Analysing change over time: repeated cross sectional and longitudinal survey data

UK Data Service

Author: UK Data Service
Updated: February 2015
Version: 1.0

Contents

Figures

1. Introduction

Quantitative analysis of change over time requires data with a time element to it. There are several different kinds of data with a time element and it is important to understand the differences between these kinds of data to know how to analyse them. This guide outlines the main types of these kinds of data available via the UK Data Service. It gives a brief overview of the key elements that you must consider when using the different kinds of data and some of the methods commonly used to analyse these kinds of data.

This guide is an introductory guide to the main types of data available and some commonly-used methods for studying change over time quantitatively. There are many books and articles about these and other methods. For more detail about analysing change over time, see the references in the bibliography.

2. UK Data Service data for analysing change over time

There are three main types of survey microdata available from the UK Data Service which can be used for analysing change over time[1]. Microdata are individual or household level data (as opposed to data aggregated by region or country). The three main types are:

- Repeated cross-sectional survey data: data in which the same (or similar) information is asked to a different sample of individuals each time - the samples can then be compared over time

- Longitudinal survey data: data in which the same information is asked of the same group of individuals over time (with new respondents added to maintain numbers)

- Retrospective survey data: data in which respondents are asked about their memory of the past e.g. 'What was your father's job when you were 14 years old?'

The following sections outline the main features of these different kinds of data.

2.1 Repeated cross-sectional surveys

Cross-sectional survey data are data for a single point in time. *Repeated* cross-sectional data are created 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. See Section three for examples of analyses of repeated cross-sectional data.

You can find the largest and most commonly used of these datasets in the Key Data section of the UK Data Service website under the UK Surveys tab. Alternatively, you can explore all repeated cross-sectional data in the UK Data Service Data Catalogue by selecting UK Surveys as the Data type.

Example: British Social Attitudes Survey

The British Social Attitudes Survey (BSA) is a repeated cross-sectional survey conducted in most years since 1983. It is designed to provide evidence about attitudes to a range of topics about Britain and the way the country is run. A new sample of respondents is selected every time the survey is run but many of the same questions are asked at each time-point. In this survey, one person in each selected household is interviewed.

The individual level data are available to be downloaded from the UK Data Service into a statistics package like SPSS, R or Stata.

Because repeated cross-sectional data take a different sample of a population over time, they are used for analysing population or group changes over time (also known as aggregate

[1] The UK Data Service also supplies macrodata which is data aggregated by country or region, and census data, both of which can also be used for examining change over time.

change over time). They cannot be used to look at individual change[2]. 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 years[3].

2.2 Longitudinal survey data

Longitudinal studies involve information directly gathered in a survey of households or individuals. Surveys following individual persons over time can be of two types, panel data or cohort data. There are a number of studies of these kinds in the UK[4].

2.2.1 Panel surveys

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 wave 3, 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. Household panel surveys rely on an initial random sample of households or individuals which is randomly selected and subsequently followed over time, with or without households being replaced as they drop out of the sample.

You can find the largest and most commonly used of these datasets in the Key Data section of the UK Data Service website under the Longitudinal studies tab or explore all longitudinal data in the UK Data Service Data Catalogue by selecting 'Cohort and longitudinal studies' as the Data type.

[2] Longitudinal studies are used to analyse change over time – see Section 2.2 for more details.
[3] Many repeated cross-sectional surveys include some retrospective questions which give information on past experiences or characteristics – see Section 2.3.
[4] See the website http://www.closer.ac.uk/ for an overview of the main longitudinal datasets in the UK

Example: Understanding Society

Also known as the UK Household Longitudinal Survey, Understanding Society is the largest household panel survey in the world, with about 40,000 households and 50,994 individuals followed yearly since 2009. This size allows for detailed longitudinal analysis to be carried out on subgroup of the UK population. All adults aged 16 and more are interviewed, with a special questionnaire filled by children aged 10 to 15.

The sample is made of a general population sample, a booster sample for ethnic minorities, as well as an innovation panel in which new questions and methods are experimented at each wave. It also incorporates respondents from the former British Household Panel Survey (BHPS) that has been running since 1991.

Data from Understanding Society can be linked with the respondents' administrative records in areas such as education, health or work and pensions or transport. Small area geographic identifiers can be used under special licence, thus enabling detailed spatial analysis using common political and administrative boundaries. Detailed Health data is directly collected by nurse for about one third of the sample received - 20000 respondents.

