

Beyond 2011 - Future Population Statistics for the United Kingdom

Nolan, Frank
Office for National Statistics, Methodology Directorate
Segensworth Road, Titchfield
Hants, PO15 5RR, United Kingdom
E-mail: Frank.Nolan@ons.gov.uk

1. Introduction

The UK has a long tradition of Census taking dating back to 1801. In agreeing to proceed with the 2011 Census of England and Wales, the Cabinet Office minister Francis Maude noted that this could be the last Census of its type.

The challenge for traditional Census taking countries has increased with the 2011 Census round. This period has coincided with an economic climate that is not conducive to the large concentrated costs of a national census. Additionally the balance between the Census providing a wide range of variables to satisfy user needs and the perception of intrusiveness and the requirements for compulsory completion (including the possibility of fines for people not completing the Census) is pitting the increasing demands of users against the concerns and refusal of respondents.

The perception is that the method of providing socio demographic statistics through Census taking is out-dated. It uses methods that have changed little in the 200 year history in the UK, and indeed are fundamentally little different from those of the earliest census takers. Additionally the outputs from the Census are too infrequent and too slow in comparison with the modern-day paradigm of instant information. This is despite the introduction of modern technologies including internet collection, scanning and recognition and delivery of outputs through the internet.

The solution offered is that of following the lead of other countries (such as the Scandinavians) in using administrative data to deliver socio demographic and population statistics. This solution solves many of the perceived drawbacks of the traditional census –

- Respondent burden
- Cost
- Frequency
- Efficiency

Some countries the production of population and socio demographic statistics through means other than Census taking is more complex because of a lack of appetite for any the development of population registers.

This paper will look at preliminary work in thinking about how the UK could develop a social demographic statistics system with less reliance on a decennial census. The first section looks at methodological and definitional issues. The second section looks at user requirements, noting that these are becoming more comprehensive given access to and skill in modern technologies. The next section describes the problem of determining what quality is required and the issue of measurement of quality with administrative data sources and combined data. There is a brief discussion on the parallels between population statistical collections and those of business statistics. And the last section outlines the current work plan that ONS is proposing to provide a recommendation in 2014 – the Beyond 2011 programme.

2. Methodology challenges

The traditional census (as seen in such countries as UK, Canada, US, Australia, New Zealand) is a survey of households and collects information on several dimensions:

- Individual
- Family
- Household
- Dwelling.

So there is often reference to the Census of Population and Dwellings. The first phase of the census is an enumeration, or listing, of dwellings and an attempt to determine the households within a dwelling (if there was more than one). Methodologically this initial work is actually developing the population frame of households, with the secondary process of completing the census questionnaires, determining the population of individuals. There is then a process (from the question on relationship to head of household) of determining families.

Social sample surveys, such as the Labour Force survey, have often used this same frame approach, built around a household. The household (which can be wider than the family) is seen as an economic unit in which people live that provides something in the way of shared income and consumption. The household is now commonly defined around shared living, shared meals or such like. The household is important in that members of the household are not independent. Collectively they can rely on a shared or common income.

The more recent Censuses have attempted to reduce costs by using mail in the enumeration process. Firstly it was relatively easy to have households post completed questionnaires back to the statistics office. This process is only successful if the management information system (or Questionnaire Tracking system) is fast in recording receipted questionnaires, so that field staff are not wasting time chasing questionnaires to be returned when they have already been posted back.

The next development has been that of post-out. With both post-out and post-back, enumerators are simply involved with collecting questionnaire which have not been posted back (or reminding the householders to post back). The success of post out is dependent on having a complete household frame. Statistical offices have used address lists (perhaps derived from postal lists or other administrative data) as a proxy for the household frame. The challenge is that the address lists are not a complete list of households, but also they record addresses at which there are not households (so both under coverage, over coverage and out of scope issues).

An added problem with address lists is the need for careful enumeration of non private dwellings – hospitals, armed forces barracks, hotels, ... The people living in these dwellings are generally not family members, but are institutionalised for some common reason – sickness in hospitals, defence forces for barracks, etc. And these can contain large numbers of people – often many more than the four person nuclear family. The non private dwellings are important in terms of population estimates at a local area level as they can contain a significant proportion (with distinctive characteristics) of the local population.

Successful administrative based population systems are generally based on systems which require compulsory registration of dwellings, of people, and of work places. The linking of these three registers can provide data on individuals and dwelling (and the people within them – perhaps even families). Common identifiers link people to work and people to dwellings. People common to an address may be able to be grouped into families or households.

Some countries have been reluctant to develop population registers. Politicians are wary of the perception of a “Big Brother” society with everyone numbered. And the cost and risk of failure in developing such an electronic based system are both high.

Without this commitment the statistical office is left with the difficult task of approximating the household frame using some address list, or further development of this. A key element of scientific method is inference; to use the available data and scientific method to verify (or not) a hypotheses. The household frame needs to be robust to have confidence in such inference. Address lists generally do not provide a complete list, and the complexity of relations between individuals, families and households are unlikely to be derived from address lists. So address list generally miss homeless people, do not recognise people who have two homes, etc.

