Using data from the 1991 census

F Azeem Majeed, Derek G Cook, Jan Poloniecki, David Martin

The 1991 census for England and Wales provides a substantial amount of data on demography, ethnicity, housing tenure, employment status, and other social factors for geographical areas ranging in size from enumeration districts upwards. Many in the health service and in the academic community are making use of the data in the 1991 census. However, users of census data need to be aware of the problems and limitations of these data, which include the format of the data, data modification and suppression, sampling error, and underenumeration. An important innovation of the 1991 census was that the census form included a question on the postcode of respondents; this allowed the Office of Population Censuses and Surveys to produce a postcode-enumeration district look up table which overcomes many of the problems previously encountered in trying to assign postcodes to enumeration districts. The new look up table also includes the grid reference of postcodes, and this will improve the geographical referencing of census data.

Introduction

The 1991 census provides the most detailed data on the demography of England and Wales. It also provides a substantial amount of data on deprivation, housing, and employment and was the first census to contain questions on ethnicity and chronic illness. Many in the health service will make use of the information obtained from the census; purchasers will use the data to help them improve their purchasing plans and their planning of health services, providers to help them tailor the services they offer to their catchment population, and health services researchers to compare measures of health need and the use of health services with demographic, social, and ethnic factors.^{1,2} In this paper we discuss the data available from the census and describe the main limitations of census data (see box 1 for a list of definitions). More detailed guidance is available in publications from the Office of Population Censuses and Surveys (OPCS).^{3,4}

Data available from the census

The census is carried out every 10 years by OPCS, and aims to collect data on the entire population. (However, there was a problem with underenumeration in the 1991 census; this is discussed later in this article.) The 1991 census contained about 20 questions including questions on housing, race, limiting long term illness, car ownership, and occupation. The questions were collected for all 113 196 enumeration districts (the basic geographical building blocks of census data and the smallest areas for which census data are made available) in England and Wales. The data in these questions were then analysed by OPCS to produce a number of predefined tables, available for different geographical areas including enumeration districts, electoral wards, and health and local authorities. Users of census data are confined to using the information in these tables unless they are prepared to commission OPCS to produce additional tables. Census data have been purchased by health authorities and are available to NHS employees. Census data

Box 1—Definitions of census terms

Enumeration district—The smallest geographical area for which census data are published
Local base statistics (LBS)—Tables of census data published for electoral wards and all larger areas
Small area statistics (SAS)—Tables of census data published for enumeration districts and all larger areas
Data modification—The random addition of +1, -1 or 0 to cells in SAS tables, and the random addition of +2, +1, 0, -1 or -2 to cells in LBS tables
Data suppression—The suppression of data in SAS and LBS tables for areas with small populations
Importing zones—Enumeration districts or electoral wards that contain data from neighbouring areas with small populations for which data have been suppressed
Special enumeration districts—Enumeration districts which were expected to contain 100 or more people living in communal establishments (such as hospitals or prisons) at the time of the 1991 census
Shipping enumeration districts—A special category of enumeration district which contains data for people living on ships at the time of the 1991 census
Part-postcode unit—The intersection of a postcode and an enumeration district
Pseudo-enumeration district—The enumeration district in which most people at a postcode live

are also freely accessible to employees of academic institutions.

The census data published in the available tables are in two formats: local base statistics and small area statistics.⁵ The local base statistics contain 99 tables and are published for geographical areas down to electoral ward level. The small area statistics are a cut down version of the local base statistics and are published for all geographical areas down to enumeration district level. Because the small area statistics are available for enumeration districts, there are fewer tables in them, and the tables that are present usually contain fewer categories (for example, broader age bands) than the equivalent tables in the local base statistics; this is done to maintain confidentiality and to ensure that individuals cannot be identified. However, the small area statistics are the only source of census data for enumeration districts and the data in them have to be used for all analyses that require census data for small areas. Most of the commonly used tables are present in both the local base statistics and small area statistics, and a restricted range of tables in the small area statistics is not usually a problem. Some of the more commonly used tables from the small area statistics are listed in box 2.

Both the local base statistics and small area statistics are available on disks for IBM personal computers, and the data in them can be accessed through several computer programs. The most commonly used program, SASPAC,⁶ can extract census data from local base statistics and small area statistics tables and save the data in a variety of file formats (such as dBase IV or Lotus 1-2-3) to facilitate transfer of the data to other computer programs. People who are inexperienced in working with census data often find census tables confusing and have difficulty in extracting data from them. Advice on what tables to use and how to extract

Department of Public Health Sciences,
St George's Hospital
Medical School, London
SW17 0RE

F Azeem Majeed, *lecturer in public health medicine*
Derek G Cook, *senior lecturer in epidemiology*
Jan Poloniecki, *lecturer in medical statistics*

Department of Geography,
University of Southampton,
Southampton SO17 1BJ
David Martin, *lecturer*

Correspondence to:
Dr Majeed.