More information is available on the UKDS website or directly at https://www.understandingsociety.ac.uk/about/

Panel data provide better opportunities to track individual level change than repeated cross-sectional data. You can use panel data to track individual changes in income, health, family composition etc.

Panel data provide 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 and allow the researcher to distinguish between *age and cohort effects* on change (see 2.2.2 Cohort surveys for more about age, period and cohort effects).

2.2.2 Cohort surveys

In cohort surveys, respondents are followed from an identical point in their life onwards, often from birth. These surveys prove very useful to study child development over time, change in marital circumstances or health for instance among people belonging to the same generation.

You can find the largest and most commonly used of these datasets in the Key Data section of the UK Data Service website under the Longitudinal studies tab or explore all longitudinal data in the UK Data Service Data Catalogue and selecting 'Cohort and longitudinal studies' as the Data type.

The Millennium Cohort Study

The MCS 2000 is a survey of about 19,517 children born in 2000/01, and is the fourth[5] of its kind in the UK. It aims at being representative of all children born in 2000, hence its name and will follow them throughout their lives. The MCS includes data about among other parenting; childcare; cognitive development; health; parents' employment and education; income and poverty; housing, ethnicity. Five waves (also called 'sweeps') of data have been collected so far, at age nine months, three, five, seven and eleven, the last one in 2012. In order to attain usable sample size, additional data, also known as booster samples have been collected for children living in disadvantaged areas, from minority ethnic background, and those living in Scotland, Wales and Northern Ireland.

More information on the MCS can be found on its UK Data Service page or at the Centre for Longitudinal Studies.

The value of cohort studies is also that they allow the researcher to distinguish between *age and cohort effects* on change. Age effects are differences that happen as people age, and cohort effects are differences due to the different times at which cohorts were born for example. For example, there may be differences in health over the life course which remains the same for people born at different times. However, some differences are due to individuals being born into society at a different point in time, such as opinions about sex before marriage which people born earlier in the 20th century are likely to have different views about compared with cohorts born more recently.

Differences between age groups can reflect both age related effects such as life-course position and maturation, but also cohort differences, differences in the historical, social, economic, cultural, and technological contexts in which different generations have grown up and lived through.

Example of cohort surveys includes the National Child Development Study, the 1970 British Cohort Study, the Millennium Cohort Survey (MCS 2000) and the Longitudinal Survey of Young People in England (LSYPE).

2.3 Retrospective longitudinal survey data

Although less common, retrospective longitudinal data can also be collected by asking respondents in cross-sectional and longitudinal surveys about past events in their life. Most of the time these types of data are part of existing panel surveys. Retrospective data tend to be less reliable than data collected directly from respondent due to imprecision of recall, but they can provide useful insight and complement to information that is current at the time of interview. Examples include employment or partnership histories.

[5] The earlier cohort studies are the 1946 Medical Research Council (MRC) National Survey of Health and Development (NSHD) (not held by UKDS), and the 1958 National Child Development Study, 1970 British Cohort Study, both available from the UK Data Service. For more about the 1946 study, see their website: http://www.nshd.mrc.ac.uk/default.aspx

The BHPS Combined Work-Life History dataset

The British Household Panel Survey Work-Life History Data 1990-2005 is a consolidated dataset derived from the main BHPS, which gathers complete employment information about respondents.

It combines information provided at each wave by respondents about their current employment and job statuses as well as changes that occurred between two waves, both having been checked for consistency errors between 1990 and 2005.

It also includes retrospective labour market information which was gathered at wave 2 (employment history) and 3 (job history) of the survey. These gather all employment and job related events that occurred since respondents left full-time education.
Data is presented both as episode or calendar based. Each episode (i.e. holding a particular job, or experiencing a spell of unemployment) is characterised by a start, an end date, duration and sequence number), whereas calendar data have monthly records of a number of key labour market variables (such as occupation or industry).

The data currently covers the first 14 waves of the BHPS. The harmonised files can be merged with any other existing BHPS files.

The data and its documentation can be accessed from the UK Data Service Data Catalogue.

3. Analysing repeated cross-sectional data

3.1. Introduction

Repeated cross-sectional data refers to data 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 contains 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.

