Problems of Harmonising UK-Wide Samples of Anonymised Records from the 1991 Census

by Elizabeth Middleton [1]
Census Microdata Unit
Fáculty of Economics and Social Studies
University of Manchester Manchester M13 9PL

Introduction
For the first time in a British Census the output from the 1991 Census includes the Samples of Anonymised Records, known as the SARs. The SARs are individual level data and can be analysed in the same way as survey data using familiar statistical software packages.

The datasets have been purchased jointly by the Economic and Social Research Council and the Information Systems Committee of the Universities Funding Council. The latter has now been superseded by the Joint Information Systems Committee. The SARs will shortly be housed at the Census Microdata Unit at Manchester University who are responsible for dissemination of the data to the academic and non-academic sectors.

The Availability of Census Microdata in Other Countries
In recent years there has been a growing demand in most industrialized countries for individual level data and the release of samples of census microdata in Great Britain follows the examples set by the United States, Canada and Australia.

The United States first released their Public Use Microdata Samples in the 1960s and these hierarchical files of housing units and persons have been heavily used ever since. From the 1990 census the samples available consist of a 5% county/county equivalent file, a 1% metropolitan area file and a 3% sample of the elderly. Canada has released census microdata samples since the 1970s and this time are producing an individual file, a household and housing file and a family file. The 1980s saw the release for the first time of Australian census microdata samples. These did not contain much geographical detail but negotiations currently in progress are likely to lead to the design of two samples from the 1991 census, one with more geographical information and less classification detail than the other.

In Great Britain, two files will soon be available, a 2% individual sample and a 1% sample of households. The hierarchical file of households has geographical identifiers only for the standard regions of Great Britain while the geographical identifiers on the individual file are for areas with a minimum population of 120,000 persons (Marsh 1992). Northern Ireland SARs, similar to those of Great Britain, should be available early in 1994.

Of course the release of microdata does involve some risks of disclosure of personal information and so in all the above examples the data has been 'anonymised' to reduce the risk of identifying any individual. Some variables may be suppressed, categories of other variables may be grouped together and perturbation techniques may be used.

In Britain individual level data is also available in the OPCS Longitudinal study where microdata for the 1971, 1981 and 1991 censuses is linked together with data drawn from the National Health Service Central Register. However, access to this microdata is tightly controlled and OPCS release tables from the Longitudinal Study only after checking the tables thoroughly for any breaches of confidentiality.

In general, other European countries have placed more restrictions on the availability of census microdata but in many countries the policy of releasing microdata is under review. Spain and Luxembourg, for example, have allowed use of microdata only by government and not by academics. Microdata from the French census is available to academics under special arrangements which may include working on the premises of INSEE, the National Statistical Institute. In Italy, academics can access microdata by special request. In fact, microdata from the 1981 Italian census was used by researchers at the University of Newcastle upon Tyne, in collaboration with the Regional Government of Tuscany and the Italian Census Agency, for work on disclosure risks. This research proved useful when British academics made the case for the release of the SARs in Britain (Marsh et al 1991).

In Sweden a Freedom of Press Act allows all governmental information to be open to the public, subject to restrictions on access to personal information, and academics can request permission to access census microdata. Microdata in Norway is not publicly available but academics can obtain access to anonymised census microdata.

Neither Denmark nor the Netherlands hold a conventional census (Langevin 1992). Instead, similar information is obtained from administrative files and registers. In Germany, recommendations have recently been made for release of data from the German Microcensus, a survey conducted annually with 1% of the population. These recommendations involve both suitable anonymisation of records and contractual commitments of recipients of the data (Knoche 1992).

Censuses in the United Kingdom
In the British Census, census forms for England, Scotland and Wales differ in some small respects: in Wales there is a Welsh language question, in Scotland there is a Gaelic language question and also a question about the lowest floor level of accommodation. The census forms are all processed in the same way by the Office of Population Censuses and Surveys and the General Register Office for Scotland.

However, in Northern Ireland there is a different constitutional framework and the Northern Ireland census is the responsibility of the Northern Ireland Civil Service. The Northern Ireland Census Office is part of the Department of Health and Social Security. Access to the data has to be negotiated separately with the Northern Ireland authorities and partly for this reason, Northern Ireland data has often been neglected in the past.

This year the ESRC is currently funding more work on the Northern Ireland Census. At Queen's University, Belfast, work is underway to produce Local Base Statistics and Small Area Statistics and the ESRC is also negotiating with the Northern Ireland Census Office for the purchase of Northern Ireland SARs.

Northern Ireland SARs
At the time of writing, the statistical specification for the Northern Ireland SARs is almost finalised and is as similar as possible to that of the British SARS. As for Great Britain, two files are to be produced.

(i) The first is a 2% sample of individuals at private addresses and residents of communal establishments. The geographical scheme identifies ten areas, chosen so that the population in each area is greater than 120,000. The SAR areas are amalgamations of District Councils with similar characteristics.