BMJ 1995;310:1511-4

Box 2—Examples of data available from commonly used small area statistics tables

Tables derived from 100% of census records

Age and marital status of residents (table 2)	Age breakdown of population
People living in communal establishments (table 3)	Number of people living in nursing and residential homes
Ethnic group (by age) of residents (table 6)	Ethnic breakdown of population
Economic position of residents (table 8)	Unemployment rates
Residents (by age) in households with limiting long term illness (table 12)	Number of people with chronic illness
Household composition and housing (table 42)	Number of overcrowded households
Household space type: tenure and amenities (table 58)	Number of households without a car
	Number of households that are not owner occupied

Tables derived from 10% sample of census records

Socioeconomic group of households and families (table 86)	Number of people living in households where head of household is unskilled
Social class of household (table 90)	Number of people living in households where the head of household is in social class V

TABLE I—Examples of sampling error for data derived from the 10% sample of census records (using data from areas in Wandsworth Health Authority with high or low percentages of unskilled households)

Area	No of people living in households	No of people with census records coded	No of people living in households where head of household is unskilled	Percentage of people living in households where head of household is unskilled (95% confidence interval)
Enumeration district	646	74	20	27.0 (17.4 to 38.6)
Enumeration district	394	32	0	0.0 (0.0 to 10.9)
Electoral ward	11 585	1 125	71	6.3 (5.0 to 7.9)
Electoral ward	7 476	699	6	0.9 (0.3 to 1.9)
Wandsworth Health Authority	172 656	17 342	535	3.1 (2.8 to 3.4)

TABLE II—Approximate 95% confidence intervals for the difference between a sum from a table containing modified cells and the corresponding actual value

No of modified cells	Local base statistics	Small area statistics
5	±3	±2
10	±4	±3
20	±6	±4
50	±10	±7
100	±15	±10
200	±20	±15
500	±35	±25

Source: 1991 census user guide 48.⁸

data from them (such as unemployment rates) is published by OPCS in *Key statistics. Definitions and cell numbers*.⁷

Data available for a 10% sample of records

Most tables of census data are derived from an analysis of all census records. However, some of the responses to the questions in the census were only coded for 10% of the population and the corresponding census tables (some of which are shown in box 2) are derived from an analysis of this 10% sample of census records. These tables are based on questions that are classified as "hard to code," such as occupation. Because they are based on a 10% sample of records, the data in these tables are estimates and will be subject to sampling error.

Table I illustrates the effect of sampling error on the estimate of the percentage of people in different geographical areas living in a household where the head of household is in socioeconomic group 11 (unskilled manual), using data from the Wandsworth Health Authority. As the table shows, for small areas such as enumeration districts the 95% confidence intervals can be very wide for percentages derived from tables based on the 10% sample of census records. For larger areas such as electoral wards and health authorities, the 95% confidence intervals are much narrower.

Problems with using small area census data

DATA MODIFICATION

To maintain confidentiality and prevent inadvertent disclosure of census data about individuals, OPCS adds +1, -1, or 0 randomly to the numbers contained in each cell of the small area statistics tables for enumeration districts (except for cells that contain true zeros).⁸ For the local base statistics tables for wards,

this procedure is carried out twice, so that each number in each cell in these tables is altered by adding a number between -2 and +2.⁸ Data in tables based on the 10% sample of records are not modified in this way as the sampling of only one in 10 records is thought to be sufficient to maintain confidentiality.

These modifications introduce errors into the row and column totals in census tables and have three main effects. Firstly, the same information may be contained in more than one census table and because the process of modifying data is carried out separately for each table, counts of variables that are contained in more than one table may differ from each other. For example, the number of people from ethnic minorities may not be the same in different tables. This will lead to differences in rates calculated for ethnic minorities (and for other groups of the population such as people living in non-owner occupied households) depending on which table the estimate of the ethnic minority population is taken from. However, these differences are likely to be significant only when areas with small populations are being investigated.

