

**The research value of the census of population:  
experiences from the UK**

Angela Dale  
Cathie Marsh Centre for Census and Survey Research,  
Faculty of Economic and Social Studies,  
Manchester University,  
Manchester M13 9PL

International Seminar on  
Ensuring the Significance of the Population Census Today

Held at Kyushu University, Fukuoka, Japan  
October 26<sup>th</sup>, 1999

---

The Cathie Marsh Centre

*C C S R*

---

for Census and Survey Research

Phone: +44 (0)161-275-4721

Fax: +44 (0)161-275-4722

Email: [angela.dale@man.ac.uk](mailto:angela.dale@man.ac.uk)

<http://les1.man.ac.uk/ccsr/>



## **The research value of the census of population: experiences from the UK**

In the UK the census has been a source of innovation in terms of methods of data collection and in the range of data outputs. It now provides enormous research potential for social scientists. This paper aims to highlight the complementarity of different census outputs and, more specifically, the way in which microdata can be linked to other data sources – not necessarily from the census - and thereby greatly enhance the research value of the data.

### **1. Introduction**

The first census of population was conducted in Britain in 1801. It was planned as a response to growing concern about the population explosion and the ability of the government to feed the growing population. These concerns were fuelled by the publication in 1798 of Malthus's famous *Essay on the Principles of Population*. The agreement of Parliament was necessary for the first census to be taken, and the arguments in support of a census included the requirement to know the size of the workforce, the numbers needed for an effective militia, as well as the level of food in relation to population size. An additional reason, well ahead of its time, was the prospect that a census might encourage the social sciences to flourish. Since it began, the census has been an essential tool used by central and local government to decide needs and priorities and to allocate resources. However, in recent decades, particularly since the advent of computers, the census has become a very important resource for social scientists – bearing out the prediction made by its proponents 200 years earlier!

### **2. The development of the census and its products**

#### **2.1 The role of the census in methodological developments**

The census has not only provided a resource for social scientists, it has also played a role in developing innovative methods of data collection and processing. The requirement to count the population at a single point in time led to the development of self-completion schedules. As early as 1841 the census recognised the need to move to a system of self-completion. This introduced the idea that ordinary citizens could be given the responsibility of providing accurate information about themselves – an idea which was revolutionary to 19<sup>th</sup> century social science (Marsh, 1985).

The 1851 Census was the first to ask exact age, marital status, relationship to head of household and to collect detailed occupational information. The latter information formed the basis for the social class system used in Britain for the next 150 years and, with newly available information on mortality, opened the way to research on occupational mortality (Nissel, 1987; Marsh, 1993a). The 1851 Census was also important in producing detailed reports for districts and sub-districts of England and Wales as well as maps showing population density (Nissel, 1987).

More recently, the census has been important in developing automated methods of recording data, including electronic reading of characters (optical character recognition, OCR) and the use of this to automate coding of industry and occupational information supplied as free-form entries.

## **2.2 Aggregate Small Area Statistics (SAS)**

The 1961 Census was the first to use a computer to process the returns and this marked the point at which a standard set of statistics was made available for small geographical areas. However, it was not until 1971 that complete national coverage was achieved for small areas – thereby allowing comparisons on a consistent basis across the country. The production of local statistics for small areas has grown in volume and complexity with each successive census (Cole, 1993). These statistics represent tabulations of up to four variables, for very small areas – the smallest in 1991 were enumeration districts with an average of about 150 households. For example, a tabulation may consist of age (in 21 groups) by sex and marital status (2 groups), giving 84 different cell counts. A total of 82 separate tables were published from the 1991 census, each available for a range of different geographical areas. These tables are referred to as Small Area Statistics (SAS) throughout this paper.

The aggregate data form the back-bone of the census output; they are planned with great care well before the census is taken and are the most widely used form of output, particularly by local and central government. In part this wide usage reflects the fact that, for many years, these were the only readily available electronic form in which census data were available. However, it also reflects one of the unique aspects of the census – *that it provides data for small geographical areas which is comparable across the entire country*. This makes these data of immense value in assessing the relative needs of different local government areas across the country, and thus the amount of revenue that they should receive from central government. The aggregate statistics for small areas are also very widely used *within* local government for planning purposes (e.g. to assess the care which needs to be given to elderly people, or the numbers of children due to begin school in the coming years) and for resource allocation. Within academia, the data are widely used by population geographers concerned with the social and demographic characteristics of particular areas and the way in which they are changing over time.

## **2.3 Migration flows**

Ever since the 1961 census, questions have been included not only on current place of residence but place of residence one year previously and, in 1971 only, place of residence 5 years previously was additionally asked. These data are used to identify movement of population within the country. In particular, they allow central government to identify trends – for example, movement from the city to more rural areas – and local authorities to identify the source of flows into and out of their areas. However, the nature of these data and the formats in which they have been made

available, have made them very hard to analyse with the result that their use has been limited to those with considerable expertise. Plans for improved forms of output and access from the 2001 Census should ensure that data on migration flows become much more widely used.

## **2.4 The Longitudinal Study**

One of the other important innovations following the 1971 Census was the establishment of the ONS (then OPCS) Longitudinal Study (LS). Planning for the LS started in the late 1960s with concern about the adequacy of social statistics available to government and, in particular, information on occupational mortality and fertility (Dale, 1993). The Longitudinal Study was designed to link together a sample of records from the 1971 Census (about 500,000 individuals) with registration information from deaths and other events. By linking details of individuals' socio-economic and demographic characteristics from the census with their subsequent mortality (including cause of death), the relationship between the two could be analysed using individual level data with accepted statistical methods of risk assessment. This could also be extended to the analysis of fertility by linking registration data on births.

