

JOINED UP LABOUR MARKET DATA

Executive Summary

This note summarises the results of a project to explore the characteristics of different sources of labour market data – primarily the Census and the Labour Force Survey – in order to be able to provide guidance about preferred sources for analysis.

There are a wide variety and number of factors which will tend to lead to differences between Census and Labour Force Survey estimates of key labour market indicators (even after adjustment to common “sampling frame” bases). Many of these are inter-related, and the separate effects are not quantifiable. But some issues are clear:

- Whilst the Census is a rich source of information, particularly relating to small areas, the data it provides relate to a point in time (April 2001). The further away from this point in time for which users require data, the less relevant the Census will be.
- Most of the differences between the Census and LFS as data collection instruments suggest that the LFS is likely to provide higher quality estimates (less biased with lower non-sampling variability) of the key LM indicators than the Census.
- In particular, the effects of the difference between the interviewer-administered LFS and the self-completed Census may be considerable - the absence of an interviewer, the differences in questioning and the resultant use of editing rules on the Census will all affect estimates of levels. The Census is likely to under-estimate employment (and hence employment-related variables) because Census respondents are less likely to observe the guidelines for completing these questions, specifically the “1-hour” and unpaid family worker criteria. At this stage any estimate of the extent of this underestimation is inevitably speculative, but the facts that the numbers of people in these “marginal” groups have fallen since 1991, and that the Census employment questions appear to be improved, suggest that the Census will understate UK employment compared with the LFS by between ½ and 1 million (2% to 4%), taking account of full-time students.
- Conversely, this difference in the nature of data collection is likely to lead to the Census overestimating unemployment relative to the LFS, for example because there are no questions to help unpick the criterion of “active job search” as there are on the LFS. The improvements to the Census questions since 1991 suggest that the source difference will be reduced, leading to the Census over-estimating UK unemployment compared with the LFS by between 100 and 250 thousand (8% and 18%), again taking account of full-time students.

- In general therefore the LFS should be considered the preferred source of national and regional labour market data. It is a high quality social survey. Its characteristics are well-known; it follows international conventions and definitions; and a long time series is available. However, there are some exceptions to this – see next bullet.
- For some analyses the Census is likely to be the preferred source. These include:
 - The analysis of national and, particularly, regional labour market data relating to small population sub-groups. For example, the LFS will be the preferred source of regional estimates of unemployment. But it will not support robust estimates of regional or local authority level unemployment by sex and age and ethnicity – such data can only be produced from the Census. In general, the LFS - in particular, the local area LFS database - is the preferred source for analyses of groups numbering 6,000 or more – though in many English and Welsh UAs/LAs this threshold is smaller¹.
 - The analysis of labour market data below local/unitary authority level (eg ward)
 - The analysis of key LM indicators in comparison with other Census data which are not collected on the LFS – for example, caring responsibilities – at all geographical levels.
- The SHS (and other large household surveys, such as the General Household Survey and the Expenditure and Food Survey) all collect labour market data. They should be used in preference to the Census and to the LFS only for analysis of LM indicators in terms of data not collected on the Census or the LFS.
- Further work is required, once Census data are available, to assess what guidance should be given in relation to the issue of coverage. On the one hand, LFS estimates for the same population (private households, plus NHS accommodation) will be higher quality than those from the Census. But the Census provides total population estimates. LFS-comparable estimates of the total population could be derived by modelling or by making allowance for the known LM characteristics of those living in non-private households, based on the Communal Establishments Survey conducted in 2000-01. Alternatively the “best” total population labour market estimates may be a combination of LFS private household data and Census communal establishment data.
- Further work is required to evaluate the strengths and limitations of Census labour market data about households, compared with corresponding data from the LFS and the FRS.

¹ See Annex D of the 2002 issue of [LFS User Guide](#), volume 6 (Local Area Data), for a list of English and Welsh authorities for which the analytical thresholds are between 1,000 and 6,000.

