

Information paper

Census strategic development review

Alternatives to a Census:

Linkage of existing data sources

Office for National Statistics, November 2003

© Crown Copyright

Contents

	Page
1 Introduction	3
2 Benefits and challenges in using data from other sources	5
2.1 Benefits	5
2.1.1 Reduction of respondent burden and compliance costs	5
2.1.2 Reduction of collection costs	5
2.1.3 Improvements to data detail	5
2.1.4 Improvements to data in coverage	6
2.1.5 Improvement in the frequency and timeliness of outputs	6
2.2 Challenges	6
2.2.1 Public opinion and legislation	6
2.2.2 Quality of data sources	7
2.2.3 Content	7
2.2.4 Connectivity issues	7
2.2.5 Timing issues	8
2.2.6 Changes in datasets	8
3 The ability of administrative sources to meet requirements	8
3.1 A benchmark count of people and housing	8
3.2 Population structures: information on households and families	9
3.3 Population characteristics and multivariate analysis	10
3.4 Further uses of alternative sources	11
4 Conclusions and recommendations	12

Census strategic development review

Alternatives to a Census: Linkage of existing data sources

1 Introduction

Over the last thirty years a number of factors have encouraged support for the increased use of existing alternative sources in counting the population of the UK and understanding its characteristics:

- confidence in the ability of a traditional census to obtain all information periodically needed from the whole population has declined;
- the pace of change in society has quickened, requiring government to measure characteristics and structures of the population more frequently to support policy and decision making;
- the amount of electronically held data about events, persons and services by government agencies and other bodies has increased, particularly those captured in the administrative and monitoring processes associated with the running of public services; and
- computing capability has expanded allowing ready linkage and analysis of datasets on a larger scale than could have been envisaged previously.

The share of the population in groups which respond comparatively poorly to the Census has increased over the last two decades. Strategies that counter differential response rates have not been effective in countering this trend. This has especially affected our ability to enumerate the population of places such as the inner cities and some population sub-groups, such as young males.

In a rapidly changing society, for some users, the 10-year gap between censuses is too great. The population is increasingly mobile and population structures are increasingly less stable. If policy and decision making are to be founded on timely, relevant information then government needs to measure the state of the population more frequently, to keep up with the pace of change, and understand the more complicated social and family structures.

In the private and public sectors the delivery of services, and the targeting of those services to

specific population groups, is underpinned by databases which hold detailed information on individuals. Stronger information management has accelerated growth in this area. Person-level data held by government covers a range of subjects and an increasing proportion of the population.

The question of whether these data can be linked to support a statistical analysis of the population is not new. Following attempts by Lord Moser when head of the CSO to establish a population register for the UK, Philip Redfern, then Deputy Director of the Office of Population Censuses and Surveys, judged that “this approach, when fully developed, provides an efficient tool for the statistical study of the inter-relationship between the many social and economic factors affecting society and the study of changes through time.”¹ In 1985 David Rhind proposed “integration of data from the many existing data sources to give an ongoing, ‘cradle to grave’ inventory of people and land.”²

The potential benefits in the approach are great both in relieving public burden and in creating a statistical basis for better management of the society in which we live. However, for such data sharing to take place, there are legal and public perception issues that would need to be addressed, and systems would need to be underpinned by practices which safeguard the data and ensure its appropriate use.

The technical challenges of linking data sources are also significant depending upon the approach to linking. There are a number of approaches that might be adopted:

- linkage of aggregated data, via geography or population profile;
- linkage of non-identified individual data—statistical matching of people with the same characteristics, but not necessarily linking data for the same person; and
- linkage of specific individuals using identifiable data.

The choice between these options depends upon the extent and nature of legal protection, what public attitudes will allow and the resources that may be allocated.

