Published as: Mitchell, R., Dorling, D. and Martin, D. (2001) Making Sense of the Census: A new means to access census data from 1971-1991 and on to 2001, Chief Scientist Office, *Research Matters*, Issue 16

Making Sense of the Census: A new means to access census data from 1971 – 1991 and on to 2001

Richard Mitchell, Research Unit in Health, Behaviour and Change, University of Edinburgh.
Danny Dorling, University of Leeds
David Martin, University of Bristol

Introduction

To run Scotland well, Scotland needs to be understood. To understand it, we need to study its growth and change and to do that, we need appropriate sources of data and a means of accessing and analysing them. Only one social survey gets anywhere near covering everyone who lives in Scotland (and the rest of Britain): the UK decennial census. The nation has just endured another census and the arguments over whether the questions were appropriate, too difficult or even necessary are fading again. The job of processing the data is well underway. In this stage of 'anticipation', social scientists, planners, policy makers and all other user groups may be turning their attention to what they would like to do with the census data when they arrive, and how they might be usefully joined to data from past censuses.

Data from all censuses are a *potentially* rich source of information about changes in the nature of the country, about who lives where and about what has happened and might happen to them. In this age of computer assisted analysis, digital census data are currently available for the 1971, 1981 and 1991 censuses. Even before the 2001 data are released, data from 1971-1991 should be the primary choice for researchers interested in how Britain and the British are changing, if only to provide the context for more specific study, but they are often ignored. Why? Ask anyone who has tried to gain access to or analyse these data – the census can be a nightmare.

Those who wish to use census data, especially to look at change over time, must overcome three big obstacles (1) the censuses have a complex access system, (2) variables, geography and access systems are different for every census and (3) the 1991 data for Britain are significantly tarnished by an undercount of about 1.2 million people which always raises questions about comparisons between censuses. However, help is at hand. A collaboration between researchers at the Universities of Edinburgh, Leeds and Southampton, designed to enhance the value of all census data (1971-2001) by removing these three obstacles and make the data easily comparable through time, is nearing successful completion. When the 2001 data are released we want you to be ready to maximise the utility of the 2001 data by being able to compare them with previous years.

Let's consider those three obstacles in a little more detail.

Complexity Of Access

The first and perhaps most damaging obstacle to using the census is the complexity of the access system. The data from all censuses are structured through 'cells' within 'tables' (a particular count of interest will be referenced via its cell and table number). Although great steps forward have been made with the advent of CASWEB (an online graphical user interface to the 1991 census data), it remains a fact that the complexities of cells, variables, table numbers and obscure census definitions render the true power of the census unobtainable for all but a small number of initiates. Attempts to access the 1981 or 1971 data strip away the thin veneer of user-friendliness which CASWEB provides for the 1991 data.

There are men and women who have acquired the knowledge and (considerable) technical skills required to 'read' the different censuses but they are few in number and tend to be employed in technical rather than substantive research programmes. The fundamental component of obstacle 1 is that access to census data is not organised in a manner similar to the typical researcher's lines of enquiry. Census-users want to think in terms of topics, questions and concepts. The census is structured in cells and tables which may (or may not) cover those topics and questions, but which certainly do not make their contents obvious. Who, but an initiate would know that SAS table 9 covers the relationship between ethnic group and economic activity amongst the over 16's? Pre-prepared and published abstracts and tables help the non-expert census user but are inevitably rather limited in the opportunities they can convey for research. In a nutshell, those who wish to know something about particular areas or types of people must become adept enough to convert their aspirations to topics, tables and cells.

Inconsistency Through Time

The second obstacle becomes rapidly apparent should you need to describe how a group or place of interest has *changed through time*. Any two censuses have two different access systems, two sets of definitions (of variables and geography). The 2001 census will be no exception. Even amongst the adept, the number of specialists able to assist falls rapidly when a temporal dimension is brought into the analysis. Including census data from the 1970s in an electronic analysis is the preserve of perhaps less than 50 specialists in Britain. But the ability to conduct analysis through time is probably amongst the most desirable of opportunities which a succession of decennial censuses offers.

The inconsistencies vary from those which cause minor inconvenience, such as changes in the way census areas are labelled, to those which cause major problems, such as radical change of areal boundaries which make comparison of two areas which are substantively the same, statistically impossible. We should stress that some of these inconsistencies are certainly not any particular group's "fault". Places grow and change, governments reorganise administrative areas, the topics which each generation wants to see in the census will change over

time. However, change in the census questionnaires, in the administrative geographies and in the access systems do make it almost impossible for a social scientist or policy maker to answer basic questions about how the country has changed.