### **2.4.1 The structure of the ONS Longitudinal Survey**

The 1971 Census was the first full census to ask exact date of birth, rather than age. Date of birth forms the basis on which the 1 per cent sample of LS members is drawn – individuals having one of four birth dates in the year (ie a 4 in 365 chance of inclusion) are included in the sample. As registration records for births and deaths also record full date of birth these, too, can be selected for the relevant birth dates and linkage established between census and registration records *for the same individuals* (Figure 1). No new data collection is required, although the quality of the study depends on having very high and accurate levels of birth and death registration and an accurate process of matching these records with records from the census. The study has continued since the early 1970s and now contains linked census data for 1971, 1981 and 1991 with plans to include records from the 2001 Census. Data from events (e.g. deaths, births) has been continuously added to the records of LS members over this time period. A detailed description of the LS and an assessment of the quality of the data is given by Hattersley and Creeser (1994). Similar studies linking census records to registration records can be found in a number of Scandinavian countries – Denmark and Finland pioneered linkage of census information to mortality records to enable the analysis of occupational mortality (Dale, 1993).

The linkage of registration data to census records means that the LS is kept under conditions of immense security. Individual researchers are not allowed access to the LS database and all research is done by a small set of approved staff working within the Office for National Statistics. Only *results* of analyses are allowed to go out of ONS (for example, tables or results from statistical modelling) and these are carefully checked to ensure that they do not reveal any information that could lead to the identification of an individual. However, this tight security also means that considerable

individual detail can be retained on the database that allows the analyses described in the following sections.

#### **2.4.2 The research potential of the LS**

The LS provides a unique basis for calculating standardised mortality rates (SMRs). Socio-demographic information is obtained from a sample of individuals drawn from the census. Mortality information comes from death records and therefore the two sources are collected independently but can be linked for the *same individuals*. This provides much better quality data than occupational information collected at death registration. Analysis of the LS has produced important insights into mortality differences by social class and their persistence over time (Fox and Goldblatt, 1982; Bartley et al, 1996). It has also allowed mortality differentials to be analysed using alternative social classifications and for different subsets of the population – including minority ethnic groups (Harding and Maxwell, 1997).

The linkage of records for the same individuals from successive censuses also provides a unique basis for establishing individual change over time. This has facilitated longitudinal studies of ethnic differences in women's employment patterns (Holdsworth and Dale, 1997); analysis of socio-demographic variation in the movement of elderly people into institutions (Grundy, 1992; Grundy and Glaser, 1997) and analysis of the process of change in particular urban areas (Lyons, 1996).

The LS also provides unique data on fertility. By linking information from the baby's birth registration to census records for the mother, the LS allows analysis of birth intervals (Werner, 1984, 1988a, 1988b; Ni Bhrolchain, 1987) and of birth weight by the socio-economic characteristics of the mother.

The LS also holds detailed geographical information for the members of the study which, within the procedures for maintaining confidentiality, is available for analysis. This allows, for example:

- Analysis of migration data that extend much beyond that available from any one census and which also includes information from registration records.
- Analysis of geographical variation in health using multilevel modelling methods that assess the role of place having controlled for individual-level variation (Wiggins et al, 1998)
- Analysis of the association between specific causes of mortality (e.g. cardiovascular disease) and environmental factors such as water-hardness and air pollution (Britton, 1990).
- The impact of internal migration of elderly people to popular retirement areas such as the south coast of England (Grundy, 1987).

The fact that the Longitudinal Study is maintained under very tight security inside ONS means that variables can be held in a disaggregated form. In the case of geography, much more detail is available on the LS than on census output that is released to the public. Although the analysis that is released must not contain too much detail, the

researcher using the LS can decide how to aggregate particular variables and does not have to accept a classification imposed *a priori*. The LS has been described as providing data in a *safe setting*, in contrast to the Samples of Anonymised Records which provide *safe data* (Marsh et al, 1994).

More details of the ONS Longitudinal Study can be found on the web site of the ESRC-funded Support Unit at the Institute of Education, London:

<http://www.cls.ioe.ac.uk/Ls/lshomepage.html>

## **2.5 The Samples of Anonymised Records**

### **2.5.1 The structure of the SARs**

The most recent innovation from the census has been the Samples of Anonymised Records (SARs) from the 1991 Census (Marsh, 1993c). Two files were released by the Census Offices for Great Britain and Northern Ireland:

#### *The Individual SAR*

- two per cent sample of individuals in households and communal establishments; this comprises 1.1 million records containing information on:
  - individual characteristics, e.g. age, sex, employment status, occupation and social class;
  - accommodation (e.g. availability of bath/ shower; tenure )
  - limited information about the individual's family head - sex, economic position and social class
  - limited information about other members of the individual's household (such as the number of persons with long-term illness and numbers of pensioners).

The finest geography is the local authority (the main unit of local government), with a minimum population size of 120,000. Sub-threshold areas are grouped with neighbouring areas.

#### *The Household SAR*

- one per cent hierarchical sample of households and individuals in those households; this comprises 215,000 households and the 542,000 individuals enumerated in them. The full range of census variables is available, with standard region as the lowest level of geography. (Great Britain was divided into 12 standard regions in 1991.)