Aggregated data linkage will stimulate little controversy in terms of public opinion but suffers from not offering the possibility of multivariate analysis. The identification of new trends in society may depend upon such analysis. Concern is often expressed also about aggregation bias in social statistics. "Aggregate data are often easier to obtain than data on individuals, and may offer valuable clues about individual behavior...problems of confounding and aggregation bias, however, are unlikely to be resolved in the proximate future."³ Aggregation often loses information about gross flows, so that measures of net change can obscure the magnitude of change taking place, or the concentration of change in a few participants.

Linkage of non-identified individual data may have a more positive public perception because of the anonymity of the data. The approach has shortcomings however. The process of anonymising data while retaining its usefulness can become complex. Anonymised data may be re-identified if they are linked to other data which is identified.

There are also operational reasons why identification of data may be important. In February 2002 a report commissioned by the Department of Health Policy Research Programme in support of the White Paper entitled 'Saving Lives: Our Healthier Nation' presented the following arguments:⁴

- **"Avoiding double-counting-***To fulfil their function, registers need to be complete. The only way to be sure of finding all cases is routinely to use several independent sources...as there will be overlap between these sources, duplicates must be eliminated to avoid counting the same people twice or more. Duplicates can be recognised only if patients can be identified.*"
- **"Validation-***If authorities are to act on results derived from analysis of registers, they must have assurance that the register is correct. The most effective way of validating completeness is to compare from time to time the information on the register with another independent source of cases. To do this, cases in both sources must be identifiable in order to know whether the sources agree. Although both sources might have 50 cases, are they the same 50 cases?"*

The two arguments are closely related to coverage and are applicable to most situations where data linking is envisaged.

With anonymised data there are limitations to the validity of the linked data when statistical matching is used. The validity of a statistical match is dependent upon knowing that the relationship of one variable to another can be determined from a common set of variables shared by a donor and a recipient. Demonstrating the degree of association between the synthetic matching variables and the other variables is likely to be an exhaustive process for just one variable and will inevitably be subject to challenge, for some analyses.

Exact matching on a unique key, or on a set of variables which when combined are unique, is probably simpler than statistical matching and less subject to dispute. Although the presence of a common unique identifier across all alternative data sources is unlikely, an analysis of previous census data and the experience of One Number Census and Longitudinal Study matching show that the task of matching on a combination of variables is manageable.

Resolution of inconsistencies in definition and content is more confident when data is linked at person level. The usefulness of such data in estimating coverage will also be enhanced. The body of this report examines the use of alternative sources linked at an individual level and their potential to replace or enhance the Census in providing the required information; specifically information relating to:

- population units: a benchmark count of people and housing, with key characteristics such as age and sex;
- population structures: information on households and families; and
- population characteristics: consistent and comparable data, with the range of topics allowing multivariate analysis, giving rich information down to small areas and population sub-groups.

Section 2 of the report sets out the benefits and challenges associated with the use of alternative sources in these contexts.

2 Benefits and challenges in using data from other sources

2.1 Benefits

The benefits of using data from alternative existing sources generally fall into six categories:

- reduction of respondent burden and compliance costs;
- reduction of collection costs;
- improvements to data detail and breadth of topics;
- improvements to data coverage;
- improvement in the frequency and timeliness of outputs; and
- increased possibility of measuring the reliability of results.

These are discussed in turn below.

2.1.1 Reduction of respondent burden and compliance costs

There are clear reasons why we would not want to collect information via a traditional census exercise if it is already available from another source and is similar in terms of accuracy and relevance:

- many surveys, including the Census, are at the limits of what it is reasonable to expect people to answer, and the contents may be too demanding for some groups of people; and
- the inclusion of topics covered in other data sources involves duplication of effort by the respondent and the agencies using the data.

2.1.2 Reduction of collection costs

Scandinavian countries moving away from a traditional census generally report significant savings. In Finland the upkeeping costs of the register-based census system for a ten year period have been reported as 40 per cent of the corresponding costs of the former traditional census.