Introduction

1. This paper is part of a National Statistics project called “Joined Up Data”. The aim of this project is to provide clear, detailed guidelines for users on the choice and use of preferred sources of data, comparing the 2001 Census with other available data from around the same time.
2. Previous work identified the Labour Force Survey (LFS) as the main existing source of labour market data, though also noted that a few other surveys provide information about particular domains – the Family Resources Survey (FRS) (for households), the Scottish Household Survey (SHS) (for data about Scotland) – whilst modelled estimates of unemployment are also available, based on LFS and administrative (benefit) data. The aim of this paper is to suggest likely differences between Census and survey (mainly LFS) estimates, identifying both the possible causes and the likely size of the differences. It uses this information to identify the preferred source for the key labour market variables – employment and unemployment – in different circumstances, and offers general guidance about the use of Census and LFS data.
3. The approach taken in this paper starts with the assumption that Census/survey differences will be the effects of the different methods of data collection and processing. These different methods are likely to have an impact – though not an identical impact – across the full range of key indicators. Over and above this there are likely to be a range of topics specific to different indicators which will affect their comparability between sources.
4. Both the LFS and Census are rich sources of labour market data, in different respects. The LFS is a large interviewer-administered household survey, collecting a wide range of data about people’s labour market activities, together with demographic data. The survey uses internationally-agreed definitions of unemployment and employment, and these data are published every month. LFS data are available at national, regional and local authority levels. The Census collects a range of information on the whole population at a particular point in time. This makes it possible to provide information at all levels of geography, and enables comparisons to be made between area of any size.

Likely differences – discussion

5. Annex A sets out a typology of differences between the Census and surveys. This forms the basis for the initial part of this section.

5.1 Mode of data collection

- 5.1.1 **Interviewer-respondent interface**: the absence of an interviewer to explain concepts and definitions to Census respondents suggests that Census results will be more biased and more variable than LFS results.

This is likely to be directly related to the complexity of concepts and definitions in any particular question. Hence we might expect the absence of an interviewer to lead to minimal differences between Census and LFS estimates of many of the classificatory variables (age, sex and so on), but greater differences for labour market indicators themselves – for example employment and unemployment, both of which are strictly defined in a way that may well not seem intuitive to an individual, especially if they don't make full use of the Census notes for guidance.

5.1.2 Computer-assisted interviewing: the LFS' use of computer-assisted interviewing (CAI), with its computerised checks in the course of the interview, will tend to provide data that are less biased and less variable than from the Census. This will affect the full range of data collected, but is likely to have greatest impact in relation to those parts of the Census form that are most difficult to follow, and where the LFS used computer-assisted coding and classification (this latter point is picked up later). It is difficult to be sure which parts of the Census are relatively most complicated, but it is probably safe to assume that any routing which involves jumping questions, or questions which have scope for incorrect multiple ticking, are most prone to error in self-completion data collection. On this basis it is likely that Census estimates of headline indicators of employment, unemployment and inactivity, and about employment status and the number of people at the workplace, will all be adversely affected to a certain extent.

5.1.3 Layout effects: In a sense this is an extension of the previous points. On the Census the layout of questions and guidance information is all-important; on the computer-assisted LFS there is no real concept of layout – the corresponding concepts relate to the computerised questionnaire and the interviewers' ability to explain and probe. It is very difficult to draw a relationship between such issues and data quality except to restate the likelihood that these issues will tend to lead to less biased LFS data with less non-sampling variability.

5.2 Differences in question wording

In general the LFS and Census questions use identical wording – the Census followed the LFS which was already established as the 'lead' survey for the purposes of harmonisation. But there are differences in the question about number of people at the workplace. In 2001² the LFS asked "how many employees were there at the place where you worked?" whilst the Census asked "how many people work/ed for your employer at the place where you work/ed. These questions are therefore exploring slightly different concepts – the 2001 LFS results will tend to include people working in contracted-out services, such as security guards, catering staff and cleaners.

² Note that the LFS adopted the Census question wording from March 2002.

Comparisons of these two data series should recognise this difference, which will tend to lead to higher estimates of average workplace size on the LFS than from the Census³.

