



Economic and Social Data Service

Analysing Change Over Time: A guide to ESDS microdata resources

ESDS Government

Author: Anthony Rafferty
Updated: Pierre Walthery
Version: 1.4
Date: September 2011



Contents

1. Introduction	2
1.1 Overview	2
1.2 ESDS Data for Analysing Change	3
1.3 Accessing the Microdata	7
2. Repeated Cross-sectional Data	9
2.1 Introduction	9
2.2 Limitations of Cross-sectional Research	9
2.3 Example Uses of Repeated Cross-sectional Data	10
2.4 Combining Repeated Cross-sectional Data	2
2.5 ESDS Supported Resources for Repeated Cross-sectional Analysis	6
2.6 Using the LFS as a Repeated Cross-section	18
2.7 The GHS Time Series Dataset (1972-2004)	20
2.8 Joining Repeated Cross-sectional Datasets using Stata and SPSS	21
3. ESDS Panel Data	25
3.1 Introduction	25
3.2 Concepts in Panel Data Methods and Research Examples	25
3.3 Non-response Bias and Response Error Bias	38
3.4 Seam Effects in Panel Data	39
3.5 ESDS Panel and Cohort Datasets	41
3.6 The GHS-Longitudinal/EU-SILC	50
3.7 Understanding Society	50
3.8 Non-UK Longitudinal Datasets	51
3.9 Organising Data for Panel Analysis using Stata	51
3.10 Worked example: Using <i>vector</i> and <i>loop</i> commands in SPSS to create a measure of attitude stability	56
Appendix A. Bibliography and Further Resources	61
Appendix C. Pre-prepared Longitudinal LFS Files – UK Data Archive Serial Numbers	68
Appendix D. Variables Contained in the GHS Time Series Dataset	71

1. Introduction

1.1 Overview

The Economic and Social Data Service (ESDS) is a national, UK data service providing access and support for an extensive range of key economic and social data, both quantitative and qualitative, spanning many disciplines and themes. ESDS provides an integrated service offering enhanced support for the secondary use of data across the research, learning, and teaching communities. These services are currently grouped as:

- ESDS Government
- ESDS Longitudinal
- ESDS International
- ESDS Qualidata

This introductory guide gives an overview of ESDS data resources for the analysis of change over time. Its content originates in part in a [one-day seminar](#) held in October, 2006 at the Cathie Marsh Centre for Census and Survey Research (CCSR) at the University of Manchester. The aim of the guide is to introduce the reader to some of the methodological, technical, and data management issues surrounding the analysis of change. This is undertaken using examples from major ESDS-supported datasets.

When conducting analyses across several years of datasets, it is critical to ensure that differences observed over time reflect real variation, as opposed to artefacts of changes in survey methodology or question design between years. An in-depth knowledge of your data and good data management skills are therefore both extremely valuable. With this in mind, the current guide focuses upon three key topics: 1) ESDS surveys for analysing change over time, 2) The construction of multiple year datasets, and 3) data manipulation techniques. Examples are given using both Stata and SPSS software. Those wishing to explore analytical methods for analysis over time in greater detail are referred to the further resources section in Appendix A.

The guide has three main sections. Section 1 gives a quick overview of the different surveys supported by ESDS for analysing change. Examples of repeated cross-sectional, panel, cohort, and duration data are given. Section 2 gives a basic overview of the advantages of using such data in comparison to one time point. Data management issues when combining repeated cross-sectional datasets across years to analyse change are discussed. A more technical treatment of creating

repeated cross-sectional datasets from the UK Labour Force Survey is also given in Appendix B. Section 3 gives a basic introduction to the research potential of 'true longitudinal' panel and cohort data, and duration data resources supported by ESDS, as well as techniques for the management of such data. The manner in which non-response bias can affect panel data is also discussed.

1.2 ESDS Data for Analysing Change

Data which can be used for analysis over time is commonly categorised into a number of types:

- Repeated cross-sectional data
- Panel data ("true longitudinal")
- (Birth) Cohort data (also categorised as a type of panel data)
- Duration data
- Retrospective questions

Repeated cross-sectional data

A large proportion of datasets hosted by ESDS are repeated cross-sectional in design, where a survey is administered to a new sample of interviewees at successive time points. For an annual survey, this means that respondents in one year will be different people to those in a prior year. Such data can either be analysed cross-sectionally, by looking at one survey year, or combined for analysis over time. Figure 1.1 outlines major ESDS Government repeated cross-sectional datasets.

Figure 1.1 Major ESDS Government Repeated Cross-sectional Surveys

Survey	Repeated cross-sectional	Panel Element?	ESDS Web Page Link
UK Labour Force Survey (LFS)	✓	1992 onwards	Click here
General Lifestyle Survey (GLF)*	✓	2005 onwards	Click here
Family Resources Survey (FRS)	✓	-	Click here
Living costs and Food Survey (LCF)	✓	-	Click here
Time Use Survey (TUS)	2000**	-	Click here
British Social Attitudes Survey (BSAS)	✓	1984-1986	Click here
ONS Opinions Survey	✓ (modules)	-	Click here
Annual Population Survey (APS)	✓	-	Click here
National Travel Survey (NTS)	✓	-	Click here
British Crime Survey (BCS)	✓	-	Click here
Health Survey for England	✓	-	Click here
Survey of English Housing (SEH)	✓	-	Click here

*Formerly General Household Survey ** 2005 in ONS Opinions Survey *** Formerly Expenditures and Food Survey

Although repeated cross-sectional data can be used to consider patterns of *aggregate change*, they cannot be used to track patterns of *individual/micro-level* change. By aggregate change, we refer to changes for population groups. If representative samples are present in consecutive years of a survey, we can compare changes in the behaviour or circumstances of different groups. For example, we can draw conclusions on how levels of smoking for men and women have changed over time. However, we cannot deduce how smoking behaviour for a given individual has changed over time, as different people form our sample in different years¹.

¹ Many repeated cross-sectional surveys include some retrospective questions which give information on past experiences or characteristics.

Panel Data

Panel data provide better opportunities to track individual level change than repeated cross-sectional data. In panel surveys, the same individuals are interviewed at multiple time points, referred to as *waves*. Respondents interviewed at wave one of a survey are interviewed again at wave two, and so forth. Reflecting both the cross-sectional (between individuals) and time-series elements, panel data are also referred to as 'cross-sectional time-series' data. You can use panel data to track individual changes in income, health, health, family composition etc. Panel data provides opportunities to capture the underlying dynamics of change. For example, whereas one might use repeated cross-sectional data to track changes in overall levels of income in the general population, panel data can be used to analyse changes in individual income over time, for example, to consider what factors influence the likelihood of entering or exiting poverty. Panel data allow a *dynamic analysis* to consider how past events or states influence current outcomes. They also help in controlling for the effects of unobserved characteristics (residual heterogeneity). These issues are considered in Section 3.

Figure 1.2 Major ESDS Supported Panel Datasets

Survey	Supported by	ESDS Web Page Link
The British Household Panel Survey (1991 onwards)	ESDS Longitudinal	Click here
Longitudinal Study of Young People in England	ESDS Longitudinal	Click here
English Longitudinal Study of Ageing (ELSA)	ESDS Longitudinal	Click here
Two-quarter and five-quarter Labour Force Surveys	ESDS Government	Click here
General Lifestyle Survey (GHS-L)*	ESDS Government	Click here
Families and Children Study	ESDS Longitudinal	Click here

* The GLF is gradually phased out, and the Family Resources Survey will adopt a longitudinal design

In terms of its disadvantages, the analysis of panel data can be more complex and often requires more sophisticated methods of analysis than cross-sectional data. Panel data also suffer from problems where non-response bias occurs over time, although non-response also affects cross-sectional data. For panel data, this may occur where individuals in one wave of an interview are lost or refuse interview at a subsequent wave, or as a result of respondents refusing to answer some items on a questionnaire. Panel conditioning presents another problem for longitudinal surveys where the repeated exposure to questions may bias responses. Despite these problems, panel data can often facilitate approaches to research that are not possible using cross-sectional data.

Duration data

Duration or spell data are collected where the measurement of interest is the duration between a given time point and the occurrence of an event. An example would be the time between starting a job and being promoted, or the duration of poverty. Each occurrence of a duration in a given *state* is referred to as a *spell*. For example, the time between becoming unemployed and finding a job would be a spell of unemployment, whereas the duration between becoming unwell and recovering is a spell of ill-health. Such data can be used for event history analysis. Duration data is available in ready-made formats in surveys such as the British Household Panel Survey (BHPS) Unified Work-Life History Dataset. The BHPS also contains information on marital, cohabitation, and fertility histories that can be used as duration data. You can also derive duration data by considering the duration people spend in a given state across several years of panel data. Such data is considered in more detail in Part 3.

Birth Cohort and Pseudo Cohort Studies

Birth Cohort data present a specific type of panel data in which people of the same or similar age are followed over time. Such ESDS supported studies are listed in figure 1.3.

Figure 1.3 Major ESDS Supported Cohort Studies

Survey	Supported by	ESDS Web Page Link
1958 National Child Development Study (NCDS)	ESDS Longitudinal	Click here
1970 British Cohort Study (BCS70)	ESDS Longitudinal	Click here
Youth Cohort Study (YCS)	ESDS Core	Click here
Millennium Cohort Study (MCS)	ESDS Longitudinal	Click here
Longitudinal Study of Young People in England (LSYPE)	ESDS Longitudinal	Click here

These datasets are discussed in Section 3.5.

In birth cohort data, groups of individuals with a shared birth point are followed through survey sweeps at different time points across their lives. The 1970 British Cohort Study for example, tracks a sample of individuals born in a single week in 1970. Since the parental survey administered after birth, six other major data collection exercises have been undertaken.

These surveys record information on topics such as health, education, and social and economic circumstances, and were carried out in 1975 (age 5), 1980 (age 10), 1986 (age 16), 1996 (age 26), 1999/2000 (age 30), 2004 (age 34) and 2008 (age 38). Further sub-samples have been studied at various ages.

The value of cohort studies is that they allow the examination of processes of change. By examining multiple cohorts over time, it is further possible to distinguish between age and cohort effects (Dale and Davies, 1994). By cohort effects, we refer to effects attributable to differences in the historical, social, economic, cultural, and technological contexts in which different generations have grown up and lived through. Thus, differences between age groups can reflect both age related effects such as life-course position and maturation, but also cohort differences. Period effects refer to where factors related to the historical position of the moment of observation influence findings. For example, in some time periods, unemployment rates are higher than others.

Pseudo cohort data offer another prospect for analysis over time and can be derived from repeated cross-sectional data. Whereas birth cohort studies track a group of people born at a specific date over time, pseudo cohort data represent cohorts on the basis of age groups from repeated cross-sectional data. Specific age groups in a given survey year can be considered to be represented by samples of older age groups in subsequent years. For example, respondents aged 20-24 years in a repeated cross-sectional survey conducted in 1980 are represented by those aged 30-34 years in the 1990 survey. This information can be used to track the average experiences of different age cohorts over time. Pseudo cohort studies are discussed in greater detail in Section 2.

1.3 Accessing the Microdata

To access ESDS supported datasets, all users must [register](#) with ESDS. If you are from a UK institution of higher or further education you will already have a username and password issued to you by your institution. Use this to log in using the [Login](#) link at the top of every web page. If you do not have a username and password issued by a UK HE/FE institution, you will need to [apply for a UK Data Archive username and password](#). If you do not have a personal username, please see [Login help](#) on the ESDS web site. Registered users can download/order the datasets direct from the ESDS web site (usually in SPSS, Stata or tab-delimited formats) via its [online catalogue](#) and the download/order section of the [Major Studies](#) web pages.

An increasing number of datasets are also available to registered users in the [Nesstar](#) system, which allows online exploration of data and basic exploratory analysis before choosing to download all, or a subset, of the data. Nesstar can save data into formats suitable for SPSS, Stata, SAS, Statistica, DIF (suitable for use in Excel), Dbase and NSDStat. Non-

registered users of Nesstar can still view descriptions of variables in datasets and basic frequency distributions, whereas access to more advanced functions requires registration.

All users requiring data for non-commercial purposes can download data free of charge. Where data is required for commercial purposes there is a per usage/project fee of £450 and an additional per study number fee of £50. For all CD orders there is a flat media fee of £7.50 per study number, handling fee of £2.50 and a flat rate postage and packing fee (£3 in the UK, £4 rest of EU, £5 rest of world). All packages are sent first class via Royal Mail. See [Charges](#) on the ESDS web site for more information.

2. Repeated Cross-sectional Data

2.1 Introduction

Repeated cross-sectional data refers to where a new random sample is collected at successive surveys. This means respondents at one interview are different to those at a prior and subsequent interview. This section begins with a brief overview of some of the advantages of using repeated cross-sectional microdata compared to the cross-sectional analysis of just one survey year. The remainder of the section focuses on relevant ESDS Government-supported data, and how to construct repeated cross-sectional datasets through combining annual cross-sectional survey files. Information on **major** methodological changes over time in ESDS Government datasets is given and a checklist is provided of useful things to consider in order to avoid making errors when combining data. The section concludes with examples of statistical procedures for creating repeated cross-sectional datasets using Stata and SPSS software.

2.2 Limitations of Cross-sectional Research

There are a number of limitations of cross-sectional data for a single time point:

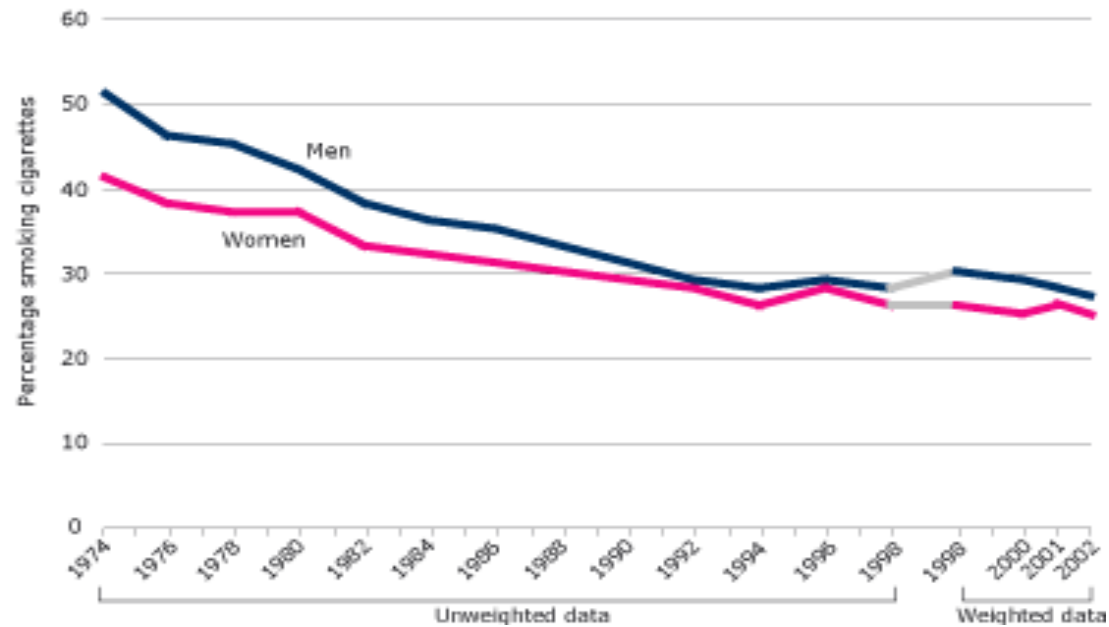
- A single cross-sectional data collection does not allow the analysis of change over time, whether at the *aggregate level* for populations or sub-groups (like repeated cross-sectional data can), or at the *micro-level* for examining individual change (like panel data can).
- Cross-sectional data provide only a snapshot at a given time point and so in some cases can lead to misleading inferences. Whether variables are influenced by systematic or sporadic fluctuations in their values in a given year can affect findings. Furthermore, many important variables are *time varying*. Cross-sectional data consequently provide limited insights into processes of social change.
- Cross-sectional data does not allow *age*, *cohort*, and *period effects* to be easily distinguished. It is therefore difficult to conclude whether the effects of age reflect differences between younger and older people in terms of maturation/life-course position, differences between older and younger cohorts in terms of their shared experiences and the historical contexts in which they have lived, or factors related to the time point the moment of observation occurred.

2.3 Example Uses of Repeated Cross-sectional Data

Change Over Time

Repeated cross-sectional data can be used to consider patterns of change at the *aggregate level*. For example, you could use information from the GHS across different years of the survey to consider changes in the number of people smoking over time (figure 2.1). From Figure 2.1, we can see that percentage of men and women who smoked generally declined between the 1970s and 1990.

Figure 2.1 Trends in Smoking Behaviour by Sex (1974-2002, GHS report, ONS²)



Although the above example uses microdata, in some cases you may be able to avoid the need for microdata by using published tables or aggregate time-series. An example of aggregate level data is the annual unemployment rate, where

² See <http://www.ccsr.ac.uk/esds/events/2006-10-30/slides/higgins.ppt>

there is one number for each year of the time-series. In many cases, some of the problems of comparability over time (considered below) will already have been addressed in such data. This may particularly be useful where changes in definitions have been made over time which require more sophisticated adjustments for comparability (such as for changes in definitions of unemployment). Some government labour market data are also seasonally adjusted for labour market fluctuations across the year.

Microdata however maintain a number of advantages over aggregate data. Firstly, they can allow you to create trends which are not readily available in pre-existing aggregate level data. They can be used, for example, to disaggregate for population sub-groups. When conducting more sophisticated analysis, the time-series analysis of repeated cross-sectional datasets can also help overcome problems of multi-collinearity that are common in aggregate level time series. Multi-collinearity occurs where there is a strong linear relationship between explanatory variables. Micklewright (1994) notes that many indicators such as income, social class, and education levels, can all move together at the same time at aggregate level, raising difficulties in establishing their independent effects. The added variation as a result of the greater number of data cases for each time point in repeated cross-sectional microdata can help overcome these problems.