Secondly, counts derived by aggregating data for small areas to larger geographical areas will differ from the counts in the actual tables for these larger areas. For example, it is possible to obtain the number of people living in households without a car in a health authority by three methods, all of which are likely to give slightly different results: by summing data for enumeration districts within the health authority, summing data for electoral wards within the health authority, and obtaining the number of people living in households without a car directly from a table of census data for the health authority (the preferred option).

Thirdly, the modified counts in tables will differ from the "true" counts for an area, but it is possible to calculate confidence intervals for the true count—these depend on the number of modified cells in the table (table II). Data modification can introduce errors into percentages calculated for enumeration districts with small populations; census data for small areas should therefore be interpreted with caution. If, however, data for enumeration districts are aggregated, the effect of data modification on percentages is substantially reduced, and much more reliable estimates obtained. Box 3 illustrates the effect of data modification on one of the small area statistics tables (table 12—long term illness in households: residents in households with limiting long term illness) for an enumeration district in Wandsworth Health Authority.

DATA SUPPRESSION FOR AREAS WITH SMALL POPULATIONS

Small area statistics are released only for enumeration districts with 50 or more residents and 16 or more households. For enumeration districts below either of these thresholds, the only data that are available are the total number of people present, the total number of residents, and the total number of resident households. All other small area statistics for these enumeration districts are combined with those for a neighbouring enumeration district, provided that the combined total number of people and households is greater than the minimum thresholds. Local base statistics are released only for those wards with 1000 or more residents and 320 or more households. As with enumeration districts, data from wards which fall below either threshold are combined with data from a neighbouring ward. Enumeration districts or electoral wards that contain data from neighbouring areas are categorised as importing zones.

SPECIAL ENUMERATION DISTRICTS

Enumeration districts that were expected to contain 100 or more people living in one or more communal

establishments (defined as an establishment in which some kind of communal catering is provided, such as nursing homes, hospitals, and prisons) on the night of the census were defined as special enumeration districts. The residents of such establishments often have different social and demographic characteristics from people living in the surrounding area. Therefore, to prevent the residents of these establishments distorting the small area statistics for the enumeration district in which they are located, these areas were treated as special enumeration districts. The total population and total number of households for special enumeration districts are published, and if the enumeration district has 50 or more residents and 16 or more households, the other small area statistics are also published. If the number of residents or households falls below either of these thresholds, then these other small area statistics for the enumeration district are suppressed. The small area statistics for special enumeration districts are not combined with those for a neighbouring enumeration district but are included in electoral ward totals.

SHIPPING ENUMERATION DISTRICTS

Every electoral ward contains one shipping enumeration district (a separate category of special enumeration district). Apart from houseboats on inland waterways (which are classified as households), vessels are treated in a similar way to communal establishments, with data on people enumerated while living on vessels published in a separate shipping enumeration district. For most areas, there are no residents of shipping enumeration districts, and no data will be present for shipping enumeration districts.

DATA AGGREGATION

Because of data modification and data suppression, statistics derived by aggregating data for enumeration districts will differ from counts obtained directly from the small area statistics or local base statistics for the larger area. For example, if enumeration district counts of people with chronic illness are summed to calculate the total number of people with chronic illness living in a health authority, this total will differ

from the total obtained directly from the local base statistics or small area statistics for the health authority. The difference arises because many people with limiting long term illness may live in special enumeration districts for which census data are suppressed, and also because of OPCS's policy of adding +1, -1, or 0 counts in enumeration district tables. To minimise the effect of this suppression and modification of data, census data from tables for the largest geographical area of interest should be used. For example, if census data for a health authority are required, then these data should be obtained from a table of census data for the health authority and not by aggregating data for enumeration districts or electoral wards within the health authority.

UNDERENUMERATION

There was a greater problem with underenumeration in the 1991 census than in previous censuses: no census data were obtained for 2.2% of the population.⁹ Although this is a relatively small percentage of the population, underenumeration was not random and was highest in inner city areas and among men aged 20-29 years; about 9% of men in this age group nationally, and nearly 20% in inner London, were not enumerated.¹⁰ Underenumeration will therefore lead to errors in census data, especially for inner London, and to a lesser extent for other inner city areas.¹¹

MIGRATION

The census is carried out every 10 years and because of inward and outward migration, 1991 census data for small areas will become progressively more inaccurate with time. OPCS does attempt to estimate the effects of internal migration within England and Wales by using data from family health services authority age-sex registers, and of international migration by using the international passenger survey. However, these estimates are unreliable at small area level.¹²