Differences are also likely to arise as a result of the more detailed battery of questions asked in the LFS. For example, the survey uses a series of questions to establish that an individual is not working, not on a government training scheme, has no other job that they are away from, and is not doing unpaid work for their own business or a business owned by a relative. All of this information is used in order to route people to the questions which concentrate on the “seeking” and “available” criteria of the definition of unemployment. By contrast, the Census relies on separate instructions to the form filler to define work as they answer the (yes/no) question “last week were you doing any work?”.

It seems certain that LFS and Census data will be affected as a result of this. The likely outcome is that relatively more Census respondents will be asked the questions about “seeking” and “availability”, and hence likely that the Census will produce higher estimates of unemployment.

5.3 Context effects

The fact that the LFS interview is concentrated on labour market issues – people’s employment characteristics, job search behaviour and so on – might well lead to respondents providing more considered information than on a multi-purpose data collection such as the Census, in the sense that they are conditioned during the interview to think about their work - for example, the type of organisation they work for, what they make or do in their jobs, their usual and actual hours, their holiday entitlement, the days they work, their journey to work, and so on. All of this is likely to result in LFS estimates which have lower non-sampling variability than corresponding Census estimates.

5.4 100% coverage of the Census

Census estimates will be based on data from virtually everyone in the population, removing the effects of variance, and hence supporting analysis for very small sub-groups, including geographical areas. In contrast the LFS sets out to sample about 0.8% of the population. Although LFS results are weighted (for non-response) and grossed^{4 5} up to be representative of

³ A more subtle difference between the LFS and the Census concerns the response categories to these questions: in the LFS these are 1-10, 11-19, 20-24, 25-49, 50-249, 250-499, and 500+. On the Census the categories were 1-9, 10-24, 25-499, and 500+. So comparisons will also be affected marginally for the smaller workplaces, although the LFS asks respondents to give the exact number, if in the range 1-10.

⁴ The characteristics – including the nature and purpose - of the LFS grossing system are documented elsewhere ([LFS User Guide](#), Volume 1). It is worth noting though that it is one of a number of possible alternative grossing methods. Hence for completeness we might consider that “grossing methods” are a possible source of LFS/Census difference.

⁵ LFS grossing is informed by analysis conducted following each Census – to link Census and survey data in order to identify the (Census) characteristics of survey non-respondents.

population data, the effective sample size is relatively small, and estimates for small population sub-groups quickly become unstable. In this regard the Census is extremely valuable as it provides the only useful labour market data for comparing small geographical areas.

5.5 Response patterns within households

Both the LFS and the Census allow proxy responses – information given by one member of a household by another. In general it is known that the quality of proxy data is poorer than data provided by the relevant individual, though this varies for different topics and across different proxy-subject relationships. About 30% of LFS responses are provided by proxies; the corresponding figure for the Census is not known. However, it is likely to have been of the same order of magnitude. It is likely that there will be small differences in the variability of data from the Census and LFS due to differences in proxy responding rates, but these are likely to be masked by other sources of variability.

5.6 Processing – coding and classification

The main issue here concerns occupation and industry coding, and differences between the overall quality of data resulting from different approaches to coding. On the LFS, data are coded by interviewers. For occupation they use a computer assisted coding system, whilst for industry they use a manual system aided by code books.

On the Census, automatic coding methods are used, supplemented by expert (manual) coding using computer-assisted technology for difficult-to-read Census data, and automatic links to the inter-departmental business register (IDBR) data based on Census responses about workplace addresses in cases of missing industry data. The former approach tends to display greater variance but less bias; the latter method is likely to be very consistent, but potentially biased. It is impossible to be clear about the overall effect of the different methods of coding, and whether they will result in significant differences in estimates of employment or previous employment by occupation and industry. It seems likely though that any such differences will be swamped by differences that arise from the way in which this information is captured initially – that is to say, the presence of an LFS interviewer to ensure that sufficient information is recorded in order to code industry and occupation is likely to be the most important factor.

A further classification issue relates to full time students. It is likely that Census data on economically active full time students will be presented as part of the suite of labour market data, but the number of this group in employment and in unemployment will not be part of the ‘headline’ figures for employment and unemployment. This means that users have to be careful in comparing like-with-like from the Census and the LFS.