(ii) The second file is a 1% hierarchical sample of households and the individuals in those households. Here the geographical identifier is just that of Northern Ireland itself.

Sampling methods are similar to those used for the British SARs but, while the British samples are drawn from the 10% of census records which are fully coded, in Northern Ireland the samples are drawn from the 100% fully coded data.

Confidentiality issues are particularly important in Northern Ireland because of the risks of disclosure of information about the security forces. However, since there is no fine detail geographical information in the SARs, no additional measures to reduce the risk of disclosure were considered necessary. The same broad-banding of categories is used in both the Great Britain and Northern Ireland SARs. For the individual sample, as in the Great Britain SARs, officers in the armed forces are grouped with Police Officers, Prison Service Officers, Fire Service Officersand Customs and Excise Officers. Other ranks are grouped together in a separate category.

Harmonisation of United Kingdom SARs
Census reports for the whole of the United Kingdom have not been available in the past and so a harmonised United Kingdom SAR dataset would be extremely useful. SAR data files containing data from both Great Britain and Northern Ireland would allow such reports to be produced easily.

However there are some problems in producing such a dataset and they fall into the following categories:

Differences in Questions on Census Form
The starting point for comparison of variables in the Northern Ireland and British SARs is of course the census form. Questions on both forms are, in general, worded in the same way but there are a few exceptions.

The wording of the Qualifications question is quite different. In Great Britain only qualifications normally obtained after the age of 18 are asked for but in Northern Ireland qualifications obtained at school level are also recorded. The subject of the highest level qualification is recorded on the British form but not on the Northern Ireland form. Other small differences in the wording of questions do not present serious harmonisation problems.

In Great Britain the ethnicity question was introduced for the first time for the 1991 Census and is not included in the Northern Ireland Census. Northern Ireland have some additional questions: a fertility question (to be answered by all married, widowed, separated and divorced women), an Irish language question, a question on religion and additional amenities questions on water supply and domestic sewage disposal.

The question on religion was voluntary and was introduced partly to assist monitoring of Northern Ireland's fair employment legislation under which employers of 25 or more people have to register to say how their workforce is balanced between Catholics and Protestants.

Differences in Coding
Even when the wording of questions on the census form is the same, the way in which information is coded can vary. Information can be lost when being transferred from form to computer at the coding stage.

An example of this arises in the coding of the relationship to head of household question. While Great Britain codes this using 16 categories, Northern Ireland codes the answers to the relationship question using just 10 categories. As shown in Table 1, such relationships as 'Cohabitant of son or daughter' or 'Boarder/lodger' would be coded separately in Great Britain but would be classed as 'Other unrelated' in Northern Ireland.

Table 1. Differences in Coding

Relationship to Head of Household	
G.B.	N.I.
0 Head of Household	0 Head of Household
1 Spouse	1 Spouse
2 Cohabitant	2 Cohabitant
3 Son/daughter	3 Son/daughter
4 Child of cohabitant	
5 Son/daughte-in-law	8 Son/daughte-in-law
6 Cohabitant of son/daughter	
7 Parent	4 Parent
8 Parent-in-law	7 Parent©in©law
9 Brother/sister	5 Brother/sister
10 Brother/sister-in-law	
11 Grandchild	6 Grandchild
12 Nephew/niece	
13 Other related	9 Other related
14 Boarder, lodger etc.	
15 Joint head	
16 Other unrelated	10 Other unrelated

Differences in Definitions
The Relationship to head of household is used to identify families within a household and differences exist between the census definitions of the family used in Northern Ireland and Great Britain.

In both censuses, a family unit has a maximum of two generations with the younger generation never married and having no partner or children. There is no age limit for a child. However, in Northern Ireland a cohabiting couple is not regarded as a family unit as it is in Great Britain. A cohabiting couple would therefore be classified in the Northern Ireland Census as 'family type' not applicable or, if the couple had children as a 'lone parent with children, with others' family type. The Great Britain definition does in fact conform to United Nations recommendations for 1990 censuses in the European Community (UN 1987).

Again, the definition of a dependent child differs in the two censuses. In the British Census a person age 16-18, never married, in full time education and economically inactive is regarded as a dependent child. In the Northern Ireland Census, the equivalent age banding is 16-19. However, in the interests of harmonisation, the Northern Ireland Census Office has agreed to change the age banding to 16-18 for the Northern Ireland SARs so that this difference in definitions will not be a problem.

Differences in Processing

Clerical/Computerised Allocation of Families
In Great Britain allocation of individuals to families is done using a complex computer algorithm whereas in Northern Ireland the number of families in a household is determined manually when coding from the census form.

The computer algorithm identifies sixty different family types and every individual in the household isallocated to a family with its particularfamily number and family type. The sixty family type categories are reduced to tencategories for the SARs, as shown in Table 2.