Age, Cohort, and Period Effects

Repeated cross-sectional data can also help (but often not fully solve) the interpretation of age as a variable in your analysis. The effects of an age variable may reflect factors related to age, period effects, and cohort effects:

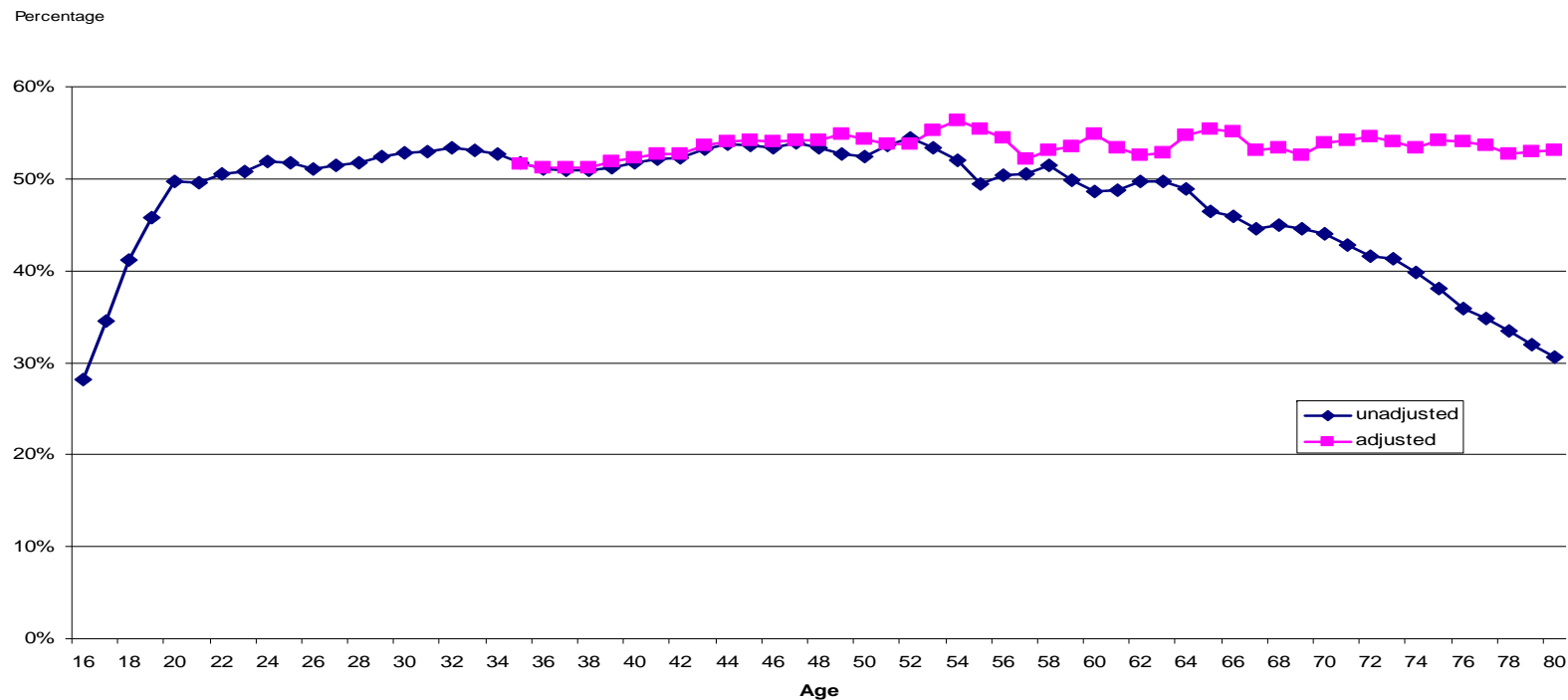
- Age: The time between birth and the observation period, Age may substitute for maturation/physical development; increases or decreases in intellectual capacities; personality development; life-stage etc.
- Period effects: Period refers to the moment of observation, although period effects may reflect the influences of longer term processes such as industrialisation; urbanisation; economic trends; gradual changes in educational standards etc (see Hagenaars, 1990:317).
- Cohort Effects: "A set of people born in the same period" (birth cohort) or a set of people who have experienced a particular basic event (such as marriage, labour market activity) in the same period (Ryder, 1965).

In terms of birth cohorts, people from different cohorts grow up in different cultural, technological, and socio-economic circumstances, or were at different ages when shared historical periods occurred. For example, consider two cohorts of

men, one born in 1938 and the other 1917. Given that one cohort would be young children and the other of age of army conscription, these two cohorts are likely to have had different experiences of the second world war (1939-45).

Figure 2.2 presents the percentage of men smoking in the 1970s by age. Those under 20 years old are less likely to smoke. However, this finding could have at least two explanations. Firstly, younger people may be less likely to smoke than older men. Alternatively, it could be that people in more recent cohorts smoke less, so these patterns reflect cohort differences. Using cross-sectional data at one time point, it is difficult to establish the extent to which this pattern reflects age or cohort effects.

Figure 2.2 Percentage of Males Smoking by Age in the 1970s



Through constructing pseudo-cohorts, repeated cross-sectional data can help to distinguish cohort and age effects by comparing age groups from pseudo cohorts for different years. To illustrate this, figure 2.3 gives the hypothetical example of three cohorts of people, born in 1950, 1960, and 1970, for the year in which each cohort is aged 25 years.

Figure 2.3 Examples of Different Birth Cohorts at Age 25 Years

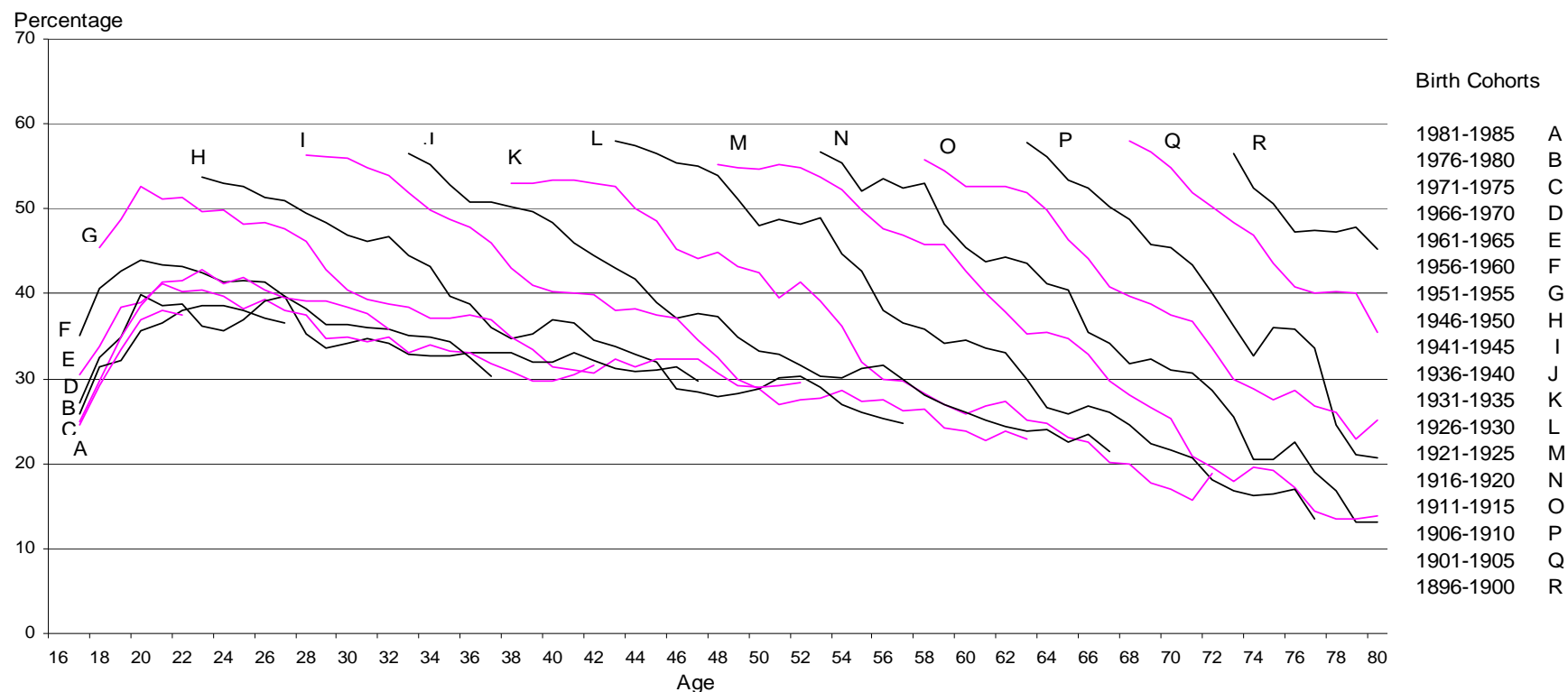
Cohort	Equivalent Age Comparison (at 25yrs)
1950	25 years of age in 1975 survey
1960	25 years of age in 1985 survey
1970	25 years of age in 1995 survey

By drawing comparisons between people aged 25 years in the three separate surveys stated, we can begin to distinguish the effects of age and cohort. Using the General Household Time Series Dataset (1972-2004), discussed below, figure 2.4 presents the percentage of male smokers at the time of interview by pseudo-cohort³. The different lines labelled A to R represent birth cohorts ranging from 1981-1985 (cohort A) to 1896-1900 (cohort R).

The X-axis indicates respondents' age whereas the y-axis the percentage smoking. By vertically comparing different cohorts at a specific age point, we can see the extent to which cohorts varied in their smoking behaviour at a given age. For example, at 30 years of age, more recent cohorts smoked much less than older cohorts (compare cohort A with cohort I). Indeed, at every age, men smoke less than the earlier cohort. Although such information is useful, we can also see some of the limitations of this approach. Firstly for more recent cohorts who are younger (to the left of the diagram), we do not have information about smoking behaviour during older age. Similarly, for the oldest cohorts (to the right of the diagram), we lack information about their smoking behaviour when they were younger.

³ Analysis by Melissa Coulthard (ONS), age adjusted. See the presentation at <http://www.ccsr.ac.uk/esds/events/2006-10-30/>

Fig. 2.4 Percentage of Male Smokers by Pseudo Cohort and Age (adjusted)



The distinction between period effects and cohort effects requires some understanding to interpret the data and is often not clear. For example, it could be argued that differences in smoking partly reflect differential exposure to period effects between different cohorts. Younger cohorts have grown up in a period of reduced public acceptability of smoking, increased knowledge and information on the health risks of smoking etc.

There are many other ways in which repeated cross-sectional data can be analysed. They can be used for regression models which have variables to indicate survey year to control for period differences between years. Repeated cross-sectional data may also be appealing for the study of sub-populations where sample sizes are small in individual cross-sectional datasets

(e.g. ethnic minority groups). Larger sample size can also be desirable to increase the precision of statistical estimates by reducing standard errors⁴. Further methods that can be used on repeated cross-sectional data include time-series, multi-way tables, and log-linear models. Appendix A contains information on further resources on these methods.

The above examples focus on patterns of aggregate-level change. Aggregate level change can be distinguished from *individual (micro-level)* which, in the context of household survey data, represents changes over time for specific individuals. Although repeated cross-sectional data can demonstrate aggregate level change for populations or subgroups, it cannot discern patterns of individual change. To undertake the latter, measurements are required for the same respondents at more than one time point. Many repeated cross-sectional surveys do contain some retrospective questions that can be used to consider individual change. However, 'true longitudinal' panel data, where measurements are taken for the same individuals at multiple survey waves can provide a more flexible alternative for such analysis (see section 3.0).

2.4 Combining Repeated Cross-sectional Data

This section considers issues surrounding the construction of repeated cross-sectional datasets. In many cases, cross-sectional data are stored as year specific files that can be combined to form repeated cross-sectional dataset. When combining datasets, it is important to make certain you have done everything possible to make your variables and data as comparable over time as possible. This is in order to ensure that differences between years reflect real variation, as opposed to artefacts of changes in survey methodology, question design, or variable coding. In some cases, where there are big methodological changes between years, the extent to which you can draw reliable comparisons over time will be limited. In other cases, some years of your datasets may not contain the variables you are interested in.

It is therefore important to research your datasets and read the accompanying documentation thoroughly before starting any dataset construction. This is undertaken in order to map out relevant methodological changes and assess their potential impact on your research. It is also useful to map out what variables you will need in your pooled dataset (for example, do you have a year indicator and an ID variable that is unique and not duplicated in different years?). Given the resultant large size of combining several years of survey data, you might also wish to select a subset of variables so that you reduce file sizes prior to combining files.

The following checklist provides a guide to some of the things to think about prior to combining datasets:

⁴ However it must be noted that in some cases if there is sizable between year variation in the true population value of what we are trying to measure, although pooling may reduce standard errors, it will not improve the precision of estimates.

1. Are you comparing like with like? It is important to check for discontinuities in method or variables between survey years. Once you have identified a subset of variables that you are interested in, things to look out for include:

- Changes in variable definitions. Are variables measured and categorised in the same way in different years? E.g. For a categorical variable, does the value 7 represent the same category in different years of the survey? Does the variable have the same number of categories or has it changed? If it has changed, how can the categories be harmonised? E.g. has the definition of household income changed (such as to reflect changes in the social security system)? Is the definition of unemployment the same in all years?
- Changes in the way in which derived variables are produced from the raw data. Path diagrams for the derivation of variables may be particularly useful and are often available in the survey documentation.
- Changes in variable names. In some cases the names of variables will change between years of a survey. It may therefore be necessary to harmonise the names of variables. Otherwise when you combine years of a dataset, the same variable with different names will be presented in two different columns as separate variables.
- Changes in question wording. This can be assessed by looking at user guides and questionnaires.
- Changes in filtering, or question applicability. This means checking to see if the applicable group for a question remains the same between survey years. You can check this from information on question routing, typically included with the variable details in the documentation, or by looking at the questionnaire.
- Changes in sampling strategy or weighting systems between years. In some cases the sampling strategy will have changed between years affecting comparison. Changes in population weighting procedure could also create artefactual jumps between the old and new weighting surveys where weights are based on estimates from different population censuses. In some surveys, weights are revised back retrospectively although this may not cover the whole survey. A [guide to weighting](#) is available from the ESDS Government web site.

In many cases, where the categorisation of variables or questions has changed, these problems can be overcome by harmonising variable categories across years. ESDS Government has started charting comparability over time for key variables in the Labour Force Survey (LFS) and General Household Survey (GHS). [Information on the following ESDS variables](#), consistent over time, is available from the web site:

- Country of birth (GHS)
- Date of Birth (GHS)
- Education (GHS)
- Nationality (LFS)
- Geography (for early LFS)
- Socio-economic classification (Nssec) (both GHS and LFS)
- Ethnicity (both GHS and LFS)
- Country of Birth (both GHS and LFS)

These resources include syntax for you to derive the above variables. Similar variables are available for the [Sample of Anonymised Records \(SARS\)](#). ONS have also recently created a repeated cross-sectional dataset for the General Household Survey from 1972 onwards. This is discussed in Section 2.5.

2. Do you have independent or repeated samples? In datasets that contain only repeated cross-sectional data your samples in consecutive years can be assumed to be independent⁵. However, some surveys contain both repeated cross-sectional and panel elements. For example, the Labour Force Survey contains a *rotating* or *refreshed panel*. This means that, although it contains a panel element, respondents are gradually replaced with a new sample. The survey can be used both as a repeated cross-sectional, or as a panel survey. If you accidentally select a repeated sample, the assumption that individual observations for an individual in repeated wave are independent may be unrealistic. You will further have duplicated cases for the same individuals at different years. Therefore you need to ensure that you do not duplicate cases.

3. Are your datasets collected from the same or different surveys? Most often, repeated cross-sectional datasets are created by combining data from the same survey at different time points. In some cases, you may wish to combine information from different surveys. It must be noted that surveys differ in their sampling strategies and sample sizes. These in turn affect the standard errors of estimates in different surveys. For example, when using methods which take into account sample clustering (such as Stata's svyset suite of commands), the clustering will be different in the two surveys which you are using.

4. Do your datasets contain a variable indicating survey year? If each of your pre-combined datasets do not have a variable indicating the survey year or time point and contain identically named variables, it may be difficult afterwards to

⁵ Given that sampling is random, there is a small probability that some people will be re-sampled. As long as each year provides a representative sample, this will not present any problems for your analysis.

ascertain which year observations come from in your combined dataset⁶. It is therefore useful to create a variable (e.g. 'year') which indicates the survey year of the cases in each of your datasets prior to joining them together.

5. Do different years of the survey have unique individual and household identifiers? In some cases the household, individual, or other identifiers within a survey will not be unique in each year (although this is rare). In such cases you may need to create unique identifiers prior to appending files. Check your documentation to find out how identifier variables are coded.

6. How big will your combined datasets be? When combining several years of a large survey, the number of variables and cases can often increase rapidly. The size of the datasets will influence the computer processing time needed. One way of reducing processing time is to select subsets of variables that you are interested in from each survey year prior to combining datasets. As long as there is, or you have created, unique survey year-specific identifiers for each case in each year of the survey you are interested in, you can always match further variables on to your datasets at a later date.

⁶ Although if you make this mistake, you may be able to ascertain this information from the unique identifier numbers.

2.5 ESDS Supported Resources for Repeated Cross-sectional Analysis

This section outlines the general features of ESDS-supported repeated cross-sectional datasets (Figure 2.5). Where present, major methodological discontinuities that could influence levels of comparability between survey years are considered. It should be noted that this list is not exhaustive and does not consider changes in variable definitions. Readers should therefore still consult the necessary survey documentation prior to attempting any analysis.

Figure 2.5 ESDS Government repeated cross-sectional surveys: Key changes over time

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
Labour Force Survey (LFS) Link to 2009 questionnaire Link to datasets	ILO measures; Training at work; Work history; Hours worked; Earnings (1992 onwards); Education; Health and Disability.	All individuals aged 16+ in the sampled household. UK. c. 60, 000 households per quarter	1973-1983 biennial (1973 data unavailable). Annually from 1984. Quarterly from 1992, with major change to sample design. Due to these changes, it is advisable to use only from 1992 onwards when measuring over time.	Since 1984 the LFS has been weighted (grossed) to produce population estimates and to compensate for non-response among sub-groups. Additionally, the earnings data is also grossed. These weights are regularly updated in order to reflect population changes based on census projections. ONS have published re-weighted QLFS estimates in 2004, 2007 and 2009 at the UK Data Archive. For more information, see the LFS Reweighting Project Between March 1992 and November 1994, interviewing in Northern Ireland was only conducted in the spring, with no quarterly element. From December 1994 (for the December 1994 - February 1995 quarter), data gathering for Northern Ireland moved to the full quarterly cycle to match the rest of the country. In accordance with EU regulations, the LFS moved from seasonal (spring, summer, autumn, winter) quarters to calendar quarters (January-March, April-June, July-September, October-December) in 2006. The last seasonal quarter dataset issued was the Quarterly Labour Force Survey, December 2005 -

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
				February 2006 (SN 5356) and the first calendar quarter dataset was the Quarterly Labour Force Survey, January - March, 2006 (SN 5369). Users should note that there is some overlap between these two datasets. See: Madouros, V. (2006) Impact of the LFS switch from seasonal to calendar quarters: an overview of the switch of the LFS to calendar quarters and the potential effects of this change on users , London: ONS.
Annual Population Survey ⁷ See LFS				The APS combines results from five different sources: the Labour Force Survey (LFS) waves 1 and 5; the English Local Labour Force Survey as well as the Welsh and Scottish Labour Force Surveys (not currently available through ESDS); and the Annual Population Survey Boost Sample (APS (B)) ⁸ .
General Lifestyle Survey (GLS) - Formerly General Household Survey Link to 2006 questionnaire Link to datasets	ILO measures; Hours worked; Earnings; Education; Health and disability; Household and family information ; Housing tenure; Consumer	All individuals aged 16+ in the sampled household. GB. Achieved 10,283 households in 2003-04. NI covered by Continuous Household Survey which	Annually from 1971 (except for breaks in 1997/98 when the survey was reviewed and 1999/2000 when it was redeveloped). The 1971 data is not downloadable from the UKDA and is only available in ASCII.	Since 2000, a dual weighting scheme has been introduced to the GHS. The dataset contains one weighting variable for two purposes (1) to compensate for non-response in the sample (2) to gross up to match known population distributions in terms of region, age-group and sex. The 2004-2005 weighting variable is called Weight04. See Appendix D of the 2002 GHS report or the 2004 GHS report for more information. The GHS has a rotating panel from 2005 onwards (see Section 3.5). A time-series, repeated cross-sectional micro dataset for 1973 to 1982 and for 1972-2004 is available with supporting documentation from the ESDS government web site. This is discussed below in Section 2.6.