Similar files are planned for the 2001 Census, although there are proposals to increase the sample size of the Individual SAR to 3 per cent and to reduce the population threshold.

This represents the first time that census users have been able to access, on their own PCs, individual records from the census. The data have been anonymised and the amount of detail released has been carefully assessed to ensure that the chance of any individual or household being identified is negligible. The data are therefore considered to be *safe*, although an additional safeguard comes through the licensing system. This

requires all SAR users to sign a legally binding undertaking to respect the confidentiality of the data and to ensure that it does not get passed to any individual or organization which is not registered to use it.

### **2.5.2 The research potential of the SARs**

Since their release in 1993, the SARs have been extremely widely used. The hierarchical structure of the data allows inter-relationships to be made between individuals within the same household or family, thereby facilitating the detailed analysis of family and household composition (Holdsworth and Dale, 1997).

(see Figure 2).

It also allows employment behaviour to be analysed in relation to family characteristics: for example, the age and number of children, the employment status of a partner. Whilst the Longitudinal Study also allows these linkages to be made, direct access to the microdata in the SARs provides a great deal more control for the researcher. For example, the data can be held on CD-ROM and used on a desktop PC, allowing analysis to proceed much more quickly. However, a growing number of researchers are using both the LS and the SARs as complementary data sources. For example, Holdsworth and Dale (1997) developed an analysis of ethnic group differences in women's employment patterns using the 1991 SARs and then modelled change in occupational attainment over time (from 1971-19981 and from 1981-1991) using the LS.

The large sample size of the SARs has been very important in allowing analysis of small groups with sufficient numbers to produce reliable estimates (e.g. minority ethnic groups; single parents). The hierarchical data of the Household SAR has extended this to support a nationally-representative study of family composition and partnership patterns amongst different ethnic groups (Holdsworth and Dale, 1995). The Individual SAR, with its even larger numbers, has played an important role in identifying ethnic differentiation and highlighting incongruities between educational attainment and levels of unemployment (Blackburn et al, 1997) and the occupational level achieved (Heath and McMahon, 1997).

The detail available in the datafiles is extensive: over 40 variables are recorded, many with considerable detail (e.g. single year of age until age 90 and then grouping; 42 categories for country of birth; 10 ethnic groups), as well as numerous derived variables. Table 1 shows how the amount of detail on key variables has been reduced on the two files.

**Table 1 Examples of reduction in categories in the 1991 SARs by comparison with the full census coding**

	<b>Full coding</b>	<b>1% SAR</b>	<b>2% SAR</b>
Occupation	371	258	73
Industry codes	334	185	60
Educational qualification (subject)	108	88	35
Country of birth	102	42	42
Ethnic group	35	10	10

This detail, together with the large sample size, provides the capacity for analysis of particular occupational groups, or those who have a specific educational qualification (e.g. doctors or teachers). This can be important in manpower planning - for example, to provide an estimate of the extent of qualified labour in the country; where people with particular qualifications are located; and what proportion are not currently in employment.

The detailed occupational and employment status variables have also been very important in providing the 'building blocks' which allow a range of alternative social classifications to be derived, based upon different ways of grouping these building blocks. Some of these have been specifically developed to facilitate international comparisons - for example, the International Standard Occupational Classification (ISCO), the Standard International Occupational Scale (SIOPS) and the International Socio-economic Index of Occupational Status (ISEI) (Ganzeboom and Treiman, 1995; Ganzeboom et al, 1992; ILO, 1990).

Microdata files also allow the analyst to choose their unit of analysis. There is a choice of working at the level of the individual, family or household. Further choices arise over the population to be analysed - for example, whether a full age range is used or restricted groups. Thus children can be selected and an analysis conducted of the circumstances of the families in which they are living. Alternatively analyses can focus upon those of school leaving age (Drew et al, 1997) or upon the elderly (Glaser et al, 1997). Choices also arise over analysis of those living in private households or institutions. Census microdata files are unique in allowing analysis of those living in residential homes, hospitals, prisons or army quarters. These choices provide the analyst with maximum flexibility - although they also require considerable care in ensuring that the most appropriate population has been selected. This contrasts strongly with aggregate census tables where the unit of analysis and population is pre-specified and may not necessarily meet the needs of the analyst.

The fact that census data include the institutional population is of importance for a number of applications. Murphy and Wang (1996) use the SARs to make marital status

population projections for England and Wales, where sample surveys cause problems because of the omission of the institutional population. Murphy et al's (1997) study of the health of the elderly provides a further example where the inclusion of the institutional population is essential because those with poorer health are more likely to be living in institutions, particularly if they are single. Information on migration in the census can help to understand differences in the movement of elderly people depending on their marital status and other characteristics (Al-Hamad et al, 1997). The data also allow analysis of inward migration in the last year and have been used to add important evidence to the debate over the role of social housing in restricting residential mobility and thus employment opportunities (Boyle, 1995).

