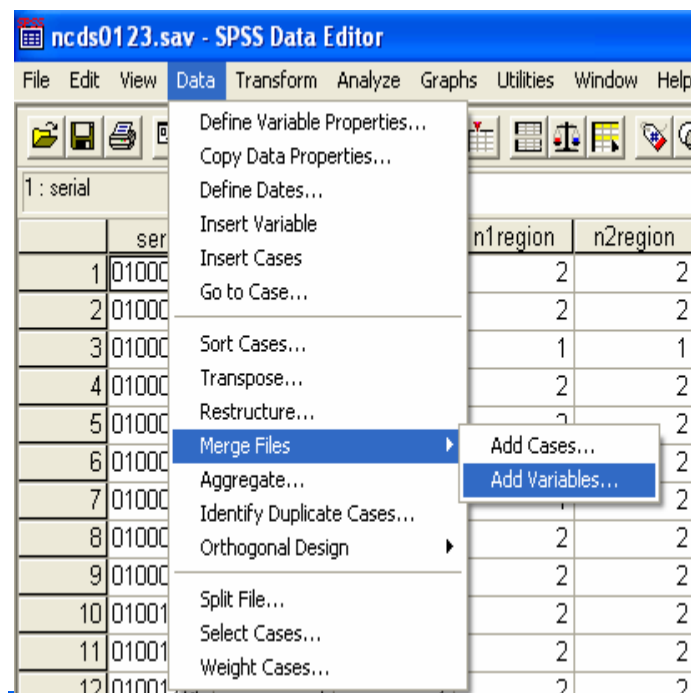
You will see a dialogue box entitled 'Add variables: read file'. Enter the name of the other file (in this case, **Ncds4.sav**)
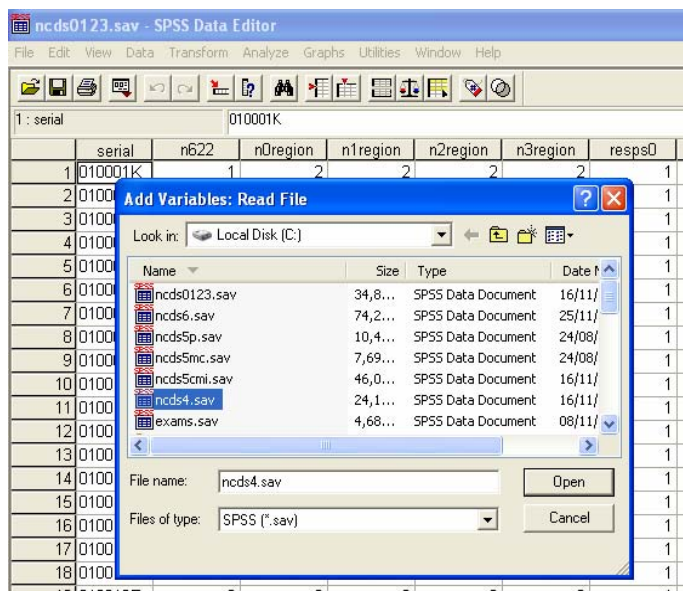
Another dialogue box appears, showing a list of all the variables in both files combined, with a separate list of 'excluded variables'. These are variables that appear in both files. Note that there is only one 'excluded' variable, since 'serial' is the only variable whose name is repeated across different datasets.
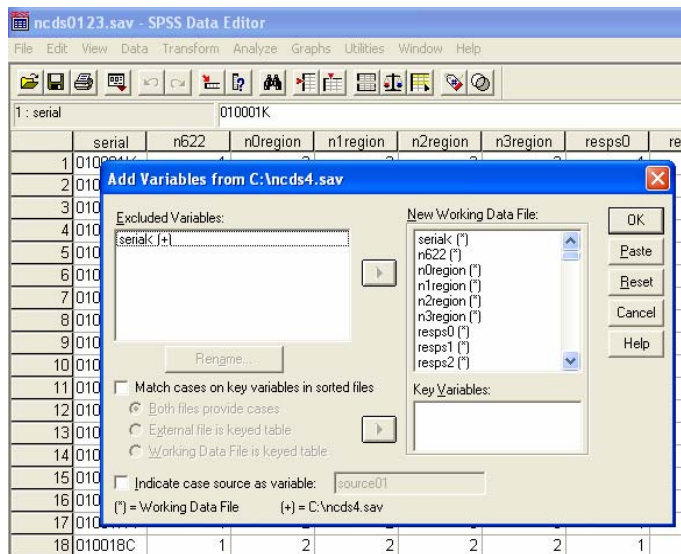
Check the box 'Match cases on key variables in sorted files', and underneath it make sure the (default) option 'both files provide cases' is checked. Then highlight the name of the link variable ('**serial**') in the 'excluded variables' box, and click the **right arrow** button to the left of the 'Key Variables' box at the bottom right. This will paste the variable name 'serial' into the 'Key Variables' box.