3.2. Limitations of cross-sectional research

There are a number of limitations of cross-sectional data (data at a single time point):

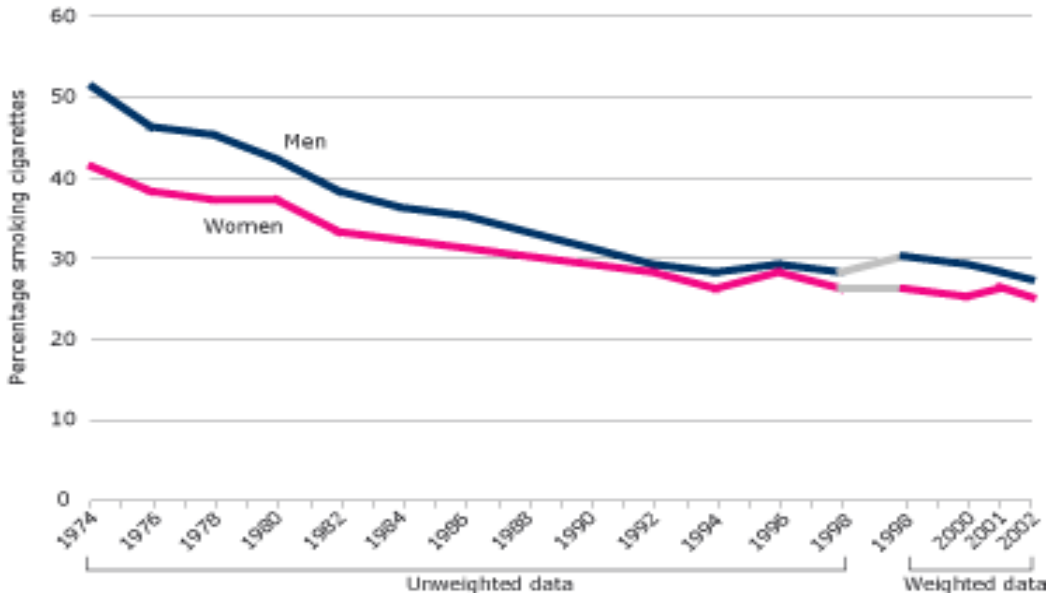
- A single cross-sectional data collection does not allow the analysis of change over time, whether at the *aggregate level* for populations or sub-groups (like repeated cross-sectional data can), or at the *micro-level* for examining individual change (like panel data can). Cross-sectional data provide 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 a*ge, cohort*, and *period effects* to be fully distinguished - however, see Section 3.3.2 for more information on what can be done in this respect. 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.

3.3. Example uses of repeated cross-sectional data

3.3.1. 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 General Household Survey across different years of the survey to consider changes in the number of people smoking over time (Figure 1). From Figure 1, we can see that percentage of men and women who smoked generally declined between the 1970s and 1990.

Figure 1: Trends in Smoking Behaviour by Sex, 1972-2002



Source: General Household Survey report, ONS. Data: GHS 1972-2002.

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 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.

3.3.2. Age, period and cohort 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:

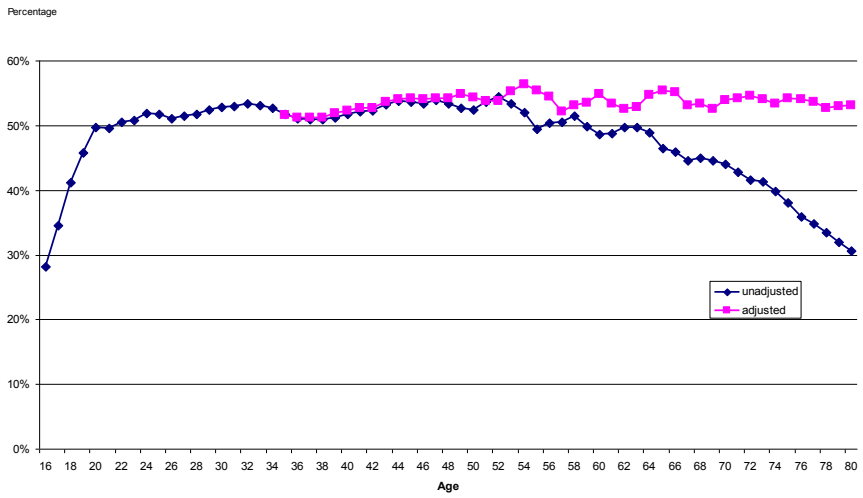
- Age: The time between birth and the observation period, Age may substitute for maturation/physical development; increases or decreases in intellectual capacities; personality development; life-stage etc.