5.7 Editing

Another effect of the absence of interviewers in Census collection is that respondents can tick any box on the form, or can miss boxes that should have been ticked. To deal with this, editing rules are used to provide ‘cleaned’ data. In the current context this particularly affects the derivation of estimates of unemployment. Any deviation from the LFS practice of treating as unemployed only those who explicitly say that they are looking for work, available to start or waiting to start a new job, will lead to potential differences in estimates.

The editing actually used in the derivation of “Activity Last Week” is quite involved. But one of the effects is that people will be classed as unemployed if their responses are as shown in the following table.

Unemployment criteria	Nature of Response
Working last week	<ul style="list-style-type: none">• No• Missing response• Multi-ticked (ie Yes and No)
Looking for work	Yes
Available to start	<ul style="list-style-type: none">• Yes• Missing response• Multi-ticked (ie Yes and No)
Waiting to start work	Any value

In other words, people who don’t respond to “working last week” or “available to start”, but merely say that they are “looking for work”, will be classified as ILO unemployed. This potentially includes people already working, looking for another job.

In general, the fact that data problems arise in the absence of an interviewer, compounded by the nature of the editing rules, are likely to lead to the Census over-estimating unemployment relative to the LFS.

6. We mentioned earlier that there were likely to be factors associated with particular indicators, as well as the overarching causes of difference described above. The main such factors are discussed next.

6.1 Geography

Geography is probably the most important characteristic within this project. Quarterly LFS estimates of the key distributions are limited, for sub-national geographies, by sample sizes. The survey supports estimates of the key indicators (employment, unemployment, inactivity), separately for men and for women, for all English Government Office Regions, and for Wales, Scotland and Northern Ireland. But at the level of unitary/local authorities, LFS estimates are less robust. ONS only publish quarterly key indicators at this level for the larger authorities – where there is a sufficiently large sample – and anyway, no male/female split is shown.

In recent years the LFS sample has been boosted in smaller geographical areas, resulting in a product known as the Local Labour Force Survey (LLFS). This provides more detailed information about the key indicators, for local and unitary authorities (as well as other geographies, such as education authorities). However these results are only available on an annual basis because of the nature of the boost element of the sample⁶. This is in stark contrast to Census estimates, which are based on 29 April 2001. This point is returned to in 6.2, below.

ONS also produces small area (UA/LAD) estimates of unemployment using a model. This model uses data from the annual LFS together with numbers claiming Jobseekers' Allowance and produces estimates of 'annual' ILO unemployment levels and rates. Experimental model-based estimates are currently available for the years 1995/96 to 1999/2000. This model has been developed because the annual LFS – prior to the survey boost, and hence the LLFS - provides estimates of ILO unemployment which are sufficiently statistically reliable to be published, for only about a quarter of the UALADs in Great Britain.

More work needs to be carried out before the model-based estimates can be routinely produced as an on-going series. The annual LFS for 2000/01 was the first year of the boosted data for England. The following year a boost was also introduced in Wales. It is envisaged that these boosts will continue, and that a boost will soon be introduced in Scotland. Methodological work is planned to assess the impact of these boosts on the model-based estimates. In addition the model-based estimates will not, in general, be the same as the estimates from the annual LFS, so if these estimates of ILO unemployment are used with the annual LFS estimates of employment and inactivity, the sum of these will not add to the relevant population totals. It is intended that further work should be carried out to develop a multivariate model to estimate unemployment and one of the other statuses (the third will then be derived from subtracting the two model-based estimates from the population total)⁷.

The model-based figures produced so far have used the annual LFS estimates, which are based on pre 2001 census population data. The annual databases will be re-weighted next year as part of the programme of re-weighting all LFS data using population estimates based on the 2001 census. The annual databases will be re-weighted in the autumn of 2003. When this has been done, it will be necessary to re-run the model to obtain new model based estimates, consistent with the new population estimates.

Comparisons of the modelled estimates with Census data will therefore be affected by timing differences – an annualised figure compared with an estimate at a point in time – as well as the fact that the latest comparable

⁶ “Methodology for the 2001/02 annual local area Labour Force Survey”, by David Hastings, [Labour Market Trends](#), January 2003.