We cannot assume similar savings in England and Wales, however. Some costs of an administrative based census may be hidden and they are certainly not insignificant. The infrastructure in countries that have moved to an administrative based census has qualities which make the transformation less complicated and easier to accomplish.

In April 2002 the government's Performance and Innovation Unit published guidelines on privacy and data sharing. The guidelines split the costs of data sharing into three categories:

- legal costs:
 - legal advice;
 - consultation; and
 - planning a legislative slot.
- sharing costs:
 - standardisation to make sharing effective;
 - measures to counteract any deterioration in quality;
 - measures to mitigate the effects of voluntary compliance levels decreasing as awareness of more data sharing by government increases; and
 - sharing uncompensated costs of data providers.
- safeguard costs
 - use of new technologies to increase privacy, for example encryption;
 - new infrastructure to limit access to data by staff, for example security at entry, file, row and column level; and
 - training of staff to ensure that new safeguards are effective.

Most of the costs outlined above tend to be associated with developing the framework in which data can be shared. Once systems have been established it seems likely that the costs will be less than those associated with a traditional census operation. One important area to investigate, however, will be matching. If a high level of clerical intervention is required then expenditure may increase significantly. The availability of associated systems, such as a national address register, may affect costs such as these.

2.1.3 Improvements to data detail

The business case for including topics in the 2001 Census evaluated:

- the relative importance of the topic statistics;
- the logistical issues of fitting questions on a form;

- the effect on response rates in increasing the complexity of the questionnaire; and
- the effect on response rates in increasing the sensitivity of the questions.

These considerations mean that a census cannot always collect the information or the detail that users would like. A question on income was strongly argued in 2001 but was not included because there was evidence that it might have affected response. The question on health simply asked whether a person considered their health over the previous 12 months to be 'good, fairly good or not good'. Electronic health records, whilst still developing in format and context, contain considerably more detail than was collected in the 2001 Census.

2.1.4 Improvements to data in coverage

Other sources will allow us to improve population coverage by including those people who engage with at least one activity or agency of local and central government but who do not engage with the census. Italy and Canada use administrative sources as part of their coverage adjustment methodologies. The US is planning their use in this context.

2.1.5 Improvement in the frequency and timeliness of outputs

The logistical and financial constraints associated with a traditional census model prohibit running a census on a frequent basis. For some of the information collected, the population that it relates to will have changed considerably since census date. The analysis of up to date information is becoming increasingly important as the rate of change in society increases. The ability to measure change is particularly significant in the context of policy evaluation.

It may be possible to balance the need for regular updates against cost and logistics by using other sources. This is the experience of countries that use the approach. The register based system in Finland has made it possible to produce, annually, most of the traditional census statistics.

Timeliness is a separate consideration. A traditional census has to go through a number of processes before results are disseminated. Population statistics for 2002 were available in Sweden by February 2003. The register-based approach is continuous and does not rely on the 10-yearly big count of a census. This allows faster production of statistics, and more frequent assessment of levels and change of population.

2.2 Challenges

These can be summarised in the following headings and are discussed in turn:

- public opinion and legislation;
- quality of data sources;
- content;
- connectivity;
- timing issues; and
- changes in datasets.

2.2.1 Public opinion and legislation

The Citizens Information Project, which is pursuing the aim of creating a population register including necessary identifying and contact information to support public service delivery, has carried out some research into public perception of the proposal. This indicates a certain amount of support for the notion that data sharing should take place to avoid needless duplication, improve the efficiency of government and prevent fraud. Some respondents expressed surprise that such data sharing did not already take place within government.

A census based upon existing alternative sources would need to win similar public support. A number of interest groups and individuals would certainly raise objections to proposals for increased data sharing. Concerns about increased data sharing are based on legitimate arguments, and legislative safeguards to address such concerns would be essential.