- 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.
- 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.

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. The 1938 cohort would be young children at the start of the Second World War (1939-45) and the 1917 cohort would be the age of army conscription so these two cohorts are likely to have had different experiences of the Second World War (1939-45).

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



Data: General Household Survey

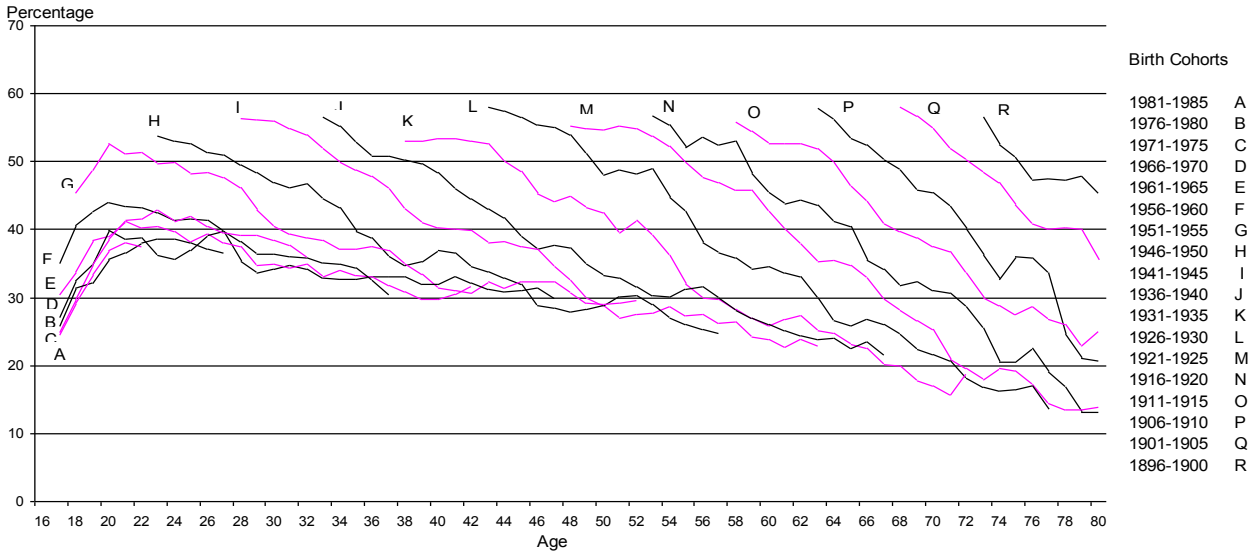
Through constructing *pseudo-cohorts*, repeated cross-sectional data can help to distinguish cohort and age effects by comparing age groups from pseudo cohorts for different years. To illustrate this, Figure 3 gives the hypothetical example of three cohorts of people, born in 1950, 1960, and 1970, for the year in which each cohort is aged 25 years.

Figure 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 4 presents the percentage of male smokers at the time of interview by pseudo-cohort[6]. The different lines labelled A to R represent birth cohorts ranging from 1981-1985 (cohort A) to 1896-1900 (cohort R).

Figure 4: Percentage of Male Smokers by Pseudo Cohort and Age (adjusted)



Data: General Household Time Series Dataset, 1972-2004

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.
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.

3.3.3. Other analyses

There are many other ways in which repeated cross-sectional data can 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 samples sizes are small in individual cross-sectional datasets (e.g.

[6] Analysis by Melissa Coulthard (ONS), age adjusted

ethnic minority groups). Larger sample size can also be desirable to increase the precision of statistical estimates by reducing standard errors[7]. Further methods that can be used on repeated cross-sectional data include time-series, multi-way tables, and log-linear models. More advanced usage of change over time using cross-sectional data include time series analysis. Time series analysis is specifically preoccupied with the analysis of aggregate change. This includes detangling its different components, namely: trends, seasonal effect, cyclical effect and random variations. Time series analysis can consist in modelling these trends using autoregressive regression model (also called ARMA), which are made necessary given that units of observations are not independent from each other. A major application of time series analysis consists also in forecasting future trends based on the observation of the past.

3.4. 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:

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.

[7] 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.

- 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 UK Data Service web site.

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 independent[8]. 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 ascertain which year observations come from in your combined dataset[9]. 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.

[8] 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.
[9] Although if you make this mistake, you may be able to ascertain this information from the unique identifier numbers.

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.