⁷ Since April 2004, all archived Annual Population Survey datasets are also available under [Special Licence](#). The 'Special Licence' APS dataset contains approximately 550 variables and it contains more detail than would be available under a normal End User Licence (EUL) (e.g. unitary authority/local area districts and individual ages). Special Licence users need to read the [Guide to good practice: microdata handling and security](#) and agree to abide by its requirements.

⁸ The APS (B) ceased to exist at the end of 2005 so APS data from January 2006 onwards contains all the above data apart from APS (B) data.

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
	durables; Pensions.	is similar to and modelled on the GLS.		
Family Resources Survey				<p>Since 1992, the FRS has used one weighting variable for two purposes (1) to gross to population (2) to compensate for non-response. However, the 1994-1995 to 2001-2002 datasets were re-released due to the inclusion of a new (interim) grossing factor introduced to make adjustments to the FRS for low-income households in Scotland. These datasets contain two weighting variables: Gross1 is the original variable and Gross2 is the new variable. The 2002-03 dataset contains Gross2 only.</p> <p>A new grossing regime for the survey was issued with the 2003-04 FRS data. This new regime consists of both an enhanced set of control totals and incorporates data on a post-Census basis. Grossing for Northern Ireland was not affected by these changes. See The New FRS Grossing Regime report on the DWP web site.</p> <p>On the 2004-05 data GROSS2 (the previous, interim methodology, using pre-Census control totals) is not included. The 2004-05 data includes only GROSS3. GROSS3 can be applied to sample estimates so that analyses reflect the overall UK population. GROSS3, has thus been back cast over the FRS series from 1994-95. For expanded information on FRS grossing go to 2004-05 User Guide 1: Grossing factors within the FRS.</p>
British Crime Survey (BCS) Link to 2008-2009 questionnaire	Levels of crime Attitudes towards and fear of crime.	One randomly selected individual (aged 16+) in each sampled household.	1982, 1984, 1988, 1992, 1994, 1996, 1998. Annually from 2000. Significant methodological	The BCS has been weighted since 1982. The survey has a number of different weights which should be applied in different circumstances. There are three main reasons for weighting the BCS (1) to compensate for unequal selection probabilities (2) to compensate for differential response rates (3) to ensure that quarters are equally weighted for analyses that combine data

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
Link to datasets	ILO measures; Violence at work.	England and Wales 1984 and 1992 onwards. GB in 1982 and 1988. Achieved 37,931 cases (non-victim form); 15,446 cases (victim form) in 2003-2004.	changes from 2001 onwards – a larger sample, moved to continuous fieldwork with a different reference period to previous years, a ‘spliced’ sample design, new questions.	from more than one quarter. All weights include a component for unequal selection probabilities. However, weighting to compensate for differential response and to equally weight quarters are included in some weights but not in others. In 2001 the survey methodology changed considerably and calibrated weights were introduced (older datasets do not have calibrated weights). See the BCS 2004/05 report for more information on calibrated weights. For general information on weighting of the BCS see the 2003-04 Technical Guide Vol.1 , section 7.
Scottish Crime Survey (SCS) Link to 2000 questionnaire Link to datasets	Views on social issues, levels of crime, fear of crime, experience of Victimization, contact with the police	SCS: One randomly selected adult (aged 16+) in each sampled household. All children aged 12-15 in the sampled household. SCJS: 16,003 adults (16+) in 2008-09.		In 1982 and 1988 the Crime Survey in Scotland formed part of the British Crime Survey (BCS) - the Scottish part of the 1988 BCS was also known as the Scottish Areas Crime Survey. In 1993 the first independent Scottish Crime Survey was run and repeated in 1996, 2000 and 2003. The 2000 survey had an ethnic boost. Achieved 5,041 interviews with people aged 16 and over in 2003. From 2004 the SCS was re-launched as the larger Scottish Crime and Victimization Survey (SCVS). In 2008, the SCVS changed its name to the Scottish Crime and Justice Survey (SCJS) with a larger sample size and changes to the sample design (see SCJS 2008-09 documentation). In 2008-09, this survey contained 16,003 cases, adults aged 16 and over.
British Social Attitudes (BSA) Survey	Attitudes towards many topics	One randomly selected individual (aged 18+) in	Annually from 1983 (except in 1988 and 1992).	The BSA survey has been weighted since 1983. The 2004 survey has a sample design weight (Wtfactor) used to compensate for unequal selection probabilities (because only one person per household is interviewed). The BSAS 2008 User Guide explains

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
Link to 2008 questionnaire Link to datasets	including economic issues and policies, education, sex and gender issues.	each sampled household. GB. Achieved 3,199 <i>individuals</i> in 2004.		this in more detail. See the Survey Question Bank web site for more details of question modules.
Scottish Social Attitudes (SSA) Survey Link to 2007 questionnaire Link to datasets	Attitudes towards many topics including health, transport, religion, party identification.	One randomly selected individual (aged 18+) in each sampled household. Scotland. Achieved 1,508 respondents aged 18 and over in 2007.	Annually from 1999.	
Young People's Social Attitudes Survey Link to 2003 questionnaire Link to datasets	Attitudes towards many topics including local area, politics, family life Economic activity (not full	All young people aged 12-19 living in households of British Social Attitudes Survey respondents GB. Achieved 663 interviews in 2003.	1994, 1998 and 2003.	

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
	ILO)			
Living Costs and Food Survey (LCF) - formerly the Expenditure and Food Survey Link to datasets	Expenditure: regular household bills, food expenditure, large items of expenditure, ownership of consumer durables; Income of individuals. ILO measures; Earnings; Hours worked.	Each individual aged 16 or over in the household keeps diary records of daily expenditure for two weeks. Simplified diaries are kept by children aged between 7-15. Achieved 6,432 households in Great Britain; 616 households in Northern Ireland in 2003-2004.	Annually from 2001 (replaced the FES & NFS). Renamed Living Costs and Food Survey (LCF) in 2008.	The design of the EFS is based on the FES, although the use of new processing software (SPSS) by the data creators has resulted in a dataset which differs from the previous FES structure. The most significant change in terms of reporting expenditure, however, is the introduction of the European Standard Classification of Individual Consumption by Purpose, or COICOP, in place of the codes used in the FES and NFS, which were unique to the two surveys. An additional level of hierarchy has been developed for the EFS to improve the mapping to the previous FES and NFS codes. Whilst the NFS and FES series are now finished, users should note that previous data from both series are still available.
Family Expenditure Survey (FES) Link to 2000 questionnaire Link to datasets	Expenditure; Income. ILO measures; Earnings; Hours worked.	Exactly the same as the Expenditure and Food Survey (EFS) above. NI incorporated	Annually from 1957-2000. Datasets unavailable for 1957-1960 and 1964-1967. Replaced by EFS in 2001.	

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
		in FES since 1968. Achieved 6,000 households in 2000/1.		
English Housing Survey (EHS) formerly the Survey of English Housing (SHE) Link to 2007-2008 questionnaire Link to datasets	Tenure; housing costs; history and moving intentions; Attitude questions revised/rotated annually. ILO measures; Earnings.	Household reference person in the sampled household. Additional interviews with private renters or their partners. England: Achieved 19,640 households in 2002-2003.	Annually from 1993. In April 2008, the SHE merged with the English House Condition Survey (EHCS) to form the English Housing Survey (EHS)	
National Travel Survey Link to 2008 questionnaire Link to datasets	Vehicle ownership; Other travel behaviour, such as use of public transport and time spent travelling.	All individuals in sampled household including children. Both parent and child answer questions for those aged under 11. GB.	1965 (dataset unavailable), 1972, 1975, 1978, 1985. Annually from 1998.	The UK 2000 National Travel Survey data was first launched in April 2002. The data has now been reweighted to use the 2001 Census results and improved editing procedures have been adopted. Detailed tables and full technical report have been made available for the first time in October 2003.

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
	ILO measures; Earnings; Travel to work.	Achieved households 7,437 in 2002; 8,258 in 2003 and 8,122 in 2004.		
Family Resources Survey (FRS) Link to 2008-2009 questionnaire Link to datasets	ILO measures; Earnings	All individuals aged 18+ in the sampled household (although some questions about benefits etc are asked about those aged 16-18). GB. Achieved 25,085 households in 2008/09. NI was included in for the first time in 2002-03	Annually from 1992 but data only available from UKDA from 1993 onwards.	
Health Survey for England (HSE) Link to questionnaires	Health and disability; Health Behaviour. Household Reference	In 2003 up to a maximum of three households per address were selected. Up to 2	Annually from 1991. Sample size greatly increased from 1993 onwards. Since 1995 children aged 2-15 have	The modules to date have been: 1993 cardiovascular disease; 1994 cardiovascular disease; 1995 asthma, accidents and disability; 1996 asthma, accidents and special measures of general health (Euroquo, SF36); 1997 children and young people; 1998 cardiovascular disease; 1999 ethnic minority groups; 2000 older people and social exclusion; 2001 respiratory disease and atopic conditions, disability and non-fatal accidents;

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
Link to datasets	Person only: ILO measures; Income (not specifically earnings from employment).	children aged 0-15 were interviewed in each household, as well as up to 10 adults aged 16 and over. Information was obtained directly from persons aged 13 and over. Information about children under 13 was obtained from a parent with the child present. England. Achieved 18,553 individuals in 2003. NI covered by NI Health and Wellbeing Survey	been interviewed (aged 0-15 since 2001). IN 1997 and 2002, the sample was boosted for children and young people (aged 0-24). In 1999 and 2004 the sample was boosted to include a representation of ethnic minority groups. In 2000, elderly residents of care homes were included.	2002 children and young people; 2003 cardiovascular disease; 2004 ethnic minority groups. For further information go to the ESDS Government HSE web pages , the Department of Health HSE web pages , National Centre for Social Research web pages . Weighting variables are year specific owing to the variable sample design and the survey topic. For example, in 2000 weights are added for different probabilities of selection in care homes - see the 2000 User Guide . In 2002, no weights need to be applied if only using the adult sample. If using the boost sample (on its own or together with the adult sample) a sample design weight which accounts for unequal probabilities of selection needs to be applied (tablewt). In 2003 non-response weights were introduced for both adults and children. There are four sets of non-response weights in total: a household level weight (hhld_wt) and three sets of individual level weights, the interview weight (int_wt), the nurse weight (nurse_wt) and the blood weight (blood_wt). The appropriate weight variable should be used for analysis done using data from the relevant sections. There is an extra weight (child_wt) to compensate for limiting the number of children (aged 0-15) interviewed in a household to two. The variables int_wt and nurse_wt for children aged 0-15 include both the child selection weights and non- response weights. See p2 of the 2003 user guide for more information.
Scottish Health Survey	General health, prescribed	The 1995 survey covered those	1995, 1998, 2003 and 2008	There have been three surveys to date: 1995, 1998 and 2003-4. In 1995, 7932 individuals completed an interview and 6958 were visited by a nurse. In 1998 the survey covered those aged 2-74.

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
Link to 2003 questionnaire Link to datasets	medicines, smoking and eating habits. Physical measurements such as height and weight. Blood and saliva samples. Economic activity (not full ILO), SOC, SEG, SIC,	aged 16-64. One adult was randomly selected for an individual interview. In 2003 a total of 8,148 adults and 3,324 children (including 391 aged 0-1) were interviewed. Of these, 5,444 adults were visited by a nurse and 2,224 children (including 254 aged 0-1).		<p>One adult and up to two children were selected. Parents completed the interview for those aged under 13, with the child present. Only children aged 8 and over completed the self-completion. In 2003-4, all adults (no upper age limit) and up to two children are included. In 1998, 9047 adults and 3892 children completed an individual interview, of these 7455 adults and 3211 children were visited by a nurse.</p> <p>For further information about the SHS go to the ESDS Government SHS web pages, National Centre for Social Research web site, Scottish Centre for Social Research or see the ESDS Government Guide to data sources for Scotland (the guide gives details for further resources, such as reports and publications using the SHS).</p> <p>The 1995 survey focused on cardiovascular disease. The 1998 survey has a wider range of topics, including asthma and accidents. The 2003-4 survey also focuses on cardio-vascular disease. Household information, demographic information, education, parental history. The 2008 survey focuses on drug abuse, alcohol and smoking; nutrition; physical fitness and exercise; specific diseases and medical conditions.</p>
Welsh Health Survey 1998 Link to 1998 questionnaire And the new survey from 2003/04 onwards	General health specific illnesses, use of and satisfaction with health service, self-perceived	New WHS (from 2003/04) Adults aged 16 and over and children 0-15 in the same private households. 2008: 13,313	Two surveys have been carried out: 1995 and 1998. The new WHS has been conducted in 2003/04, 2005/06, 2007 and 2008. Note that the results from the	<p>For further information about the WHS go to the ESDS Government WHS web pages, National Centre for Social Research web site and the Welsh Assembly Government web site.</p> <p>In 1998, around 30,000 individuals completed a postal questionnaire for adults aged 18 and over.</p> <p>The 2003-5 survey involves a household interview and a self-completion questionnaire (for all members of the household,</p>

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
Link to 2008 questionnaire Link to datasets	health, lifestyle. Demographic information, carers.	adults and 2,653 children	new WHS are NOT comparable with those from the earlier surveys because of differences in questionnaire and survey methodology. See the documentation (2008).	including children) which is collected by the interviewer. Wales. The achieved sample size for 2008 is around 13,313 adults and 2,653 children.
National Food Survey (NFS) Link to 2000 questionnaire Link to datasets	ILO Measures; Earnings Food consumption and expenditure	Household member who does most of the food shopping. UK. Achieved 6,000 households in 2000/1.	Annually from 1940 but data is only available from UKDA from 1974 onwards. Replaced by EFS in 2001	1940-1949: survey of the 'urban working class'. 1950: widened to cover the population of GB as a whole. 1994: extended to cover 'eating out'. 1996: household food part of the survey extended to cover Northern Ireland.
Time Use survey Link to 2000 questionnaire Link to dataset	Work/leisure balance; Gender differences in childcare;	All individuals aged 8+ in the sampled household.. UK. Achieved 6,500 households in 2000/1.	2000. Plans for a small 'pre-coded' time use module on the Omnibus survey in 2005 and another full survey in 2010.	
Omnibus survey (now ONS Opinions Survey)	Various – see Omnibus homepage .	One eligible person aged 16+ in the sampled	Carried out in 2/3 months each quarter since 1990.	Since April 2004 the Omnibus Surveys have been running in the field for 12 months. Each month's questionnaire consists of two elements: core questions, covering demographic information, are asked each month together with non-core questions that

Survey	Topics	Sample	Measurement over time	Other Key Changes Over Time
Link to April 2004 questionnaire Link to datasets		household. GB. Achieves around 1,800 individuals per month.	From January 2008 the <i>ONS Omnibus Survey</i> changed its name to the <i>ONS Opinions Survey</i> and became part of the <i>Integrated Household Survey (IHS)</i> .	<p>vary from month to month. In April 2005, a new weighting system was introduced on the Omnibus which supplies weights to correct for non-response bias. Applying these weights will gross up the data by age and sex and by region to the population control totals used on the Labour Force Survey. As well as accounting for the unequal probability of selection, these weights will correct for certain types of non-response bias and should improve precision for most variables. Weights will be supplied at person level in each survey month and if required will be available at a household level and on a quarterly or annual basis.</p> <p>As a result of becoming part of the <i>Integrated Household Survey (HIS)</i>, certain classificatory variables were altered to harmonise with the rest of the surveys that form the HIS (see detailed breakdown of the changes contained within the documentation for 2008 studies onwards). However, In January 2010, the OPN component was dropped from the IHS due to only one individual per household being interviewed, while the IHS requires questions to be asked of all household members. This process significantly increased the length of the OPN interview and, therefore, OPN reverted back to interviewing one household member, but still contains questions harmonised to the IHS</p>

2.6 Using the LFS as a Repeated Cross-section

The UK-Labour Force Survey contains both panel and repeated-cross-sectional elements. The structure of this dataset is consequently more complicated than some other ESDS Government supported datasets. Since 1992 in England, Scotland and Wales, and 1994 in Northern Ireland, the Labour Force Survey has been conducted quarterly to obtain seasonal labour market estimates. A longitudinal element was introduced in 1992, with respondents being interviewed for five consecutive quarters with 20% of the sample being replaced each quarter. This design was implemented to obtain stable employment estimates on a quarterly basis. The datasets therefore represent a 'rotating panel' with a fifth of the panel being refreshed at each quarter.

Fig 2.6 UK Quarterly Labour Force Survey Sample design

	Spring	Summer	Autumn	Winter	Spring +1
W1	12k	12k	12k	12k	12k
W2	12k	12k	12k	12k	12k
W3	12k	12k	12k	12k	12k
W4	12k	12k	12k	12k	12k
W5	12k	12k	12k	12k	12k

Figure 2.6 outlines the sample structure of the quarterly LFS. The purple boxes (the dark diagonal boxes for those reading in black and white) represent one group of people, followed through successive quarterly waves (W) of the survey. In the present example, the sample group who are highlighted in wave 1 in the first quarter are in wave 2 at the second quarter and so forth until they exit the survey after being interviewed in the fifth quarter. In each quarter, a group (of 20 per cent of the sample) pass beyond their fifth wave and so exit the sample. After their fifth interview, respondents are replaced or 'refreshed,' by a new sample of respondents who are in the first wave of their interview. The diagram represents the intended sample rather than the achieved sample. As a result of sample attrition (discussed in Section 3.3) respondents may be lost from the sample between waves, reducing sample sizes in later waves. Unlike some panel surveys, the LFS does not follow respondents when they change addresses. The LFS samples addresses so that the new residents at a given sample address become part of the survey sample, but do not represent part of the original panel.

The LFS can be used in a number of ways:

- as a cross-section to study one specific quarter
- files can be combined for repeated cross-sectional analysis across survey years
- quarters can be pooled to create repeated cross-sections to increase sample sizes
- the quarterly element can be used as panel data.

In terms of creating repeated cross-sections for the LFS, it is important to make sure that the samples you are combining represent separate, unique individuals. Appendix B describes a number of ways to achieve this goal. Section 3.2 discusses the use of the panel element of the quarterly LFS.