Then click **OK.** This will merge the two files ncds0123.sav and ncds4.sav.
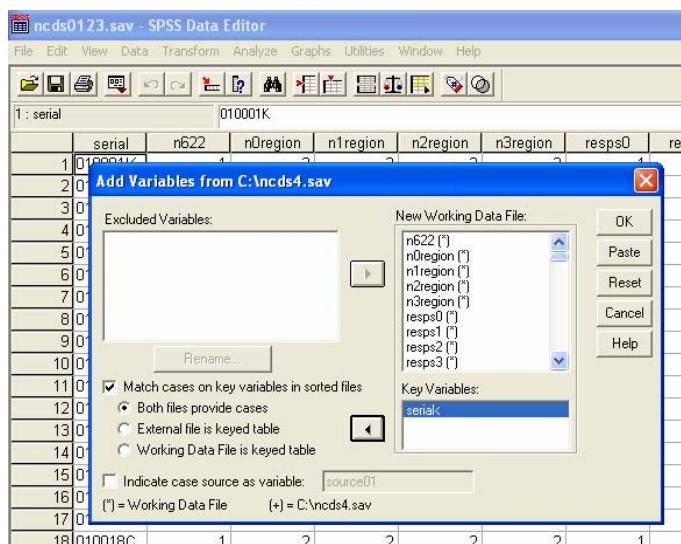
You can then save and name the combined dataset as, say NCDS01234.sav **(File >
Save As)**
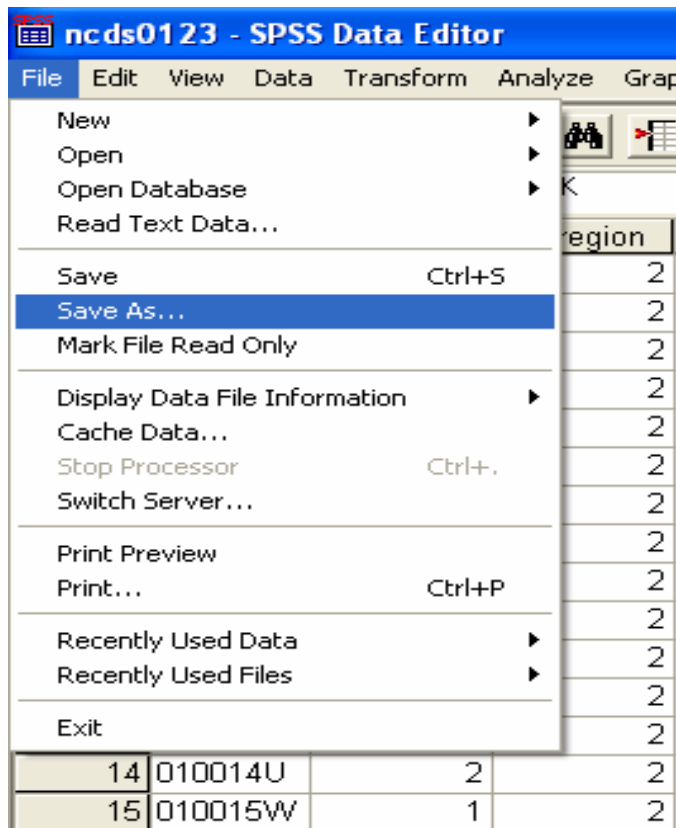
You can continue to merge further datasets (e.g. NCDS5cmi.sav, NCDS6.sav) with
this combined file in the same way as above.

Note that at each cross-sectional time point, there will be some NCDS members who
were not present at that survey. This will mean all the variables corresponding to that
survey will be system-missing for that person in the combined longitudinal dataset,
but variables for other time points will have valid values (see first section above:
"What are we doing…").

Merging Child Data for NCDS5

The NCDS5 survey was unique in that, for a 1 in 3 sample, the mother was asked a detailed questionnaire about her children (the 'mother' interview), then asked a separate series of questions about each child in turn (the 'your child' interview), then tests were administered to each child (the 'child' interview).

The contents of each of these questionnaires can be summarised as follows:

Mother Interview

Family life
Pregnancy and birth
Health history
Separations from mother
Experience of being "in care"
Pre-school experience
Schooling history
Experience of day care

Your Child Interview
(mother answering questions about each child in turn)

Motor and Social Development
Behaviour Problems Index
Temperament
Home Environment

Child Interview/Assessments
(conducted directly with each child, tailored to specific age ranges)

Peabody Picture Vocabulary Test
McCarthy Scale of Children's
Abilities: Verbal Memory Subscale
Peabody Individual Achievement Tests:
Math Subscale
Reading Recognition Subscale
Reading Comprehension Subscale
Weschler Intelligence Scale for Children: Digit Span Subscale
Perceived Competence Scale
Self-Perception Profile
Plus, an interviewer evaluation of:
Testing Conditions
Child Temperament
Home Environment
Child Height and Weight Measurement

The distribution of numbers of children in families interviewed is summarised below:

Number of children listed in mother interview, or undergoing assessments

Number of Children in Family	Frequency	Percent
1	1383	53.4
2	821	31.7
3	301	11.6
4	69	2.7
5	12	.5
6	1	.0
8	1	.0
Total Mothers	2588	100.0

There was also a fourth set of data where the interviewer filled in a series of observations about the home environment, and the mother's perceived relationship with the specific child: whether the mother talked to the child, responded to them, caressed them or slapped them, etc. (the 'home environment' variables).

The results of these four sets of data were stored in such a way that each child has its own separate record in the data file corresponding to that instrument. Unlike all the other NCDS data files, where the case is uniquely defined as the cohort member, these four sets of data have duplicate serial numbers. So for example, a cohort member with four children is liable to generate four records corresponding to the mother interview, four corresponding to the 'your child' interview, and four records for the 'child' assessments (it's slightly less clear-cut than this, as some children refused the interview or weren't at home, etc., despite being listed in the mother interview).

The particular child within the family is distinguishable by their 'person number' in the household grid (usually 3,4,5 etc., since the mother and father will normally be persons 1 and 2).

Users who have experience of NCDS data from the UK Data Archive in its older format may remember that these four datasets were previously available as four separate files (mother.sav, ychild.sav, child.sav and hoenv.sav). Great care was needed in merging these datafiles with each other and then with the rest of the NCDS longitudinal data. This was because one had to make sure each child of each cohort member was linked not only to their parent's record, but to their own record in the other child-based files, and not to one of their siblings.

We have now made this process much simpler by combining all the NCDS5 child-based files into one composite file, ncds5mc.sav. Like its constituent data files, it contains more than one record per cohort member, but the records for each child can be distinguished by the variable 'person'. Note that the unit of analysis in this file is the *child*, not the cohort member, so that a frequency count of any variable will yield a total of 4,287 (i.e. the number of children involved in the NCDS5 assessments) rather than 2,588, which is the total number of mothers. Note also that the mother is not

necessarily the cohort member. In the case of male cohort members, it would normally be their partner who was interviewed.