⁷ “Development of improved estimation methods for local area unemployment levels and rates”, by David Hastings, Nick Maine, Gary Brown and Marie Cruddas, [Labour Market Trends](#), January 2003.

modelled estimate will relate to 1999/2000. Leaving these issues to one side, comparisons should not be undertaken until the LFS data have been re-weighted during autumn 2003 and the model re-run on the new basis. The model-based estimates of unemployment will not be comparable with the Census until late 2003.

The second way in which geography is important operates at a higher level. This is the result of differences between the Censuses conducted in England & Wales, Scotland, and Northern Ireland. Such differences were relatively small, but may impact on estimates of economic activity by religion – in Scotland and Northern Ireland Census questions asked about the religion the respondent was brought up in, as well as the current religion, whilst in England & Wales the Census question simply asked “what is your religion?”. The LFS (in Great Britain) asks two questions: “what is your religion even if you were not currently practising?” ... “do you consider that you are actively practising your religion?” A further difference is that the category “none” is placed at the end of the LFS question, but was the first response category in the Census question.

However, these questions were not introduced to the LFS until 2002, so any comparisons will be affected by both timing and question-difference issues.

6.2 Timing

This issue was mentioned earlier. One of the strengths of the Census, from an analysts’ perspective, is that it collects information about people at a single point in time – so the latest Census was a snapshot of the characteristics and whereabouts of people on a single day (29 April 2001). LFS estimates, on the other hand, are based on continuous interviewing. Estimates of the key distributions are published every month, for the latest three-month (13 week) period.

This is likely to have a major impact on the comparability of Census and LFS data. The nearest LFS quarter to Census day was the three-month period March-April-May 2001. But the effects of seasonality on labour market variables is known to be considerable. Differences in the seasonal patterns in March and April are very unlikely to be exactly balanced by the seasonal pattern in May. To the extent that there is an imbalance, seasonality will tend to exacerbate differences between key indicators from these two sources⁸.

6.3 Coverage

The Census aims to enumerate 100% of the population. It aims to do so by collecting information about people living in communal (non-private) households, as well as private households. In comparison the LFS concentrates on private households, although its sample design is bolstered

⁸ Because seasonally-adjusted LFS data for the latest 3-month period are published every month, we can derive monthly estimates of seasonality for key LM indicators from the LFS and use these to inform LFS/Census comparisons.

to ensure it collects information from people living in NHS accommodation, and about students living in halls of residence – in all it misses about 1.5% of the population. So the Census provides unbiased information in terms of types of accommodation. This is an important point to note in making comparisons of headline estimates from the two sources.

Household level data

7. Like the LFS, the Family Resources Survey is an interviewer administered household survey, which used harmonised concepts and definitions. It has a smaller sample than the LFS - it is designed to produce annual estimates, chiefly of household/family income – for example, the Households Below Average Income series is derived from the FRS. ONS and DWP (which sponsors the FRS) are currently working on a project to identify and explain reasons for differences between estimates of the number of children in workless households. As this work develops it will become clear whether the LFS or FRS should be used in preference for such analyses, in different contexts.

Data for Scotland

8. The Scottish Household Survey (SHS) collects information about the employment status of the highest income householder only, so does not support labour market estimates of Scotland as a whole. Hence the SHS does not ‘rival’ the LFS as a source of labour market data about Scotland.

What can we tell from previous Censuses?

9. The 1991 Census included questions about individuals’ labour market behaviour. These did not set out to follow the ILO conventions as closely as in the 2001 Census, but nevertheless work published⁹ in 1994 comparing Census and LFS estimates identified the following important differences:
 - The 1991 LFS found about 5% more people in employment than the Census, probably because in the Census some people ignored small amounts of paid work which would be included in the LFS. Some students with jobs were recorded as inactive.
 - The largest differences in employment status were for part-time employees and those on government schemes.
 - Eight percent fewer people were classified as unemployed in the 1991 LFS than the Census (although in the Census people were not asked whether they were available for work, nor how recently they had looked for work).
 - The LFS estimates of the number of unemployed men were 19% *lower* for men, and 11% *higher* for women, than the Census estimate. The pattern for men is attributed to the absence of Census questions establishing the individual criteria of unemployment. The pattern for

⁹ “Economic activity results from the 1991 Labour Force Survey and Census of Population”, by Frances Sly, Employment Gazette, March 1994.

women was thought to be a result of a general misconception that the relevant Census question was interpreted as asking about *main* economic activity status, although in fact it asked for several boxes to be ticked if they applied (eg employee *and* looking after the family/home).