2.7 The GHS Time Series Dataset (1972-2004)

See www.esds.ac.uk/findingData/snDescription.asp?sn=5664

ONS has developed a combined repeated cross-sectional dataset for the General Household Survey (GHS: Time Series Dataset). This was constructed through combining each annual round of the GHS into one dataset from 1972 to 2004. The GHS Time Series Dataset contains over 40 variables with information on demographics, households, education, employment, and health for a combined sample of over 800,000 individuals. The dataset also contains birth cohort variables that can be used for pseudo-cohort analysis. The procedures used to create this dataset are as follows:

- Prior to creating the dataset, ONS undertook a consultation to assess what variables should be included.
- The identified variables were examined to consider whether they had changed in each year of the GHS.
- Syntax was written to merge the annual year files and to create key identifier variables such as survey year and individual and household identifiers. Variables were recoded, and renamed where possible to reflect 2003 coding and labelling.
- Syntax was written for each variable to re-categorise and rename them in a manner consistent over time.
- Derived variables including birth cohort and age were next constructed.
- Quality assurance was undertaken at key stages of the dataset creation. On completion of the final dataset, the dataset was matched back to the original GHS data files to ensure that the variables had been re-categorised correctly and overall numbers were the same. Initial data exploration was also undertaken for each variable to check for potential outliers and inconsistencies.

A full list of variables contained in the GHS Time-series dataset is given in the [User Guide](#).

2.8 Joining Repeated Cross-sectional Datasets using Stata and SPSS

The majority of statistical software packages contain commands for combining data files. In this section, basic examples of how to combine datasets using STATA and SPSS version 14 are provided. A more technical example of creating pooled data using the LFS is given in Appendix B.

Appending Datasets using Stata

Stata offers a number of different commands for combining datasets⁹. In Stata, the distinction is drawn between *appending* and *merging*. With the **append** command, the cases from the appended dataset are added below the last case of the dataset in memory. Using the **merge** command, added variables can be matched on by an individual, household, or other identifier, adjacent to already existing cases. The merged variables can be found in the data viewer to the right of the last variables of the original dataset. Given that in repeated cross-sectional data, cases from different survey years represent unique individuals, it is the append command that is used to join datasets from different years. The merge command can still be useful for adding on extra variables from the same survey years to your data. The merge command however is more fully utilised in the construction of panel datasets (see Section 3.7).

- **For pooling different years of cross-sectional data, we use the *append* command:**

```
use datafile1.dta
```

```
append datafile1 using datafile2
```

```
save mgland2.dta
```

When either merging or appending, the dataset in memory is referred to as the *master* dataset, whereas the one being called to memory is the *using* dataset. Thus in the above command syntax, the file not presently in memory (in this case, datafile2) is called upon by the *using* part of the command. The master data file (datafile1) does not have to be stated within the syntax as it is assumed to be the one currently held in memory. You can append further datasets without having saved your data by repeating the same process, although you may wish to save intermediate files to check for errors.

⁹ An ESDS Government Introductory Guide to Stata can be found at www.esds.ac.uk/government/resources/analysis/

There are a number of things to consider when checking your appended datasets:

- Where a variable is contained in one dataset that is not present in the other, its value for the latter dataset will be coded as missing.
- If a variable has changed its name between years, it will be presented as two separate variables. You may therefore wish to consider harmonising some (**or all**) of your variables in each year to make them consistent over time prior to appending your datasets. Alternatively, you can change the names of variables to reflect which year of data the information comes from, so as to maintain original information. The *rename* command in Stata can be used to achieve this. You can then create cross-year derived variables to give information for all respondents using the **generate** and **replace** commands.
- If the coding of categorical variables has changed between years (e.g. 7 in one year represent something different in another), the variable may look correct but it will contain errors. Again, you need to harmonise your coding prior to merging, or rename variables to reflect which year of the survey they came from.
- After appending, you can check to see that the id number uniquely identifies individuals by typing: `isid <idvar>`. The *duplicates* suite of commands in Stata also provides a way to more closely examine duplicated cases. Type `help duplicates` into Stata to find out more.
- When joining together large data files the amount of memory used can rapidly increase. The *compress* command can be used to reduce the amount of memory occupied by your data: `compress`

Creating a unique identifier in Stata

If household or individual id numbers are not unique in each year of a survey, you can derive a survey year specific individual identifier for yourself. A simple way to do this is to first establish the format and length of personal identifiers within the given survey then create a new variable which adds the current identifier variables values to a number which prefixes the survey year to the front of this identifier.

E.g. If the individual id variable for each year of a dataset takes on a four digit value (between 0001 and 9999) we can create a year specific value by creating a six digit number which has the last two digits of the relevant year followed by the original id number. So for the year 1991, we add 910000 to the id number. In Stata we can create a unique identifier (here called 'yrid'):

```
generate yrid = 910000+id
```

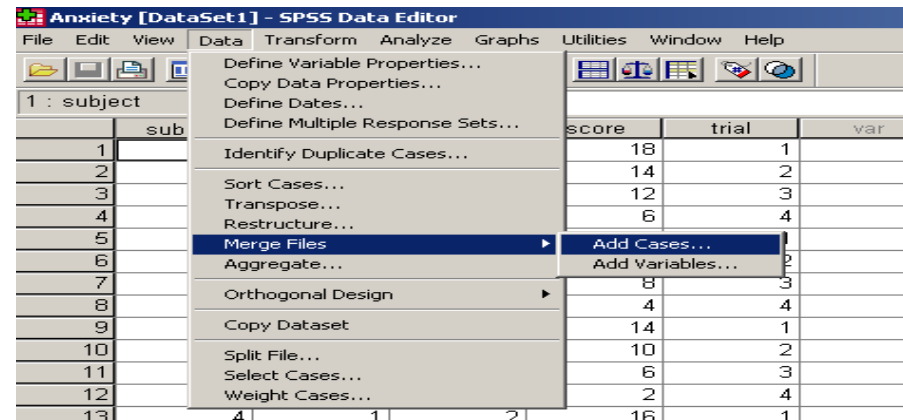
e.g. A respondent whose id number was 4321 will now have a year specific id of 914321.

We can repeat this process for every survey year by simply changing the front two digits of the added number. Be sure that the number you prefix has sufficient digits to represent every year of your combined dataset. For example, using 1 digit (i.e. 1-9) will be insufficient if you are combining data from before and after 2000, as 91 will equal 1, and 2001 will equal 1, potentially leading to duplicated id values for these two years. Using two digits will make sure you do not create a 'millennium bug' in your id variable.

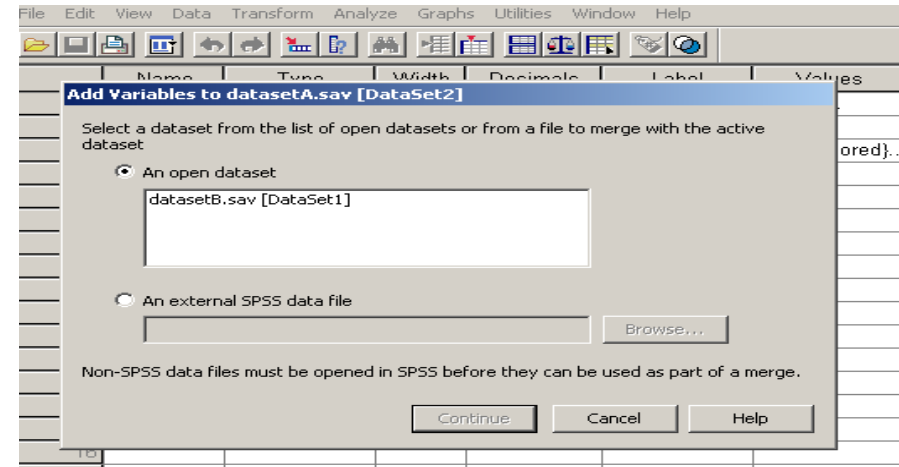
Joining Files using SPSS Version 14

SPSS Version 14 allows users to open more than one dataset at the same time. Datasets can therefore be combined either by joining two datasets currently in memory, or by holding one dataset in memory and calling a second data file.

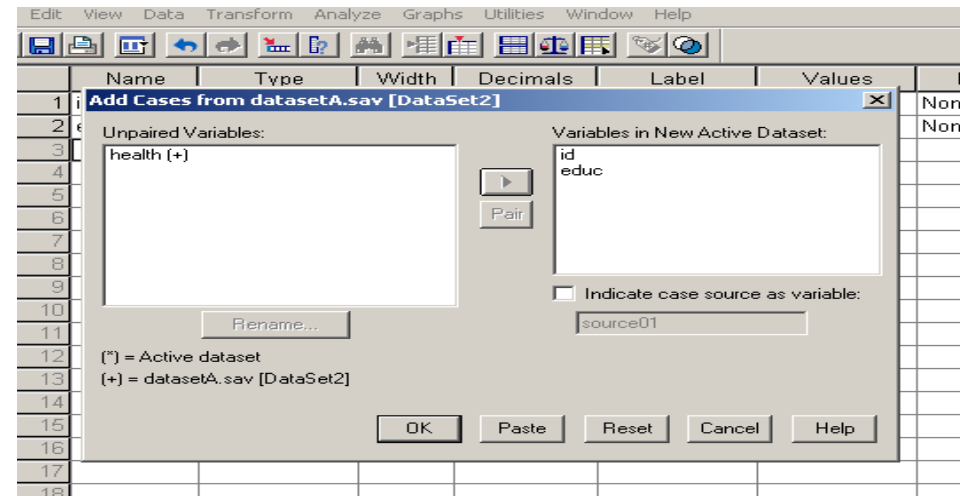
- Select *merge files* from the *data* menu.
- The *add cases* option is the equivalent to the *append* command in Stata. This is used to add the cases from a data file to the bottom of another dataset.
- The *add variables* command is equivalent to the *merge* command in Stata.



- To select a second dataset that you already have open, select the *add another dataset* option and click on the name of the dataset.
- Alternatively, the *external SPSS data file* option can be used to specify a file to be joined that is not currently in memory.
- Click continue once you have made your selection.



- The Next window indicates what variables are contained in the dataset to be merged on, but which are not in the active dataset.
- You can select these unpaired variables or omit them. If the variable is identical to one in the active dataset but has a different name, you can also rename variables.
- Selecting the *indicate case source as variable* option will create a variable in the combined dataset indicating which dataset cases originated from.
- Click OK.



3. ESDS Panel Data

3.1 Introduction

Panel data provide repeated measurements for the same individuals at multiple time points. Respondents interviewed in one wave of a survey are interviewed again at wave two, and so forth for consecutive waves. The first part of this section gives an introduction to some of the basic concepts in panel data analysis. This is followed by an introductory discussion of non-response and response error, which can be a cause of sample bias in panel datasets. Following this, the main features of ESDS supported panel surveys are outlined. The final section of the chapter considers data manipulation techniques for panel data using Stata and SPSS.

3.2 Concepts in Panel Data Methods and Research Examples

Panel data includes a cross-sectional dimension (across different individuals (i) at a given survey wave) and a time-series dimension (for the same individuals across survey waves/time points (t)). For this reason, panel data is sometimes referred to as 'cross-sectional time-series data'. In **balanced panels**, there are full observations for each case at every time point, whereas **unbalanced** panels may contain missing information for some individuals at time points¹⁰. A number of advantages of panel data are commonly cited:

- allows analysis of **individual/micro-level change**
- facilitates the modelling of **durations** as opposed to just cross-sectional stock or proportions
- allows temporal ordering of variables in order to help establish claims on causal relationships between variables
- can be used to control for the effects of **omitted variables** or **residual heterogeneity**, which can be a source of bias in regression models
- can help to identify **state dependence** where current behaviour is dependent on the occurrence or durations of earlier outcomes. This can be undertaken using **lagged variables** in **dynamic panel models**.

The following sub-sections discuss these concepts.

Using Panel Data to Study Micro-level Change

¹⁰ See Wooldridge (2002) for a discussion of balanced and unbalanced panels.

One motivation for using panel data is its capacity for the analysis of individual/micro-level change. Micro-level change refers to where the characteristics or circumstances of individual respondents differ over time. One of the simplest ways to consider individual level change is to cross-tabulate the characteristics of respondents in one wave of a survey with their characteristics in a subsequent wave. Using the panel component of the Families and Children Study (FACS), figure 3.1

cross-tabulates the working hours of a representative panel of lone parents in 1999 and 2000, and in 2000 and 2001. The numbers in the tables represent column percentages, indicating the percentage of people who were working a given number of hours in one year (we will refer to as $t-1$, read as "time point minus one"), by their working hours in a subsequent year of the survey (t). Such information might be used to explore whether low hours work provides a stepping-stone to greater hours of employment.

From this table, it can be seen that lone mothers working fewer than sixteen hours per week have a higher risk of exiting the labour market, compared to those working a greater number of hours. They are also more likely to move into work of more than 16 hours per week in a subsequent year, compared to those who were not in paid employment. The question remains why are some lone mothers who work fewer than sixteen hours per week more likely to exit the labour market, whilst others more likely to increase their level of participation. To answer this, we might use some of the multi-variate techniques for panel data analysis discussed below.

Figure 3.1 Working Hour Transitions (FACS, Lone Mothers, 1999-2001)

	Column Percentages		
	1999		
	Working +16 hrs	Working < 16hrs	Not Working
2000			
Working 16+ hrs	89	27	9
Working <16hrs	2	46	3
Not Working	9	27	88
<i>Un-weighted Base</i>	567	111	928
	2000		
	Working +16 hrs	Working < 16hrs	Not Working
2001			
Working 16+ hrs	92	35	15
Working <16hrs	1	35	3
Not Working	7	30	82
<i>Un-weighted Base</i>	562	86	830

Other ESDS supported datasets that contain panel components can also be used to examine employment transitions. The two-quarter and five-quarter UK Labour Force Survey (LFS) datasets contain ready-made **flow variables** summarising movement between employment states across year quarters. Information for the same people for these two time points is obtained from two successive quarters of the survey. Figure 3.2 summarises the different combinations of employment states respondents can occupy. Around 72.7 per cent of respondents were employed in both quarters (EE), whereas 1.1 per cent were employed in the first quarter and unemployed in the second (EU). Overall, the percentage of spells that represent transitions between states is fairly low.

Figure 3.2 Employment State Transitions (Two Quarter UK-LFS)

Category	Description	Level		Flow per cent
		Unweighted	Weighted	
Entering				
working age	Age 15 quarter 1, working age quarter 2	269	180,327	
EE	Employed both quarters	34,549	26,324,784	72.7
UE	Unemployed quarter 1, employed quarter 2	411	357,047	1.0
NE	Inactive quarter 1, employed quarter 2	554	478,703	1.3
EU	Employed quarter 1, unemployed quarter 2	471	391,109	1.1
UU	Unemployed both quarters	784	681,758	1.9
NU	Inactive quarter 1, unemployed quarter 2	336	305,351	0.8
EN	Employed quarter 1, inactive quarter 2	620	459,783	1.3
UN	Unemployed quarter 1, inactive quarter 2	404	313,667	0.9
NN	Inactive both quarters	8,944	6,905,055	19.1
Leaving				
working age	Working age quarter 1, above working age quarter 2	210	140,777	
Total		47,552	36,538,361	100.0

Beyond simple cross tabulations, panel data provides more sophisticated ways to look at change over time. Such data can be used in multivariate analyses to analyse changes in the value of a continuous or discrete (categorical) dependent variables between measurement time points. Jenkins (2000), for example, outlines a number of longitudinal approaches that have been used to study the dynamics of poverty. By poverty dynamics, we refer to the study of the duration, and movements

into and out of poverty. In such studies, poverty is typically defined by drawing a poverty line at a specific level of income (say 60 per cent below the median population income, adjusted for household composition), where people with incomes below this line are defined as living in poverty.

In one modelling strategy, referred to as a *transition probability model* (see Jenkins, 2000), the probability of moving either out of poverty or into poverty is modelled. In such models, the transition probability of moving out of poverty is the probability that a person i is not poor at a survey time point t , given that they were living in poverty at a preceding time point $(t-1)$. Conversely, a model of entry into poverty is the probability of a person living in poverty at time point t , given that they were not poor at time point $t-1$. The probability of such transitions can be modelled as a function of a set of constant and time varying covariates which predict the likelihood of a transition. Such models thus represent a multivariate extension to analysing the types of changes witnessed in the above examples of transition tables. Presenting research examples, including using data from the British Household Panel Survey, Jenkins (2000) gives an overview of the value of longitudinal perspectives to poverty analysis, covering further technical issues that may be of interest to the reader.

Modelling Durations

In the above modelling strategy, transitions between states are modelled without considering how long a person has been in poverty. There are some potential limitations with such an approach. Take for example two people (person A and person B) who were in poverty at $t-1$ and had moved out of poverty by time point t . Taking the above approach, these two events would be considered as identical for modelling purposes. However, it could be the case that whereas person A had only been living in poverty for one year, person B had been in poverty for ten years. In terms of understanding the *rate* at which different people exit a given state such as poverty, it may be of interest to examine *durations* to see how long it takes different people to exit a state. Another example is youth and elderly unemployment. Although unemployment is more common amongst younger adults (in terms of stock counts), younger people may leave unemployment more quickly than the more elderly unemployed who on average have longer durations. Thus, a cross-sectional analysis can give a partial and sometimes misleading picture.

There are a number of techniques for modelling spell durations prior to transitions, referred to collectively as **event history analysis**. Event history analysis can be used to analyse why certain individuals are more at risk of experiencing events than others. This is estimated through assessing the duration of time between becoming at risk for a given event and experiencing an event. Examples would be the time between becoming unemployed and finding a job, or getting married and divorcing. The risk of experiencing a certain event by a given point in time is predicted by a set of covariates. Such models are also referred to as 'survival analysis', 'duration analysis, or 'failure-time analysis' (Allison, 1984; Blossfeld, 2001).

In event history analysis, the term *state* refers to categories of the outcome/dependent variable. The *failure event*, denotes a positive occurrence of the outcome variable (e.g. a move out of poverty). *Right censoring* occurs where a given spell ends for a reason other than a failure event, or the time of the spell ending or failure event is unknown because a spell continues beyond the observation period. At each point in time, each individual exclusively occupies one state of the dependent variable. Thus, event history analysis can be defined as the analysis of rates of occurrence of the event during the risk period (Yamaguchi, 1991). The risk period refers to the time during which a given individual is 'at risk' of experiencing the failure event. A person must occupy a certain state to be at risk. For example, a person must be employed in order to be at risk of being promoted, or married to be at risk of divorce. Data already in an appropriate format for event history analysis can be used such as the British Household Panel Survey Combined Work-Life History dataset. However, you can also derive your own event history data from panel data by creating variables indicating the duration a person occupies a given state across different panel waves, and whether such spells end with a 'failure event' of interest or are right censored.

Event history modelling is based around the analysis of either the survival function or, more commonly, the hazard function. The survivorship function denotes the probability of the non-occurrence of an event at a given time point (Blossfeld, 2001). The 'hazard rate,' as coined by Barlow (1963), gives the probability per time unit that a case that has survived to the beginning of a time interval will fail in that interval. In life table approaches, this rate is computed as the number of failures per time unit in the respective interval, divided by the average number of surviving cases at the mid-point of the interval. The hazard rate or hazard function $h(t)$ expresses the instantaneous risk of experiencing an event at $T=t$, given that the event did not occur prior to t , where T again denotes the duration of non-occurrence of an event.

