

New Methodologies for the 2001 Census in England and Wales

Andy Teague, Office for National Statistics

Abstract

This paper looks at the significant innovations that are being introduced into the 2001 Census in England and Wales. Whilst the traditional concept of collecting information from each person and household in the country at a fixed point is being adhered to, in virtually every other respect the Census in 2001 will be very different from those that have preceded it.

Keywords: Differential Underenumeration, One Number Census, Quality.

Introduction

A Census has been taken every ten years in Great Britain since 1801, with the exception of 1941. Naturally, through developments in society and technology, a number of changes and innovations have been made and introduced along the way. These are well documented in Mills (1987) which covers the developments up to and including 1981. The 1991 Census followed a similar model to the 1981 Census. However, the 2001 Census, will be very different in a desire to improve quality.

In 1992, the United Kingdom Census Offices [the Office of Population Census and Surveys (for England and Wales), General Register Office for Scotland, and Census Office, Northern Ireland] carried out a Policy Evaluation and Reappraisal (OPCS, 1992). This looked at the needs for Census-type information over the coming decades. It concluded that the need was still there and, following a review of the possible options for collecting the information, that a

traditional Census was the best approach to collecting the information in 2001. A number of other countries, (particularly in Scandinavia) have, in recent years, abandoned the notion of a traditional Census in favour of an administrative register-based system. In the UK, however, it was concluded that it was highly unlikely that registers of sufficient 'quality' could be developed by 2001.

The UK Census Offices then set about planning the next Census. Quality was, and still is, at the heart of this planning (see Jones (1997)). An extensive research programme was set up to look at various options within the scope of a traditional Census - to issue and retrieve a form from every household and person in the country on Census night. There were a significant number of drivers for change - experience with the 1991 Census, society, technology and customer needs and expectations. This paper will describe the outcome of that research programme and outline the innovations to be introduced in the 2001 Census. I will explore in turn - the content of the Census and the Census form, collecting the data, processing, output and quality. I will do this with respect to the England and Wales Census, although in almost all respects the approach to the Census in Scotland and Northern Ireland will be very similar.

Content of the Census and the Census Form

The number and types of questions included on the Census Form changes from Census to Census in response to new policy initiatives, other user needs, society in general and public perception. The look and feel of the Census Form has also changed in response to design techniques. The next Census in England and Wales will be no different in this respect. The Government published its proposals including the questions to be asked for the 2001 Census in a White Paper in March 1999. In addition to the questions asked in the 1991 Census, there will be new questions on provision of care, general health, time since last employment (for those not currently working) and lowest floor level of accommodation. The 2001 Census form will

also contain a number of revisions to topics and questions included in previous Censuses, notably those on ethnic group, qualifications and relationships within the household.

A strong case was made for a question on income to be included in the 2001 Census. Consultation with users about requirements for information from the Census indicated widespread support for the inclusion of a question on income. But the strength of such requirements had to be balanced against the possible disquiet about the acceptability of such a question in a compulsory Census, the doubts about the reliability of the information collected, and the availability of alternative sources of the information. The Government consequently did not include provision for a question on income in the Draft Order laid before Parliament on 10 January 2000. The Order was subsequently approved.

The Government also proposed that a question on religion should be included in the 2001 Census. However, to do so, requires a change to be made to the primary legislation – the Census Act 1920. Lord Weatherill introduced a Bill in the House of Lords to this effect in December 1999 with government support. Subject to this Bill receiving Royal Assent, a question on religion will be included in the 2001 Census.

The challenge, of course, is to meet new and changing user needs whilst maintaining a degree of comparability from one census to the next. This is perhaps best illustrated by the ethnic group question where changes in user needs and society have led to new categories on Irish, mixed ethnic group, and Black British and Asian British for the question in England and Wales. But this has been done in such a way that allows broad comparability with the 1991 Census question - see illustrations below.

Further information on the development of questions for the 2001 Census is given in Moss (1999). Details of questions asked in previous census are given in Mills (1987), OPCS/GRO(S) (1992) and OPCS/GRO(S) (1977).

A primary consideration for the 2001 Census has been to minimise the burden on the public. Questionnaire design is an obvious pre-requisite to achieving this aim. Major changes have been made to the look of the Census form since 1991. Whilst this partly reflects the requirements of new processing technology (automatic imaging and recognition - see Processing the Data below), the 2001 Census form incorporates recommendations made by form design experts early in the testing programme. These included ensuring that the questionnaire is clear and easy to follow by minimising routing and superfluous instructions. Testing has been undertaken and determined that, even without detailed instructions, proposed questions can be understood and produce reliable results. Those shown to require substantial explanation, such as proficiency in English, have not been proposed for inclusion.