The case for legislation to support the statistical use of these data will need to:

- guarantee the certain and proper protection of all records obtained by the National Statistician for statistical purposes;
- affirm and reinforce the authority of the National Statistician to obtain individual records;
- affirm access to administrative records for statistical purposes;
- define the proper situations where statistical records can be used for research purposes; and
- define the conditions for matching statistical and administrative records for statistical purposes.

2.2.2 Quality of data sources

Eurostat defines the quality of statistics with reference to the following criteria:

- relevance of statistical concepts;
- accuracy of estimates;
- timeliness and punctuality in disseminating results;
- accessibility and clarity of the information;
- comparability of statistics; and
- coherence.

These criteria can only be examined with reference to individual data sources. The general use of alternative sources raises broader issues of compatibility, accuracy, and coverage.

E-government initiatives seek to set standards for compatibility by regularising data concepts and formats. The introduction of BS7666 to create a common format for addresses is an example of this approach. Standardisation of person data such as date of birth is also being undertaken.

The trend towards common data formats is likely to continue but will not entirely remove the need for remedial measures to ensure comparability. Within the BS7666 address structure there is the potential for 'free-text' in the Building and Sub-building fields, enabling different users to describe addresses using different syntaxes and abbreviations. Such variations will need to be ironed out by agencies using the data.

Compatibility of data sources is also a question of definitions and classifications. In discussing data sharing the Performance and Innovation Unit have judged that "the most useful indicators of data quality are likely to be good metadata - descriptors of the data capture, validation and management processes - from which potential new users can assess the appropriateness of the database for their desired purposes." It will be necessary to assess compatibility on a case by case basis.

The coverage of an administrative data source is necessarily restricted to the population of interest to the agency using the data and depends upon how well that agency is able to capture that population and its whereabouts. A recent National Statistics review of international migration statistics carried out by ONS and

the Home Office judged that National Health, Inland Revenue and Department of Works and Pensions data might contribute to these statistics but had distinct deficiencies in coverage.

A Department for Works and Pensions press release of 27th March 2003 revealed that over two million poor households were failing to claim means-tested benefits worth as much as £4.5 billion a year. This exemplifies the limits to the coverage of administrative sources, which in that particular data source was due to people not registering although they are qualified to do so.

Other data sources may have over coverage. Investigations by the Cabinet Office show the extent to which duplication of an identity underpins fraud. In other circumstances over coverage may be legitimate. Students have been allowed to register on the Electoral Roll at their place of study and at 'home'.

As well as absolute under coverage and over coverage there may be localised under coverage and over coverage in a data source. Where the registration process is voluntary and is not associated with the receipt of regular services there is a probability that address information will be out of date. Students at university do not necessarily re-register when they return 'home'. As well as visiting doctors less frequently, young men are slower to re-register on moving.

2.2.3 Content

Countries moving towards alternatively sourced censuses have similar problems when attempting to meet the content requirement. Information in these data sources reflects the fact that they are not collected for statistical purposes and are particularly weak with regard to a person's cultural characteristics such as ethnicity. The Netherlands 2001 'virtual census' uses a register-based approach to meet the Eurostat requirement for a limited set of key tables. Despite this narrow focus, the Central Bureau of Statistics relies heavily on survey information to supplement register data. Norway also uses survey data to supplement register-based data on families.

2.2.4 Connectivity issues

We bring together information from a variety of administrative sources to build a composite picture. Confidence in that picture depends upon our ability to demonstrate that the separate parts of the picture are truly connected. Where 'exact' matching of individual records is possible, confidence may be high. Exact

matching in this context means the matching of two observations by a unique identifier such as a National Insurance number, or a set of identifiers which, in combination, provide a unique reference.

It is certain in the short and medium term that no single unique identifier will exist on all possible data sources. Matching in a census system is likely to be probabilistic, or judgmental.