Using postcodes to assign census data to individuals

One of the most common ways in which census data are used is to estimate the social and ethnic characteristics of people on the basis of their postcode. This is usually done by assigning a postcode to an enumeration district and using the census data for this enumeration district to characterise people living at the postcode. This process requires a method to assign postcodes to enumeration districts. After the 1981 census, the most common method was to use the grid references of postcodes and enumeration districts. However, Gatrell¹³ and Reading and Openshaw¹⁴ suggested that this method is often inaccurate, with about 40% of people being assigned to an incorrect enumeration district. This applied in England and Wales; in Scotland, postcodes are mapped directly to enumeration districts, with no boundary problems.¹⁵

To attempt to overcome this problem and improve the geographical referencing of data from the 1991 census, the postcode of each household and community establishment was collected on the census form and put on the computer record. This allowed OPCS to produce a new look up table (the postcode-enumeration district directory) by directly linking the postcode on the census form to the enumeration district in which that postcode fell.^{16 17} However, some postcodes lie in more than one enumeration district, which can make it difficult to assign a postcode to an enumeration district. To help overcome this problem, the look up table contains some additional information: the "pseudo-enumeration" district of the postcode (the enumeration district in which most people at that postcode live); and the number of households in each postcode-enumeration district intersection (known as

Box 3—Effect of data modification on small area statistics

SAS table 12—Long term illness in households: residents in households with limiting long term illness. Data taken from an enumeration district in Wandsworth Health Authority

Age	Total persons	Males	Females
All ages	35	17	18
0-15	2	1	1
16-29	3	2	1
30-44	3	2	1
45-59	6	3	3
60-64	4	2	2
65-74	8	4	4
≥ 75	9	3	6

Cells in SAS tables are modified by randomly adding +1, -1, or 0. The number of people aged 65-74 with long term illness is based on two cells, each of which has been modified by adding a number between +1 and -1. The number of people aged 65-74 with long term illness (8) therefore differs from the true total by ± 2 . The total number of people with long term illness in this enumeration district (35) is the sum of 14 cells and differs from the true total by ± 14 ; the true number of people with long term illness living in this enumeration district therefore lies between 21 and 49. However, extreme results are unlikely, and there is a 95% probability that the true total lies between 29 and 41.

Key messages

- The 1991 census provides data on demography, social deprivation, employment, housing, chronic illness, and ethnicity for enumeration districts, electoral wards, and health and local authorities
- Census data have an important role in helping to plan health services and in monitoring how health services are being used, and they are freely available to both NHS employees and employees of academic institutions
- Census data for small areas contain inaccuracies, but the effect of these errors is substantially reduced by either aggregating data or by using data for larger areas
- Data from the 1991 census were used to produce a new postcode to enumeration district look up table, which overcomes many of the problems encountered with older methods of assigning postcodes to enumeration districts

part postcode units). The table also contains the Ordnance Survey grid reference of the postcode, but this is only given to a precision of 100 metres. However, the mapping of postcodes should improve in the future as a result of the Ordnance Survey's Address-Point initiative, which aims to allocate a grid reference to every address in England and Wales to an accuracy of one metre.¹⁸ The main limitation of the look up table is that it contains only the postcodes that existed at the time of the 1991 census and will become progressively more out of date with time, as new postcodes are created (for example, because of the construction of new residential estates).

ASSIGNING ENUMERATION DISTRICTS TO INDIVIDUALS

Assigning an enumeration district to an individual living at a postcode which relates to only one enumeration district is straightforward. However, if the postcode relates to more than one enumeration district then assigning an enumeration district can be done in two

ways: by using the number of households in a part postcode unit to determine the probability of individuals living in each part postcode unit (this is the preferred option); or by assigning all individuals living at a postcode to its pseudo-enumeration district (box 4).

Although using the number of households in a part postcode unit as the weighting factor is the preferred method of assigning social and ethnic characteristics to individuals on the basis of their postcode, it still has some limitations because the number of households in a part postcode unit is only an approximate measure of the number of people living in that unit. This is because the number of people per household is not constant within or between enumeration districts. Moreover, in some enumeration districts, some people live in communal establishments such as nursing or residential homes; the number of households in a part postcode unit is increased by between one to four to take into account these individuals. Another limitation of the look up table is that it gives the probability that an individual falls into a certain category (for example, lives in a household without a car). Whether the individual has been correctly assigned to a category depends on how representative the individual is of the enumeration district in which he or she lives. However, when estimated data for individuals are aggregated to larger populations, the effect of this limitation on the estimated values is reduced.¹⁹

Conclusions

The 1991 census offers valuable data to academics and to people working in the health service. However, users of census data should be aware of its limitations, chiefly underenumeration, data modification, and data suppression, the effects of which are reduced substantially when census data are aggregated. The new postcode-enumeration district directly produced by OPCS overcomes many of the problems previously encountered in trying to assign postcodes to enumeration districts and will improve the geographical referencing of census data.