The Census form for 2001 will also have a page-per-person style as opposed to matrix style used in previous censuses. While there was little difference in overall public response to the two types of form, the former was cheaper to produce and thought to be more acceptable in households of unrelated persons when answering sensitive questions.

Questions on Ethnic Group - 2001 Census and 1991 Census

2001 Census - Ethnic Question (for England & Wales)

8 What is your ethnic group?

♦ Choose one section from a to e, then
✓ the appropriate box to indicate
your cultural background.

a White

British Irish

Any other White background,
please write in

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

b Mixed

White and Black Caribbean

White and Black African

White and Asian

Any other Mixed background,
please write in

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

c Asian or Asian British

Indian Pakistani

Bangladeshi

Any other Asian background,
please write in

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

d Black or Black British

Caribbean African

Any other Black background,
please write in

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

e Chinese or other ethnic group

Chinese

Any other, *please write in*

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

11 Ethnic group

Please tick the appropriate box.

If the person is descended from more than one ethnic or racial group, please tick the group to which the person considers he/she belongs, or tick the 'Any other ethnic group' box and describe the person's ancestry in the space provided.

White	<input type="checkbox"/>	0
Black-Caribbean	<input type="checkbox"/>	1
Black-African	<input type="checkbox"/>	2
Black-Other	<input type="checkbox"/>	
<i>please describe</i>		
<input type="text"/>		
<input type="text"/>		
Indian	<input type="checkbox"/>	3
Pakistani	<input type="checkbox"/>	4
Bangladeshi	<input type="checkbox"/>	5
Chinese	<input type="checkbox"/>	6
Any other ethnic group	<input type="checkbox"/>	
<i>please describe</i>		
<input type="text"/>		
<input type="text"/>		

Collecting the data

It was estimated that the 1991 Census counted some 98% of the population. This included some 1.5% of people 'imputed' in households known to exist but from which no Census form was received. While the overall level of coverage was high in comparison to international standards, the level of coverage varied considerably from area to area and from subgroup to subgroup. This so-called 'differential underenumeration' led to a rethink about the entire approach to conducting the Census.

There have been significant societal changes - households and people are away from home more frequently, and there has been a significant increase in the number of one-person households and entry-phones into buildings. This makes the enumerator task of collecting forms that much more difficult.

Since 1991, the research programme on Data Collection has focussed on the issue of maximising coverage and particularly on reducing the bias in the resulting count. This meant that our efforts and resources had to be targeted at those households and people with the

greatest propensity to not comply with the Census - single person households, households in multi-occupancy, young men, students, the elderly and ethnic minorities.

A number of initiatives have been made. Firstly, the 2001 Census will be conducted on an entirely resident basis. In previous censuses, people have been counted where they are on Census night and where they are resident if different. People away from home on Census night - some 1.5 million in 1991 - were required to have information about them supplied twice. However, there was some evidence that people away on Census night understandably felt they had already complied with the Census and their details were not recorded at their home address. Further, the information on the number of 'visitors' (that is non-residents) to an area was not extensively used. So the first step was to reduce this burden on the public. In 2001, people will only be asked to supply information at their home (resident) address.

Secondly, the public will be asked to post their Census forms back. Pre-paid return envelopes will be provided. In tests, over 80% of responding households have posted back their forms. It should be emphasised that enumerators will still deliver the Census forms and 'follow-up' those households who have not returned a form. This 'follow-up' procedure will start some four to five days after Census Day. In tests, this has shown to be the best strategic approach, allowing enumerator resources at 'follow-up' to be targeted at those areas which are likely to have a lower postback response - inner cities, areas of high elderly population, areas of high ethnic minority population, areas with high levels of multi-occupied buildings etc. Of course, this has a downside in that there is less enumerator contact with households to assist in completing forms. But we have planned for that too in that a public enquiry line will be available (a local rate phone number will be used) together with another initiative for 2001, which I will now outline.