A number of different specifications of the hazard rate can be implemented. In parametric specifications, a particular shape is imposed on the baseline hazard rate (see Yamaguchi, 1991 for examples). Alternatively semi-parametric techniques may be used which do not specify an underlying shape to the hazard rate (Cox, 1972). A disadvantage of parametric models is that they hold strong assumptions about the shape of the hazard function which may not be realistic (Hosmer and Lemeshow, 1999; Blossfeld, 2001). The Cox Proportional Hazards Model, or 'Cox Model' (Cox, 1972), represents the most widely used semi-parametric approach to survival modelling (Yamaguchi, 1991). In the Cox model as available in common software packages such as SPSS and Stata, the baseline hazard is not explicitly modelled, and so does not require any potentially constraining distributional assumptions. The Cox model has no intercept as this is assumed to be contained in the baseline hazard function. Event history analysis can further be categorised into continuous time and discrete time methods. For discrete time methods, you will need to organise your data into 'long format' with multiple rows for individual cases, with each row representing different discrete units of analysis time (see Jenkins, 1995 for further details). So, if a given spell for a

respondent lasts for 18 months and you wanted to create a long format file representing individual months, you would create 18 rows of data for the spell. This can be achieved easily in Stata using the **stspl** command.

Omitted Variables and Residual Heterogeneity: Fixed, Between, and Random Effects

A major application of panel data analysis has been to attempt to control for missing variables or other causes of *residual heterogeneity* within regression models. When estimating a regression model, there may often be unmeasured variables that influence the value of a dependent variable. Such omitted variables could further be correlated with explanatory variables. If the latter is true, the estimates of the effects of an explanatory variable will to some extent be biased, as it will 'pick up' some of the effects of the omitted correlated variable. In a cross-sectional study, we might try to use a variable which proxies for an unobserved variable (or instrumental variables¹¹) (see Wooldridge, 2002), although in many cases this may not be possible. A number of **panel regression** techniques have been developed to attempt to deal with the problem of omitted variables and residual heterogeneity. The most commonly used techniques are **fixed effects**, **between effects**, and **random effects**.

One way to understand the issue of unobserved effects is to consider the difference between an experimental study and a regression analysis that uses a random sample of survey data. In experimental studies, such as clinical trials which examine the effectiveness of a new drug, individual participants are randomly assigned to either an experimental (treatment) group which receive an intervention of interest (e.g. a new drug), or one or more control groups which do not receive the intervention under assessment (e.g. they receive a dummy pill or an alternative treatment). The effectiveness of the intervention can be assessed by drawing statistical comparisons between the experimental and control groups on relevant outcome variables. In the case of a new headache pill, this might involve an assessment of whether there are significant differences in the duration and frequency of headaches in the experimental and control groups. In this study, our *dependent variable* would be some form of measure of headaches, whereas our main *independent variable* (the value of which is experimentally manipulated) would be whether a person has received the new treatment or is a member of the control group. In social research, random assignment is used in psychology experiments, and in policy evaluation studies. An example policy evaluation would be where one group of a population is randomly assigned to an experimental group receiving a new set of interventions, and compared on one or more relevant outcome variables to another group that does not receive the new policy intervention.

¹¹ See x for a discussion of instrumental variable approaches

The key value of random assignment is that other variables beyond the experimentally manipulated independent variable(s) are randomly distributed between the experimental and control groups. Consequently, it can be assumed that such factors exert equal influence on the outcome variables in the two groups, and so are *experimentally controlled* for. In this sense, random assignment can also control for the effects of variables which are unmeasured in data, but which influence an outcome variable.

In practice, random assignment often presents an impractical, impossible, or unethical approach to answering research questions. Regression analysis using samples of populations of interest provides an alternative approach. Instead of experimental control, in multivariate regression, we attempt to *statistically control* by providing estimates of the effects of an independent variable on a dependent variable, after statistically holding the effects of all other explanatory or control variables in a model constant. However, the problem of omitted variables may mean that there are important explanatory factors which are not controlled for in our models. Some times this may be through a lack of data collection on certain topics and so important factors remain unmeasured. In others, it may be because the variables of interest are difficult to measure in the first place.

We can more formally consider the omitted variable problem by considering the structure of a linear regression equation. In Ordinary Least Squares (OLS) regression, the expected value of a dependent variable (Y) is modelled as a linear function of a set of explanatory and control variables.

$$Y_i = \beta_1 + \sum_{j=2}^k \beta_j X_{ji} + \varepsilon_i$$

Wooldrige (2002) uses an example where wages are modelled as a function of educational attainment, often defined as years of schooling, and a set of other explanatory or control variables (for example employment experience, age, sex, and marital status). Instead of experimentally manipulating the values of the independent variables on the right hand side of the equation, the effects of the independent variables are estimated from values obtained from the population sample, by estimating regression coefficients (β) for their effects. In the case of a model of wages, the regression coefficient for years of education would give an estimate of how having one extra year of schooling affects the level of wages people receive, after holding the effects of the other explanatory and control variables constant.

We can extend the linear regression equation to panel data by adding subscripts for time to several of the terms in the cross-sectional linear regression equation, to represent the same individuals at multiple time points. Thus for a panel sample of

individuals at a number of time points (survey waves), the dependent variable (Y) for individual (i) at a time point/survey wave (t) can be considered as a function of the sum of (Σ) the effects of observed explanatory and control variables (X) as indicated by the coefficients (β_j), and where (β_1) represents the intercept and ε is the disturbance term (error):

$$Y_{it} = \beta_1 + \sum_{j=2}^k \beta_j X_{jit} + \varepsilon_i$$

The value of t depends on which survey wave a case for a specific individual represents. So if we have five waves,¹² t can range between 1 and 5. This example model assumes that the error term for a given individual at consecutive time points is uncorrelated, and is commonly referred to as a **pooled regression** model.

The assumptions made regarding the error term are of central importance in panel regression. One standard interpretation of the error term is that it represents the variance in the value of the dependent variable (Y) which is not explained by the covariates, possibly attributable to measurement error or 'noise'. However, if an important explanatory or control variable is omitted from the model, some of the variance in the dependent variable Y which is not accounted for by the covariates, will be attributable to the unmeasured variable(s). In such situations, part of the variance represented by the error term may be attributable to missing variables, and if the values of these missing variables were known, this would reduce the amount of variance in the dependent variable incorporated into the error term, providing a better fitting model that explains more of the variance in the data. If missing variables are correlated with a given independent variable, its inclusion in the model will also change the size of the estimated effect for this variable. Its absence from the model means that our estimates for the correlated variable may be biased.

For models where important variables are omitted, we can write a panel regression equation to represent the 'true' model which includes the unmeasured variables by adding terms for these variables (Z) and their coefficients (γ):

$$Y_{it} = \beta_1 + \sum_{j=2}^k \beta_j X_{jit} + \sum_{p=1}^s \gamma_p Z_{pit} + \varepsilon_{it}$$

In the above equation, j and p index the different observed and unobserved variables, and k and s equal the total number of observed and unobserved variables respectively. Note that j starts at 2 as β_1 is used again to denote the intercept. The sum

¹² In a 'balanced panel,' the number of waves (and so the range of t) is identical for all respondents. However 'unbalanced panels' may occur for a number.

of the effects for all the unobserved variables are often simplified for presentation and represented as alpha (α), the **unobserved effects**:

$$\alpha_i = \sum_{p=1}^s \gamma_p Z_{pi}$$

Thus, using alpha to summarise all of the effects of the unobserved variables:

$$Y_{it} = \beta_1 + \sum_{j=2}^k \beta_j X_{jit} + \alpha_i + \varepsilon_{it}$$

The question remains whether the influence of the unobserved effects (alpha) is important to the rest of the model, i.e. whether we can estimate unbiased coefficients for our independent variables of interest in the absence of controls for the unobserved variables. This will largely depend on whether the unobserved variables are correlated with a given observed variable.

In the case of a study of the effects of education on earnings, one missing variable in our model could be innate ability. We might expect that this variable is both a predictor of earnings and is also correlated with level of educational attainment¹³. Thus a model which does not include controls for such factors (or unobserved effects more generally) might give a biased estimate of the effects of education on earnings. In a cross-sectional study, we can attempt to control for the effects of omitted variables through including measured variables in our models which proxy for omitted variables.¹⁴ In terms of a proxy for ability a test score could be used, obtained from a childhood ability test.

However, by utilising information for the same individuals at different survey waves, panel data provides a number of strategies for controlling for omitted variables. **Fixed effects** models can be used to control for omitted variables that differ between individual cases (i) but are constant over time¹⁵ (t) (Greene, 2003). Information at successive waves of a survey is used to eliminate the individual fixed effects (α_i) from the regression equation. There are different fixed effects approaches including: (1) the within group method, (2) the first difference method, and (3) dummy variable approaches.

¹³ Another factor might be early childhood conditions such as poverty and material deprivation.

¹⁴ The use of other instrumental variables provides an alternative approach.

¹⁵ See Wooldridge (1997) for an example of the use of fixed effects regression to estimate the effects of children on women's wages.

(1) In the within-group method, average values for the dependent and independent variables across all the panel waves for a given individual are used to eliminate the unobserved effects (alpha) from the regression model. The unobserved fixed effect is assumed to be the same across sample waves for individuals so that its mean value across waves is identical to the individual values for each wave. By subtracting the mean value, the fixed components over time (including alpha) of the equation are eliminated.

So for the panel regression equation:

$$Y_{it} = \beta_1 + \sum_{j=2}^k \beta_j X_{jit} + \alpha_i + \varepsilon_{it}$$

The corresponding mean regression equation can be written as:

$$\bar{Y}_i = \beta_1 + \sum_{j=2}^k \beta_j \bar{X}_{ji} + \alpha_i + \bar{\varepsilon}_i$$

The line above several of the variables (above Y , X , and e) in the second equation indicate that this symbol denotes the mean value for individuals across the different waves of the survey. Note that the sum of the unobserved effects (alpha) does not have a hat. This is because the unobserved effect in the fixed effect model is constant over time. Thus, its average is identical to each of its individual values in each wave (consider in basic arithmetic the average of 2, 2, 2, 2 = 2+2+2+2/4 =2). Because the average of the fixed effect equals the individual values, we can remove (eliminate) it from the first equation by subtracting the averaged equation.

So subtracting:

$$\bar{Y}_i = \beta_1 + \sum_{j=2}^k \beta_j \bar{X}_{ji} + \alpha_i + \bar{\varepsilon}_i$$

From:

$$Y_{it} = \beta_1 + \sum_{j=2}^k \beta_j X_{jit} + \alpha_i + \varepsilon_{it}$$

Gives the following where the unobserved effect (a) is eliminated:

$$Y_{it} - \bar{Y}_i = \sum_{j=2}^k \beta_j (X_{jit} - \bar{X}_{ji}) + \varepsilon_{it} - \bar{\varepsilon}_i$$

A problem with fixed effects approaches is that because of the mathematical manipulation, all variables which are time constant and so have identical values in each wave of the survey are also eliminated from the regression equation, as well as the fixed effects. This may be unwanted if you wish to include important time constant variable in models (such as sex), and may be also problematic for variables which do not have much variation in value for individuals between years.

(2) The first difference approach uses repeated information at different time points for respondents to eliminate the fixed effects. As the fixed effects are constant across time points, subtracting values for consecutive time points similarly eliminates the fixed effects like the within-group method. Again, as a consequence, other time constant variables are eliminated from the model.

(3) In Dummy variable approaches, a time constant dummy variable is added for each cross-sectional unit (e.g. respondent). This can lead to a high number of variables, requiring one dummy indicator per respondent, and so affect the power to detect significant effects (Baltagi, 2001).

Beyond fixed effects approaches, another set of methods, known as **between effects**, attempt to control for omitted variables that change over time but are constant between cases (Baltagi, 2001). Between effects are the equivalent to taking the mean of each variable for each case across time and running a regression on the collapsed dataset of means. This leads to a loss in information and so is less used than fixed effects or random effects (discussed next) (see Greene, 2003). Time dummies provide an alternative for controlling for omitted variables that change over time but are constant between cases (Wooldridge, 2002).

Random effects are used to control for omitted variables that may be constant over time but vary between cases, whereas others may be fixed between cases but vary over time. Random effects models assume that each of the unobserved variables (Z) are randomly drawn from a defined distribution. A critical assumption of random effects models is that the

random error term is not correlated with the covariates (Mundlak, 1978). If this assumption is violated, then the introduction of random effects can be a cause of bias in itself. Fixed effects models thus produce more consistent estimates, whereas random effects models are more efficient. The Hausman test (Hausman, 1978) is used to assess the null hypothesis that the coefficients estimated by the more efficient random effects estimator are the same as the ones estimated by the more consistent fixed effects estimator. This is taken as a test of the assumption that the random effects error term is independently distributed from the covariates (Hsiao, 2003). If the difference between the random and fixed effects model is insignificant, then the more efficient random effects model is preferred. The manner in which random effects are calculated¹⁶ and assumptions about the underlying population distribution can also influence estimates. Further information on these topics can be found in the further resources section (Appendix A).

State Dependence

State dependence refers to where the experience of a past event or state influences the probability of experiencing an event or state again in the future. A common finding in studies of unemployment is that people who have experienced past unemployment are more likely to experience further unemployment. There are a number of potential explanations of this finding. Firstly, in terms of state dependence, there may be something about the very experience of unemployment that increases the risk of future unemployment. For example, after experiencing unemployment, people might lower their job expectations and accept poorer job offers (like lower paid, unstable work) than they would have done if they had never been unemployed. Employers might also consider a history of unemployment in job applications as a negative signal, and so adversely select against people with a history of unemployment in recruitment decisions.

An alternative explanation is that some people have individual characteristics related to poorer employability and job stability, which increase the risk of unemployment. Although some of these factors may be observed in our data and so controlled for, other factors may be unobserved, thus raising the omitted variable problem. Blumen et al (1955), for example, argues that potentially unobserved characteristics such as 'individual instability', may cause persistent patterns of job turnover. Other unobserved factors which have been contended to influence the risk and duration of unemployment include 'motivation' and 'innate ability'. If such unobserved factors were adequately controlled for, then the observed relationship between past and future employment might diminish or no longer exist.

¹⁶ For example, statistical software may use Gauss-Hermite quadrature in the estimation of some random effects models (such as xtlogit in Stata), where the number of quadrature points specified may influence estimates. Type 'help quadchk' into Stata to find out more.

Panel data can help inform debates over the effects of state dependence and residual heterogeneity or omitted variables. Firstly, to estimate the effects of past unemployment on future unemployment, we can include **lagged variables** for the dependent variable (unemployment), to estimate whether the experience of unemployment in a previous time period (e.g. $t-1$) predicts the likelihood of being unemployed in a subsequent panel wave (t). Such models are often referred to as **dynamic panel models** and can also include lags of independent variables. From the above discussion of the omitted variable problem, we could also include in our model controls for residual heterogeneity, such as fixed effects, to consider how the experience of past unemployment predicts present unemployment after controlling for such factors. Undertaking this approach, Arulampalam et al (2000), for example, uses the British Household Panel Survey to estimate the following model:

$$y_{it}^* = x_{it}' + x_{it}'\beta + \lambda y_{it-1} + v_{it}$$

Where y_{it} denotes the individual propensity for unemployment, x_{it} represents the independent variables affecting y_{it} , β is the coefficients associated with each x , and v_{it} is the unobservable error term. The error term v_{it} is composed of two components, an individual specific fixed effect (alpha), and random error (NB. random error not random effects):

$$v_{it} = \alpha_i + e_{it}$$

This simply represents an alternative way of writing the fixed effects error term discussed above. The term λy_{it-1} represents the lagged dependent variable, denoting the experience of a different occurrence of unemployment at a previous time point. When including such a variable, it is important to ensure that the lagged variable represents a different occurrence and not just an earlier period of the same spell as witnessed at time point t .

Another technical issue to be handled in panel data models is the problem of **initial conditions** where the start of a process generating the values of a dependent variable is not recorded in the data. In the example of assessing the effects of state dependence on unemployment, the initial conditions problem arises when the start of the observation period does not coincide with the start of the process generating individuals' unemployment experiences. One solution to this problem is to use an instrumental variable approach (see Arulampalam et al, 2000 and Wooldridge, 2002).

3.3 Non-response Bias and Response Error Bias

The collection of panel data comes with the practical challenge of maintaining contact with respondents between interviews, and of ensuring consistency between information collected at different waves. The following section discusses some potential sources of bias in panel data. Panel datasets are subject to two main sources of possible error: non-response bias and response error bias (misclassification):

- *Non-response bias* occurs due to different groups of people, for example by age or sex, having different probabilities of dropping out of the survey between interviews, or through the occurrence of respondents refusing to respond on certain items. Dropout can occur where people move addresses between waves and are not traced.
- *Response error bias or misclassification* arises when respondents give incorrect answers to survey questions, for example due to misunderstanding or a lack of knowledge. It can also occur through coding errors made by interviewers (such as when coding occupation). When individual responses are linked across waves to determine transitions over time, these errors can lead to an apparent change of category when the true situation is no change of category. Therefore the number of people changing categories is likely to be over-estimated in such cases.

Non-response biases can be considered as missing data mechanisms. Rubin (1976) and Little and Rubin (1987 quoted in Couchley and Oskrochi (2001a)) classify such mechanisms into three different types (also see Little and Rubin (2002)):

- Missing completely at random (MCAR), if the pattern of missing data is independent of both observed data unobserved data.
- Missing at random (MAR), if conditional on the observed data, the missing pattern is independent of the unobserved data.
- Non-ignorable missing (NIM), if the missing data is neither MCAR nor MAR.

MCAR and MAR can be considered as independent right-censoring and therefore not involving sample bias (Couchley and Oskrochi, 2001a), although the assumption of MCAR and MAR in many contexts may be unrealistic. For example, when studying labour market behaviour, employed participants who experience promotion or get a new job may have a higher probability of dropping out of panels because such events may result in moving house or location. If such moves cannot be traced, then this will lead to sample attrition. Thus through sample drop out, estimates of the incidence of promotion or movement between jobs may be under-estimated as some transitions are not observed due to sample attrition. Non-ignorable missing data presents a more pertinent problem than MCAR and MAR as it indicates sample selection bias. Little

(1993) classifies techniques for detecting the nature of missing data (selection models and pattern-mixture models). Readers further interested in these issues are referred to this text, as well as the paper by [Couchley and Oskrochi \(2002\)](#) which gives specific reference to the BHPS. Wooldridge (2002, Chapter 12) also gives a more detailed discussion of issues of sample attrition in panel data analysis.

Weighting systems are commonly available in many longitudinal datasets which attempt to weight for non-response bias. It is therefore important to consult information on weighting in user guides for datasets prior to conducting your analysis. More general information on the imputation of missing data can be found on the [Missing Data web site](#).