4. Analysing longitudinal data

4.1. Introduction

Longitudinal data are data in which the same information is asked of the same group of individuals over time. Additional respondents are normally added in each wave to replace respondents who have dropped out since the last wave.

This section provides a brief overview of a few common statistical techniques used to analyse change over time using longitudinal data. Longitudinal data have a number of advantages over cross-sectional data but also have additional features that need to be considered when analysing it (see Section 4.2 below).

As with cross-sectional data longitudinal analysis can be descriptive, that is focusing on simply describing change, or predictive, trying to model it alongside the factors that influence it. The choice of a suitable technique will depend on the focus of the study i.e. whether it is on trends in change or the occurrence of events, on the number of time points (two vs three or more), the units of observations (individual vs aggregate), as well as the measurement level (continuous vs discrete/categorical) of the data.

A key advantage of longitudinal data is that it enables researchers to analyse individual/micro-level change. One application of this is that the researcher can examine state dependence - where past event or states influences the probability of experiencing an event or state again in the future. So for example, a common finding in studies of unemployment is that people who have experienced past unemployment are more likely to experience further unemployment.

Another major application of panel data analysis has been to attempt to control for missing variables or other causes of residual that is unobserved 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.

4.2. Working with longitudinal data

There are a number of extra factors that need to be considered when analysing longitudinal data.

4.2.1. Independence of observations

In longitudinal analysis, observations do not equate individual respondents anymore. Instead, they record a measurement of a given outcome of interest for this respondent at a given time point. From the point of view of statistical analysis, this means that observations are not independent from each other anymore, as they are assumed to be in cross-sectional surveys. In particular, 'errors' (unobserved heterogeneity) are correlated since observations are made from the same individuals. As a result, the assumptions on which common techniques rely, for instance as is the case of traditional regression do not hold anymore and special techniques need to be used to take this into account.

4.2.2. Long vs wide format

Longitudinal data can appear in two distinct formats in statistical packages:

- In the 'long' format, each measurement occasion is recorded as an observation – that is a line in the data file. Imagine a four wave panel survey conducted every other year between 2002 and 2004. Each respondent has four lines of data, each one corresponding to one survey year. Occasion measurements can be matched to the corresponding respondents by way of an auxiliary identification variable which has the same value within respondents. The long format is usually required for event history or duration analysis, or for creating longitudinal graphs.
- Alternatively, in the 'wide' format, each wave of data is stored in a new variable, there is therefore one line or row of data per respondent. In the same example as above, a respondent´s job status in 2002 would be recorded in a variable JOBSTAT02 in 2002, JOBSTAT04 in 2004, and so on. In some surveys, a letter prefixes variable names identifies waves, where in others a suffix number is used for that purpose. In most surveys available at UKDS, data is presented in the wide format, and datasets can be converted – the technical term is *reshaped* -- into the long format from within mainstream statistical packages such as Stata, R or SPSS.

In addition, given that some longitudinal studies involve many waves of data, each wave tend to be available as a distinct file. Users can then match several or all waves together by merging their respective files. This is made possible by a unique Identification variable across datasets.

4.2.3. Balanced vs unbalanced panels

A further distinction commonly made is between balanced and unbalanced longitudinal datasets. Balanced panel or cohort 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 difference between balanced and unbalanced sample size can be substantial and matters for the statistical power of many analyses, given that some analysis only use respondents with complete records.

4.2.4. Bias and attrition

There are biases that occur in all surveys but some that apply specifically to longitudinal data (where individuals or households are followed over time).

Bias

Longitudinal datasets are subject to two main sources of bias: non-response and response error (misclassification).

Non-response bias

Non-response bias occurs due to different groups of people, for example by age or sex, having different probabilities of attrition (i.e. 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

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). This can happen in a cross-sectional survey but 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.

Attrition

Missing data[10] are traditionally classified as

- 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, if the missing data is neither MCAR nor MAR.

Although the assumption of MCAR and MAR in many contexts may be unrealistic, they can be considered as not involving sample bias. 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.

Using weights to minimise bias and attrition

Most large scale surveys available at the UK Data Service use weighting to compensate for various causes of unequal sampling probabilities: design (for instance when groups in the population are oversampled), non-response (certain type of respondents are less likely to answer the survey). These weights are cross sectional in that they are intended to improve the representativeness of the estimates produced for a single point in time so that they match known characteristics of the population of interest for instance the whole of the UK population.