A community liaison programme has been established. This is more than part of the publicity strategy for the 2001 Census. The objective is to establish links at national and, more

particularly, local levels to 'sell' the benefits of the Census to different communities, and give assistance where needed. This programme is in its early stages of development but it is intended that a partnership approach of this type will eventually be cascaded via the Census Area Managers - the top level of the Census field force. The Director Census is personally advocating this approach.

To assist in identifying households and ensuring enumerators find them all, good maps are essential. Two developments have been made. Firstly, enumerator maps will be produced using new GIS technology based on Ordnance Survey's product Addresspoint™. This will provide enumerators with a customised map on a single sheet of paper - the days of enumerators juggling several maps in the rain should be long gone (apart from the rain, that is!). Secondly, a list of addresses will provide a start point for enumerators but their instructions will still be to call at every address in their area to try and make contact to establish how many households and people there are behind the front door. In areas of high level of multi-occupancy (that is an address with more than one household), enumerator workloads (that is the number of addresses an enumerator is expected to cover) will, on average, be half the size of those in 'easier' areas.

Students are another particularly difficult group to enumerate in a Census. There are distinct advantages, from a user perspective, in counting students at their term-time (as opposed to vacation) address. It is the term-time population of an area that is used in allocating resources from central to local government. To achieve an accurate term-time count the Census needs to be carried out in term-time and this is proposed for the 2001 Census. Students will still need some encouragement however! The publicity strategy for the Census is, of course, vital. This raises awareness of the public to the purpose of the Census and reminds them to return their form. While the main messages will be just this, more specific targeting towards difficult to enumerate groups will be needed.

All in all, the approach to collecting Census data will be to target those groups and areas, which are most susceptible to be missed. Once collected, of course, the next major task is to process the 30 million or so Census forms.

Processing the data

Technology has advanced significantly during the 1990s, particularly in the area of automatic scanning and recognition of documents. Systems which were considered risky a decade or so ago are now commonplace. The Census will be able to take advantage of the new technology and, in particular, this will reduce the significant clerical effort used in previous censuses.

Census forms will be imaged and OMR and OCR technology used to automatically capture the data from the images. Automatic and computer-assisted coding software will be used to code responses to pre-defined classifications. This approach will allow all responses to questions on topics like occupation and industry of employer to be coded. This is a substantial improvement on previous censuses in Britain, when only 10% of answers to such questions were coded because of the high costs involved. The capture and coding of Census data will be carried out on behalf of the Census Offices by Lockheed Martin under contract.

In an exercise as large and complex as the Census, it is inevitable that errors will occur. These are mainly caused at the form completion stage - through answers being missed or inconsistent answers given (such as someone who describes themselves as married but are only 10 years old). It is simply not practical to check every form with every household so it is traditional for editing routines to be used in processing. Editing and imputation techniques, similar to those used in previous Censuses, will be used to 'rectify' and estimate such data. The approach to be taken in 2001, which builds on the original Fellegi and Holt (1976) principles, is described in Vickers and Yar (1998).

In summary, answers to questions which are considered to be inconsistent with one another are assessed in a series of logical rules. An automatic decision-making process using other information on the form decides which answer is most likely to be incorrect based on the principle of making least change to the data. The answer to the question is then either set to missing or, where there is only one appropriate answer that answer substituted. Returning to my 10 year old married person example, either the age will be set to missing or marital status set to 'single'.

Imputation of missing answers follows by a route whereby the answer to the question is copied from a similar person or household (called the 'donor' person or household) in the same or nearby geographical area. Unlike 1991, it is intended that when a Census form contains several missing answers, one 'donor' person or household be used. This has been demonstrated to maintain the statistical integrity of the marginal and joint distributions better than using several 'donors' to supply several missing answers for one person or household record.

One Number Census

Following the editing and imputation of returned forms comes the estimation of numbers and characteristics of households and people missed entirely by the Census. Despite the initiatives outlined above, it is inevitable that some people and households will not be counted, and some sub-groups of the population to a greater propensity than others. Before I describe the substantial innovations for 2001 in this respect, I will return to the experiences in 1991.

It is traditional in many Census taking countries for the Census to be 'followed-up' by a Post Enumeration Survey (PES) to estimate the number of people missed by the Census. This was the approach taken in 1991 in Britain with the Census Validation Survey (as the PES was then known). This Survey had a dual purpose - to measure both coverage (the number of people

counted) and the quality of answers to Census questions. Some 20,000 addresses were sampled. Further details are given in Heady et al (1994).