#### *Detailed geography*

The Individual SAR, with a population threshold of 120,000 provides more geographical detail than any other comparable microdata file in the UK – although much less than the aggregate small area statistics and less than the LS. The ability to identify geographical areas at a relatively detailed level allows 'place' to be included in the analysis by the use of multilevel modelling methods (Goldstein, 1995). For example, in an analysis of unemployment one might want to ask whether the area in which someone lives has an effect on their probability of unemployment *additional* to any individual characteristics such as age or level of educational qualifications. The geographical areas available in the 2 percent Individual SAR (local government areas) can be aggregated to approximate to local labour markets and multilevel modelling then used to identify the effect of a particular type of labour market on an individual's probability of unemployment. Fieldhouse and Gould (1998) have conducted an analysis based on the SARs to establish the role of local labour markets in unemployment amongst different ethnic groups, *over and above* the effect of individual characteristics such as age and qualifications. The results showed that there were, indeed, area-level effects. By including in the analysis explanatory variables which related to the local labour market, drawn from the 100 per cent small area statistics, Fieldhouse and Gould were also able to identify how the particular characteristics of the area impacted. For example, areas with high level of unemployment were particularly disadvantageous to Black Caribbeans. In contrast, Asian groups were less affected by the local unemployment level. Thus using multilevel models which include information about the individual and their local labour market, provides a powerful means of analysis that capitalises on the linkage between microdata and aggregate tabular data. In the following section we explore in more detail different ways by which record linkage may enhance microdata.

More information about the SARs can be obtained from the web site of the ESRC-funded Census Microdata Unit at the University of Manchester:

<http://les1.man.ac.uk/ccsr/cmu/>

### **3. Linking microdata**

#### **3.1 Exact record linkage**

The extent to which microdata can link with other data sources, not necessarily from the census, is only now becoming fully recognised. We have already seen that the census microdata in the ONS Longitudinal Study can be linked to records drawn from birth and death registration using exact matching techniques. Linkage has been extended to other registration records collected by ONS which contains exact date of birth (e.g. cancer registration) and can, in principle, be expanded to take in other data sources (e.g. marriage records) as long as the matching keys are available. This one-to-one exact record linkage is unique to the ONS Longitudinal Study and can only take place for records recorded for a large section of the population and which contain exact date of birth. Whilst it produces a unique data source, it is very resource intensive and the data need to be kept under high security. Maintaining the confidentiality of the LS members is always of paramount concern in any discussion of extending the database.

However, exact record linkage is being increasingly used by national statistics offices as a way of avoiding the cost of new data collection. The public acceptability of linkage varies between countries and is always approached with great caution. In some European countries, (e.g. Netherlands) a census of population is no longer taken; instead, the necessary population information is obtained from administrative records using exact linkage methods.

#### **3.2 Linkage of aggregate data**

All microdata allow records to be matched on the basis of aggregate characteristics. This can be illustrated by describing a measure of social standing that has been added to each person in the SARs and the LS who has information on occupation and employment status. This measure is a social status score that has been derived by Prandy et al (1992) using a range of different data sources. However, the score is unique to each occupation coded in the ONS Standard Occupational Classification. The score can be added to any dataset that contains detailed occupational information. Thus all individuals with the same occupation and employment status (e.g. all managers of small shops) are given the same occupational score. In a similar way aggregate earnings information has been added to individuals who provide occupational details. Mean hourly earnings were derived from an employer-based survey of incomes in the form of a large table broken down by occupation, age, sex, full or part-time working and region. The availability of all these matching keys on the microdata files, coded in the same way in both data sources, allowed this 'earnings score' to be matched to all individuals who reported an occupation. Details of the addition of aggregate data to individual records in the SARs and the LS can be obtained from their respective Web sites.

The role of aggregate census statistics in multilevel models, mentioned in the previous section, is another example of linkage of aggregate data to individual records. In this

example, a model is developed which recognizes the importance of both individual and area-level characteristics in explaining the probability of unemployment. Explanatory variables can be included at both levels – individual and area. Individual explanatory variables come from the SARs whilst area-level explanatory variables come from the aggregate Small Area Statistics. Similar approaches have been used in modelling women’s employment patterns (Ward and Dale, 1992) and the incidence of ill health (Gould and Jones, 1996). This example is developed further by the addition to the SARs of an area-level classification, discussed in the next section.

#### **4. Complementarily between SAS and SARs**

There is considerable scope for exploiting the complementarily of the various forms of output from the Census. It is evident from the preceding paragraphs that there is a wide range of outputs, from aggregate tables for very small areas (SAS) using 100 percent of the data to microdata samples of 1 or 2 percent of the population which contain very extensive detail for each individual (SARs). Both types of output ensure the confidentiality necessary for release outside the census offices. The Small Area Statistics (tabular 100 percent data) ensure confidentiality by restricting the amount of detail in a table, while samples of microdata rely heavily on the small sampling fraction and restricted geographical detail for confidentiality.

However, both sets of data are taken from the Census database, with a common set of questions, which relate to the same point in time. This allows relationships between the two to be drawn in a variety of different ways, all of which provide scope for analysis which extends that which would be possible with either data source taken alone. In the following sections we explore some of these relationships.