Judgmental and clerical matching depend upon understanding the data and identifying attributes which when combined will automatically and confidently identify matches. When ordering foreign currency at a post office for example you will normally be asked for your postcode, house number, surname and initials. Street name is also asked as a quality check. For the purposes of this matching exercise the automatic limitation of possible matches to a location, the subsequent limitation of matches to a name and the qualitative judgements of an operator are sufficient to allow the transaction to continue.

Probabilistic matching techniques are not dependent upon an understanding of the data. All possible combinations of observations in two data sources may be examined. Within that examination all attributes of the data are scored according to the probability that their correspondence or close correspondence is significant.

It is impossible within the scope of this report to describe all variations on the various approaches to matching. The National Statistics Methodological Series includes an exhaustive work on the subject.

2.2.5 Timing issues

Even in synchronous datasets, where the data have been collected in the same time frame, there will be differences in common information. Addresses in the NHS Central Register (NHSCR) are unlikely to be as up to date as addresses held by benefits agencies.

When datasets are not synchronised these problems are magnified. As well as making matching more difficult, the process of judging which of the conflicting pieces of information to use is complicated. When faced with the same individual and two different addresses which should we use? The AREX experiment, carried out by the US Census Bureau and discussed

further in **section 3** of this report, based its selection of address on geocodability, currency and quality. We would have to make similar judgements.

2.2.6 Changes in datasets

Our relationship with the agencies that collect administrative data has two major weaknesses:

- we may be limited in our ability to influence the scope or to control the quality of information that external agencies collect or the way in which those data are processed; and
- we probably cannot safeguard the continuity of a specific data source beyond the policy that requires its creation. In a period of deregulation, many longstanding administrative data sources have ceased.

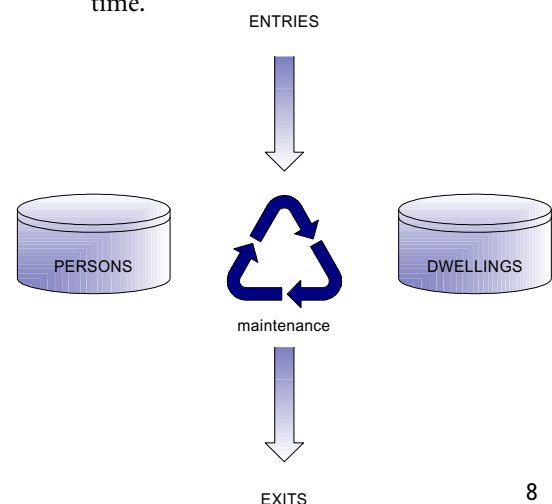
Some countries have dealt with this problem via legislation. Finland notes that, even so, it still has problems with administrative sources changing or ceasing to exist.

3 The ability of administrative sources to meet requirements

3.1 A benchmark count of people and housing

The key sources of information in countries that have moved towards the use of alternative sources for the production of benchmark statistics are a population register and a dwellings register. A robust model is totally register based and has the following features:

- a population register where persons are uniquely identified;
- a register where all dwellings are uniquely identified and to which persons have been linked; and
- processes which allow, as a minimum, the changes in the relationship of a person to the living unit to be maintained over time.



The maintenance function is particularly important in this model. As well as recording the entries, exits and changes (births, deaths and marriages) that we associate with the registration services it crucially maintains the link between a person and a dwelling which is broken when a person moves.

Denmark has produced register-based population statistics since 1981. At the core are the Central Population Register (CPR) and the Building and Dwelling Register. The CPR is continually updated with key life events. When children are born they are given a unique personal identification number and are linked to their parents. Every person is linked to a unique 'dwelling' in the Buildings and Dwellings Register and is required to register a change on moving.

Although often cited similarly, the Norwegian system has not reached the robustness of the Danish model because of weaknesses in the dwelling register. Increased use of administrative data linkage was enabled in Norway by the introduction in 1964 of a unique person identifier in a CPR and its subsequent use in other registers. At the time there was not a comprehensive 'dwelling' register since older dwellings within multi-dwelling buildings were not uniquely identified.