Table 2. Great Britain SARS Family Type

1. Married Couple	- no children
2	- dependent child(ren)
3	- non©dependent child(ren)
4. Cohabiting Couple	- no children
5.	- dependent child(ren)
6.	- non-dependent child(ren)
7. Lone parent	- dependent child(ren)
8	- non-dependent child(ren)
9. Non-family household	- one person
10	- two or more persons

It is worth noting that the computer algorithm cannot identify families amongst a group of household members unrelated to the head of household and in such cases the census forms must be processed manually.

In Northern Ireland, the number of families in a household is determined at the coding stage by examination of the census form. The household is classified as a non-family household or as one of the household types shown in Table 3. The Northern Ireland census database does not hold information about which family an individual belongs to.

Table 3. Northern Ireland SARS Household Family Type

1. One-family household	- married couple, no children, no others
2.	- married couple , no couples with others
3.	- married couple, with children, no others
4.	- married couple, with children, with others
5.	- lone parent, with children, no others
6.	- lone parent, with children, with others
7.	- lone grandparent, with grandchildren, no others
8.	- lone grandparent, with grandchildren, with others
9. Two-family household	- related, no others
	- related, with others
	- not related, no others
	- not related, with others
13. Three of more family household	

This difference in processing means that the only way of harmonising family type is to use the Northern Ireland classification of Household Family Type and to derive this classification for Great Britain households in the household SARfile. In the Northern Ireland individual sample, variables such as Social Class of Family Head and Economic Position of Family Head will not be available. Instead similar information will be available for the Head of Household.

Calculations of Distance
The Great Britain SARs hold information about the distance travelled by people to work and also on the distance of move of migrants.

Calculations of distance to work are carried out in different ways for the two censuses. The Central Postcode Directory used in Great Britain is not currently used by the Northern Ireland Census Office but distances are calculated using the Northern Ireland grid square reference marked on the respondent's census form and the grid square reference of the employer. The employer's grid reference is only known for large employers employing over 25 persons and so the Northern Ireland SARs will hold distances to work for people working for large employers only. The distance of move of migrants is found using the postcode of the previous address of the migrant and so cannot be found without using the Central Postcode Directory. This means that there are currently some unresolved problems in including this in the Northern Ireland SARs.

Clerical/Computerised Editing and Imputation Procedures
In Great Britain a computerised editing system is used to identify inconsistencies and missing values in the data and then to impute valid, consistent answers (Mills 1991). In Northern Ireland, only clerical processing is used. Some inconsistencies are resolved manually at the coding stage and clerical procedures give rules for handling missing answers.

The differences between these two approaches can be seen by comparing the procedures for treating missing values for the number of cars in a household. In Great Britain analysis of previous census results shows that a good indication of the number of cars in a household is given by the number of people in the household, the tenure of the household, and whether the accommodation is in a permanent or non-permanent building.

Missing values are imputed by reference to these other factors. However, in Northern Ireland the missing value would be coded as zero.

This means that when making comparisons between Northern Ireland and Great Britain, it is important to understand the effects of the different imputation procedures.

Conclusion

Although there are some harmonisation problems, it is felt that the production of a harmonised United Kingdom SAR dataset would be of great benefit to census data users. Since there is little good quality data available for the whole of the United Kingdom, the UK SARs would provide a new research resource to enable work to be carried out in such areas as deprivation and discrimination.

Even within the United Kingdom, small differences in census definitions and processing lead to difficulties in making direct comparisons of some social characteristics of the regions. Hopefully, the increasing adoption of international standards will reduce the amount of effort required to be spent in the future on harmonisation problems and result in better quality comparative statistics.

References

KNOCHE, P (1992). Factual Anonymity of Microdata from Household and Person Related Surveys. The Release of Microdata Files for Scientific Purposes. *Proceedings of International Seminar on Statistical Confidentiality, Dublin, organised by EUROSTAT and ISI.*

LANGEVIN, B; BEGEOT, F; PEARCE, D (1992). Censuses in the European Community. *Population Trends*, 68.

MARSH, C; SKINNER, C; ARBER, S; PENHALE, B; OPENSHAW, S; HOBCRAFT, J; LIEVESLEY, D & WALFORD, N (1991). The case for samples of anonymised records from the 1991 census. *Journal of the Royal Statistical Society(A)*, 154, 2.

MARSH, C; TEAGUE, A (1992). Samples of anonymised records from the 1991 Census. *Population Trends*, 69.

MILLS, I; TEAGUE, A (1991). Editing and imputing data for the 1991 Census. *Population Trends*, 64.

UN (1987). Recommendations for the 1990 Censuses of Population and Housing in the ECE Region. *United Nations Statistical Commission for European Statisticians, Statistical Standards and Studies*, 40.

1. Presented at IASSIST/IFDO 93 Conference held in Edinburgh, Scotland. May 1993.