##### **4.1 Area classifications**

A similar matching method to that described in section 3.2 has allowed extra geographical information to be added to the SARs. However, the geographical information added comes from the 100 percent Small Area Statistics, in the form of an area-level classification which describes the locality in which the individual lives. Different classifications have been added to each of the SARs but both classifications are based on aggregate Small Area Statistics from the census. For example, the classification added to the 1 percent Household SAR has been derived by the Office for National Statistics (Wallace et al, 1995; Wallace and Denham, 1996) from the 100 per cent small area statistics for wards (a ward is the lowest level of administrative geography in Britain, with an average size of about 1,000 households). This classification has 14 different categories into which each ward in the country is classified. Thus every ward is allocated a value from 1-14 with descriptive material which gives the defining characteristics of each category. The categories of the classification are given below:

- 1                    Suburbia
- 2                    Rural Areas

3	Rural Areas with mixed economies
4	Industrial & Manufacturing Towns
5	Middling England
6	Prosperous wards
7	Purpose-built, Inner City estates
8	Established Owner-Occupiers
9	Armed Forces bases
10	Metropolitan professionals
11	Deprived City Areas
12	Lower Status Owner Occupiers
13	Mature Populations
14	Deprived Industrial Areas

The addition of this classification to the SARs has to be done by ONS as it requires accessing confidential information about the ward in which the sample member lives. The availability of this classification provides an additional geographical dimension to the SARs<sup>1</sup> which considerably enhances their research potential and provides an example of the way in which aggregate small area census statistics have been used to produce a classification which is linked to individuals in the census microdata files.

#### **4.1.1 The research potential of area classifications**

The addition to the SARs of a variable which describes the type of area in which an individual lives allows the multilevel models described in section 2.5.2 to be extended by fitting cross-classified hierarchical multilevel models where individuals (level 1) are cross-classified by SAR areas (geographical contexts) and area classification (residential contexts) both at level 2 (Jones *et al* 1997, Goldstein 1994). These methods are being used in an ESRC-funded project (<http://les1.man.ac.uk/ccsr/research.htm/>) to investigate the variation in unemployment between individuals, taking into account not just the local labour market in which they live, but also the type of locality within that area – as identified by the area classification (Fieldhouse and Tranmer, forthcoming). The scope for multilevel modelling is extended even further by combining these data with Small Area Statistics from the main census database (Tranmer and Steel, 1998).

#### **4.2. Using SAS to improve the precision of estimates from the SARs**

There is considerable further scope for exploiting the complementarity between the SAS and the SARs. One of the disadvantages of census microdata is the lack of precision in estimates based on small samples, particularly when used at the level of local authority. For example, a 2 percent sample for Manchester local authority represents only 8,000

---

1

The ONS classification has been added to the ward of each household in the Household SAR and GB Profiles has been added to the enumeration district of each individual in the Individual SAR. Full details are given in SARs Newsletter No.8, accessible from the website: <http://les1.man.ac.uk/ccsr/>

individuals and for small authorities this may be as low as 2,500 individuals. However, microdata can provide important cross-tabulations which are not available with the 100 per cent SAS. For example, detailed economic activity rates by age and sex for each ethnic group are required for labour force forecasts in multi-ethnic areas but can only be obtained from SARs (Bradford City Council, 1996). For subgroups such as the elderly in rented accommodation, or children in one-parent families, the SARs allow detailed individual-level analyses not possible with the SAS.

However, small numbers result in large confidence intervals around any estimates. To overcome the problem of small cell size Bradford Council (1996) supplemented SAR data for Bradford with SAR data from larger regions containing Bradford – West Yorkshire and England and Wales – until the sample reached a minimum of 100 for a given cell. Nonetheless, marginal distributions of the sample data will not be exactly the same as the 100 percent data - due in part to sampling error and also because the SARs do not contain imputed records for households missed by the census. The precision of estimates can be improved by using the univariate or bivariate distributions of key variables obtained from the 100 percent Small Area Statistics. Simpson (1998) describes how the marginal numbers in each age category obtained from the sample data can be scaled to give numbers consistent with the numbers obtained from the 100 percent count available for age using Iterative Proportional Fitting (IPF).

### **4.3. Synthetic estimation of non-census data**

Census data are often used to estimate a relationship at a local level that has been established using a national sample survey. The link between the two data sources is made using variables common to both. Estimates of non-census characteristics in the local area are based on the assumption that the national relationships also hold locally. This is termed *synthetic estimation* and is often carried out with 100 percent SAS tabulations –for example when disability rates by age and sex have been estimated from a national survey and are then applied to the local area using census statistics to give the age-sex structure. However, in some applications the use of census microdata is essential. An example comes from work by Charlton (1998) who used a national survey to predict rates of serious illness which were then imputed to local authority areas using the SARs. The variables used to predict the probability of serious illness in the national survey – age, sex, employment position, tenure, social class and family type – are also available in the census. It was necessary to use the microdata samples (SARs) rather than the 100 per cent Small Area Statistics for two reasons. Firstly, because the regression model was non-linear and, secondly, because the model contained a large number of predictor variables which were different for each age-sex category and could only be obtained using census microdata. The predictions from the model were therefore applied to each individual in the 1991 Individual SAR and the probability of each individual having a serious illness was estimated. When added, these probabilities provided an estimate of the total number of individuals with a serious illness in each of the SAR areas. In this example, use of the SARs allowed relatively accurate synthetic estimates to be made for local authority areas with a minimum population of 120,000.

Ideally, estimation would have been made at even smaller areas but this was not possible because microdata files were not available for smaller geographical areas.

#### **4.4 Micro-simulation of whole populations**

A final example of the way in which the aggregate small area statistics (SAS) and the SARs can complement each other comes from attempts to simulate individual records for the small areas in the SAS. A key reason for attempting this microsimulation is to overcome the limitations of tabular data and to try to re-create the benefits of individual-level data at very small areas – for example, to allow the use of multivariate statistical methods for individuals or households within small geographical areas.