In addition to these, longitudinal panel data require weights meant to tackle the sources of bias: identified above. Computing and using these weights is complex because the reason why people drop off may change as more waves are added to the survey, and weights need to be computed for each wave of data (i.e. for respondents dropping out at any wave up to that one). They also need to be combined with traditional cross-sectional non-response weights. Typically, longitudinal weights compensate for attrition, based on the inverse probability of dropping out of the sample between two waves by giving more importance to remaining respondents which share characteristics with those who left the sample.

4.3. Example uses of longitudinal data

4.3.1. State comparison and transitions

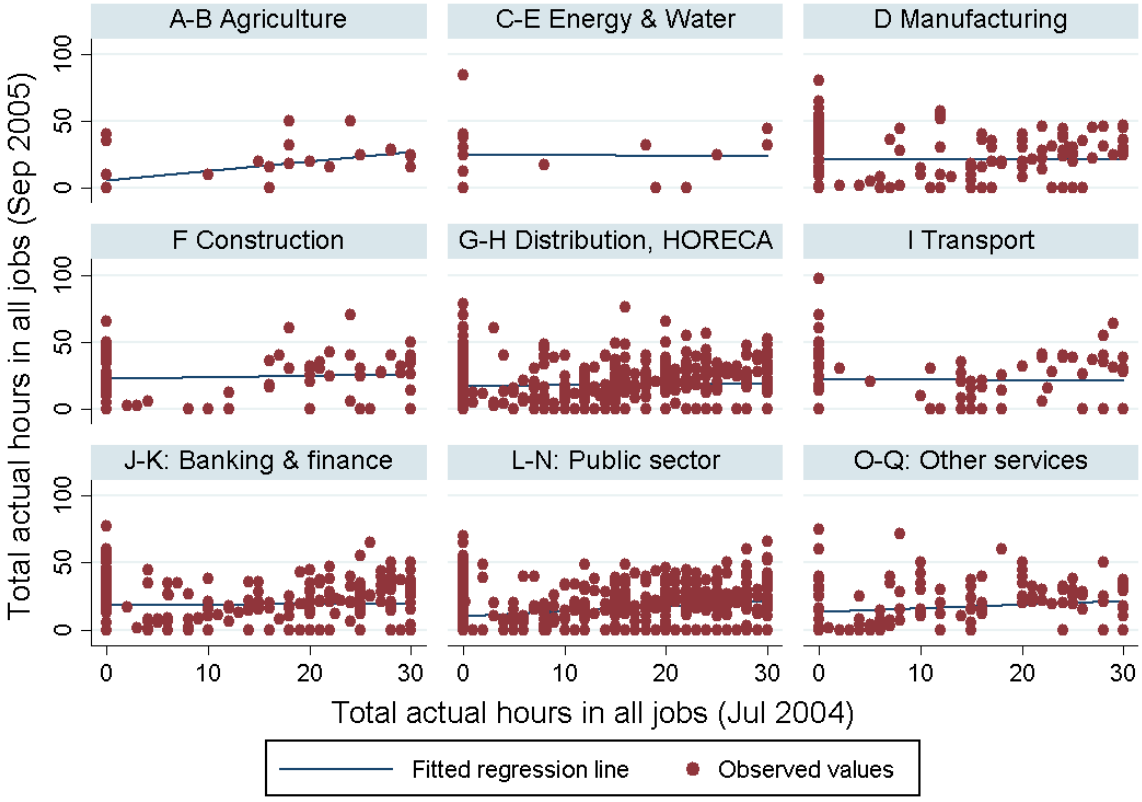
A straightforward way to describe change over time with longitudinal surveys is to compare data between waves. This can be done by plotting the data, either as a scatter plot between two time points or as a curve/line if three or more are available. For example, you could look

[10] This website at the London School of Hygiene & Tropical Medicine gives an overview of the issues around missing data and some suggested solutions: http://www.missingdata.org.uk

at the differences in household earnings between time t and t-1 for each individual and plot these differences as scatter plots, perhaps making the comparison between variations for different subgroups.

Figure 5 is a scatter plot which represents the number of hours a week worked by respondents of the Five Quarter Longitudinal Labour Force Survey between July 2004 and September 2005 disaggregated by industry. Those working 0 hours are those temporarily sick or on holiday. Figure 5 also illustrates how data can be grouped according to relevant categories (here by industry) and includes a fitted regression line which is an approximation of the trends in working-time variability between the two periods of observation.