To remedy this situation the Norwegian authorities set up 'The Dwelling Address project' to identify all dwellings and establish links from the CPR to them. Initiatives associated with the 2001 Housing Census resulted in 55 per cent of all persons living in multi-dwelling buildings being linked to a unique dwelling. By the end of 2002, coverage had increased to 65 per cent, largely through updates provided by removal companies. Work is currently underway to seek further improvement. This work includes additional surveys to collect address information for persons lacking a dwelling number.

The Central Bureau of Statistics in the Netherlands produces population statistics from a population register and dwelling information from a dwelling register, but the registers are not linked as closely as they are in the Danish model. In the Netherlands system full address details are maintained on the person register rather than a unique reference to a dwelling on the dwelling register.

The three systems cited have different characteristics, but all are capable of producing benchmark population and housing information. The situation in this country is

different. We do not have a population register and cannot rely on other sources to accurately produce a population benchmark. Key data sources overstate the population, as ineligibles are not removed when they should be. The largest alternative data source we have is based upon GP registration lists, but numbers on these lists are commonly much higher than the underlying population estimate in many local authorities. Electoral rolls are unreliable for similar reasons.

The most accurate records we have on the population tend to be those associated with benefit receipt. Child benefit payment records offer a count of children under 16. State retirement pension records offer a count of people of retirement age. These data sources only refer to a partial population and may suffer from having out of date address information. Lists are also inflated by people entitled to benefits who are living abroad.

Each of our sources falls short of offering benchmark population figures. The situation with address and property information is more encouraging, however. The Acacia partnership, a consortium of government agencies that include Ordnance Survey (OS), HM Land Registry (HMLR), Registers of Scotland (RoS), Valuation Office Agency (VOA), Improvement and Development Agency (IDeA), and the Royal Mail may eventually deliver an improved addressing infrastructure.

If this addressing infrastructure could be enhanced so that multi-occupancy was fully identified and the characteristics of all properties were recorded we would have developed a powerful statistical source which could act as an independent measure of completeness for population figures. Such a source would also be a reliable address link between other data sources. The biggest challenge to creating this source is the identification of multiple occupation of a single address by more than one household. The Norwegian 'Dwelling Address Project' demonstrates the extent of this challenge.

3.2 Population structures: information on households and families

The identification of households and families is a significant part of the traditional census. Countries producing statistics based upon administrative sources have generally been successful in identifying 'households' from their registers, but have often used survey data to complement register data for the identification of families.

The two most commonly used definitions of a 'private household' are the 'dwelling' based definition and the stricter 'housekeeping-unit' definition[#]. A 'dwelling register' when fully developed is analogous to a household register if the 'dwelling' based definition is used. Registers do not support the 'housekeeping unit definition'. The household-dwelling concept is used in Denmark, Finland, Sweden and Norway. Improvements to the definition of our address base outlined in **section 3.1** would allow us to approach a 'dwelling' based household definition. These improvements would also increase our ability to determine some characteristics of population structure by where people live. This would require classification of all properties and not only those residential properties identified in data collected by the valuation office agency.

Register-based systems generally carry a high level of relationship information in their population registers. One weakness tends to be families that extend beyond those explicitly defined in statutes. Norway has a highly developed population register, but cannot fully identify families. "The number of cohabiting couples... is almost impossible to count. It is difficult to define clearly when two people who share a common dwelling should count as a cohabiting couple, and even if this problem could be overcome it would require a lot of resources to do the counting. We therefore have to rely on survey data." Sources available in England and Wales cannot approach the comprehensiveness of register based systems. In the long term we would have to link registration events to a population register to achieve comparability. We may also have to supplement register-based data with survey data to ensure representation of all population structures.