There are records which have missing values for variables corresponding to some element of the child-based questionnaires. This is because not all children listed by the mother in the 'mother' or 'ychild' interview were present for the 'child' assessments, or agreed to do them.

To merge this combined dataset, ncds5mc.sav, with any other NCDS data file (e.g. ncds4.sav) it is necessary to use a slightly different procedure from the normal 'merge files' process outlined in the earlier part of this document. We have to tell SPSS to conceptualise the ncds4.sav file as a 'look-up table', whose values for all other variables are spread identically across all cases with that particular NCDS serial number.

ncds4.sav   **ncds5mc.sav**

serial	person	n516220	n516221	n516223
010030S	3	1	18	4
010030S	4	2	9	3
010030S	5	2	14	6
010030S	6	2	29	5

serial	status	n4113	n4118	n4120
010030S	8	137	1	2
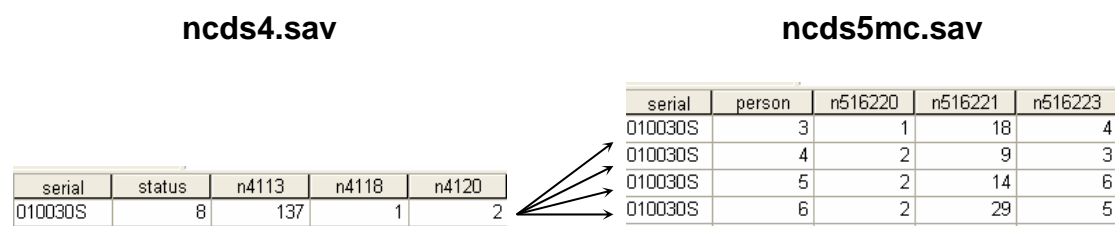
In the above diagram, the variables 'status', n4113, n4118 and n4120 are from the NCDS4 data file, and apply to the cohort member whose serial number is 010030S.

The variables 'person', n516220, n516221 and n516223 are from the ncds5mc file, and apply to each of the four children of that cohort member, who are distinguished by their person numbers. When we merge the two files, we need to ensure the values of the ncds4 variables (which are not child-specific) are attached in exactly the same way to each of the four children, as in the diagram below:

result of merging, using ncds4 as look-up table

serial	status	n4113	n4118	n4120	person	n516220	n516221	n516223
010030S	8	137	1	2	3	1	18	4
010030S	8	137	1	2	4	2	9	3
010030S	8	137	1	2	5	2	14	6
010030S	8	137	1	2	6	2	29	5

The values of the variables 'status' n4113, n4118 and n4120 are replicated four times in this case, but the values of 'person', n516330, n516221 and n516223 remain distinct for each child. In this merged file, the unit of analysis is the child, not the cohort member.

To do this merge in SPSS, as in the normal file merge procedure outlined earlier, one can use the 'syntax' approach or the 'drop-down menus' approach.

(a) Syntax approach

** open the ncds4 data file and file ncds5mc.sav, sort by serial and save.
get file='c:\ncds4.sav'.
sort cases by serial.
save outfile='c:\ncds4.sav'.

get file='c:\ncds5mc.sav'.
sort cases by serial.
save outfile='c:\ncds5mc.sav'.

** spread the values of ncds4 variables identically over all the children in each family.
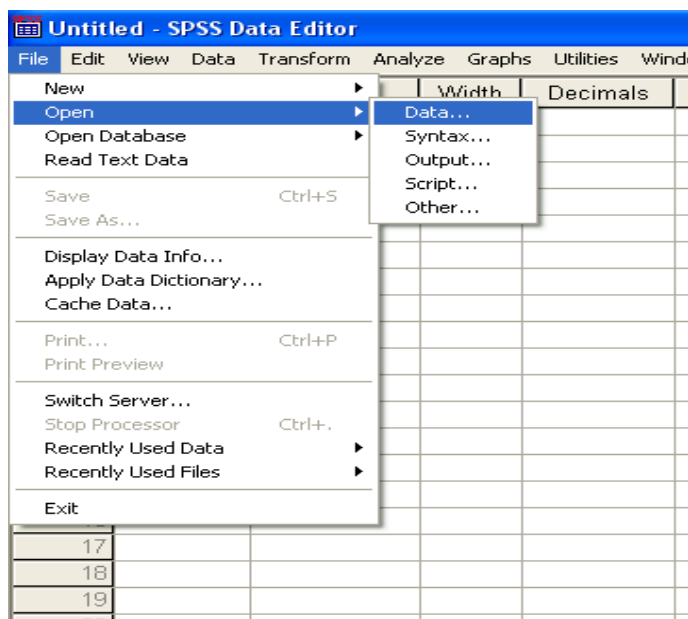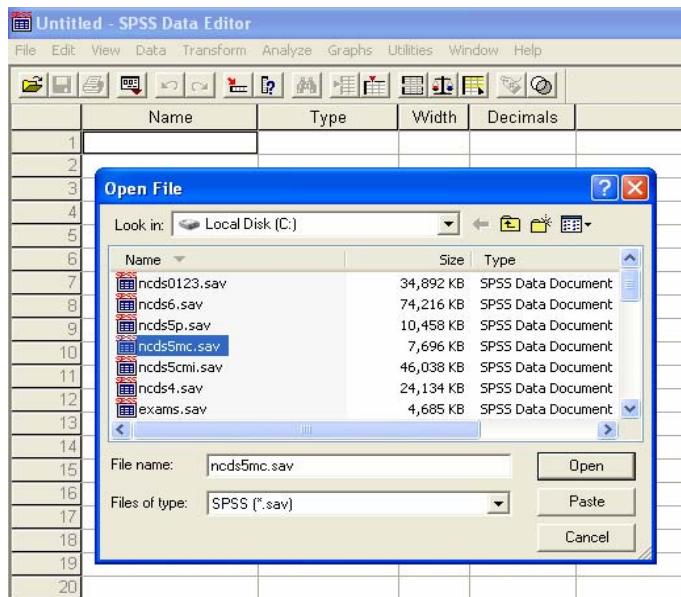match files file=*
  /table='c:\ncds4.sav'
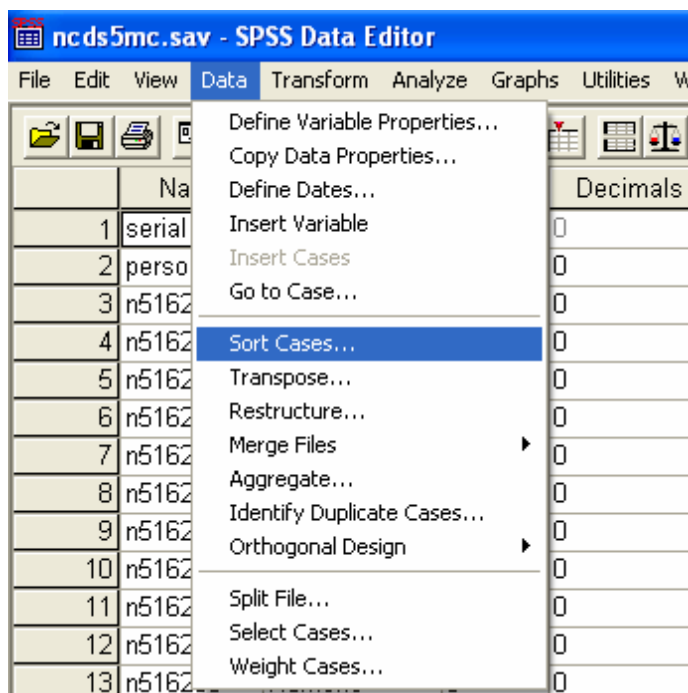 /by serial.
save outfile='c:\ncds4_ncds5mc.sav'.