The method locates a sample of microdata from the SARs that best match the aggregate tabular characteristics of the areas selected in the SAS. Williamson et al. (1998) describe and compare algorithms for searching the SARs for sub-samples that match chosen SAS tabulations. The aim is to develop an algorithm that provides the best fit between the samples of microdata and the small area tabular data. Williamson et al. validate the simulation by recreating tabulations that were not used in the simulation itself and report encouraging results.

In practice, microdata sub-samples cannot recreate exactly the characteristics of individuals in a SAS table. In part this is because of differences between samples of microdata and Small Area Statistics with respect to coverage and data modification. In addition, small areas which are very distinctive (for example those largely composed of hotels and hostels) are more difficult to simulate than ‘average’ areas, whose households will be similar to many represented in the national SARs.

These micro-simulation methods are extremely computer-intensive and are not, at the moment, widely used in research applications. They are, however, invaluable in developmental work (Martin, 1998) where individual records are needed to develop methods for defining geographical boundaries for census data. Once the methods are available they can be applied by the Census Offices to the full census database within the safe confines of the national statistical office.

With the increased availability of computational power, micro-simulation methods may become more widely used to overcome the confidentiality restrictions which prevent the release of individual records at small areas. However, there are also developments to provide secure on-line access to the 100 per cent census database that, if feasible, might provide the benefits of microdata whilst also retaining confidentiality (Rees, 1999). An alternative approach which is also being developed (Dale and Elliot, 1998) is assessing the feasibility of providing safe microdata samples at relatively small areas.

#### **4. Conclusions**

It is evident from this discussion that the range of data sources available in the UK Census of Population now provides enormous research potential. In particular this paper has highlighted the way in which microdata, either from the LS or the SARs, can be linked to complementary data sources and thereby greatly enhance the research value of the microdata. In the case of the LS this is most importantly through direct matching with registration data – thereby creating a much more extensive dataset. For Samples of Anonymised Records, we have shown that there is enormous scope to exploit the complementary relationship with the 100% Small Area Statistics. Linkage between the SARs and the SAS does not involve direct matching of records but is based on the fact that these two sets of data based on the same data source. This allows relationships between the two to be established in a variety of different ways, all of which provide scope for analysis which goes well beyond that which would be possible with either data source taken alone.

Future developments will be facilitated by plans for a One Number Census for 2001 (Brown et al, 1999). Imputation of missing individuals and measures to protect confidentiality will take place before establishing a common census database from which all forms of output will be drawn. This will be a major step forwards in increasing the scope for linkage between different census outputs.

#### **Acknowledgements**

This paper draws on material in *Analyzing Census Microdata*, to be published by Arnold in summer 2000 and includes contributions from Ludi Simpson, Ed Fieldhouse and Clare Holdsworth.

I am grateful to ESRC for support through the Census Programme (Award No.H507 25 5140)

#### **References**

- Al-Hamad, A., L. Hayes, and R. Flowerdew (1997) "Migration of the Elderly to Join Existing Households: Evidence from the Household SAR," *Environment & Planning A* 29, No.7, 1243-55
- Bartley, M., Carpenter, L., Dunnell, K. and Fitzpatrick, R. (1996) "Measuring inequalities in health", *Social Health and Illness*, 18, pp 455-475.
- Blackburn, R. M., A. Dale, and J. Jarman (1996), "Ethnic Differences in Attainment in Education, Occupation and Lifestyle," in *Employment, Education and Housing among Ethnic Minorities in Britain*, V. Karn, ed. London: HMSO,

Blane, D, Harding, S, and Rosato, M (1999) "Does social class mobility affect the size of the socio-economic mortality differential? evidence from the Office for National Statistics Longitudinal Study", *Journal of the Royal Statistical Society Series A*, 162, Part 1, pp 59-70

Bradford Council (1996) *Forecasts of the labour force: technical report*. Corporate Services, City Hall, Bradford BD1 1HY.

Britton M. (Ed.) (1990) *Mortality and Geography: decennial supplement*, London: HMSO

Brown, J.J., I. D. Diamond, R.L. Chambers, L. J. Buckner and A. Teague (1999) A methodological strategy for a one-number census, *J.R.S.S. (A)*, 162, pp.247-267

Boyle PJ 1995 Public housing as a barrier to long-distance migration *International Journal of Population Geography* 1 147-64

Charlton J (1998) Use of Census sample of Anonymised Records (SARs) and survey data in combination to obtain estimates at local authority level. *Environment and Planning Series A*, 30, pp.775-784

Cole, K. (1993) The 1991 Local base and Small Area Statistics in Dale, A. and Marsh, C. (eds) *The 1991 Census User's Guide*, London: HMSO

Dale, A. (1993) The OPCS Longitudinal Study in Dale, A. and Marsh, C. (1993) *The 1991 Census User's Guide*, London: HMSO

Dale, A. and Elliot, M. (1998) A report on the disclosure risk of proposals for SARs from the 2001 Census, CCSR Working Paper No 5

Dale, A. and Marsh, C. (1993) *The 1991 Census User's Guide*, London: HMSO

Dressler W. W. (1994) 'Social status & health of families', *Social Science and Medicine* 39, 1605-1613.

Drew, D., Gray, J. and Sporton, D. (1997) Ethnic differences in the educational participation of 16-19 year olds, in V. Karn (ed) *Ethnicity in the 1991 Census: Employment, education and housing among the ethnic minority populations of Britain*, London: HMSO

Fieldhouse, E. and Gould, M. I. (1998) "Ethnic minority unemployment and local labour market conditions in Great Britain", *Environment & Planning A*, Vol 30, No 5, 833-853.