Unfortunately, the Census Validation Survey (CVS) failed to find many of the people missed by the Census. When the CVS estimates of the Census undercount and the Census Counts themselves were analysed (particularly by looking at the proportions of males to females), it was evident that a more reliable indicator of the national figure would be provided by the existing series of population estimates based on the 1981 Census. Further, the CVS (mainly because of its size) was unable to provide much breakdown of the undercount across the country, although some information was used to 're-base' the local population estimates on the 1991 Census, constrained to the national estimate. This was unsatisfactory all round, as insufficient information was available to guide Census users on the level and nature of the undercount, particularly at the local level. Only broad guidance could be provided. Further, several figures on the total population of the country were actually produced, so adding to the confusion.

The Census Offices, in response to the problems in 1991, have tackled this in earnest. A separate research programme was set up in 1996 to evaluate different methodologies and to plan the approach well in advance. The objective of the One Number Census programme (as it is known) is to estimate the level of underenumeration and to integrate this with the Census counts so that all Census outputs sum to One Number - the national estimate of the population on Census Day. The methodological research has included looking at administrative records and whether these could be used to aid the estimation of the Census undercount. It was clear however that no such records were available to the required quality supporting the conclusions arrived at in the early 1990s. The approach is therefore to use a post-enumeration survey but one which concentrates exclusively on coverage and which is much bigger to give the statistical resilience that was not there in the 1991 Census.

A Census Coverage Survey of some 300,000 households will be carried out some three to four weeks after the Census. It will encompass an intensive re-enumeration of some 20,000 postcode units (average size 15 households) across the country. The sample will be designed to produce direct estimates for some 100 'design groups' - average population size 500,000 people. The sample will be stratified by a 'hard to count' index so that estimates can be separately made according to the likely level of underenumeration. The Survey will comprise short doorstep interviews of around 10-15 minutes each.

The information from the Survey will be combined with that from the Census and estimates of underenumeration made at the design group level using a combination of dual system and regression-based estimators. These will then be cascaded using synthetic estimation techniques to the local/unitary authority level (average size 120,000 population) to provide the new base of local population estimates by age and sex.

The final step of the ONC process is to estimate the probabilities of households and people being missed at the local level by type of household and person. Imputation of households and people according to these probabilities will follow to produce a fully adjusted Census database. This final process will be constrained to the estimates produced at the local/unitary authority level. Further information on the methodology for the ONC is available in ONS/GRO(S)/NISRA (1999) and at the ONS website (<http://www.ons.gov.uk>).

The end result of this innovative census processing stage will be a database comprising both individuals and households counted by the Census and synthetic households and people representing those estimated to have been missed. It forms the database to be used to produce the output.

Output Policy and Production

The effort and cost of taking a Census is only worthwhile when the results meet needs, and are delivered effectively. A number of innovations will be made to the output production process to improve the products and services available.

A major innovation is the timetable for release of Census Output in that national and local data will be produced concurrently. In previous Censuses, results have been published area by area which meant that local versus national comparisons could not be made until all the country's results were available. National results from the 1991 Census were available some five months after the first local results. The actual timetable for the release of results from the 2001 Census is uncertain at this stage largely because of the degree of change in the processing methodologies. These are about to be rehearsed and until that is done and the product portfolio finalised, it will not be possible to set a definitive timetable.

The second major change from 1991 is that virtually all products to be produced will be electronic. A print on-demand service is proposed but this is to be just that. The primary means of dissemination will be electronic. This brings with it new challenges in visualising the data on screen. Advances in data manipulation and visualisation (particularly using GIS software) make this possible in a way for the new, casual and experienced user alike. To back this process up, it is proposed to bring together the various metadata products produced in 1991 (documentation, user guides, reports on quality and so on) into one electronic source.

Products to be produced essentially fall into two parts. One is a set of standard products - a pre-defined set of 'tables' of Census 'counts' for all levels of census geography, the other a service to respond to individual customer needs. Such customised needs can either be pre-defined or follow at a later stage. The customised service must be faster and cheaper than it has been for previous Censuses and the Census Offices are aiming for this. An extensive consultation process is underway with Census users to tailor the products and services to meet needs.

For the 2001 Census, a key objective has been to improve the ability of users to compare Census data with other government sources of statistical information. Harmonisation of concepts, definitions and classifications improves comparability and where possible harmonised questions and classifications have been used. This will enable the Census to be used more effectively as the benchmark for which it is intended.