We thank Ivor Evans and Andrew Frost of Wandsworth Health Authority's information department for their help.

Box 4—Obtaining weighted average of census data for individuals by using their postcode

Postcode	Enumeration district	Proportion of people living in households without a car			In enumeration district	
		Pseudo-enumeration district	No of households	Weight	enumeration district	In postcode (estimated)
SW8 4JD	ANFM12	ANFM12	5	1.00	0.76	0.76
SW8 4HX	ANFM12	ANFM12	17	0.85	0.76	0.75
SW8 4HX	ANFM16	ANFM12	3	0.15	0.68	0.75

Consider two postcodes, SW8 4JD and SW8 4HX, for which we wish to know the likelihood that people living at these postcodes are living in households without a car. The intersection between postcode SW8 4JD and the enumeration district ANFM12 is known as a part postcode unit. Because SW8 4JD lies entirely within enumeration district ANFM12, it produces only one part postcode unit, and the people living at this postcode have a 0.76 probability of living in a household without a car.

Postcode SW8 4HX lies in two enumeration districts, ANFM12 and ANFM16, and hence produces two part postcode units. There are 20 households in this postcode, of which 17 (85%) lie in enumeration district ANFM12 and three (15%) in enumeration district ANFM16. The probability of living in a household without a car at this postcode is therefore estimated as 0.75 $(= (0.76 \times 0.85) + (0.68 \times 0.15))$. An alternative approach is to use data from the pseudo-enumeration district for the postcode.

- 1 Majeed FA, Chaturvedi N, Reading R, Ben-Shlomo Y. Monitoring and promoting equity in primary and secondary care. *BMJ* 1994;308:1426-9.
- 2 Majeed FA, Cook DG, Anderson HR, Hilton S, Bunn S, Stones C. Using patient and general practice characteristics to explain variations in cervical smear uptake rates. *BMJ* 1994;308:1272-6.
- 3 Office of Population Censuses and Surveys. 1991 census definitions. Great Britain. London: HMSO, 1992.
- 4 Dale A, Marsh C, eds. The 1991 census user's guide. London: HMSO, 1993.
- 5 Office of Population Censuses and Surveys. 1991 census user guide 3. Local statistics and small area statistics. London: OPCS, 1991.
- 6 SASPAC user manual. London: London Research Centre, 1992.
- 7 1991 census user guide 47. Key statistics. Definitions and cell numbers. London: OPCS, 1993.
- 8 1991 census user guide 48. Local base statistics/small area statistics. Modification of counts for confidentiality. London: OPCS, 1993.
- 9 Balarajan R. Public Health and the 1991 census. *BMJ* 1994;309:287-8.
- 10 Simpson SN. Coverage of the Great Britain census of population and housing. *Journal of the Royal Statistical Society* 1994;157:313-6.
- 11 Dorling D, Simpson S. Those missing millions: implications for social statistics of undercount in the 1991 census. *Radical Statistics* 1993;55:14-35.
- 12 Marsh C. Methods used in rolling forward estimates of population. In: Dale A, Marsh C, eds. The 1991 census user's guide. London: HMSO, 1993:161-2.
- 13 Gatrell AC. On the spatial representation and accuracy of address-based data in the UK. *International Journal of Geographical Information Systems* 1989;3:335-48.
- 14 Reading R, Openshaw S. Do inaccuracies in small area deprivation analyses matter? *J Epidemiol Community Health* 1993;47:238-41.
- 15 Elliot P, McGale P, Vincent TJ. Description of population data and definition of areas. In: Draper G, ed. The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966-1983. London: HMSO, 1991:17-23. (Studies on medical population subjects No 53.)
- 16 1991 census user guide 26. ED/postcode directory. London: OPCS, 1993.
- 17 Martin D. Postcodes and the 1991 census of population: issues, problems and prospects. *Transactions of the Institute of British Geographers* 1992;17:350-7.
- 18 Barr R. Census geography. II. A review. In: Dale A, Marsh C, eds. The 1991 census user's guide. London: HMSO, 1993:70-83.
- 19 Majeed FA, Cook DG, Poloniecki J, Griffiths J, Stones C. Sociodemographic variables for general practices: using census data. *BMJ* 1995;310:1373-4.

(Accepted 13 April 1995)