ESDS Government provides a [guide to weighting](#) which can be downloaded from their web site.

3.4 Seam Effects in Panel Data¹⁷

A good example of response error bias resulting in misclassification is a 'seam effect', often seen in employment or health histories. The effect occurs where two survey reporting periods join and information for the period either side of the join is collected at different interviews. The effect manifests itself in an unfeasibly large number of transitions in status (e.g. transitions from one type of employment to another) at the join. Of course, not all transitions observed at this join will be spurious but spikes in the number of transitions occurring at joins are a good indicator that some of the observed change is not real. Seam effects are especially prevalent where the reporting period is shorter than the gap between data collections (e.g. where survey interviews are annual but reference periods are monthly). Here, respondents have a tendency to report the same status for *all* the reference periods between interviews, thus producing annual spikes even though the reference period is a month. For a comprehensive introduction to the topic, see Annette Jäckle's paper '[The Causes of Seam Effects in Surveys](#)'.

A seam effect is therefore an artefact of the survey process rather than a measure of the true, underlying rate of transition and may be the result either of simple recall error – the respondent providing incorrect information on their status and/or duration in that status – or a coding bias – where the interviewer(s) receive the same information but record it differently – or an error in data entry. Survey methodologists have developed two broad responses to the problem of seam effects: *event history calendars (EHCs)* and *dependent interviewing (DI)*:

¹⁷ Many thanks go to Jack Kneeshaw, ESDS Longitudinal, for providing this section.

- The *event history calendar* method features the use of a calendar in the survey interview in order that respondents may use sequential and parallel (and visual) cues to improve their accuracy of recall (Belli et al, 2007). Calendars allow respondents to order events (sequential cues) and to match events in one life domain to another – or to match personal events to external, national events (parallel cues).
- *Dependent interviewing* is a technique that employs responses from the previous data collection (i.e. fed-forward data) to remind respondents what they have told interviewers previously – e.g. "According to our records, when we last interviewed you, on <date>, you were receiving <source>, either yourself or jointly. For which months since then have you received <source>?"

Both of these techniques have been applied with some success in major longitudinal survey settings around the world – though their application also continues to be experimental and refinements to the techniques continue to be made. In the UK, the life history interviews in wave 3 of ELSA (data not yet available via ESDS, 30.6.08) employed a 'life grid' method by which respondents supplied information on a series of rows relating to children, partners, accommodation and work, with a calendar running across the top of the grid. DI was also incorporated into the design of ELSA, not least in an effort to maximise accuracy/consistency between the Health Survey for England data collection (wave 0) and wave 1 of the study. The major birth cohort studies have increasingly used fed-forward data – on questions relating to fertility, housing, partnership and job histories – as has the LSYPE (e.g. parental employment histories).

Mainly for reasons of maintaining longitudinal comparability¹⁸, the BHPS has used fed-forward data more for sample management/questionnaire routing than on substantive questions. However, DI was introduced in the BHPS at wave 16 – see the ISER working paper '[The Introduction of Dependent Interviewing on the British Household Panel Survey](#)'. In addition, the [ISMIE](#) project – an ESRC-funded project undertaken by the UK Longitudinal Studies Centre (ULSC) – used a sub-sample of the BHPS to investigate the effects of various DI designs on response. [Data from the ISMIE project](#) are available via the ESDS web site.

¹⁸ DI relies heavily on computer-assisted interviewing and it was therefore not considered feasible to introduce DI whilst the BHPS used paper and pencil interviewing (PAPI) from wave 1 through wave 8.

3.5 ESDS Panel and Cohort Datasets

This section goes on to outline the general features of major panel and cohort datasets supported by ESDS. These are:

- Two and Five Quarter Labour Force Surveys
- British Household Panel Survey
- National Child Development Study
- 1970 British Cohort Study
- Millennium Cohort Study
- English Longitudinal Study of Ageing
- Longitudinal Study of Young People in England
- Families and Children Study

Two-Quarter and Five-Quarter Labour Force Survey Panel Datasets

ESDS Page: www.esds.ac.uk/findingData/lfsTitles.asp

Documentation: http://www.esds.ac.uk/doc/6457/mrdoc/pdf/lfs_vol1_background2009.pdf

List of Variables: www.esds.ac.uk/doc/5548/mrdoc/excel/list_of_variables.xls

Overview: The general features of the UK Labour Force Survey were outlined in Section 2.4. Since 1992, The Labour Force Survey interviews respondents for five consecutive quarters with twenty per cent of the sample being replaced each quarter. The dataset thus represents a 'rotating panel,' with a fifth of the panel being refreshed at each quarter. A series of datasets is available from ESDS Government which contains two and five quarter panels. The two-quarter datasets link data from two consecutive waves, while the five-quarter datasets link across a whole year (for example summer 1999 to summer 2000 inclusive), containing data from all five waves. The samples are restricted to working age populations (15 to 59 years of age for women and 15 to 64 years of age for men), and respondents who receive full interviews at each wave of the sample. These datasets are available from winter 1992/93 onwards. The LFS for Northern Ireland did not become quarterly until winter 1994/5. Consequently, two and five quarter datasets prior to winter 1994/95 do not contain the Northern Ireland sample.

In accordance with EU regulations, the LFS moved from seasonal (spring, summer, autumn, winter) quarters to calendar quarters (January-March, April-June, July-September, October-December) in 2006. The last seasonal five-quarter longitudinal dataset issued was the *Labour Force Survey Five-Quarter Longitudinal Dataset, March 2005 - May 2006* (SN 5469), and the first calendar five-quarter dataset was the *Labour Force Survey Five-Quarter Longitudinal Dataset, July 2005*

- September 2006 (SN 5549). A [paper on the implications of changing to calendar quarters](#) and [more information about the changes](#) can be obtained from the ESDS Government web site.

Sample Attrition and Weighting: Issues around attrition have been addressed by ONS in the longitudinal LFS datasets through the inclusion of longitudinal weights. A weighting system (*LWGHT*) compensates for attrition, ensuring that the gross flows are consistent with the changes in stocks. If you wish to use two or five wave panel data it is strongly recommended that you use the ONS datasets, rather than create your own. If you need to combine data to produce your own longitudinal series you should note that your dataset will be subject to uncorrected attrition. Problems of response error bias however remain unresolved and subject to continued investigation.

The extent to which attempts are made to track respondents who cannot be contacted for interview or change addresses affects the level of attrition in panel surveys. The LFS does not follow people who move address. Consequently, those who do move are not sampled and instead the new resident at the address is included in the sample. The LFS is orientated towards obtaining reliable point estimates of labour market activity, and so achieving a large sample size. Tracking individuals across addresses is highly resource intensive. A trade off is thus made between devoting resources to reducing attrition and maintaining a large sample size.

British Household Panel Survey

ESDS Page: www.esds.ac.uk/findingData/bhpsTitles.asp

ESDS guide: www.esds.ac.uk/longitudinal/access/bhps/L33196.asp

United Kingdom Longitudinal Studies Centre (ULSC) Page: <http://www.iser.essex.ac.uk/bhps>

Overview: The BHPS was designed as an annual survey of each adult (aged 16 years and over) member of a nationally representative sample of more than 5,000 households, making a total of approximately 10,000 individual interviews. The same individuals are re-interviewed in successive waves and, if they leave their original households, all adult members of their new households are also interviewed. Children are interviewed once they reach the age of 16. There is also a special survey of household members aged 11-15 included in the BHPS from Wave 4 onwards. Between 1997 and 2001 the BHPS included a further sample for the *European Community Household Panel* (ECHP). This component has not been continued beyond Wave 11. A major development at Wave 9 was the recruitment of two additional samples to the BHPS in Scotland and Wales. The target sample size in each country was 1,500 households. At Wave 11 an additional sample from Northern Ireland was added to increase representation to the whole of the United Kingdom. The target sample size of this sample is 2,000 households.

An index and thesaurus of variables for the BHPS can be found at <http://www.iser.essex.ac.uk/bhps/documentation/volb/index.html>. When using the BHPS to conduct longitudinal analysis, it is necessary to check that the definitions and coding of variables is identical between years. Information on changes to variables and other modifications to the data for the latest wave are detailed in Volume A of the documentation (Introduction, Technical Report and Appendices) - see Appendix 4.

Sample Attrition and Weighting: For the purposes of panel analyses, only cases which responded at all waves are generally of interest to creating balanced panels (although unbalanced panels may contain records where drop out has occurred). The longitudinal respondent weights (wLRWGHT) select cases who gave a full interview at all waves in the BHPS files (where *w* is replaced by the specific wave suffix). At each wave, cases are re-weighted to take account of previous waves' respondents lost through refusal at the current wave or through some other form of sample attrition. Thus the longitudinal weight at any wave will be the product of the sequence of attrition weights accounting for losses between each adjacent pair of waves up to that point, as well as the initial respondent weight at wave one. It should be noted that response also includes the deceased, people who have moved into institutions, or otherwise gone out-of-scope. These fail to give an interview not through non-response but due to a terminating event that results in their leaving the population of interest (Taylor et al, 2007).

The longitudinal respondent weights are calculated in two stages. First, all respondents at both waves including those with "terminal events" are weighted to adjust for the attrition of cases whose final status was indeterminate, in that it was not known whether these cases were still eligible for interview or had left the population of interest. These include people who had moved from their previous wave address and were subsequently not traced for interview, as well as refusal households where the interviewer was unable to determine who was still resident and eligible. The second adjustment weights up the cases interviewed at both waves to take account of those who refused an interview, were proxied or were unable to give an interview at Wave Two (those with terminating events were not included since all non-respondents in this group were known to be ineligible). In addition to those interviewed at all waves, original sample member (OSM) children enumerated in a respondent household at all waves before they reached 16, and respondents thereafter, will have positive longitudinal respondent weights.

When applying longitudinal weights in the BHPS, it is therefore necessary to apply the weight from the **last** wave of the panel dataset you have constructed. This is in order to use the weight that accounts for adjustments in previous years of the survey. Thus, if you are analysing waves 1 to 6 (1991-1996), you should use the 1996 weight (fLRWGHT). If you use an

earlier weight your data will be incorrectly weighted. Further information on weighting can be found in the [BHPS technical report](#).

BHPS Combined Work-Life History Datasets

See ESDS Page: www.esds.ac.uk/findingData/snDescription.asp?sn=3954

Overview: The 'BHPS Combined Work-life History Dataset' provides employment spell data that can be used amongst other things for event history analysis. The BHPS collects extensive information on respondents' labour market status at three points: at the time of interview at each wave of the panel; throughout the period between 1 September a year before the interview date; and retrospectively from when the respondent first left full-time education. Because the retrospective information has been collected in two tranches (one focusing on employment status (e.g. whether employed or unemployed), the other on occupational information (i.e. what occupation a person was employed in), there are four different types of labour market history information, located in different files in the BHPS database. The Combined Work-Life History Datasets constitute 'reconciled' files, giving single continuous records derived from the above sources of information.

The combined datasets were created as follows. The first part of the exercise is to take 'current status' information for each wave and combine it with the inter-wave history, and then to combine the waves creating a continuous record from September 1990 to the latest wave. The second stage incorporates the lifetime employment status history collected at Wave 2, and the lifetime occupational history collected at Wave 3, combining each of them with analogous information drawn from the combined panel file. This results in employment and occupational histories that stretch from labour market entry to the latest wave. The third stage combines these two extended lifetime histories into a single record which contains both employment status information and occupational information. Variables from the standard BHPS waves datasets can be added to supplement your analysis (see Halpin, B. (1997) [Unified BHPS work-life histories: combining multiple sources into a user-friendly format](#), Technical Papers of the ESRC Research Centre on Micro-Social Change, 13, Colchester: University of Essex.

Given that multiple records exist recording information for the same time point, there can be considerable disagreement between records for a given time point in reported employment status and occupations. The effects of such overlaps are *seam effects* (see Section 3.4 above). In some cases, changes may represent real changes in labour market status, whereas in other they may represent response error (recall bias), or coding bias. For example, if a respondent in one wave fails to remember the dates or job in which they worked, this is recall bias. If the same job was coded differently by interviewers at two waves of the survey, this would be coding bias. The BHPS Combined Work-life History dataset maintains all transitions in

the dataset (whether real or spurious). This underlying structure is represented by what is referred to as 'splits'. The dataset therefore contains variables which allow users to define rules regarding what transitions are considered as real or spurious, in order to create 'episodes' which are super-imposed over the split structure and deemed to constitute real transitions. Further details can be found in [BHPS Work-life History Files, Version 2](#) by Brendan Halpin.

Gillian Paull (Institute for Fiscal Studies) has constructed a Work-Life History Dataset for the BHPS, details of which can be found in her report '[Biases in the reporting of labour market dynamics](#)'. This paper also includes a detailed examination of issues of seam effects and recall bias in work-life history data. Couchley and Oskrochi (2001b) further discuss work to create a unified work-history dataset for the BHPS. This working paper is available from the [Applied Statistics \(University of Lancaster\) web site](#).

National Child Development Study (NCDS)

ESDS Page: www.esds.ac.uk/findingData/ncdsTitles.asp

ESDS guide: www.esds.ac.uk/longitudinal/access/ncds/133004.asp

Centre for Longitudinal Studies Page: www.cls.ioe.ac.uk/studies.asp?section=000100020003

Overview: The National Child Development Study (NCDS) originated in the 'Perinatal Mortality Survey', which examined social and obstetric factors associated with still birth and infant mortality among over 17,000 babies born in Britain in a single week in 1958. Surviving members of this birth cohort have been surveyed on eight further occasions in order to monitor their changing health, education, and social and economic circumstances - in 1965 (age 7), 1969 (age 11), 1974 (age 16), 1981 (age 23), 1991 (age 33), 1999-2000 (age 42), 2004 (age 46) and 2008 (age 50, data not yet available 30.6.08). Data from the 1958 Perinatal Mortality Survey and the 1965, 1969 and 1974 surveys are held at UKDA under SN 5565. The more recent sweeps are held under SNs 5566, 5567, 5578 and 5579. During the 1991 survey, a special study was also undertaken of the children of one in three cohort members. There have also been surveys of subsamples of the cohort, the most recent being in 1996 (age 37) when the basic skills of a representative sample of 10 per cent of cohort members were assessed.

1970 British Cohort Study

ESDS Page: www.esds.ac.uk/findingData/bcsTitles.asp

ESDS guide: www.esds.ac.uk/longitudinal/access/bcs70/133229.asp

Centre for Longitudinal Studies Page: www.cls.ioe.ac.uk/studies.asp?section=000100020002

Overview: The 1970 Birth Cohort Study (BCS70) has been developed on lines similar to NCDS, originating in the British Birth Survey of over 17,000 babies born in Britain in a single week in 1970. Subsequently, seven further major surveys have monitored the changing health, education, social and economic circumstances of the surviving cohort members - in 1975 (age 5), 1980 (age 10), 1986 (age 16), 1996 (age 26), 1999-2000 (age 30), 2004 (age 34) and 2008 (age 38, data not yet available 30.6.08). These studies are also held at UKDA - under SNs 2699 (five-year follow-up), 3723 (ten-year follow-up), 3535 (sixteen-year follow-up), 3833 (twenty-six-year follow-up), 5558 (thirty-year follow-up) and 5585 (thirty-four-year follow-up). Further information about other BCS70 datasets may be found under GN:33229. As in NCDS, sub-samples have also been studied, the most recent being in 1991 (age 21), when paralleling the NCDS survey at age 37, a 10 per cent representative sample was assessed for basic skills difficulties.

The Centre for Longitudinal Studies (CLS) at the Institute of Education (formerly the Social Statistics Research Unit located at City University) has been responsible for the National Child Development Study since 1985, when the study was transferred from its previous home at the National Children's Bureau. The 1999-2000 follow-ups were the first combined wave in the NCDS and BCS70 series, and took place when NCDS cohort members were aged 41/42 and BCS70 cohort members were aged 29/30.

For the second edition of the study (January 2003), the depositor supplied an updated version of the data file and some extra documentation, including a revised user guide and three CLS Data Notes, which cover longitudinal linkage for BCS70 datasets, and pregnancy histories and household grid variables in the combined NCDS/BCS70 1999-2000 dataset.

Millennium Cohort Study (MCS)

ESDS Page: www.esds.ac.uk/findingData/mcsTitles.asp

ESDS guide: www.esds.ac.uk/longitudinal/access/mcs/133359.asp

Centre for Longitudinal Studies Page: www.cls.ioe.ac.uk/studies.asp?section=000100020001

Overview: The original objectives of the first MCS survey, as laid down in the proposal to the Economic and Social Research Council (ESRC) in March 2000, were:

- to chart the initial conditions of social, economic and health advantages and disadvantages facing children born at the start of the 21st century, capturing information that the research community of the future will require
- to provide a basis for comparing patterns of development with the preceding cohorts (the *National Child Development Study*, held at the UK Data Archive (UKDA) under GN 33004, and the *1970 Birth Cohort Study*, held under GN 33229)
- to collect information on previously neglected topics, such as fathers' involvement in children's care and development
- to focus on parents as the most immediate elements of the children's 'background', charting their experience as mothers and fathers of newborn babies in the year 2000, recording how they (and any other children in the family) adapted to the newcomer, and what their aspirations for her/his future may be
- to emphasise intergenerational links including those back to the parents' own childhood
- to investigate the wider social ecology of the family, including social networks, civic engagement and community facilities and services, splicing in geo-coded data when available

Additional objectives subsequently included for MCS were:

- to provide control cases for the national evaluation of Sure Start (a government programme intended to alleviate child poverty and social exclusion)
- to provide samples of adequate size to analyse and compare the smaller countries of the United Kingdom

The first sweep (MCS1) interviewed both mothers and (where resident) fathers (or father-figures) of infants included in the sample when the babies were nine months old, and the second sweep (MCS2) was carried out with the same respondents when the children were three years of age. Further information about the MCS can be found on the CLS web page link above.

English Longitudinal Study of Ageing (ELSA).

ESDS Page: www.esds.ac.uk/findingData/elsaTitles.asp

ESDS guide: www.esds.ac.uk/longitudinal/access/elsa/I5050.asp

Institute for Fiscal Studies Page: www.ifs.org.uk/elsa/

Overview: The ELSA study is a longitudinal survey of ageing and quality of life among older people that explores the dynamic relationships between health and functioning, social networks and participation, and economic position as people plan for, move into and progress beyond retirement. The study was created to:

- Describe health trajectories, disability and healthy life expectancy in a representative sample of the English population aged 50 and over, and to

- Examine the relationship between economic position and health
- Investigate the determinants of economic position in older age
- Describe the timing of retirement and post-retirement labour market activity
- Understand the relationships between social support, household structure and the transfer of assets

Wave 1 of the survey covers March 2002-March 2003, whereas the interview period for wave 2 was March 2004- March 2005. Wave 0 was collected in the 1998, 1999, and 2001 as part of the Health Survey of England and pre-dates the ELSA project. The current deposit on the data archive comprises Waves 0, 1, 2 and 3 of the survey. Further files are also planned for future deposit. The ELSA dataset includes several files. Data and supporting documentation are accessible through the following link:

A [guide to ELSA](#) is available from ESDS Longitudinal.