(b) Drop-down menus approach

First read in the combined child-based dataset, ncds5mc.sav, as the active file (**File > Open > Data > ncds5mc.sav**).

Then **Sort** the link variable before merging files (**Data > Sort Cases > by serial**)
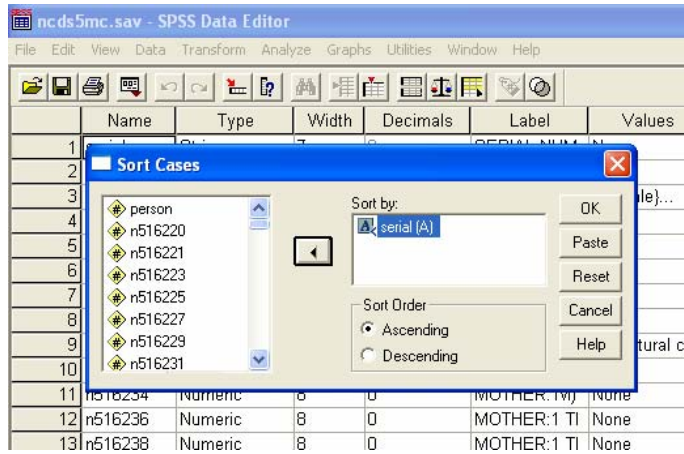
Click **Data** > **Merge Files** > **Add Variables**
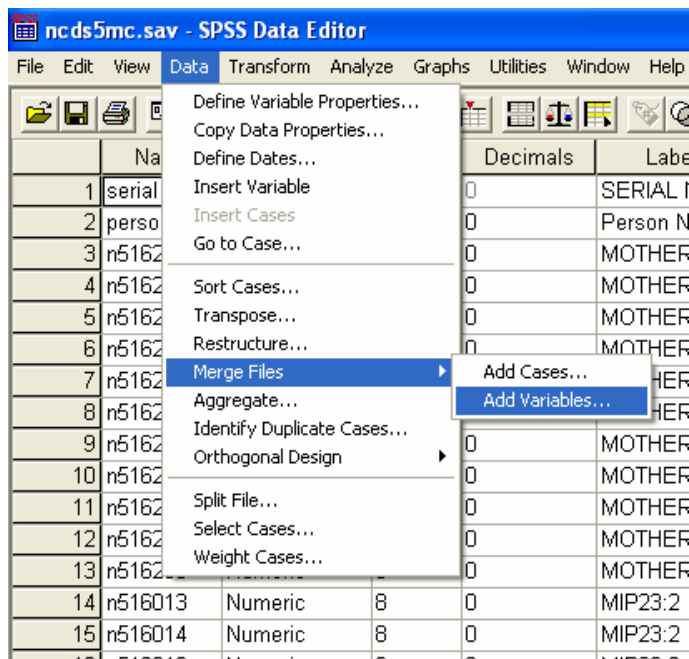
As before, you will see a dialogue box entitled 'Add variables: read file'. Enter the name of the other file (here **NCDS4.sav,** which you should also have previously 'sorted' as above).
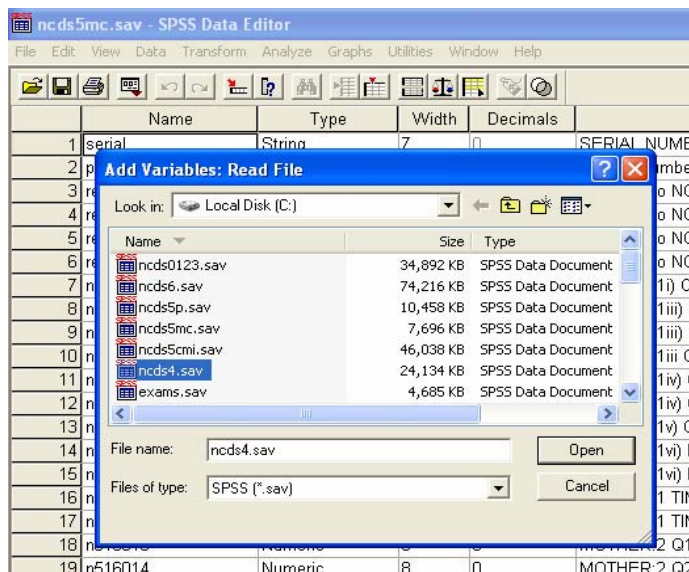
Another dialogue box appears as before, showing a list of all the variables in both files combined, with a separate list of excluded variables, which should just contain 'serial'.
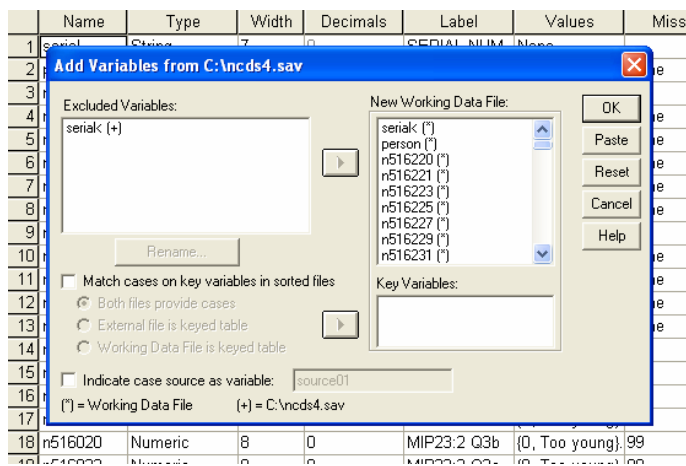
Check the box 'Match cases on key variables in sorted files', but this time make sure the checked option underneath it is 'external file is keyed table'. Then, as before, enter the name of the link variable ('**serial**') in the 'Key variables' box on the lower right by clicking the **right arrow** button to the left of that box. Then click **OK.**
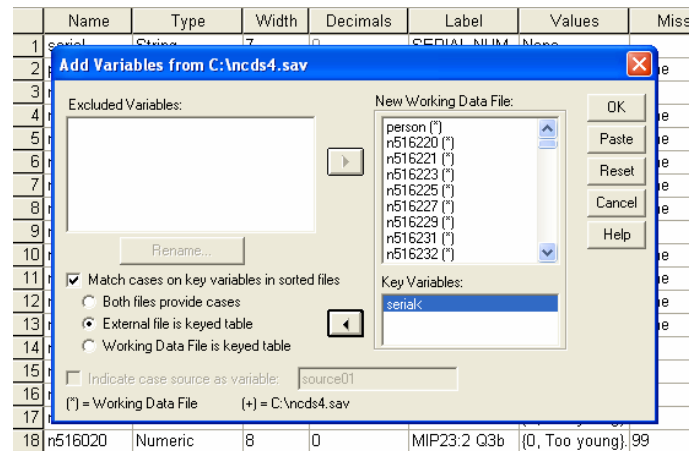
Note that the resulting merged file will still only contain as many cases as there were in 'ncds5mc.sav'. All the additional cases in ncds4.sav will not have been added, because this latter dataset was merely used as a 'look-up' to spread data across cases with the relevant serial numbers in ncds5mc.sav. But all the ncds4 variables will have been added for those cases (assuming the corresponding cohort member was present at NCDS4 – if not, the ncds4.sav variables will all have missing values).