The Missing Million

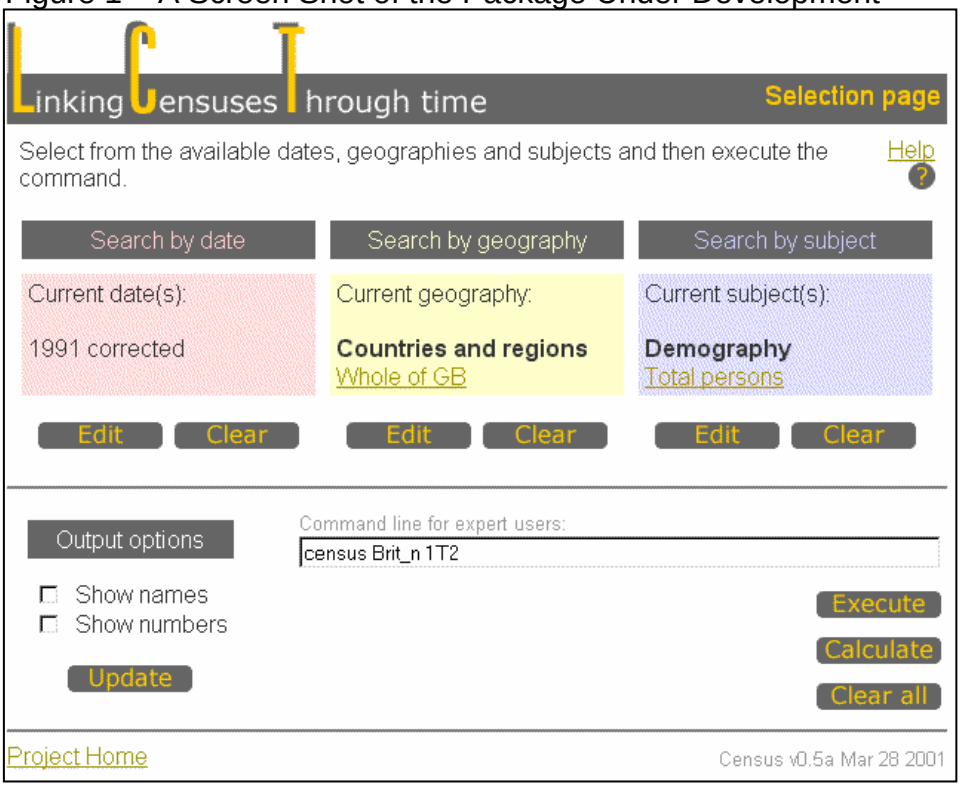
The third obstacle for attempting analysis through time is that in 1991, the UK census missed about 1.2 million people. Detecting, assessing and decoding change over time becomes difficult when one cannot be sure whether the changes seen are due to people missed in 1991, or genuine differences in society. This is of particular significance for those most interested in the more marginalised members of society. It was these groups whose number was captured least accurately.

The New Access Package

The advent of the 2001 census prompted consideration of how to maximise the utility and value of the data which will stem from census night. In preparation for the release of those new data, the ESRC has funded a project to tackle all three of these obstacles and to prepare (as far as possible) enhanced access to those new data *and* to data from the other 'electronic' censuses (i.e. those for which data are available in an electronic form). The project "Linking Censuses Through Time: Easy Access To Data 1971 – 2001", is now almost complete. It has created a new kind of interface to census data, structured around topics and questions, rather than data cells and tables. The package provides access to data from 1971 to 1991 (with scope to include the 2001 data when they are ready) and crucially, makes available 1991 census data in which every datum has been corrected for the 1991 undercount. The missing million have been found.

We hope that the package will be available to the user through the internet and, provided they have registered to use the data, should provide access to census data which are linked through time which are correct counts (as far as possible) and for which no specialised knowledge is required. If you can surf the web, you can compare groups and places from 1971 to 1991.

Figure 1 – A Screen Shot of the Package Under Development



The package will have two user modes; easy and expert. The easy-mode user chooses *what* topics they want their data to cover, *where* they want the data to describe (in terms of what types of areal units they would like their data provided for) and *when* they want their data to describe, in terms of the years they would like to cover (including the option to use the 1991 data which have been corrected for the under count). Comprehensive help files guide the novice user through the process.

The easy-mode user can choose to extract any number of census counts which describe demography, migrants, country of birth and ethnicity of residents, economic activity, household tenure and amenities, living arrangements, illness rates, household composition, residents in communal establishments, industry of employment, occupation, travel to work and car ownership and qualifications. They can then specify the type of areal units which these counts will describe. There are about 30 areal units available (all at ward level or above) and they include various country and regional boundary sets, historical sets, administrative sets, electoral sets, functional sets (including labour market areas and travel to work areas), and health and education sets.

The package is constructed in such a way that when the user choose to access a variable from different censuses, its definition is held as constant as possible from census to census, as are the areal unit definitions. This means that any changes seen in the numbers of people or households included in the variable

are, as far as possible, due to real change in society rather than an artefact of the data. The user can see how the variables are defined at each census. The package also includes a calculator function to allow the user to derive new variables (such as rates or change indicators) during the extraction process. For those who really understand the census, the package also has an expert user mode. This allows the extraction of any variables, from any census for any of the available areal unit sets via a line command which can be directly entered through the package interface. Expert user mode requires knowledge about cell and table references, but again – help files guide the user as far as possible.

Once the data selection has been made, the package extracts the required variables and allows the user to download a labelled file to their own PC for analysis. Now the user can exert time and effort on the substantive interpretation of their data, rather than their extraction. In the coming months, debate over the nature of Scotland's growth, funding, health and wealth will develop. We hope that this new means of accessing census data will help to make the debate as informed as possible.

More information about census data, including details of how to register for access can be found at www.census.ac.uk

This project was funded by the ESRC under the auspices of the Census 1999-2001 development programme