A solution is to further develop the address list. This can be by validating the list through field work, perhaps using enumerators – introducing cost again. Alternatively there can be the use of visual tools (Goole Earth). But there is again difficulty of looking in the house or talking to occupants to determine whether it is in scope and how many households or families live there.

So it is difficult to go from an address register to the key population element – the individual. And there is not necessarily and one-to-one relationship between individuals and households. These different populations (such as weekday or weekend) are important for users, especially if there are material changes in their totals particular populations at different times.

Of course many administrative systems will be based on transactions with people. Combining many administrative datasets together could provide an estimate of the population using some decision rules (recognising that some people address links will be out-of-date). But there is a need to understand the quality of the population or household frame, as this determines the validity of the inference from the data.

3. What do users want?

The work of reviewing the UK system of population statistics will be undertaken through the Beyond 2011 programme. This is a four year programme, starting in April 2011 (just after the 2011 UK Census day of March 27), to establish and test models for meeting future user requirements for population and wider socio-demographic statistics.

A significant first phase of this work will be consulting with users about their future needs. There is already a good understanding of much of this need through the comprehensive discussions that have taken place around the 2011 Census.

Already user expectation is around keeping the current census type outputs with socio demographic variables at individual level to produce fine level geographies. This allows multivariate analysis. The traditional justification of a census collection is to provide outputs for small populations, both in terms of small areas and in terms of domains such as age or ethnicity. Additionally the Census provides the benchmark for household sample surveys, including those conducted by the private sector, which often use Census totals for post stratification.

Increasingly access to the micro data has been provided to users. Using a data enclave (or micro data laboratory), micro data provides researchers with greater levels of access for more complex analysis, while maintaining the confidentiality and security of the micro data through

binding the researchers in both access protocols and output checking (disclosure controls). And the Census provides a rich source of micro data.

A challenge in Census work, which will be also important in the Beyond 2011 programme, is around the different needs of the individual countries in the United Kingdom. The Census has already highlighted different needs around topics like religion, education, and language. With the some statistics being a devolved matter (i.e. the responsibility of the different individual countries in the UK), there is the risk that future population statistics could be produced by different types of systems. Certainly the administrative systems in the different countries are different and there again may be differences in having a system driven through administrative data. But there is a strong user demand for harmonised UK outputs. The UK Chief Statisticians have agreed to work together on the Beyond 2011 programme, as they have done with the 2011 Census.

A key requirement is the need to adhere to the regulations of the European Community. This is currently around a decennial set of statistics across several domains that could be satisfied by a Census or administrative or register based data. Given the legal imperative of European regulation, providing these outputs must be a base for future systems.

Users will also want to be convinced about continuity – the production of good time series. The major risk with administrative data is that being not primarily for statistics, the base of the collection may be changed because of different policy needs. The influence of the statisticians in the UK is less than in some other countries where the Director General may have a statutory role in changes to administrative systems.

And there will be requirement to be able to provide statistics for the range of different and changing geographies. With no common geographic building blocks and with some local administrative boundaries changing because of population changes, any future system needs to take careful consideration of the geographies that can support the outputs.

But the user will also expect change around frequency and speed – the way of the modern world. Results need to be instantaneous – data today about today. And they need to be frequent – available for every year or every quarter; certainly not every decade!

The Statistical office is left to determine priorities from what will be conflicting and impractical demands. Low level granularity, high frequency, wide ranging interconnected accurate data for today will not all be achieved.

4. The accuracy of data required

Statistical Offices are required to produce outputs that are fit-for-purpose. This is recognised both in Codes of Practice and Statistical Laws. The Statistical Law in the UK requires

The Board .. to promote and safeguard the quality of official statistics...[where quality] includes their impartiality, accuracy and relevance, and their coherence with other official statistics.

So there is a demand for accurate data. This is difficult to measure especially for data from administrative systems. The Beyond 2011 programme is based on measuring the success and quality of any alternative systems by making a comparison with the 2011 Census, the Census method being seen as a gold standard in terms of accuracy.

There has been considerable work in the UK to look at quality of statistical outputs. This includes the use of metadata included Basic Quality Information and Summary Quality

Reports associated with ONS statistical outputs. These fit the requirements of the UK Code of Practice for Official Statistics to ensure that *users are informed about the quality of statistical outputs*, .. A current challenge with this work is looking at how to measure the quality of Population estimates which are derived from Census data (as the base), survey data (for external migration) and administrative sources (for births deaths and internal migration). This work becomes more relevant if there is no Census.

Looking ahead, if there is no future census, what comparator is available to determine accuracy? All good quality administrative data is likely to be included in the model, and the data that is excluded is necessarily poor quality. But what is good quality?

The measure of quality is further complicated if there is a requirements for linking data, which is often the case with administrative data systems. Currently no other source of data is as comprehensive as the Census. Alternative models will necessarily mean matching data (perhaps at record level) between administrative sources and possibly surveys as well. Although technology provides greater potential for matching, there will still be accuracy considerations, especially depending on the match variables available (not always name and address).