Longitudinal Study of Young People in England

ESDS Page: www.esds.ac.uk/findingData/lstypeTitles.asp

ESDS guide: www.esds.ac.uk/longitudinal/access/lstype/L5545.asp

Overview: The *Longitudinal Study of Young People in England* (LSYPE), also known as *Next Steps*, commissioned by the Department for Education and Skills (DfES), is a major innovative panel study of young people that brings together data from a number of different sources, including both annual interviews with young people and their parents and administrative sources. The main role of the study is to identify, and enable analysis and understanding of, the key factors affecting young people's progress in transition from the later years of compulsory education, through any subsequent education or training, to entry into the labour market or other outcomes.

Sample boosts took place for deprivation factors and for ethnicity. Schools with 20% or more of pupils entitled to free school meals were over-sampled by a factor of 1.5, and over-sampling for ethnic minority groups was introduced at pupil level in the maintained sector. The boost method used for ethnic groups means that these boosts are representative samples of the relevant subpopulations as a whole rather than e.g. drawn disproportionately from areas or schools with high numbers of ethnic minority pupils.

Data from LSYPE have been linked to administrative data held on the *National Pupil Database* (NPD), a pupil-level database which matches pupil and school characteristic data to pupil-level attainment. Due to the potentially disclosive nature of some of these variables, the linked administrative data have not been included in the initial deposit of LSYPE data. The DfES plans

to deposit limited, non-disclosive administrative data at a later date. In the meantime, researchers requiring access to the linked administrative files should contact the DfES directly.

Families and Children Study

ESDS Page: www.esds.ac.uk/findingData/snDescription.asp?sn=4427

ESDS guide: www.esds.ac.uk/longitudinal/access/facs/l4427.asp

Overview: The *Families and Children Study* (FACS), formerly known as the *Survey of Low Income Families* (SOLIF), originally provided a new baseline survey of Britain's lone-parent families and low-income couples with dependent children. The survey was named SOLIF for Waves 1 and 2, and FACS from Wave 3 onwards.

The FACS study has become a 'true panel', whereby 1999 respondents have been re-interviewed in subsequent annual waves from 2000 to 2005, and new families added in each of these years, to allow a representative cross-section as well as longitudinal comparisons. Starting with Wave 3 (2001) the survey was extended to include higher-income families, thereby yielding a complete sample of all British families (and the subsequent name change). From Wave 4 (2002) onwards, longitudinal comparisons can now be made.

The main objectives of the survey are to:

- evaluate the effectiveness of the Government's work incentive measures in terms of helping people into work, improving living standards and improving child outcomes
- compare the living standards and outcomes for children and for families across the income distribution
- compare changes in the above across the waves since 1999
- FACS also aims to provide commentary on longer-term objectives such as the Government's Public Service Agreement to eradicate child poverty within a generation.

For the seventh edition, data and documentation for Wave 7, and some documentation pertaining to all waves of the survey, were added to the study. Further information, including links to reports and other publications, may be found on the DWP [FACS](#) web pages and the National Centre for Social Research [Families and Children Study](#) web pages.

Sample Attrition and Weighting: The FACS contains non-response weights for longitudinal research. Prior to the 2001

wave, FACs contains a representative sample of 'low to moderate income families', defined in relation to benefit receipt of Family Credit, Working Family Tax Credit (WFTC), or Working Tax Credit. Further details are provided (see above link for online documentation). For waves 2001 to 2005, a larger sample reflecting all families (and not just low-moderate income ones) is represented. Consequently, longitudinal weights reflect these sample differences. In the 1999-2005 dataset, there is a 'longitudinal weight for all families' which can be used for looking at waves 3-6 (2001 onwards) with a base of all families with dependent children for the 2001 sample (fLWAF). A further 'longitudinal weight for original families' (fLWOF) is included providing a panel weight for the original families from the 1999 dataset (i.e. low to moderate-income families).

3.6 The GHS-Longitudinal/EU-SILC

Overview The General Household Survey (GHS) is a multi-purpose annual survey dating from 1991 to present. It is a continuous survey based on an achieved sample of between 8,000 and 13,000 households in Great Britain. The 2005-2006 GHS fieldwork is the first to be undertaken under a new longitudinal survey design (GHS-L). The new design is also accompanied by a slight change in substantive emphasis, which leans towards a greater range of questions on social exclusion. The changes to the GHS aim to satisfy new EU requirements for the EU-SILC to provide a for year rotating panel sample in which a quarter of the sample is replaced in each year. Three quarters of the sample will therefore overlap between successive years. The overall sample size has been increased (from 8,700 in 2004 to 10,200 in 2005 – achieved households), although three-quarters of these are not new cases. Unlike the LFS, individual respondents are further followed when they change household address. This data is currently unavailable of Data Archive. Further details can be found through the following links:

3.7 Understanding Society

The [Understanding Society](#) study (also known as UK Longitudinal Household Study) is based at the Institute for Social and Economic Research (ISER) at the University of Essex, together with colleagues from the University of Warwick and the Institute of Education.

Key features include:

- a total target sample size of 40,000 households/100,000 individuals
- an ethnic minority booster sample of over 3,000 households
- incorporation of the British Household Panel Survey (BHPS)
- interviews from all household members, aged 10 and above
- topic coverage relevant to a wide range of disciplines and policy fields

- links to supplementary data, such as neighbourhood information
- the collection of health indicators and biomarkers
- a platform for the collection of qualitative data
- an *Innovation Panel* for methodological research.

See the above link for updates and news on the development of this survey.

3.8 Non-UK Longitudinal Datasets

A list of [Non-UK longitudinal datasets](#) is given on the ESRC Longitudinal Data Analysis for Social Science Researchers web site.

3.9 Organising Data for Panel Analysis using Stata

Panel data can be organised in *wide* and *long format*. In wide format (figure 3.3), one line represents an individual case (indicated by an id variable) and the numerical suffix to the other variables indicates which wave of the survey the variable represents information for (e.g. varx1 if variable varx from wave 1). In the below example, we have three different cases (id) and a variable named 'varx' at waves 1 to T of our dataset. Time is thus represented across the columns.

Figure 3.3 Wide Format Panel Data

Id	varA1	varA2.....	...varAT	varB	VarB.....	...varBT
0001	2	7	6	1	1	2
0002	3	5	8	2	1	2
0003	7	4	5	1	1	2

The wide data format is useful for transition tables, which can easily created by cross-tabulating different years of the same variables (e.g. varA1 varA2).

In *long format data*, cases for multiple waves for the same individual are represented on different rows of the dataset (figure 3.4). The information represented in figure 3.4 is identical to that in figure 3.3, except there is an additional variable (here called 'wave') that indicates the wave to which the case belongs. The id variable indicates which respondent the case belongs to. In fig 3.4, there three individuals (idvar 0001 to 0003) which have 3 waves of data each (wave 1 to 3) Given that only one year of data is represented on a single row, it is no longer necessary to pre-fix the other variables to indicate their wave number. In long format time is thus represented down the rows.

Figure 3.4 Long Format Panel Data

Id	Wave	varA	varB
0001	1	2	1
0001	2	7	1
0001	3	6	2
0002	1	3	2
0002	2	5	1
0002	3	8	2
0003	1	7	1
0003	2	4	1
0003	3	5	2

Long format data is the desired format for panel regression models in software packages such as Stata (see `xstset` and `xt` “cross-sectional time-series” commands) where the variable indicating time (wave) is specified when running models. Long format data is also used for discrete time event history analysis where multiple rows of data are constructed for each respondent, indicating specific time intervals.

A further distinction commonly made is between **balanced** and **unbalanced** datasets. Balanced panel datasets refer to where values for each respondent are observed for each wave of the panel. Unbalanced panel datasets occur where there is not full information for every individual for every wave. This may be because of sample attrition, or in the case of international panel datasets, because some countries began conducting a panel at a later stage than other countries. The following examples focus on balanced panel datasets¹⁹.

Creating Long and Wide Datasets

In many datasets (such as the BHPS), panel data for secondary analysis are provided as separate wave specific files. It is therefore necessary to combine datasets together prior to analysis, and often using multiple files for each wave. It is useful to consider the format of data you will require before joining files. If you require information from multiple waves to derive variables, it can often be more convenient to construct datasets in wide format first then convert them to long format. This is because in long format, you will have the information for every wave for each respondent on the same row of data, so can

¹⁹ See Baltagi and Hueng-Song (2006) for a discussion of unbalanced panel data.

use standard variable recode and replace procedures to create variables. Some software such as Stata also provides commands for switching between different formats.

Long format datasets in Stata are created using the *append* command. When creating long format balanced panel datasets you will need to:

- ensure identical panel samples are selected in each year of your data
- create variables consistent in name, coding, and routing for each of your covariates and dependent variables prior to appending files.
- create a wave identifier variable prior to merging your dataset
- join together data files using the *append* command.

The following example omits the first stage. It creates a dataset with one variable for housing tenure (*wtenure*) with the household identifier (*pid*), and a created wave indicator (*wave*) from BHPS individual file (*wINDRESP*).

```
use pid atenure using aindresp.dta
```

```
*create wave indicator:
```

```
generate wave= 1
```

```
*create a variable consistent in name over time for tenure
```

```
generate xwtenure = atenure
```

```
*(n.b your variable will need value labels adding in for some variables, recoding for consistency)
```

```
save xxatemp
```

```
clear
```

```
use pid btenure using C:\DATA\stata6\bindresp
```

```
generate wave= 2
```

```
generate xwtenure = btenure
```

```
*append the two datasets:
```

```
append using xxatemp
```

```
*finally sort your data by pid and wave so you visual inspect
```

```
sort pid wave
```

It is useful to check samples sizes to ensure you have not accidentally added unwanted cases.

Wide format panel datasets can be created in Stata using the *merge* command. It is again important to be mindful of the sample selection prior to merging to ensure you have only the cases you require. The present example omits this stage.

*load the data:

```
use pid atenure using aindresp.dta
```

*sort the datasets by the variable that will be used to identify individual cases (pid):

```
sort pid  
save atemp  
clear
```

*load the second dataset, sort, and then merge:

```
use pid btenure using C:\DATA\stata6\bindresp
```

```
sort pid  
merge using atemp  
sort pid wave
```

A system variable “_merge” is created when using the merge command. This tells you from which datasets individual cases come.

```
ta _merge
```

Tabulating _merge gives a table in which 1 represents cases present in the master dataset only, 2= cases present in the ‘using’ dataset only (that called upon by the ‘using’ command, and 3 represents cases that are present in both datasets. This variable is important for checking you have selected the appropriate cases from each wave. You can also use the _merge variable to drop cases that you do not require. So for example, if we needed to drop cases which were not present in both waves:

```
drop if _merge ~=3
```

If you plan to add successive waves to the datafile, then you will need to either rename or drop the `_merge` variable, otherwise an error message will appear during merging telling you that the variable `_merge` already exists.

When creating either long or wider format datasets, it is important to closely monitor the sample size whilst files are being merged. Useful Checks:

- If you are creating a balanced wide format panel dataset and have selected the appropriate identical samples in each wave prior to merging, the number of cases following the merge should not increase.
- If you are creating a balanced panel dataset in long format and have selected identical samples in each wave prior to merging, use a variable without missing values (such as an id variable) and divide the number of observations after merging files by the number of waves. This number should represent your original sample size.

Converting between long and wide data formats

Stata provides commands for shifting between long and wide format data using the **reshape** command. The syntax for reshaping from long to wide format is:

```
reshape wide stubnames, i(varlist) j(var) [options]
```

The stub names portion of the command is replaced by the list of variables to be reshaped. When the data is reshaped, Stata will add a suffix to the variables depending on the value of the wave variable. Thus `xvar` from wave 1 will now be called `xvar1`. It is therefore important to rename variables which already end in a number prior to reshaping, as if you shift back into long format, Stata will take this numerical ending as part of the wave number.

The following command:

```
reshape wide varx, i(id) j(wave)
```

...will reshape this:

Id	Wave	Varx
0001	1	2
0001	2	7
0001	3	6
0002	1	3

0002	2	5
0002	3	8
0003	1	7
0003	2	4
0003	3	5

..into this....

Id	Varx1	varx2.....	...varxn
0001	2	7	6
0002	3	5	8
0003	7	4	5

This process can be reversed through the following command:

```
reshape long varx, i(id) j(wave) [options]
```

Note that the j(wave) specifies the name for the variable which indicates which wave a case comes from.

3.10 Worked example: Using *vector* and *loop* commands in SPSS to create a measure of attitude stability²⁰

Often, users of longitudinal data will want to create a variable that denotes a respondent has made a transition from one state to another and/or perhaps how long a respondent has spent in one particular state. To create such a variable, the user would need to incorporate the responses to the same question from at least two – and ordinarily more – waves of data.

One example might be that a BHPS user is interested in constructing a variable that measures attitude/opinion stability – essentially the ‘transition’ here is from one point of view to another. Suppose that the user has five waves of data in which a particular question was asked and they would like to construct a variable that indicates whether a respondent’s opinion appears to be stable (zero or one transitions) over time or whether a respondent’s opinion appears to be unstable (two or more transitions).²¹

²⁰ Many thanks go to Jack Kneeshaw, ESDS Longitudinal, for providing this section.

²¹ This research question falls within the ‘non-attitudes’ or ‘pseudo-opinions’ debate. While many opinion measures appear stable over time at the aggregate level, they may hide a good deal of ‘churn’. According to one line of thinking (see Page and Shapiro’s [The Rational Public](#)), this churn is mere ‘oscillation’ – respondents switch opinion

In SPSS, one can create such a dummy variable where 0 = stable²² and 1 = unstable by putting together information gathered at each of the five waves of data. This can be done using the *vector* and *loop* commands. Below is the syntax used to create such a dummy variable, including the matching of files and the recoding of the five original variables, in this case WOPSOCD²³:

* Retrieve individual-level data files for each wave, save only the key variable and the attitude variable in five new files called 'statew', all saved in a folder called 'nonattitudes':

```
get file='s:\bhps\spss\aindresp.sav'  
  /keep=pid,aopsocd .  
sort cases by pid .  
save out='m:\nonattitudes\statea.sav' .
```

```
get file 's:\bhps\spss\cindresp.sav'  
  /keep=pid,copsocd .  
sort cases by pid .  
save out='m:\nonattitudes\statec.sav' .
```

```
get file 's:\bhps\spss\eindresp.sav'  
  /keep=pid,eopsocd .  
sort cases by pid .  
save out='m:\nonattitudes\statee.sav' .
```

```
get file 's:\bhps\spss\gindresp.sav'
```

between 'agree strongly' and 'agree' or 'neutral' and 'disagree' – and is of no great concern. Another explanation is that aggregate stability hides a significant amount of 'non-attitudes' or 'pseudo-opinions' – respondent reports that are not oscillations around a long-term position but actually resemble random switching (e.g. successive opinion reports resemble sequences similar to those produced by die throwing or coin tossing).

²² This example is based on an opinion measure which used a 5-point Likert scale ranging from 'agree strongly' to 'disagree strongly'. So that one can identify unstable/irrational opinion 'switchers' (rather than mere 'oscillators'): (1) the scales were collapsed into 3 categories (agree/neutral/disagree) to allow respondents to switch from 'agree strongly' to 'agree' without being coded as a switcher; (2) respondents had to switch not only from one side to another (e.g. 'agree' > 'disagree') but also back again over the course of the five waves (i.e. 'agree' > 'disagree' > 'agree' or 'disagree' > 'agree' > 'disagree').

²³ This variable is based on the question "Which answer off the card comes closest to how you feel about the following statement? Major public services and industries ought to be in state ownership: strongly agree/agree/neither/disagree/strongly disagree". See <http://www.iser.essex.ac.uk/ulsc/bhps/doc/volb/wave1/aindresp10.php#AOPSOCE>.

```
/keep=pid,gopsocd .  
sort cases by pid .  
save out='m:\nonattitudes\stateg.sav' .
```

```
get file 's:\bhps\sps\jindresp.sav'  
/keep=pid,jopsocd .  
sort cases by pid .  
save out='m:\nonattitudes\statej.sav' .
```

* Match the five statew files, including adding a variable that indicates whether respondent was present in that wave (e.g. /in=w1), to be used later. Assign missing values and save merged file as multiwave.sav:

```
match files file='m:\nonattitudes\statea.sav' /in=w1  
/file='m:\nonattitudes\statec.sav' /in=w2  
/file='m:\nonattitudes\statee.sav' /in=w3  
/file='m:\nonattitudes\stateg.sav' /in=w4  
/file='m:\nonattitudes\statej.sav' /in=w5  
/by=pid .  
missing values all (-9 thru -1) .  
save out='m:\nonattitudes\multiwave.sav'  
/keep=pid,w1,w2,w3,w4,w5,aopsocd,copsocd,eopsocd,gopsocd,jopsocd .
```

* Using multiwave file, collapse the 5-value Likert measures for wopsocd into 3 values: 1=agree, 2=neither, 3=disagree. Rename the wopsocd variables, op1 to op5:

```
get file='m:\nonattitudes\multiwave.sav' .  
recode aopsocd copsocd eopsocd gopsocd jopsocd (1,2=1) (3=2) (4,5=3)  
(else=copy) into op1 op2 op3 op4 op5 .
```

* Create a vector called 'op' using the variables op1 to op5. Then compute temporary variables (e.g. #varname) called #eva (short for **ever agree**) and #evad (**ever agree then disagree**) and set all values to 0. Compute a new variable (not temporary this time) called evada (**ever agree then disagree then agree**) and set all values to 0.

* Next, use the loop command to create a temporary index variable - #i - that will run through the vector variables op1 to op5. Upon encountering a value of 1 (agree) for the first time in its loop through op1 to op5, SPSS will assign a value of 1 to the temporary variable #eva.