Fieldhouse, E. and Tranmer, M. (forthcoming), 'Spatial mismatch or residualisation? Exploring labour market and neighbourhood variations in unemployment risk using census microdata', paper to be presented at the Institute of British Geographers Conference, Brighton, January 2000

Fox, J. and Goldblatt, P. O.(1982) 1971-1975 *Longitudinal Study Socio-demographic mortality differentials*, LS Series 1, London: HMSO

Ganzeboom, Harry and Donald Treiman. 1996. "Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations." *Social Science Research* 25:201-39.

Ganzeboom, Harry, Paul DeGraaf, and Donald Treiman. 1992. "A Standard Socio-Economic Index of Occupational Status." *Social Science Research* 21:1-56.

Glaser, K. and Grundy, E. (1998) "Migration and household change in the population aged 65 and over, 1971-1991", *International Journal of Population Geography*, **4**, pp 1-17

Glaser, K., Murphy, M. and Grundy, E. 1997. Limiting long-term illness and household structure among people aged 45 and over, Great Britain 1991, *Ageing and Society*, **17**, 3-19.

Goldstein H (1994) 'Multilevel cross-classified models', *Sociological Methods and Research* **22**, p.364.

Goldstein, H (1995) *Multilevel statistical models*, 2<sup>nd</sup> Edition. Arnold: London

Gould MI and Fieldhouse E (1997) Using the 1991 Census SAR in a multilevel analysis of male unemployment *Environment and Planning A*, **29**, 611-628.

Gould MI and Jones K (1996) Analysing perceived limiting long-term illness using UK Census microdata *Social Science and Medicine* **42**, 857-869

Grundy, E. (1987) Retirement migration and its consequences in England and Wales, *Ageing and Society*, **7**, (1), 57-82

Grundy, E. 1992b. Socio-demographic variations in rates of movement into institutions among elderly people in England and Wales: an analysis of linked census and mortality data 1971-1985. *Population Studies*, **46**, 65-84.

Grundy, E. and Glaser, K. (1997) Trends in, and transitions to, institutional residence among older people in England and Wales, 1971 to 1991. *Journal of Epidemiology and Community Health*,.

Harding, S and Maxwell, R (1997) "Differences in mortality of migrants", in Drever, F and Whitehead, M (Eds) *Health Inequalities: decennial supplement*, ONS Series DS no. 15, London; The Stationery Office, pp 108-121

Hakim, C. (1998), *Social Change and Innovation in the Labour Market*. Oxford: Oxford University Press.

Hattersely, L. and Creeser, R. (1994) *The OPCS Longitudinal Study*, London: HMSO

Heath, A. and McMahon, D. (1997) 'Education and occupational attainments: the impact of ethnic origins', pp.91-113, in V. Karn (ed) *Ethnicity in the 1991 Census: Employment, education and housing among the ethnic minority populations of Britain*, London: HMSO

Holdsworth, C. and Dale, A. (1995) 'Ethnic Homogeneity and Family Formation: Evidence from the 1991 Household SAR', *CCSR Occasional Paper*, No 7. Manchester, CCSR.

Holdsworth, C. and Dale, A. (1997) "Ethnic group difference in women's employment", *Work, Employment and Society*, Vol 11, No.3, pp 435-457

International Labour Office (1990) *International Standard Classification of Occupations 1988*

Jones K (1994) Using multi-level modelling with area level data in the Longitudinal Study In *LS User Guide on Analysis Issues in the OPCS Longitudinal Study* (Edited by Creeser R.), SSRU, City University, London.

Jones K and Bullen N (1994) Contextual models of urban house prices: a comparison of fixed- and random-coefficient models developed by expansion, *Economic Geography* 70, 252-272.

Jones K, Gould MI, and Watt R (1997) Multiple contexts as cross-classified models: the Labour vote in the British General Election of 1992' *Geographical Analysis*, in press.

Lyons, M. (1996) "Gentrification, socio-economic change and the geography of displacement", *Journal of Urban Affairs*, 18:1, pp39-62

Marsh, C. (1985) 'Informants, respondents and citizens' in M. Bulmer (ed), *Essays in the History of British Sociological Research*, Cambridge: Cambridge University Press

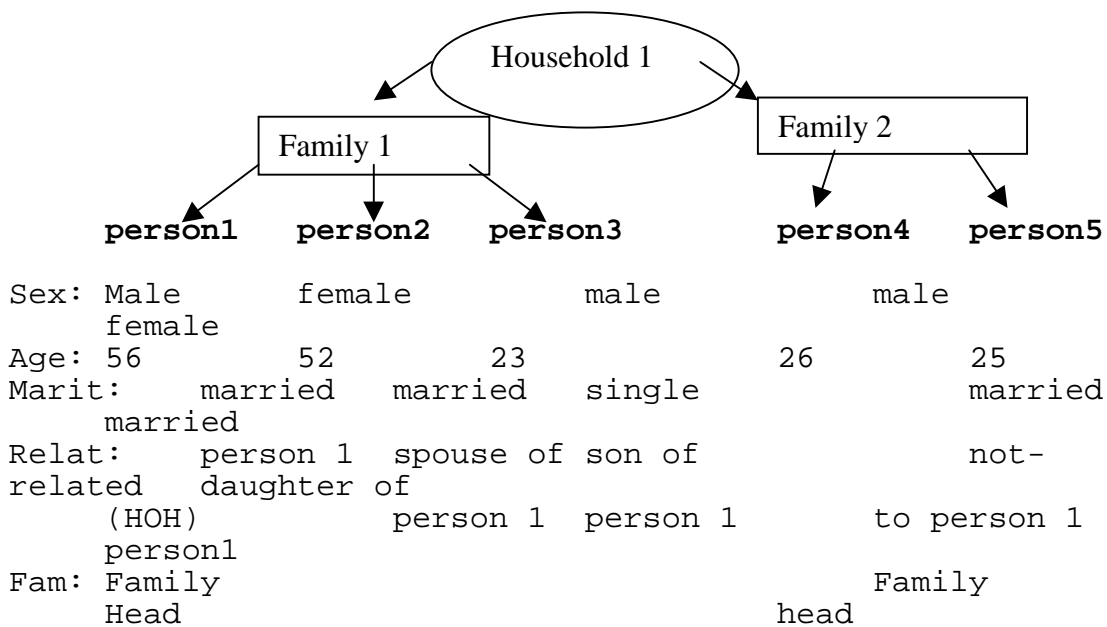
Marsh, C. (1993a) 'An overview' in Dale, A. and Marsh, C. (eds) *The 1991 Census User's Guide*, London: HMSO