Remember that the values of the added NCDS4 variables will be identical for all children of the same individual mother, because they have the same serial number (although different 'person' numbers).

You can then save and name this dataset (e.g. ncds4_ncds5mc.sav).

Merging Child Data in SPSS, Keeping Cohort Member as the Unit of Analysis

The above method succeeds in merging NCDS5 child-centred data with the cohort member's data from other sweeps, but does have the disadvantage of leaving us with a dataset where the unit of analysis is the child.

If you need the unit of analysis to be the cohort member, it is possible in SPSS to produce various summary statistics derived from the variables in the ncds5mc data file, by using the AGGREGATE procedure.

This effectively 'distils down' information on multiple children's records into one variable, which can then be associated with the cohort member.

ncds5mc.sav   **aggregated file**

The summary statistics available are:

SUM	Sum	MEAN	Mean
SD	Standard Deviation	MAX	Maximum
MIN	Minimum	PGT	% cases less than specified value
PLT	% cases less than specified value	PIN	% cases between spec. values
POUT	% cases not in specified range	FGT	Fraction greater than spec. value
FLT	Fraction less than value	FIN	Fraction between values
FOUT	Fraction not in range	N	Weighted number of cases
NU	Unweighted number of cases	NMISS	Weighted number of missing cases
NUMISS	Unweighted number of missing cases	FIRST	First nonmissing value
LAST	Last nonmissing value	MEDIAN	Median

In the example above we have the following three child-based variables in file ncds5mc:
- n516220 (sex of child, 1=male, 2=female)
- n516227 (age of child in years)
- n521223 (whether completed basal score 5/5 in PIAT maths (1=yes, 2=no)

Each variable can give rise to a number of summary variables: here, the variable n516220 in the child-based file gives rise to the three variables n516220, n516220_1 and n516220_2 in the aggregated file:

- n516220   Total number of children
- n516220_1   Percent male
- n516220_2   Percent female

Similarly, variable n516227 gives rise to two variables in the aggregated file:

- n516227   Age of eldest child
- n516227_1   Mean age of all children

To produce an aggregated file using SPSS, again we have two methods:

(a) Syntax approach

```
** open the ncds5mc data file.
get file='c:\ncds4.sav'.

** aggregate records, using 'serial' as the break variable, and asking for
summary statistics on three variables to be stored in the aggregated file
'c:\aggr.sav'.

aggregate
 /outfile='c:\aggr.sav'
 /break=serial
 /n516220 = nu(n516220)
 /n516220_1 = pin(n516220 1 1)
 /n516220_2 = pin(n516220 2 2)
 /n516227 = max(n516227)
 /n516227_1 = mean(n516227)
 /n521223 = pin(n521223 1 1).
```