The loop will continue until a value of 3 (disagree) is encountered in a later `opn` variable – here SPSS will assign a value of 1 to the temporary variable `#evad`. Finally, the loop will continue either until a second value of 1 is encountered (`evada=1`) or the agree/disagree/agree pattern is not found after the loop has completed its run through `op1` to `op5` (here `evada=0`, as per the earlier command `compute evada=0`). For our purposes, where `evada=1`, the respondent has provided an unstable set of opinions.

```
vector op=op1 to op5 .
compute #evA=0 .
compute #evAD=0 .
compute evADA = 0 .
loop #i=1 to 5 .
. if op(#i)=1          #evA=1 .
. if #evA=1 and op(#i)=3 #evAD=1 .
. if #evAD=1 and op(#i)=1 evADA=1 .
end loop if evADA=1 .
```

* As per the comments above, we now create a variable `evdad` where a value of 1 equals a set of opinions that include the pattern disagree/agree/disagree.

```
vector op=op1 to op5 .
compute #evD=0 .
compute #evDA=0 .
compute evDAD = 0 .
loop #i=1 to 5 .
. if op(#i)=3          #evD=1 .
. if #evD=1 and op(#i)=1 #evDA=1 .
. if #evDA=1 and op(#i)=3 evDAD=1 .
end loop if evDAD=1 .
```

* Using `evada` and `evdad`, we create a temporary variable called 'switchno' which indicates whether a respondent has either an agree/disagree/agree or a disagree/agree/disagree pattern, both, or neither. We then recode `switchno` to create the variable 'switcher' which has a value of 1 where either or both of `ada` or `dad` are present (i.e. unstable) and a value of 0 where this is not the case.

```
compute switchno=evada+evdad .
recode switchno (1,2=1) (else=copy) into switcher .
```

* Finally, having constructed our principal variable of interest - switcher - we undertake an exploratory frequency calculation. In doing so, we run the following syntax to select only those respondents present at all 5 waves for which we have data (e.g. everpres=5).

```
compute everpres=w1+w2+w3+w4+w5 .  
temp .  
sel if everpres=5 .  
freq switcher .
```

Appendix A. Bibliography and Further Resources

References on Repeated Cross-sectional Analysis

Dale, A. and Davies, R. (eds) *Analysing Social and Political Change: A casebook of methods*, London: Sage.
Alison, P.D. (1984) *Event History Analysis*, Beverly Hills: Sage.

Micklewright, J. (1994) "The Analysis of Pooled Cross-Section data," in R. Davies and A. Dale (eds.) *Analysing Social Political Change: A Casebook Of Methods*, London: Sage

Ryder, N.B. (1965) "The Cohort as a Concept in the Study of Social Change," *American Sociological Review*, vol. 30, pp.843-681.

References on Panel Regression

Arulampalam, W. et al (2000) "Unemployment Persistence," *Oxford Economic Papers*, vol. 52, pp. 24-50 (example of dynamic model using lag variables, discussion of initial conditions problem).

Arulampalam, W. (2001) "Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages," *The Economic Journal*, vol. 111, no. 4, F585-F606.

Greene, W. H. (2003) *Econometric Analysis (5th Edition)*, Upper Saddle River: Prentice Hall.

Hagenaars (1990) *Categorical Longitudinal Data: Long-linear, Panel, Trend, and Cohort Analysis*, London: Sage

Hausman, J.A. (1979) "Specification Tests in Econometrics," *Econometrica*, vol. 46, no. 6, pp. 1251-1271.

Hsiao, C. (2003) *Analysis of Panel Data*, Cambridge: Cambridge University Press.

Jenkins, S.P. (2000) "Modelling Household Income Dynamics," *Journal of Population Economics*, vol. 13, pp. 529-67.

Mundlak, Y. (1978) "On the Pooling of Time Series and Cross Section Data," *Econometrica*, vol. 46, no. 1, pp. 69-85.

Woolridge, J. (2002) *Econometric Analysis of Cross-Section and Panel Data*, Cambridge Mass.: MIT Press.

References on Event History Analysis

Baltagi, B.H. and Heun Song, S. (2006) "Unbalanced panel data: A survey," *Statistical Papers*, vol. 47, pp. 493-523.

Barlow, E. et al (1963) "Properties of Probability Distributions with Monotone Hazard Rate," *Annals of Mathematical Statistics*, vol. 34, no. 2, pp. 375-389.

Blossfeld, H.P. (2001) *Event History Modelling: New approaches to causal analysis*, New Jersey: LEA.

Blumen, I. et al (1955) "The Industrial Mobility of Labor as a Probability Process," *Cornell Studies in Industrial and Labor Relations*, No. 6, Ithaca: New York.

Cox, D.R. (1972) "Regression Analysis and Life Tables (with discussion)," *Journal of the Royal Statistical Society, Series B*, vol. 34, pp.187-222.

Crouchley and Oskrochi (2001a) "Using SPSS and Gauss to unify the BHPS Work History Data," CASS Working Paper No. 2001/03, Applied Statistics, University of Lancaster.

Elliot, J. (2002) "The Value of Event History Techniques for Understanding Social Processes: Modelling women's employment behaviour after motherhood," *International Journal of Social Research Methodology*, vol. 5, no. 2, pp107-132.

Hosmer, D.W. and Lemeshow, S. (1999) *Applied Survival Analysis: Modelling of time to event data*, New York: Wiley.

Jenkins, S.P. (1995) "Easy Estimation Methods for Discrete-time Duration Models," *Oxford Bulletin of Economics and Statistics*, vol. 57, pp. 129-37.

Jenkins, S.P. (2000) "Modelling Household Income Dynamics," *Journal of Population Economics*, vol. 13, pp. 529-67.

Yamaguchi, K. (1991) *Event History Analysis*, Beverly Hills: Sage.

References on Missing Data and Sample Attrition

Crouchley, R. and Oskrochi, G. (2001b) "A Random Effects Treatment of Dropout in Multi-spell Multi-state Labour Market Panel Data," CASS Working Paper No. 2001/02, Applied Statistics, University of Lancaster.

Little, R.J. (1993) "Pattern Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, vol. 88, pp. 125-134.

Little, R.J. & Rubin, D. (1987) *Statistical Analysis with Missing Data*, New York: Wiley.

Rubin, D.B. (1976) "Inference and Missing Data," *Biometrika*, vol. 63, pp. 581-592.

References on Seam effects, calendar-based interviewing and dependent interviewing

Belli, R. F. et al (2007) "Methodological Comparisons Between CATI Event History Calendar and Standardized Conventional Questionnaire Instruments," *Public Opinion Quarterly* vol. 71(4), pp. 603-622.

Jäckle, A. (2008) "The Causes of Seam Effects in Panel Surveys," *ISER working paper series*, 2008-14:
<http://www.iser.essex.ac.uk/publications/working-papers/iser/2008-14.pdf>.

BHPS User Guide

Taylor, Marcia Freed (ed). with John Brice, Nick Buck and Elaine Prentice-Lane (2007) *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. Colchester: University of Essex.

ESDS Longitudinal Resources <http://www.esds.ac.uk/longitudinal/resources/introduction.asp>

Appendix B. Pooling Labour Force Survey datasets

It is possible to combine LFS files to create repeated cross-sectional datasets. Methods for undertaking such pooling were discussed in Section 2. Given the complicated data structure of the UK Labour Force Survey, this section gives an overview of specific ways in which the LFS can be combined. Prior to reading this Appendix, it is recommended that the descriptions of the UK Labour Force Survey in Section 2 and Section 3 are reviewed.

A number of different strategies can be taken to pool LFS data. An important variable when pooling is THISWV, indicating which wave a respondent occupies. This can be used to define and merge subsets or a given survey quarter:

- If you are combining data from survey for **four** quarters, you can select those in their first or fifth wave. These will represent separate, independent samples (i.e. select w1 and w5 from quarters 1, 2, 3 and, 4).
- If you wish to combine quarters for longer periods than four quarters, you can select those in their first wave only. No matter how many datasets you merge, you will have only one record per respondent if you select only wave 1 cases from each quarter survey.
- When combining data over several years, an alternative approach is to take a particular quarter of the year (e.g. spring quarter) and to drop wave 5 responses to avoid replication with wave 1 year respondents from the previous year. In this case, four waves contained in a single (same) quarter for each year are used. This will reduce the data manipulation effort involved.

In addition to THISWV, a number of further variables are important for the pooling of datasets. The variable WAVFND in the LFS indicates the wave in which a household was first found. The LFS samples addresses and so does not follow respondents to a new address if they move address between waves. Instead, the new residents of their old address are sampled. Households which have replaced those in the original sample will therefore take a value of WAVFND>1, and so can be identified accordingly.

The population base for the LFS consists of residents of private households, residents of NHS accommodation, and students living away in halls of residence. Those in NHS accommodation are enumerated in this accommodation and can be identified using the NURSE variable (where NURSE = 1). Students in halls of residence are captured at any private address from which they are absent during term time, and are often proxy respondents as they live away. Since Summer 1996, students living away have been identified by the variable HALLRES. Prior to 1996, students were identified with their person number (PERSNO). A number greater than 90 indicated that the respondent was a student in a hall of residence.

B.II Creating unique address and individual identifiers in pooled datasets (Special Licence Only)

In many of the LFS datasets, addresses are identified with the variable **REMSERNO**, although this variable is not available in all datasets. This variable is calculated from a number of administrative variables:

In SPSS

```
compute remserno =  
(quota*1000000000)+(week*10000000)+(wlyr*1000000)+(qrtr*100000)+(add*1000)+(wavfnd*100)+hhld.
```

In Stata:

```
ge double remserno  
=(quota*1000000000)+(week*10000000)+(wlyr*1000000)+(qrtr*100000)+(add*1000)+(wavfnd*100)+hhld
```

Where week= week of interview; wlyr= year first interviewed; qtr= survey quarter, add= address number on interviewers list; wavfnd= wave first found in, and hhld = the address at which a case refers to on a specific interviewers list (this is because more than one household can live at a single address).

A derivation for an individual identifier (CASENO) is also often used, which expands the above derivation to include QUOTA (stint number when interview took place) and PERSNO (person number within household):

```
caseno=(quota*1000000000000)+(week*10000000000)+(wlyr*100000000)+(qrtr*1000000)+(add*100000)+(wavfnd*10000)+(hhld*100) +(recno)
```

N.B. in pooled datasets, these two derived variables do not necessarily uniquely identify addresses or individuals in different survey quarters. To understand why, consider the following breakdown of the manner in which REMSERNO is calculated. This shows how the above calculation assigns each of the values of the included variables to a different portion of a longer number, which contains 8 to 9 digits. The front of the number is contributed by WEEK (either 1 or 2 digits), the next portion by W1YR, then by QTR, and so forth, until HHL D contributes the last portion of the number:

Remserno 110223110=

	Week	W1yr	Qtr	Add	Wavefnd	Hhld
Max Digits	2	1	1	2	1	2
Range	1-13	0-9	1-5	1-80	1-5	1-17
e. g. number	11	0	2	23	1	10

In terms of a unique identifier in pooled datasets, this derivation, and the CASENO variable, as formulated above, have two problems:

- Only 1 digit is assigned in REMSERNO to represent the W1YR value. Thus if we have 1 digit for W1YR and are pooling data from before and after the year 2000, we can not distinguish between many pairs of years (for example, 1992 and 2002 would both have a digit 2 for the W1YR portion or REMSERNO, as would 1993 and 2003, and so forth). **As a result of this millennium bug, the above calculation does not necessarily uniquely identify where data from before and after 2000 are being pooled.**
- A second problem is that W1YR changes in terms of the number of digits it uses to record the year of first interview. For example, in 1992, W1YR records this information as four digits (1992), whereas in other waves it records it as one digit. In records where remserno has 4 digits, it will affect the value of the portion of the REMSERNO which is supposed to be determined by week, raising the potential for non-unique identifier numbers to be created. CASENO faces the same issues.

To overcome these problems:

- 1. Recode your W1YR variable so that it gives two digits, uniquely identifying the year in which a respondent was first interviewed (so 1992=92, 2005= 05 etc)**
- 2. Add an additional 'space' through an extra zero digit to the original REMSERNO derivation to allow two digits for your recoded W1YR variable.**

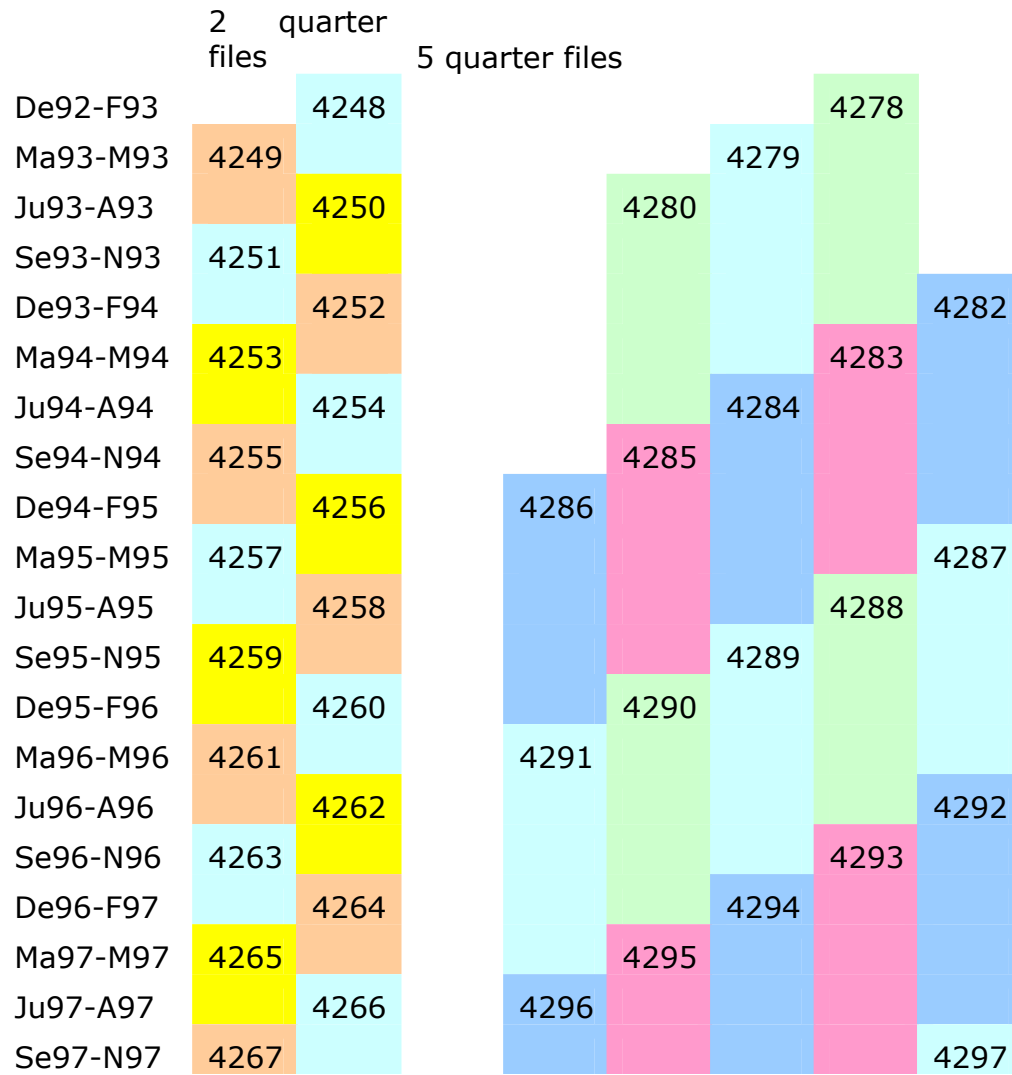
In the below illustration, the variables created in this manner are referred to as AID (address ID), and PID (Personal Identifier)

```
aid = (quota*10000000000)+(week*100000000)+(wlyr*1000000)+(qrtr*100000)+(add*1000)+(wavfnd*100)+hhld
```

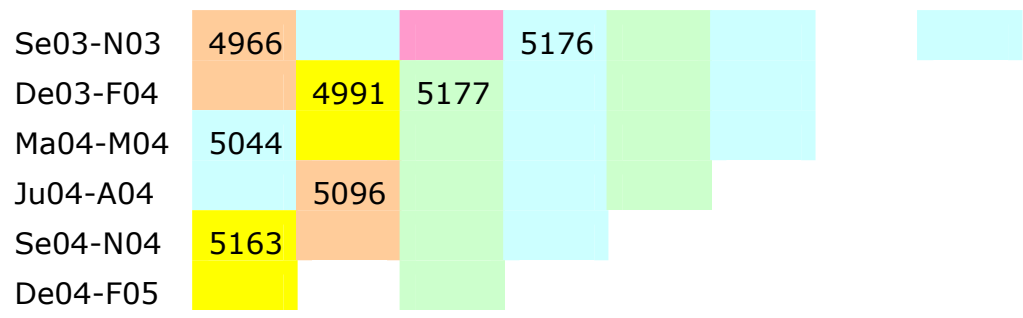
```
pid=(quota*1000000000000)+(week*10000000000)+(wlyr*100000000)+(qrtr*10000000)+(add*100000)+(wavfnd*10000)  
+(hhld*100) +(recno)
```

N.B. until Summer 1996 two different variables have captured person number within household; PERSNO and RECNO. Since Summer 1996 RECNO has become identical to PERSNO (since HALLRES was established to identify students living in halls of residence). PERSNO was intended solely to identify students living in halls of residence.

Appendix C. Pre-prepared Longitudinal LFS Files – UK Data Archive Serial Numbers



De97-F98		4268					4298
Ma98-M98	4269				4299		
Ju98-A98		4270					
Se98-N98	4271			4301			
De98-F99		4272					4302
Ma99-M99	4273					4303	
Ju99-A99		4274			4304		
Se99-N99	4275						
De99-F00		4276		4656			
Ma00-M00	4277						4657
Ju00-A00		4671					4658
Se00-N00	4672				4659		
De00-F01		4673					
Ma01-M01	4674			4661			
Ju01-A01		4675					4662
Se0-N01	4676						4663
De01-F02		4677			4670		
Ma02-M02	4678					4725	
Ju02-A02		4679		4768			
Se02-N02	4669		4807				4989
De02-F03		4724					
Ma03-M03	4796					4990	
Ju03-A03		4806				5045	



A comprehensive list of Longitudinal LFS dataset can be found on the ESDS [longitudinal LFS](#) web pages.

Appendix D. Variables Contained in the GHS Time Series Dataset

See: www.esds.ac.uk/findingData/snDescription.asp?sn=5664

Variable Period available:

Tenure 1972-2004
Demographics Country of birth 1972-2004
Father's country of birth 1972-2004
Mother's country of birth 1972-2004
Socio-economic group of HOH/HRP 1972-2004
NS-SEC of HRP 2001-2004
Age 1972-2004
Sex 1972-2004
5-year estimated birth cohort 1972-2004
Marital status 1972-2004
Household type A (Grouped) 1974-2004
Household type F 1984-2004
Labour market DV for ILO in employment - 3 categories 1972-2004
Education Education Level 1972-2004
Full Time Student Status 1972-2004
Car Ownership Number of cars 1972-2004
Health Health on the whole in last 12 months 1977-2004
Any long standing illness or disability 1973-2004
If longstanding illness limits activity 1973-2004
Illness/injury reduce activity 1972-76, 1979-2004
Consulted doctor last 2 wks (exc.hosp) 1972-2004
Number of consultations 1972-2004
Hospital outpatient attend - last 3mths 1972-2004
In hospital as day patient in last year 1992-2004
Hospital in patient in last year 1972-2004
LLSI OR NON-LLSI 1972-2004
IF LLSI or restricted activity 1972-2004

Smoking Number of cigarettes smoked per day 1972-2004

Current smoking status 1972-2004

Go without smoking for whole day 1992, 1994, 1996-2004

Like to give up smoking altogether 1992, 1994, 1996-2004

When first cigarette of day 1992, 1994, 1996-2004

Age started smoking regularly 1972-73, 1980-2004

Number of smokers in the household 1972-2004