The third major innovation in Output from the 2001 Census is the geographical base. In previous Censuses, output areas have, in England and Wales, been based on enumeration districts, which are designed as workloads for enumerators. Discussions with users after the 1991 Census signalled a desire to separate the geographical base used for collection from that used for output. Common requests were for smaller, more homogeneous, and consistently sized areas built from postcode units. For the 2001 Census, advances in technology have enabled a methodology to meet these requirements resulting in an approach that will create small geographic 'building blocks' called Output Areas.

The methodology is based around building synthetic postcode unit boundaries using Ordnance Survey's Addresspoint™ product and natural features such as road centre lines. Postcode units are then 'grouped' into output areas according to a set of rules based on a target population size, degree of homogeneity (provisionally based on tenure of household) and shape. Further, as one of the confidentiality measures to protect data about individual people and households, output area population sizes must be above a minimum threshold (the precise level is yet to be set).

Quality

The innovations and changes proposed for the 2001 Census and which I have outlined above have been introduced with the desire to improve quality. The approach to designing, and building for, quality has been there throughout the research programme. Different options

have been suggested, tested or prototyped, and evaluated. In particular, a series of tests looking at form and question design have been carried out and the results evaluated before recommending the best approach.

Other research has taken the form of simulations using 1991 data to test new methodologies. Prototype systems have been built and data simulated to test them out. The 1997 Census Test provided a major opportunity to try out new methodologies. A postback collection method was compared against a traditional enumerator delivery and collection method. Two types of form design were also tested, as were two different versions of the ethnic group question and a Census form with a question on income and one without. This approach, although it complicated the design of the test, enabled decisions to be taken on the data collection methodology and form design. It also provided a lot of information on the subsequent development of the ethnic group question and the problems of including a question on income. Details of the results of the test are given in ONS (1998).

In addition to this quality management approach to designing the Census, information has also been gathered throughout the testing programme about the quality of Census results that are likely to occur. This has been backed up by a specific Quality Survey conducted in May 1999 and currently being analysed. This was designed to measure the likely degree of respondent error to Census questions.

Census data will also be quality assured and validated as it is processed. Lockheed Martin, the processing contractor, are contracted to process Census data to pre-determined quality levels. The Census Offices will, in addition to the quality control processes put in place by Lockheed Martin, validate the data. This process will 'look' at the data in the way users might. Attempts will be made to spot patterns that might be considered unusual or beyond expectations. The strategic objective is to identify problems early so that they can be rectified (if necessary) or explained well in advance of the data being published.

Information on quality will be put together to form a Quality Report. This will form part of the Census metadata I referred to earlier.

Conclusion

The 2001 Census will in almost every respect be very different from those that preceded it. The degree of change described is necessary if the Census is to work in today's Society and to meet the needs of customers whose expectations are higher. It will be innovative and leading edge; the challenge is to integrate all the changes so that the 2001 Census is the best ever and sets the benchmark for the new millennium.

References

Fellegi, I and Holt D (1976). A systematic approach to edit and imputation. JASA, Vol 17.

OPCS and GRO(Scotland) (1977). Guide to Census Reports, Great Britain 1801-1966. HMSO, London

Mills, I (1987). Developments in Census-taking since 1841. Population Trends, 48. HMSO, London.

OPCS and GRO(Scotland) (1992). 1991 Census definitions. HMSO, London.

OPCS (1993). Report on review of statistical information on population and housing (1996-2016). Occasional Paper No 40. OPCS, London.

Heady, P, Smith S and Avery, V (1994). 1991 Census Validation Survey: Coverage report. OPCS Social Survey Report SS1334. HMSO, London.

Jones, G C (1997). Planning the 2001 Census: only four years to go. Population Trends, 88. TSO, London.

ONS (1998). 1997 Census Test Evaluation. Census Advisory Group paper, AG(98)01. ONS, Titchfield.

Vickers, P and Yar, M (1998). The development and evaluation of the donor imputation system (DIS) for the 2001 UK Census of Population and Housing, Joint IASS/IAOS Conference, Mexico.

The 2001 Census of Population (Cm 4253).

ONS, General Register Office (Scotland) and Northern Ireland Statistical and Research Agency (1999). A Guide to the One Number Census. ONS, Titchfield.

Moss, C (1999). Selection of topics and development of questions for the 2001 Census. Population Trends (forthcoming).