(b) Drop-down menus approach

First read in the combined child-based dataset, ncds5mc.sav, as the active file (**File >
Open > Data > ncds5mc.sav**).
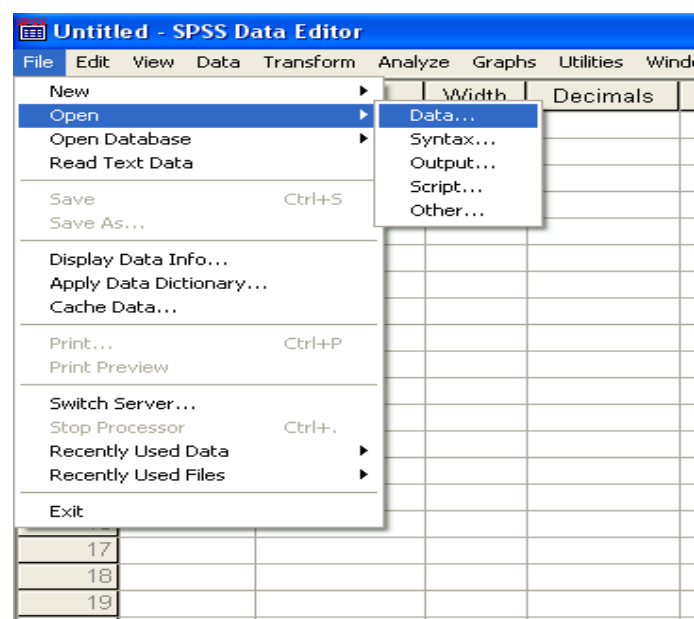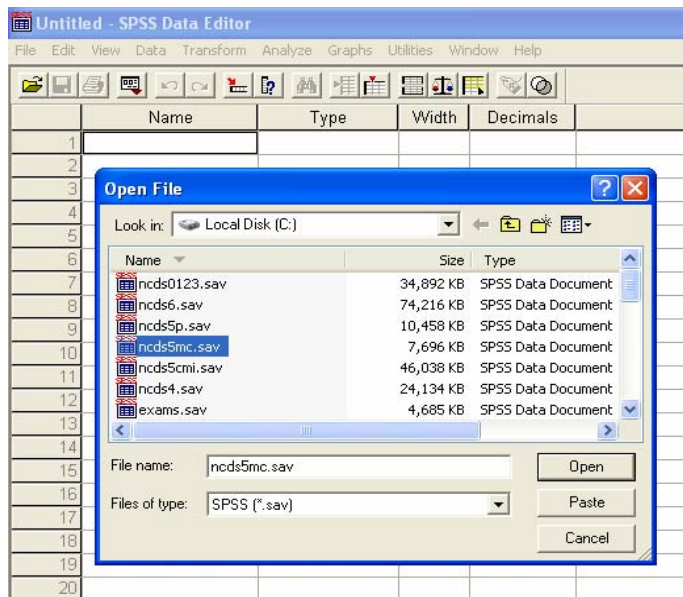
Click **Data** > **Aggregate**
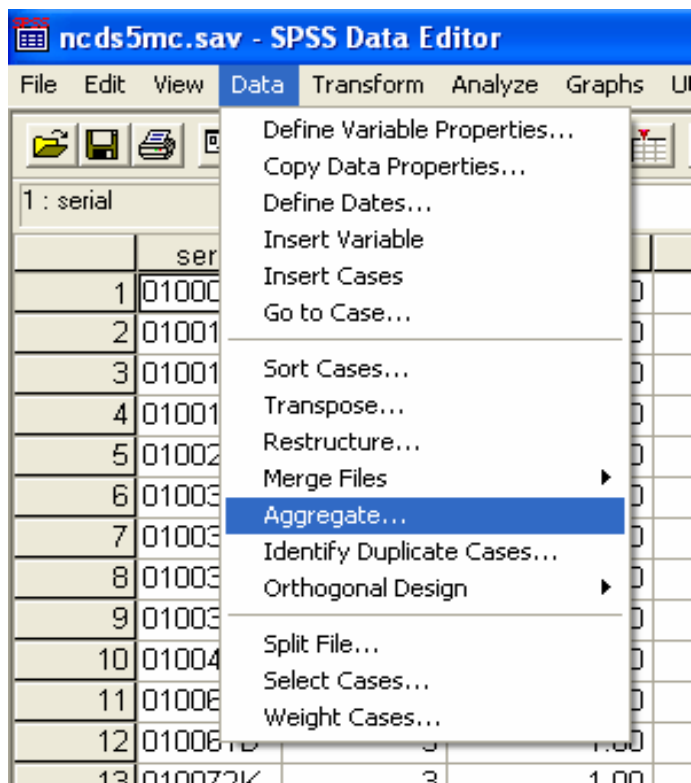
A dialogue box appears, inviting you to nominate the break variable (serial), and the aggregate variables (i.e. which ones you'd like distilled into summary statistics in the aggregated file):
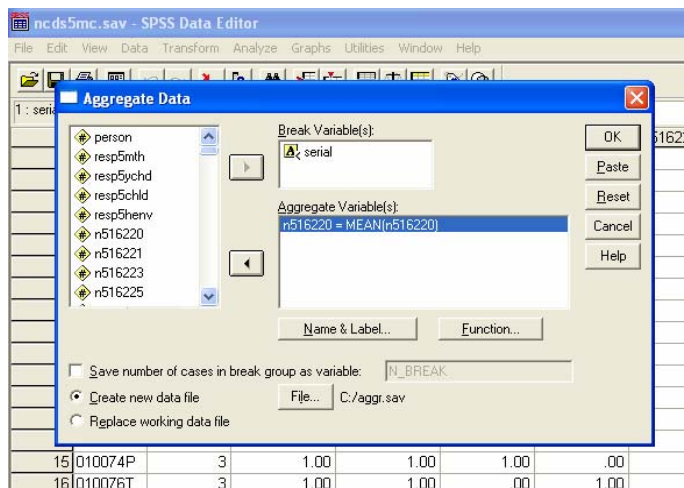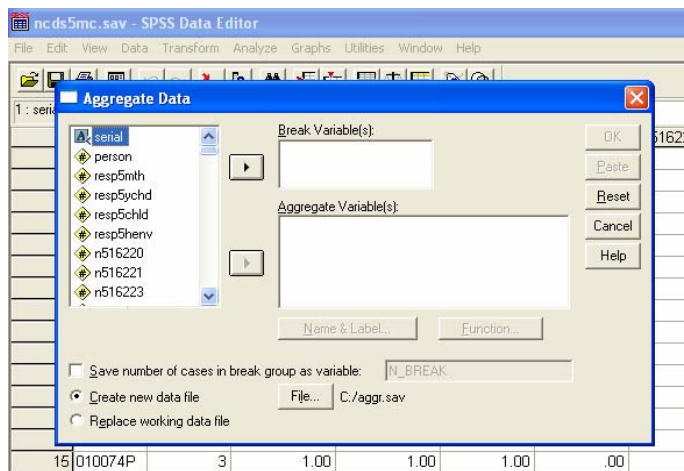
By default, SPSS assumes you want the MEAN of that variable, but if you want any of the other summary statistics, click the FUNCTION button. Another dialogue box appears, inviting you to choose other functions. In this case we choose 'Unweighted number of cases':