Marsh, C. (1993b) Privacy, confidentiality and anonymity in the 1991 Census, in Dale, A. and Marsh, C. (eds) *The 1991 Census User's Guide*, London: HMSO

- Marsh, C. (1993c) The Samples of Anonymised Records in Dale, A. and Marsh, C. (eds) *The 1991 Census User's Guide*, London: HMSO
- Marsh, C., Dale, A. and Skinner, C. (1994) 'Safe Data versus safe Settings: Access to Microdata from the British Census' *International Statistical Review*, 62, 1, 35-53
- Marsh, C.; Skinner, C.; Arber, S.; Penhale, P.; Openshaw, S.; Hobcraft, J.; Lievesley, D.; Walford, N. (1991). The Case for a Sample of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society Series A*, 154, 305-340.
- Martin, D. (1998) Optimizing census geographies: the separation of collection and output geographies, *International Journal of Geographical Information Science*, 12, 673-685
- Murphy, M., Glaser, K. and Grundy, E. 1997. Marital status and long-term illness in Great Britain. *Journal of Marriage and the Family*, 59, 156-164.
- Murphy, M. and D. Wang (1996), "A dynamic multi-state projection model for making marital status population projections in England and Wales," in *Exploiting National Survey & Census Data*, CCSR Occasional Paper 10, A. Dale, (ed.) Manchester: CCSR, University of Manchester, .
- Ni Bhrolchain, M. (1987) "Period parity progression ratios and birth intervals in England and Wales, 1941-1971: a synthetic life table analysis", *Population Studies*, Vol 41, pp103-125
- Nissel, M. (1987) *People Count: A History of the General Register Office*, London: HMSO
- Prandy, K. (1990); Revised Cambridge scale of occupation, *Sociology* Vol.24. No.4.
- Prandy, K. (1992) Cambridge Scale Scores for CASOC Groupings, Working paper, No 11, Social and Political Sciences, Cambridge.
- Robinson, V. (1996b) "Inter-generational differences in ethnic settlement patterns in Britain", In: Ratcliffe, P. (ed.) *Ethnicity in the 1991 Census*, Vol. 3 'Social Geography and Ethnicity in Britain, London: HMSO, Chapter 5, pp 175-201.
- Rees, P.H. (1999) 'The case against a third SAR' Paper presented at the Manchester Meeting on the 2001 SARs organised by the Cathie Marsh Centre for Census and Survey Research, University of Manchester, 21 May
- Simpson, S. (1988) (ed) *Making local population statistics, a guide for practitioners* LARIA

- Tranmer M and Steel, D G (1998) Using census data to investigate the causes of the ecological fallacy. *Environment and Planning Series A*, 30,
- Wallace, M. and Denham, C. (1996) The ONS classification of wards. SMPS 60. (London: HMSO)
- Wallace, M., Charlton, J. and Denham, C. (1995) 'The new OPCS area classification', *Population Trends*, 79, 15-30
- Ward, C. and Dale, A. (1992) "Geographical variation in female labour force participation: an application of multilevel modelling", *Regional Studies*, Vol 26:3, pp 243-255
- Werner, B. (1984) "Some examples of fertility analysis from the Office of Population Censuses and Surveys' Longitudinal Study", *Population Trends*, 35, pp 5-10
- Werner, B. (1988a) "Birth intervals: results from the OPCS Longitudinal Study 1972-84", *Population Trends*, 51, pp 25-29
- Werner, B. (1988b) "Spacing of births to women born 1935-1959: evidence from the OPCS Longitudinal Study", *Population Trends*, 52, pp 20-25
- Wiggins, R, Bartley, M, Gleave, S, Joshi, H, Lynch, K and Mitchell R (1998) "Limiting long-term illness: a question of where you live or who you are? A multilevel analysis of the 1971-199 ONS Longitudinal Study" *Risk, Decision and Policy*, 3 (3), pp 181-198. Routledge
- Williamson, P., M. Birkin, and P. Rees (1998) "The simulation of whole populations using data from small area statistics and samples of anonymised records," *Environment & Planning A*, 30, 785-816.

Figure 2 Diagram of relationship between households, families and individuals



**Figure 1 Diagrammatic representation of the ONS Longitudinal Study**