3.3 Population characteristics and multivariate analysis

Census data contain detailed population characteristics and support multivariate analysis. This quality is dependent upon collecting consistent and comparable data, with a range of topics giving rich information down to small areas and population sub-groups. David Wroe's 1998 paper entitled 'Beyond 2001 - Alternative to the Census' catalogued available data sources that could provide similar information. Wroe's analysis pointed to the following information being unavailable from administrative sources:

[#] "A private household may be defined as a group of persons who share the same dwelling... A somewhat stricter definition is the "housekeeping-unit concept": in addition to living in the same dwelling, household members should also have common housekeeping" Statistics On Households And Families In Member Countries Of The CES (Geneva, 10-12 June 2003)

- language;
- economic activity;
- qualifications;
- religion;
- limiting long term illness; and
- usual address one year ago.

Since Wroe reported there has been little increase in the availability of person topics from other sources. Wroe assumed that many of the key demographic variables would be available from a future population register. In this he included:

- name;
- date of birth;
- sex;
- marital status;
- country of birth; and
- term time address.

Were the population register within the Citizens' Information Project to include marital status or country of birth in the core variables, then it would form a valuable 'spine' by which to connect other sources.

Data that might be used to substitute for topics associated with a traditional census include:

- car ownership;
- income; and
- qualifications (post 1993).

We should endeavour to use administrative data on subjects such as income and benefits to enhance census data. In the judgement of The Treasury Select Committee "it is clear that a question on income would have been found useful by many users of census data and we recommend...that further consideration should be given to the inclusion of such a question in any future census."

An analysis of quality and coverage of other potential topics is required before inclusion. Car ownership can be determined from DVLA registration records, but is not the same as car use and might show significant variations at a local level.

Wroe was optimistic even in 1998 about the possibility of collecting housing information from alternative sources. Since the report there have been further improvements. Initiatives

in the Office of the Deputy Prime Minister are moving towards the creation of a national property database.

The revaluation of 21.5 million properties in England and Wales for council tax banding purposes is particularly significant. This will collect information on a number of topics associated with dwelling type, size of property and amenities. The revaluation process is due to complete in 2006 with information being available early in 2007.

The specific census topics covered in the revaluation exercise are:

- number of rooms;
- central heating; and
- floor level.

Considerable development of the administrative sources is still required and it is not possible to say whether they will be sufficiently developed to replace census questions in 2011. However, as a minimum, census responses to these questions should be compared with the administrative source after that date, which will allow assessment of the quality of any future register of housing.

It is unlikely that all possible information required by census users will be available from other data sources by 2011. Address register development is underway, although the development path is unclear. Population register development remains only a possibility.

Careful evaluation of the cost savings that could be achieved by scrapping a census against the loss of specific topics will be needed before decisions on the future of the Census beyond 2011 can be made. This should include an evaluation of those sources where, although there may be some relevant information, the quality and coverage are insufficient.

3.4 Further uses of alternative sources

In 2000 the US Census Bureau ran an Administrative Records Experiment (AREX). The experiment was planned in response to a need to produce a comprehensive assessment of the feasibility of an Administrative Record Census in 2010. Six national files and a Census 'Numident' file (a Social Security Number Master file) went through a series of processes designed to create a single source of the most

trusted administrative data. The Social Security Number (SSN) was used for exact matching.

The composite record included selection of a 'best' address when multiple addresses were found. This selection was based upon geocodability, currency and quality. The composite records included imputed demographics where missing. When all composite records had been created the resulting record set was compared with the Census database. Where an address was missing from the AREX file, Census records for that address were added to the AREX file.

Outputs from the AREX file were finally compared with outputs from the 2000 Census from five test sites in Maryland and Colorado. Although the experiment was considered too small to conclude that a census based largely on administrative records was viable, it was concluded that further research should be carried out in their use for:

- substitution of administrative records for non-response follow up;
- improving accuracy in imputing characteristics of unclassified households;
- improving the quality of address lists;
- developing and testing unduplication methods; and
- estimating and adjusting undercount within a triple system estimation.