Implications for data from the 2001 Census

10. Since the 2001 Census employment questions were broadly similar to those asked in 1991 – identifying all groups across the domains of economic activity and inactivity - we would expect to see the same pattern of Census under-estimating employment compared with the LFS. Other things being equal we might expect the improvements to the 2001 Census employment questions to tend to reduce the extent of this under-estimation, although it is difficult to quantify this with any precision.
11. The unemployment questions differed more markedly between these two Censuses. Whilst the 2001 Census questions follow the ILO definition of unemployment more closely, they do not do so as strictly as the LFS approach. Hence we would expect the 2001 Census to overstate unemployment, based on past evidence, for both men and women, though to a much smaller extent than the 1991 Census overstated male unemployment.

Census – Survey Comparisons: Main differences

Introduction

1. The main differences perceived to be relevant are set out below. (Note that issues such as the *use of documentation* is excluded because the only Census topic which might realistically benefit from respondents' use of documentation is "qualifications", but surveys don't ask for such documentation).

Mode of data collection

2. Interviewer interface dimension - Census respondents will not have the benefit of an interviewer to explain and clarify concepts. The main effect is likely to be relatively more confusion on the part of Census respondents than survey respondents, so Census results which are both more biased and more variable (ie more non-sampling bias and variance).

3. Computer-assisted interviewing – all of the surveys use CAI. This will lead to better quality (less bias, less variance) data than from the Census because of automatic routing, computer-assisted coding, immediate identification of errors, and so on.

4. Other differences - questions are developed and tailored in a certain way depending on the mode of data collection, which may lead to different results. The Census has limited space for extensive coding frames and prompts; conversely surveys can use showcards, interviewer probing, and so on. On any self-completion form respondents have to follow all the instructions as well as answering the questions, so layout and wording of everything is vitally important. The Census form is constrained by space and scanning requirements compared to survey instruments. This may have systematic effects.

Context effects

5. Answers given by respondents may vary between the Census which is very general in nature and surveys, particularly those with a specific focus eg health survey, labour force survey. Respondents to the specialist surveys might be conditioned to think in more detail about the relevant topics as the interview proceeds, and may respond differently than when suddenly faced with a certain topic. Also, the ordering of questions may have an effect on results. It seems likely that the adverse effects of these issues will be tempered by the fact that Census respondents will be able to look over the whole form before they complete it, so topics should not suddenly surprise them. But in general, we would expect context effects such as these to lead survey data having less variance than Census data.

Nature of the Census

6. The Census achieved response rates far in excess of those associated with (voluntary) surveys, because of its compulsory nature. The ONS methodology

has provided results about (virtually) everyone. As a result the Census will deliver extremely small area data, and data will be very precise (low/no sampling variability), with much less item non-response than for surveys.

Response patterns within households

7. In many cases we expect a single person to complete the Census return on behalf of others in the household. Most surveys – with the exception of the LFS - tend to discourage the use of ‘proxy’ responses, because of concerns that such proxy respondents may not know the detailed information required about the data subject. In general Census results are likely to have relatively more non-sampling (respondent) error because of the use of proxies.

Processing (especially coding/classification)

8. Interviewers are in general less accurate in occupational coding than expert office coders specially training for that job, eg. the coders who did the Census coding in 1991. However, automated coding has been used in 2001. This will lead to very high consistency, but with the possibility of systematic bias. Interviewer coding (with small workloads) tends to lead to greater variance but less bias. So there will be differences between the two but it isn't clear which will be better quality.