Perhaps there is a case to return to the use of some high quality (compulsory) sample survey to undertake some quality adjustment, such as dual system estimation. This applies the method of a Census Coverage Survey, but in this case applying the estimation to the administrative model rather than the Census collection. This is an option used by some European population systems, where it is known that there is a likely high overcount (because there is no incentive for individuals or families to deregister when migrating). Taking a sample survey does provide a means to look at measuring over and undercount and then make suitable adjustments.

5. The programme of UK work - models

The Beyond 2011 programme will be investigating a range of models or systems to derive population statistics. Critical amongst these will be the use of administrative data. This has become more accessible for statistical needs through the Statistics and Registration Service Act (2007) which has specific clauses for data sharing.

The first investigation will be around low-level aggregate models. These provide for access to aggregated administrative data that is not at the record level. This set of data is important as it circumvents problems where departments are reluctant to provide access to micro data, but which could provide aggregations.

The second investigation will extend these models to look to see if there is enough micro data that could be made available to provide unit record level data. This depends on the availability of these data and the quality of any matching required, noting the absence of a unique population identifier.

The third investigation is around the combination of administrative data along with sample surveys and perhaps some form of census. This is the most complex of the options and is an attempt to make use of all data that is available. Within this option there is the need to consider if there is a need for special measures for population that are known to be difficult to count. So while there may be one model for most of the population, there may be differences for the more difficult, but perhaps important sub populations.

Significantly, there is also an important activity to determine how to evaluate the different options. This means setting out some criteria. This will likely cover the range of quality dimensions (relevance, accuracy, timeliness, accessibility, coherence and comparability) as well as costs. But there will also be a need to test systems for public acceptability. While the public are concerned with the burden of completing questionnaires and of sensitive questions, they need to be reassured about how their data is managed, especially if there is data linking. The systems proposed must be seen to be used only for statistical purposes and not to target individuals.

While this programme is being lead by the Office for National Statistics, there are also opportunities for wider involvement of statistician in the research. The Beyond 2011 programme will work jointly with researchers through the Economic and Social Research Council to provide wider thinking on possible solutions.

6. The Parallel Business Statistics programme

Some of the same drivers that are to the fore for changes to population statistics are also evident with business statistics: lower cost, lower respondent load and greater quality. So there is a parallel programme of work in the UK looking at how increasing levels of administrative data could be used in producing business and economic statistics.

There is already a key advantage in this stream of work with there being a register of businesses, which is derived and maintained from administrative data, including registration for Value Added Tax (VAT). Business data is traditionally collected through surveys of businesses, benchmarked with a census type survey at regular intervals. With departments such as the tax office having improved computer systems there are opportunities for using data such as taxation returns for both administrative and statistical purposes. This is important for small and medium sized enterprises where business compliance costs are seen as constraining innovation and profitability.

A key difference with businesses is the criticality of large businesses in their contribution to economic statistics. And conversely the small contribution that many small businesses make. This leads to greater emphasis on data from large business, which may be multi dimensional in terms of their contributions to different regions and different industries. A possible solution implemented in other countries is the compulsory survey requirements for large business, but the use of administrative data for small businesses supplemented by small sample surveys to be used for correlations for imputation of data not available through administrative system.

7. Conclusion.

Population statistics are critical in setting and monitoring government policy. Significant levels of public funding are allocated to targeted populations and the correct allocation of these funds depends often on accurate population statistics.

The current environment is not conducive to the conducting of a traditional Census as a base for population statistics. Census-taking in the form of enumeration and questionnaires is seen as out-dated. It is criticised as being expensive, intrusive, and quality is questioned. Local organisations depending on population-based funding are increasingly challenging the methodologies applied, and using local data to challenge local estimates from national surveys (where the outcome would be of benefit financially).

The new paradigm is based around use of data already collected through administrative (or commercial) processes. This paradigm is based upon the success of some European countries using registers as a base for population statistics. But the personal identifier, which is part of the base of the European system, is not available in the UK and unlikely to be available.

The Beyond 2001 programme has four years to investigate what users see as their requirements for population and socio demographic statistics. In particular, the programme must look at the range of topics, the geographic granularity that appears in the outputs and the range of analysis that the outputs need to support. However there is also a critical and new challenge to determine what quality of data is required.

The programme is to investigate a range of data models incorporating administrative (and commercial) data, sample surveys and Censuses and combinations of these. This work is especially difficult, with a solution which is likely to be more complex and difficult for users to understand. But the goal is to provide a system that better reflects the needs of the twenty-first century.

Abstract

The tradition of Census taking in the UK extends back to 1801. The decennial collection provides information from each individual and household in the country. The Census estimates then form the basis for future annual estimates of population. Increasingly users want a greater range of statistics to be available more frequently and with greater timeliness. There is also a need to reduce the cost and the cost profile resulting from a Census collection. And alternative data sources which have the potential to provide information on the population are becoming more widely available. This paper explores the several options available and will review the models that are being evaluated based around combinations of household surveys, administrative data and possible alternative Census models. This work is incorporated in the Beyond 2011 project.