The AREX has greater significance to countries without a Central Population Register. All of the applications outlined above exist for an England and Wales census.

Further research of these uses should be undertaken although there would need to be significant differences in the approach to matching. The AREX was aided by the existence of the SSN on the national files. We would have to use multiple matching techniques in the absence of a common identifier across potential sources.

4 Conclusions and recommendations

Information on the key population units of people and housing can be effectively produced in a system based on population and address registers. There are initiatives underway which might lead to the development of a single address register but the development path is not certain. More significantly, research into the feasibility of developing a population

register has been undertaken but it remains only a possibility. The US Census Bureau's AREX experiment tested the feasibility of an 'Administrative Record Census' without a population register. While this research tends to support the notion that the approach is feasible it does not offer absolute confirmation.

Reliable information on the population structures of households and families are also dependent on the development of address and population registers, but require each to be taken further. Population registers would need to identify dwellings within addresses, and a population register would need to include family relationships, which would be a step beyond what is currently being planned. Even then, informal 'cohabiting' relationships would still only be identifiable via survey data.

Sources of information on housing and population characteristics are going through considerable development, and may well provide a wealth of data by 2011. These sources are currently not of sufficient quality and it is unclear whether they will be by 2011. Certain key topics such as means of travel to work are not available from any source. Further work is required to assess and monitor the quality of these sources as they develop.

Given that information on population units and population structures are the most important aspects, linkage of existing data sources does not offer a viable replacement for the Census in 2011. However, the potential development of a population register, and the development of an address register and other administrative sources mean that it may well be a viable replacement beyond 2011. For this to happen, research and development of a linked sources database should start as soon as possible, with an assessment of progress by 2007 to inform decision beyond 2011.

This and other development would also enable research into using administrative sources to significantly enhance a census in 2011.

An address register will be fundamental to the success of a 2011 Census particularly for:

- the identification of addresses for post-out of forms;
- the identification of property types to determine enumeration approaches; and
- the identification of multi-occupancy.

As well as pursuing the creation of such a register ONS should also:

- plan in 2011 to use administrative data as a further aid in measuring coverage of the main population unit data;
- research the potential to use administrative sources to provide additional or substitute topic information in a 2011 Census;
- research the potential to use administrative data as a substitute for non-response follow up in 2011;
- research the use of alternative sources for improving accuracy in imputing characteristics of unclassified households and individuals; and
- research the use of alternative sources in developing and testing unduplication methods.

Clearly, such use of administrative sources will require considerable public debate and changes to legislation. The Office for National Statistics, together with the Government Statistical Service as a whole, is beginning to engage with issues associated with the use of these sources in a statistical framework.

1 Philip Redfern 'Future Censuses of Population' Speech to SSRC Users Conference 1971

2 Rhind, David. 'Successors to the census of population' *Journal of Economic and Social Measurement*, Vol. 13, No. 1, Apr 1985

3 David A. Freedman 'Ecological Inference and the Ecological Fallacy' University of California 1999

4 John Newton and Sarah Garner 'Disease Registers in England' ISBN 1 8407 50286 February 2002

5 Eurostat and Statistics Finland 'European Workshop on Using Administrative Data in Population and Housing Censuses' Helsinki, Finland, 9-11 October 1995

6 Cabinet Office 'Privacy and Data Sharing: the way forward for public services' April 2002

7 Cabinet Office 'Identity Fraud: A Study' July 2002

8 Leicester Gill et alia 'Methods for Automatic Record Matching and Linkage and their Use in National Statistics' *National Statistical Methodological Series No.25* 2001

9 Christer Hyggen 'Demography of the family in Norway' First report for the project 'Welfare Policy and Employment in the Context of Family Change' December 2002

10 See www.statistics.gov.uk/census2001/pdfs/1998altcensusrep.pdf

11 Treasury Committee First Report 'The 2001 Census In England And Wales' Session 2001-02