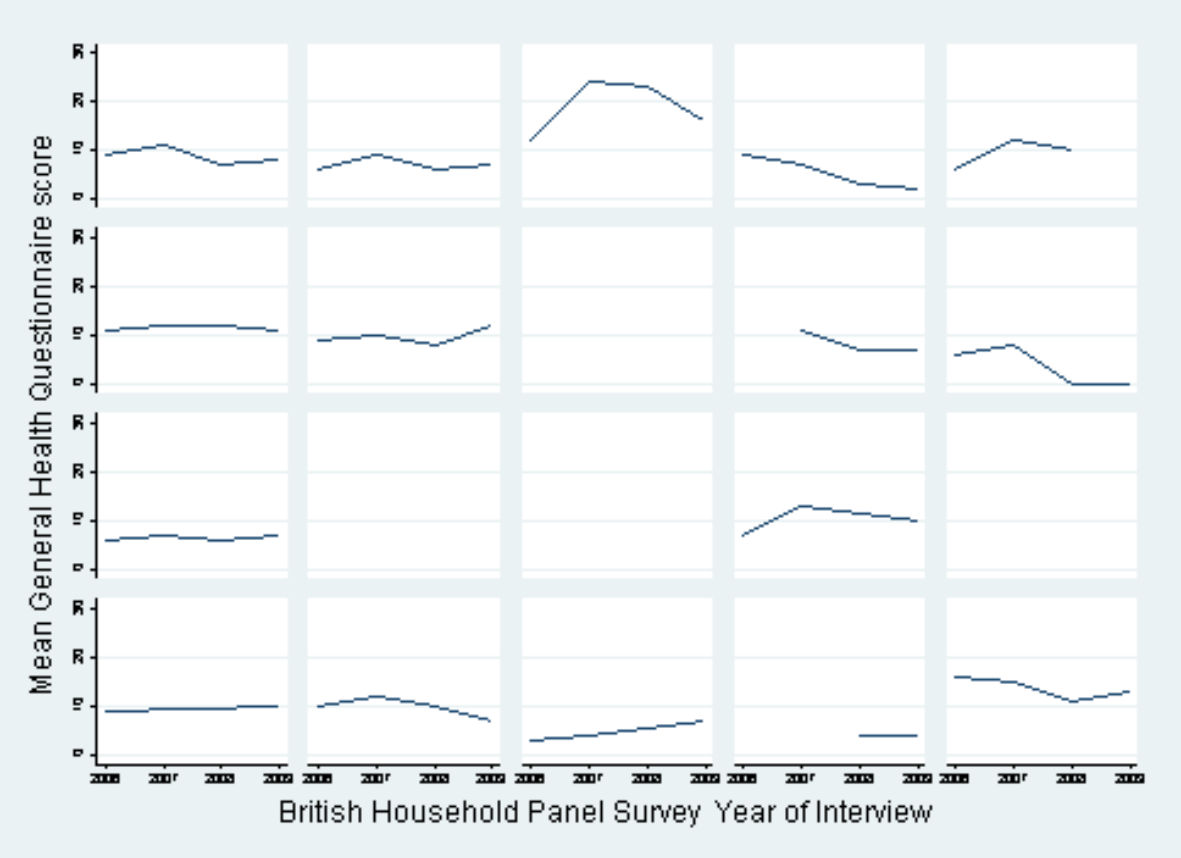
Figure 5: Change in the total number of hours a week worked between 2004 and 2005 in the UK by Industry



Data source: Five-Quarter Longitudinal Labour Force Survey. July 2004 to September 2005

Figure 6 shows the mental time trajectories of 20 random BHPS respondents between 2006 and 2009. This type of graphs requires data to be set up in long format (i.e. one line per measurement occasion, see above). Instead of individual respondents, such graphs can also be easily produced for aggregate groups (for instance by age band, by educational level).

Figure 6: Trends in mental health between 2006 and 2009, selected individuals



Data source: 20 randomly chosen respondents from the British Household Panel Survey Waves O-R

For continuous variables, statistical significance can be achieved by looking at the correlation between measurements, for example scores of a mental health test. An extension of this are the ANOVA/ANCOVA families of models where the mean value of the outcome of interest is compared between waves and the significance of the differences tested.

For categorical variables, another straightforward way to consider individual level change is to cross-tabulate the characteristics of respondents at one wave of a survey with their characteristics at a subsequent wave, and test the differences with a Chi square test.