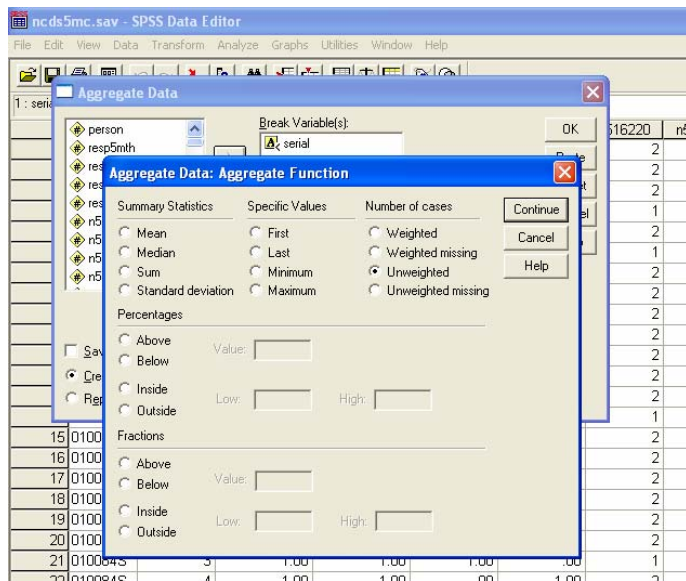
Continue to specify as many variables as you wish, and as many functions for the same variable as you wish (SPSS automatically attaches '_1'…'_2' to subsequent ones to distinguish them from the first statistic in the aggregated file):
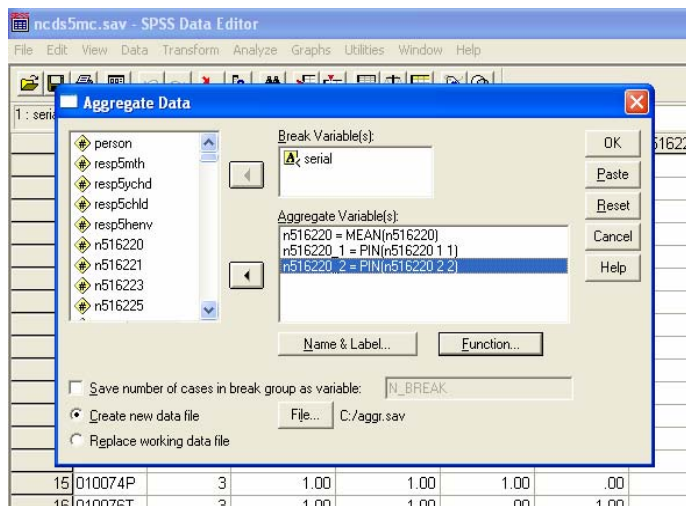
Note that, to request the percentage of cases having one specific value (e.g. '2' = girl for variable n516220), you need to request that it be in the range 'Low=2', 'High=2':
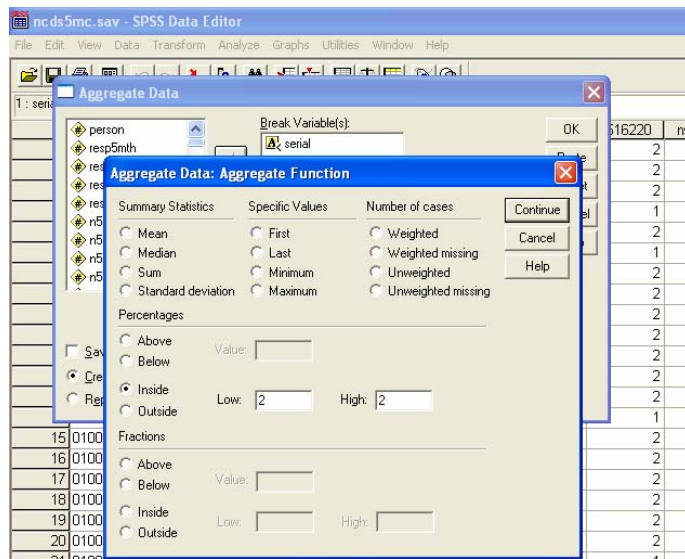
Finally, click the OK button in the 'Aggregate Data' dialogue box, and the aggregated file 'c:\aggr.sav' will be produced.

We demonstrated this process on the file ncds5mc.sav, but note that this whole process could equally have been done on a file which had already been longitudinally linked (e.g. to ncds4.sav) using the method in the previous section.

Working with smaller subsets of data

Linking datasets together longitudinally can produce an enormous file, which may be unwieldy to work with. If you know you are only going to be interested in a certain number of variables for your research, you can produce a subset very simply by running the following syntax:

save outfile='c:\ncdssubset.sav'/keep=serial var1 var2 var3 var4 var5 var6'.

If any of the variables you want to save to the subset happen to lie next to each other in the large dataset (i.e. in the order you normally see them listed on your screen), you can shorten this syntax by referring to the first and last ones in the list:

save outfile='c:\ncdssubset.sav'/keep=serial var1 to var6'.

It is possible to do this using drop-down menus, by clicking **File > Save As**
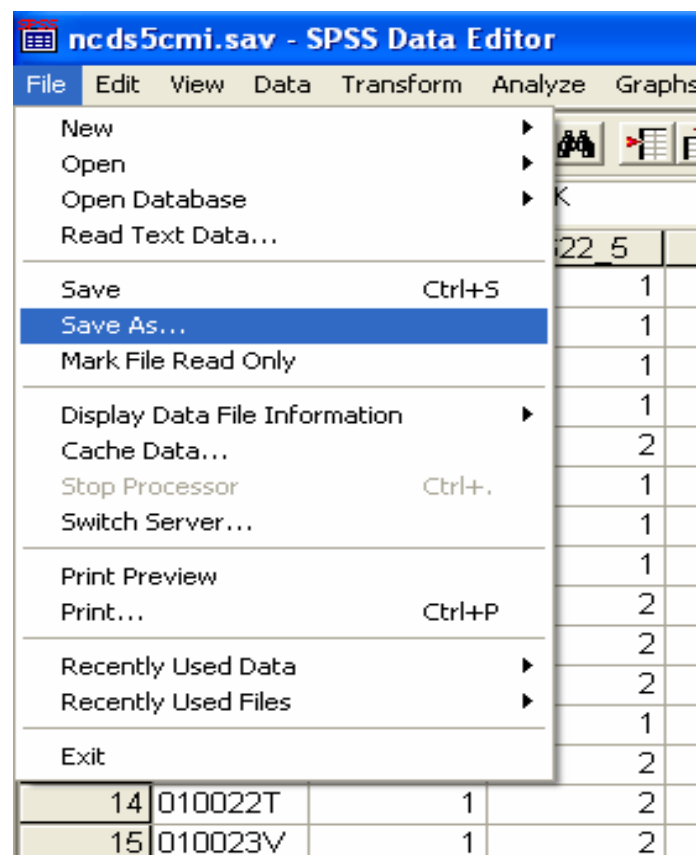
Then click the **Variables** button in the dialogue box you see:
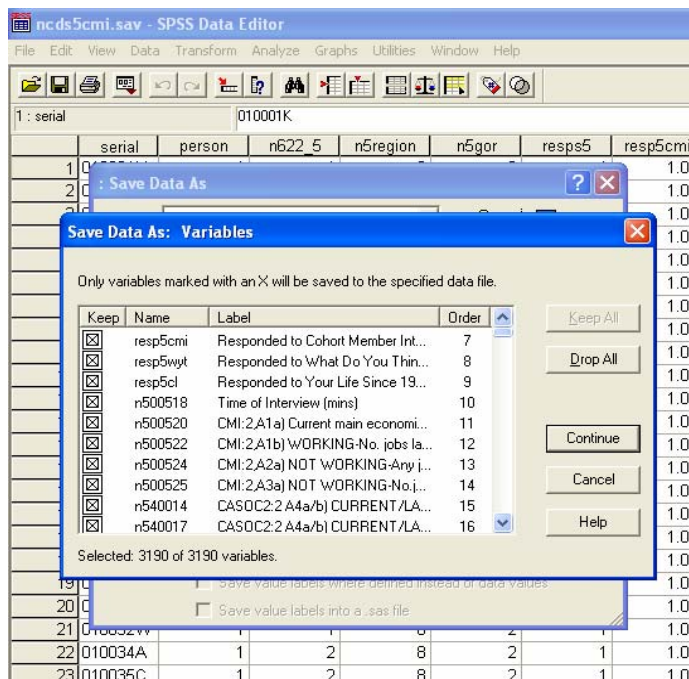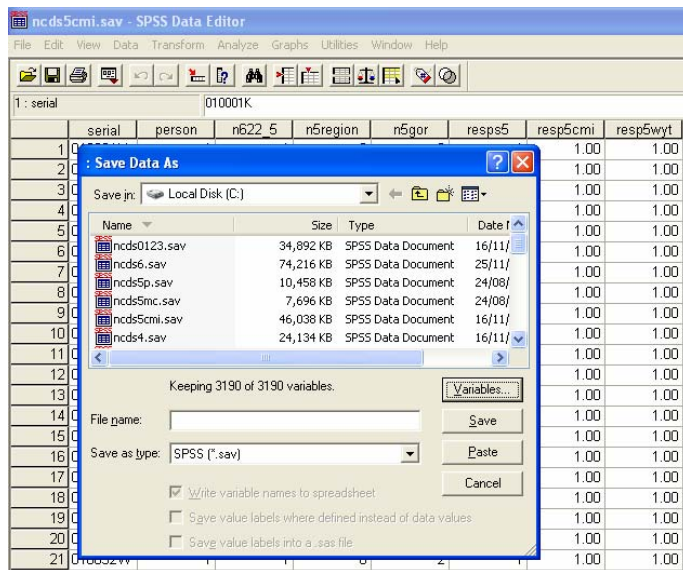
You can then check the boxes on the left next to each variable you want to keep (or 'uncheck' the ones you want to drop). Then click **Continue** and enter the name of the new data file: just the specified variables will be saved to this file.

This works where only a small number of variables need to be checked or unchecked, but can be tiresome for saving a large number which all appear consecutively. In this case, it's much better to use the syntax method explained above.