Log linear models where the transition probability from one state to another is computed are also a straightforward way to compare change between two time points.

Figure 7 presents an example of such transition table, based on change in the family composition of the cohort children in the Millennium Cohort Study. The first line of each category represents the relative (or row) percentages, whereas the second one shows the absolute cell percentage. For instance, the first cell tells us that 83.02 percent of natural fathers who were in the same household as the cohort children when s/he was 9 month old were still living with her/him by the time s/he was 7. The next columns shows that 13.4 % of father who were present initially, were not living in the same household, but were a still in touch with their child, and the group of cohort children who experienced this change represented 11.4 percent of the total sample size.

Figure 7: Change in paternal presence between age 9 months and age 6

	Present	Absent		Total
		In contact	Not in contact	
Present	83.02	13.43	3.55	100.00
	70.43	*11.40*	*3.01*	*84.84*
Absent, in contact	26.98	50.56	22.46	100.00
	2.54	*4.77*	*2.12*	*9.43*
Absent, in contact	13.45	24.93	61.62	100.00
	0.77	*1.43*	*3.53*	*5.73*
Total	73.75	17.59	8.66	100.00

Data source: Millennium Cohort Study 2000 sweep 1 and 4.

4.3.2. Modelling techniques for longitudinal data

Several modelling techniques can be used for longitudinal data. The list below is not exhaustive and is only meant to provide an overview.

Event history analysis[11]
Suppose we have data of the following kind: data about the status of individuals over time, such as whether or not they were in poverty or in employment at each time point.

Descriptive approaches are limited with these kinds of data. Take for example two people (person A and person B) who were in poverty at time-1 and had moved out of poverty by time point t. 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. It may be of interest to examine *durations* to see how long it takes different people to exit a state.

Event history analysis can be used to study 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, for example the time between getting married and getting divorced.

The risk of experiencing a certain event by a given point in time can be predicted by a set of covariates. For example, event history analysis can analyse how length of unemployment is related to age.

Event history analysis has other names in different fields of application including *survival analysis* or *duration analysis*.

Change score and binary change regression[12]

Change score regression (continuous outcome variable) and binary change regression (binary outcome variable) where the difference between outcomes at two waves are modelled with -

[11] For more about Event History Analysis, see Allison, P. (1984) *Event History Analysis: Regression for Longitudinal Event Data.* London: *Sage*, or Box-Steffensmeier, J. M., and B. S. Jones. (2004) *Event History Modeling: A Guide for Social Scientists.* Cambridge: Cambridge University Press.
[12] P.D. Allison (1990) *Change scores as dependent variables in regression analysis*, Sociological Methodology, 20, pp. 93–114.

- usually -- person level covariates.

Lagged regression

Lagged regression (continuous or categorical): This is regression where the outcome of interest at a one wave (typically the last one) is regressed against covariate including the outcome of interest at a previous wave.

Multilevel models[13]

Longitudinal data can be thought of as observations (at a number of time points) that are embedded or nested into individual respondents. Multilevel modelling techniques are therefore well adapted to studying this data structure, by providing a framework and set of tools to simultaneously model *within-person* change over time and *between-person* variations in change. Observations at each time points are therefore level 1, and respondents, level 2.

Multi-level models can:

- Describe average pattern of change (i.e. within person) over a given time period
- Estimate difference in direction and rate of change between persons over a given time period
- Assess the impact of covariates on these differences in change patterns
- To determine whether change in these circumstances over time affect within person change.

[13] Some suggested reading about multi-level models: Snijders & Bosker (2012) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling,* second edition. London etc.: Sage Publishers.

5. Further reading

This guide is intended as an overview of the different kinds of longitudinal data available from the UK Data Service and some of their key features and uses. More detail about how to analyse such data can be found from a number of books and other resources. Here are a few suggestions for further reading around these topics:

Survey data features and analysis
Applied Survey Data Analysis by Steven G. Heeringa, Brady T. West and Patricia A. Berlund, Chapman & Hall, 2010

Longitudinal data preparation and analysis
A practical Guide to Using Panel Data by Simonetta Longhi and Alita Nandi, Sage Publications, 2014

23 February 2015

T +44 (0) 1206 872143
E help@ukdataservice.ac.uk
W ukdataservice.ac.uk

The UK Data Service delivers quality social and economic data resources for researchers, teachers and policymakers.

UK Data Service

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL