SOUTHAMPTON STATISTICAL SCIENCES RESEARCH INSTITUTE (S3RI)

# Review of Methods for Estimating Populations with Administrative Data

**James Raymer, Dilek Yildiz and Peter WF Smith**

**3 October 2013 (Revised)**

The United Kingdom’s Census Offices are investigating new ways of producing population and small area socio-demographic statistics without relying on traditional census enumeration. Many of the options under consideration involve combining data from a variety of sources, such as surveys, address registers and administrative data. In particular, making better use of currently available administrative data holds much promise for improving the efficiency and accuracy of estimating population totals and characteristics. However, the issues and obstacles surrounding combining different sources of administrative data are not well understood. In this report commissioned by the National Records of Scotland, we provide an overview of the main issues concerning administrative data and some recommendations for combining administrative data to produce current estimates of local populations in Scotland by age and sex.

## 1. Introduction

### *1.1 Background*

The United Kingdom's Census Offices are investigating new ways of producing population and small area socio-demographic statistics without relying on traditional census enumeration. Many of the options under consideration involve combining data from a variety of sources, such as surveys, address registers and administrative data. In particular, making better use of currently available administrative data holds much promise for improving the efficiency and accuracy of estimating population totals and characteristics. However, the issues and obstacles surrounding combining different sources of administrative data are not well understood. In this report, we provide an overview of the main issues concerning administrative data available in Scotland and some recommendations for combining such data to produce current population estimates by age, sex and area of residence.

Administrative data suffer from a number of limitations. They represent populations in a particular register and tend to exhibit variable quality and consistency over time. Furthermore, the absence of both a complete population register and a unified personal identification system limits the scope for micro-integration amongst various registers. In this context, the use of administrative data for population estimation presents a considerable statistical challenge.

In recent years demographers, geographers and statisticians have put forward a range of approaches to construct or quality assure estimates of population stocks and flows using a combination of different sources. In this report, the relevant methodological work in demography and related fields are reviewed and some recommendations for developing and testing methodologies for producing population estimates from combined administrative data are provided. The proposal takes into account existing knowledge about data sources currently available for this research in Scotland and their quality characteristics. It also considers new sources created through micro-integration of several primary sources.

### *1.2 Scope of project*

The specific aims of this report are to review current research on combining administrative data and to set out a proposal for investigating the use of statistical modelling techniques for population estimation using administrative data.

### *1.3 Outline of report*

The remainder of the report is organized as follows. In Section 2, we describe the basic requirements of a population estimation system and the broad issues concerning the combination and inclusion of administrative data in a demographic accounting framework. This is followed in Section 3 with an introduction to the main administrative data sources in Scotland that can be used for population estimation. Here, a distinction is made between whole population data sources and subgroup population data sources. In Section 4, we present various approaches for using administrative data to estimate populations. We end the report in Section 5 with some recommendations for developing an administrative data base to estimate local populations in Scotland by age and sex.

## 2. Identifying the requirements of population estimation system[1]

Population statistics are needed for planning and making policies. They are also used for international comparisons. The minimal population statistics include totals by age, sex and location. Population statistics are often classified by attribute information, such as ethnicity, education and health. Furthermore, statistics are often provided concerning the underlying

[1] This section draws directly from the ‘Conceptual Framework for UK Population and Migration Statistics’ report written by James Raymer, Phil Rees and Ann Blake (2012).

drivers of population change: fertility, mortality, internal migration and international migration. In order to create a population estimation system, one needs to conceptualise the types of population statistics that are required, the availability of data sources, the procedures used to align the available data to requirements of population statistics and the outputs to be produced. Furthermore, providers of population statistics should be clear about the quality of the outputs.

### *2.1 Conceptual Framework*

As outlined in Raymer et al. (2012), statistical outputs are often the result of matching available data to a particular population concept. In doing so, the data are sometimes processed or combined with other information to fit a particular concept. The production of population and migration statistics is further complicated because populations are both dynamic and heterogeneous. Populations continuously change according to the addition of births, subtraction of deaths and addition or subtraction of migrants (domestic and international). These processes are influenced by the social and cultural environments, economic environments and natural and built environments in which the populations live, as well the intersections between them (Bycroft 2011). In order to understand population statistics, one must first realise that they only represent a 'snapshot' of a population at a particular time.

All types of population can be related to the actual population at time *t* in location *i*. Likewise, concepts of migration can be related to the movement of all people in and out of location *i* between two time points. To estimate particular populations, therefore, one must consider the types of entries and exits (including births and deaths, respectively) between time points *t-n* and *t*, where *n* refers to the width of the time interval (e.g., days, months or years).

The size and characteristics of a local population may vary greatly, depending on both the time of day and day of the year it is measured. In the United Kingdom, a mid-year ‘usual resident’ or ‘night time’ population (30 June / 1 July) is estimated. A ‘night time’ measure captures the population where it sleeps, whereas a ‘day time’ measure captures the population where it goes to school, work, market or takes leisure. Business travellers and visitors are usually excluded from official and international mainstream statistics on population and migration. Temporary workers may be included in official estimates of short term migrants, but not in the usually resident population.

Perhaps the most accurate information we have is on the number of live births and deaths for locations in the United Kingdom over time. This is because all births and deaths have to be registered by law. Births are published according to the age, sex, birthplace and residential location of the mother. Deaths are recorded for all persons by age, sex and residential location. Migration, on the other hand, “is a loosely defined process that represents the relocation of people during a period of time that causes them to relinquish the ties with their previous locality” (Raymer and Smith 2010, p. 703). Migration can involve people moving within a country, as well as across international borders. The factors that separate migration from other forms of mobility (e.g., daily commuting, week-day/weekend commuting, holiday visits or seasonal moves) are generally distance travelled and length of time spent in the destination (or away from the origin).

Migration data are obtained from general purpose censuses / surveys or administrative registers. The practical measures of migration obtained from these sources often do not coincide with theoretical or contextual definitions of migration (Bell et al. 2002). The reason for this is that, unlike births and deaths, there are no legal frameworks for measuring migration. In practical terms, migration can be defined as relocations between administrative areas and mobility as relocations within areas.

Geographic classifications are fundamental for understanding society and population change. There are many different ways of representing geography, depending on the data source. However, statistics usually come in the form of aggregate units, such as local authority districts, counties or regions in England and Wales or electoral wards and council areas in Scotland. Sometimes, the actual geography is not of interest but rather the area type, such as urban, rural or coastal. Geographic information is important for planning schools, hospitals, workforce, as well as for comparing different spatial patterns of residence according to ethnicity or density.

Age, sex and geographical location are considered the baseline characteristics required for population and migration statistics. For understanding change or differences between population groups, it is often useful to have more detailed attribute information, depending on the need or users. For example, for those interested in migrant integration, information on the foreign population, their levels of education and their occupations are useful. For those wishing to set migration policy, understanding the reasons or drivers for migration is important. For services provision, information on population health, number of children and economic activity are useful.

### *2.2 Error and Uncertainty*

The acknowledgement and inclusion of error and uncertainty in population and migration estimates is an important aspect in the production of official statistics. Uncertainty refers to the overall accuracy of the estimates. Bias refers to whether the estimates are over-predicted or under-predicted on average. Both measures are required to understand the quality of the estimates being produced.

Uncertainty can come into the population estimation model in many ways. At the onset, the base population taken from, for example, a recent census may contain error. The components of change may contain error. In the United Kingdom, internal migration and international migration are considered the most problematic in terms of accuracy. Whereas, registrations of births and deaths are considered highly accurate since they have a clear legal framework and long history of data collection.

The sources of bias tend to be undercounting of population groups that are less likely to fill in questionnaires or register in an administrative source. These include highly mobile groups, such as students, migrants, homeless persons and the armed forces population, as well as populations living in areas of high deprivation, unemployment or crime. Over-counting may also be an issue in some administrative sources as a result of lags in deregistration.

When producing estimates, we know that accuracy is higher for larger populations as it takes more to influence their population change. If a recent census is used as the base population for estimation, then accuracy should decrease over time as one moves further away from this benchmark. For example, we would expect the 2012 population estimates to be more accurate than the 2013 population estimates because the most recent census occurred in 2011. The availability and inclusion of historical time series data and covariate information should also improve accuracy. However, this may not always be the case as each source of additional information contains error and introducing them into the estimate could result in more potential sources of error.

In general, there are two ways of including uncertainty in population and migration estimates. The first is to present ‘high’, ‘principal’ and ‘low’ scenarios based on assumptions about the plausible ranges the estimates can take. The second is to incorporate the actual or estimated probabilities. Information that can be used to estimate uncertainty in this way includes empirical data (including time series of historical data), auxiliary data and expert judgements. This second option is preferred because it provides a quantitative figure as a basis. However, it is considerably more complicated because it involves identifying and incorporating all the main sources of error from all the information used in the estimation.

The accuracy of censuses is usually obtained by conducting a post-enumeration survey in particular areas of the country. Sometimes, two sources of estimates may be compared to identify error or biases. An example of this is the comparisons of 2011 Census counts with alternative counts derived from administrative registers, carried out as part of the post-census validity checks. For example, the Office for National Statistics found a large number of 'missing' young adult males in the census compared to what they expected. Another example is a comparison between the 2001 Census internal migration flows with the observed flows from the 2001 National Health Service Central Register, which highlights the undercounting of young adult males in the health register data.

## 3. Current administrative data in Scotland

Population estimates can be achieved in several ways. Estimates may be the result of direct counting from an administrative or population register or by taking a population census. They may also be made by adapting existing administrative registers to fit a particular population concept or by adding and subtracting relevant demographic flows from a prior population count or estimate. There are many administrative data sources that can be used to estimate populations in Scotland. We focus on the main administrative data sources assessed by National Records of Scotland (2012) in terms of their closeness to mid-year population estimates, which include the National Health Service Central Register, the Community Health Index, the Department of Work and Pensions' Customer Information System, the School Census, Child Benefit data and the Super Old Persons database.

Administrative data sources are only able to capture the persons who use their services. For example, National Health Service Central Register only collects information from people who are registered with a General Practitioner. Therefore, it is essential to clarify which population groups are included in the administrative sources. The assessment is conveyed visually by using Venn diagrams as outlined in Figure1 below. The aim of these diagrams is to illustrate how the concept of the usual resident population, i.e., our main population statistic of interest, relates to the population covered by the administrative source in a specific area *j* at a specific point in time *t*.

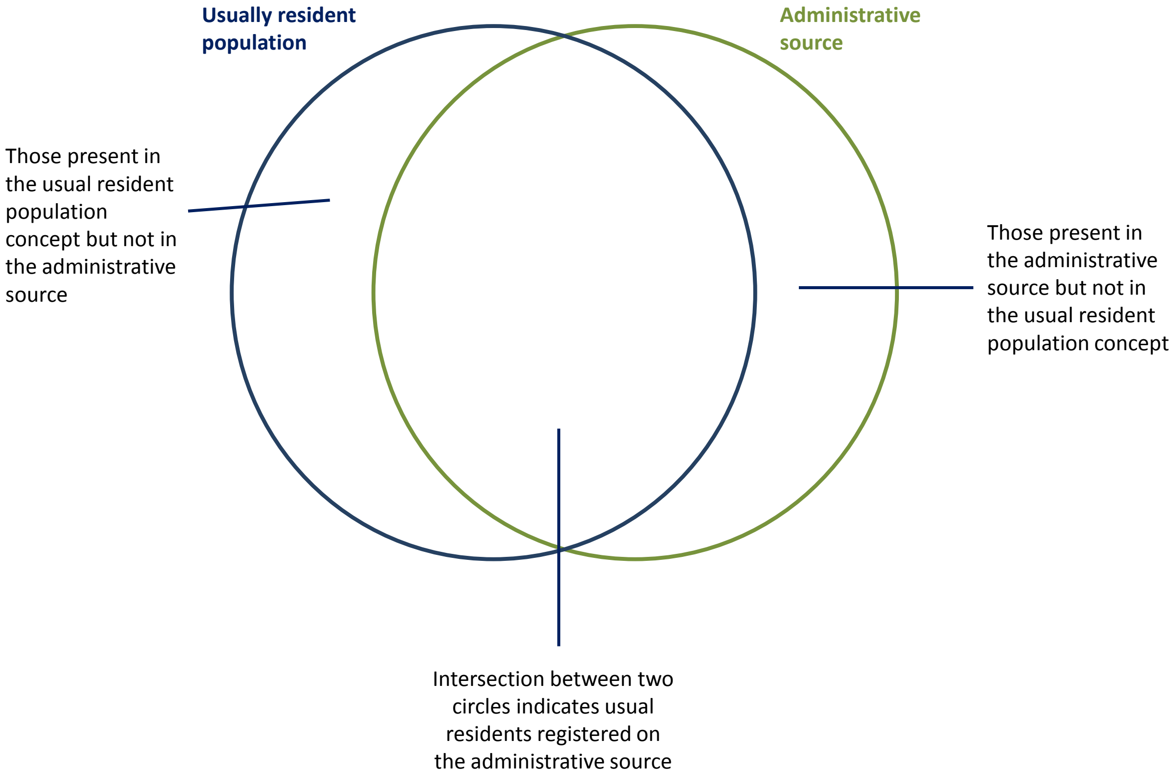
Usually resident population
Administrative source
Those present in the usual resident population concept but not in the administrative source
Those present in the administrative source but not in the usual resident population concept
Intersection between two circles indicates usual residents registered on the administrative source

*Figure 1. Relating administrative sources to the usual resident population in area j at time t* (Source: Raymer et al. 2012)

### *3.1 Whole population registers*

In this section, we describe administrative data sources which may be used to estimate the whole population in Scotland. These include the National Health Service Central Register, the Community Health Index and the Department for Work and Pensions' Customer Information System. Because the coverage of these data sources are the closest to the whole population, they could potentially be used for population estimation with surveys or other administrative registers used to fill missing data or to augment weaknesses.

### 3.1.1 National Health Service Central Register (NHSCR)

The NHSCR includes all births, deaths and persons registered with a General Practitioner in Scotland (General Register Office for Scotland 2013). The National Records of Scotland uses an extract of NHSCR which includes information on gender, date of birth, health board registration and postcode of residence. According to the National Records of Scotland (2012), the NHSCR overestimates population counts at all ages except the youngest age group (0-4), resulting in a population count that is eight per cent higher on average than the estimated mid-year population estimates. This difference is mainly due to unreported migration. Another reason for higher numbers in the NHSCR data is people who have multiple registrations although this is thought to be a small percentage (ONS 2012a). The NHSCR also has problems of underreporting or missing some population subgroups, such as immigrants who are not registered with a General Practitioner, prisoners, Armed Forces personnel and lags in the recording of newborns. Note, according to ONS (2012a), some Armed Forces and their dependents are not in the NHS Patient Register because they have their own military health services. The main differences between the population in the NHSCR and the usual resident population are presented in Figure 2.

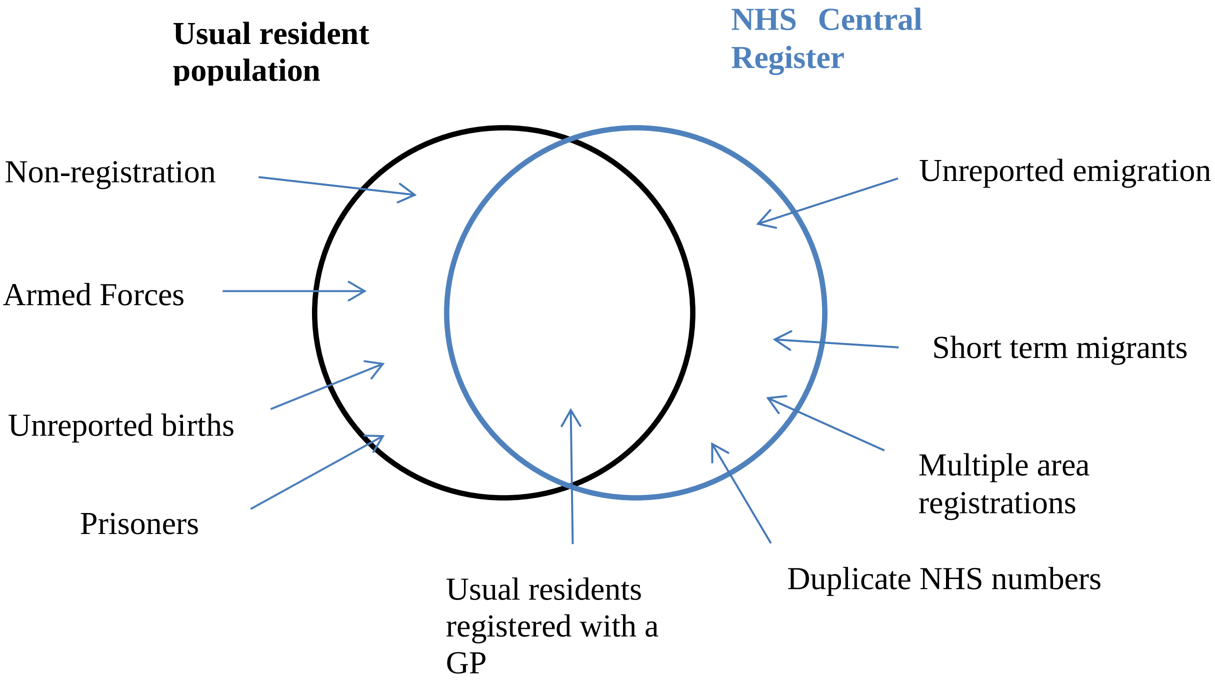

*Figure 2. The relation between usually resident population and NHSCR*

According to National Records of Scotland (2012), the NHSCR provides more accurate population counts of females than males. This situation is also consistent with the Office for National Statistics' (2012a) National Health Service Patient Register assessment for England and Wales. The NHSCR records overestimate the female population by 5.7 per cent and the male population by 10.3 per cent in comparison to the 2010 mid-year population estimates. The differences between the mid-year population estimates and the NHSCR-based population counts for females reach their highest values between ages 25 to 35 and at the oldest age group (90+). The differences for males are considerably more noticeable, more than 10 per cent between ages 25 and 54 years and above the age of 80 years.

### 3.1.2 Community Health Index (CHI)

The CHI is a unique number for everyone who is registered with a General Practitioner in Scotland or who had contact with other health services. It is a ten digit number which consists of date of birth and four other numbers. The CHI data set includes information on gender, date of birth and address. Although the CHI is highly correlated with NHSCR, it differs from NHSCR by including people who received health services but were not registered with a General Practitioner in Scotland. According to the National Records of Scotland (2012), the difference between the two databases does not cause coverage problems in the long term as they are regularly synchronized.

In the National Records of Scotland (2012) report, the 2001 and 2010 CHI-based population counts are compared with the 2001 Census and the 2010 mid-year population estimates, respectively. This comparison showed that the 2010 CHI-based population counts were five per cent higher than mid-year population estimates for the total population. The corresponding figures for females and males were 2.7 per cent and 7.4 per cent, respectively. The overestimate is greatest between ages 25 and 49 years for males and ages 25 and 34 for females. The CHI-based population counts underestimated the population totals in relation to the mid-year population estimates for ages 75 years and over. The same pattern can be seen in the comparison between 2001 CHI-based population counts and the 2001 Census, apart from the differences in the oldest two age groups.

### 3.1.3 Department for Work and Pensions Customer Information System (CIS)

The basic information of people who have been a client or customer of the Department of Work and Pensions since 1999 in Great Britain or Her Majesty's Revenue and Customs since 2005 in the United Kingdom is stored in the CIS. The CIS includes information on gender, date of birth and address.

An extract provided by DWP which only contains aggregate information on population counts by age, gender and postcode was used to compare 2010 population counts for Scotland by National Records of Scotland (2012). In comparison to the mid-year population estimates, the CIS covers the working age groups better than younger and older age groups. The CIS-based population counts are 2.3 per cent lower than mid-year population estimates. Amongst all ages, the CIS produced closer male population counts to mid-year population estimates (1.0 per cent difference for males and 3.6 per cent for females) but among working ages, the female population counts are closer to the mid-year population estimates.

As illustrated in Figure 3, the reasons for the differences in the population covered by CIS and the mid-year population estimates are due to (1) unrecorded immigration, (2) children under age 16 years, (3) spouses of immigrants who do not have a National Insurance Number, (4) unrecorded deaths, (5) unrecorded emigration, (6) short term migrants and (7) multiple National Insurance Numbers (Office for National Statistics 2012b).

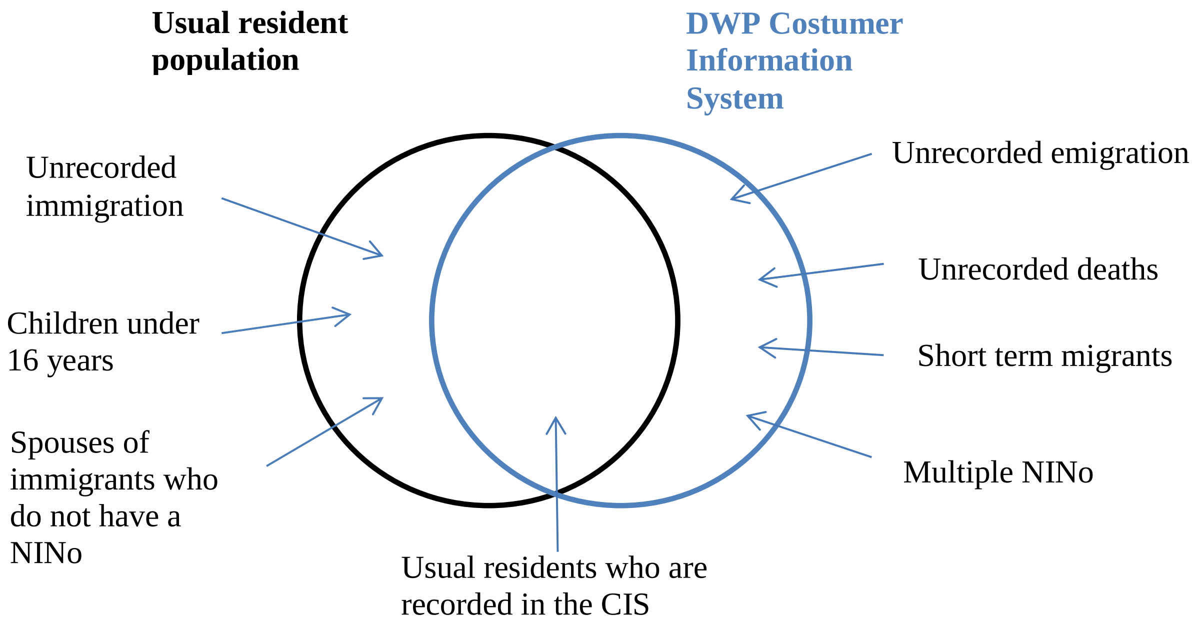
Usual resident population
DWP Costumer Information System
Unrecorded immigration
Children under 16 years
Spouses of immigrants who do not have a NINo
Unrecorded emigration
Unrecorded deaths
Short term migrants
Multiple NINo
Usual residents who are recorded in the CIS

*Figure 3.The relation between usually resident population and CIS*

### *3.2 Subpopulation registers*

In this section, we focus on administrative data sources which are designed to collect information only from a subset of the whole population. These include the School Census, Child Benefit data, Super Old Persons database and the Electoral Register.

#### 3.2.1 School Census

The School Census collects information about pupils in publicly funded schools in September each year. The children who are studying in the independent schools and who are educated at home are not included in the School Census, which  covers 96 per cent of children at national level and 82 per cent of children in the council areas where attending the publicly funded school proportions are the lowest (NRS 2012). As illustrated in Figure 4, the difference between the usually resident population in school age and the School Census coverage arises from (1) pupils attending private schools, (2) home educated pupils, and (3) young children not attending a nursery or attending a private nursery, (4) children who are aged between 16 and18 years and not attending school and (5) children of short term migrants who are attending publicly funded schools (Office for National Statistics 2012b). The most accurate data in the School Census are collected from pupils aged between 5 and 14 years. It includes information on gender, age, home address and ethnicity.

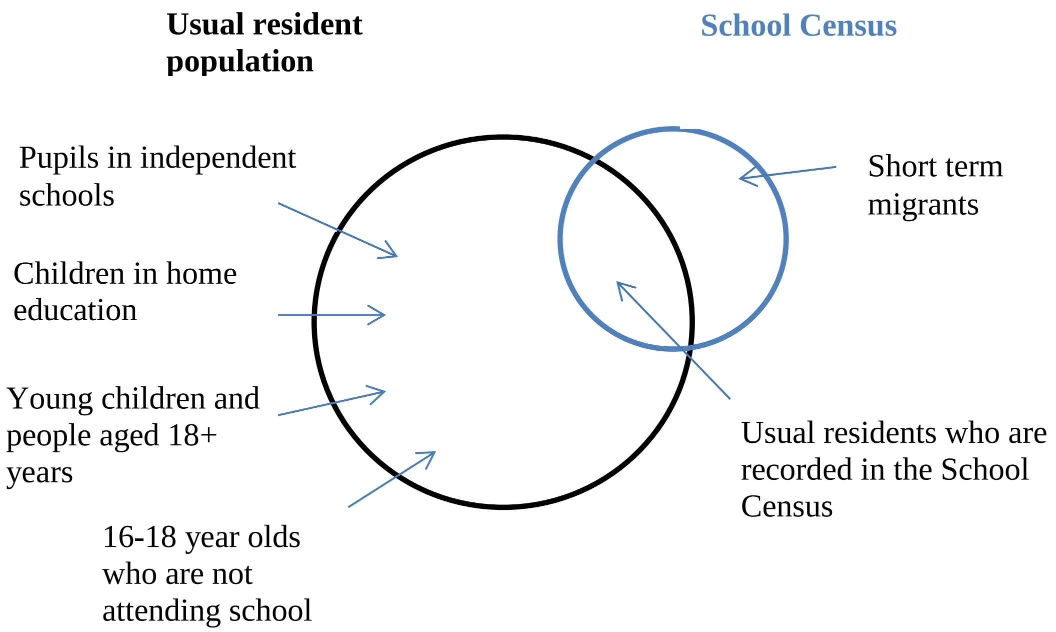
Usual resident population
School Census
Pupils in independent schools
Children in home education
Young children and people aged 18+ years
16-18 year olds who are not attending school
Short term migrants
Usual residents who are recorded in the School Census

*Figure 4. The relation between usually resident population and the School Census*

The National Records of Scotland (2012) report compared population counts based on 2010 School Census with 2010 mid-year population estimates for single ages between 5 and 14 years. Since the School Census does not cover all pupils, it underestimates population counts among all ages by 2.7 per cent (2.6 per cent for females and 2.8 per cent for males). However, it slightly overestimates the 9 and10 year old male population and the 10 year old female population.

### 3.2.2 Child Benefit Data

The Child Benefit data set captures 98 per cent of all eligible children. The persons eligible to receive child benefit are supposed to be physically present in the United Kingdom (together with their children), have a right to reside in the United Kingdom, and responsible for the child who is living with them (HMRC 2013). In Figure 5, the relationship between the usual resident population and the population in the Child Benefit database is described.

The age and zone of residence information derived from the child benefit data set is used to estimate the population count of children aged 15 years or younger. According to National Records of Scotland (2012), the child benefit data underestimated the population between 0 and 15 by one per cent on average with overestimates in 12, 14 and 15 year old ages.

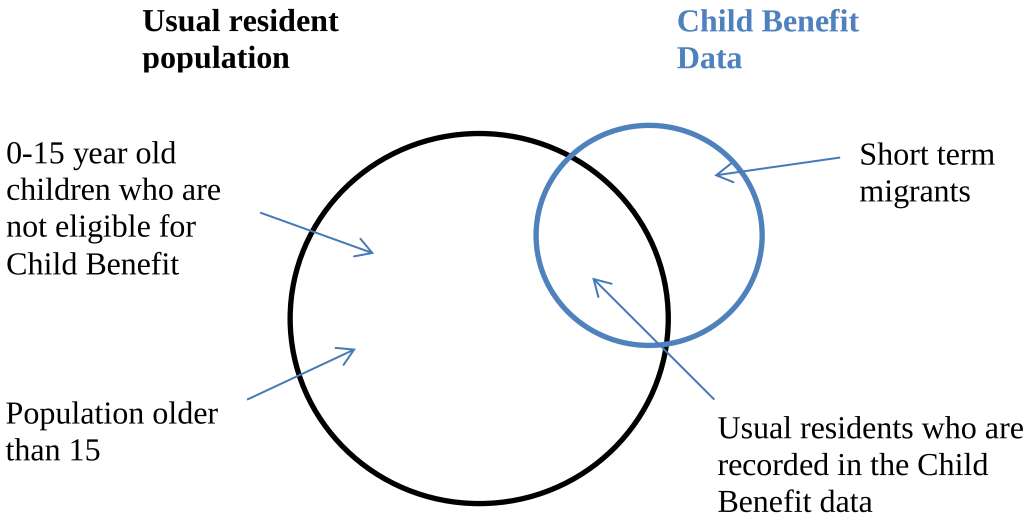
Usual resident population
Child Benefit Data
0-15 year old children who are not eligible for Child Benefit
Short term migrants
Population older than 15
Usual residents who are recorded in the Child Benefit data

*Figure 5. The relation between usually resident population and Child Benefit Data*

### 3.2.3 Super Old Persons Database

The Super Old Persons database in the Department of Work and Pensions contains information on persons aged 65 years and older. According to the National Records of Scotland (2012) report, the Super Old Persons database underestimates about one per cent of the true population. However, at the oldest age group (90+), there is an over-estimate of around 14 per cent for males.

### 3.2.4 Electoral Register

The Electoral Register only includes information of people who are eligible to vote and who opt to register. People are eligible to vote in the United Kingdom if (1) they are 18 years of age or over on polling day, (2) they are a British citizen, a qualifying Commonwealth citizen or a citizen of the Republic of Ireland and (3) they are not be subject to any legal incapacity to vote. People who are older than 18 but not eligible to vote both at general election and at elections to local authority are (1) anyone other than British, Irish and qualifying Commonwealth citizens, (2) convicted persons detained in pursuance of their sentences and (3) anyone found guilty within the previous five years of corrupt or illegal practices in connection with an election (The Electoral Commission 2013).

The relationship between the usual resident population and the population in the Electoral Register is illustrated in Figure 6. The Electoral Register does not provide age and gender information. The total population aged 18 and over by council area is compared with the mid-year population estimates in the National Records of Scotland (2012) report. The population counts based on Electoral Register differ from the mid-year population estimates by 6 per cent. However, there exist some variability around this number. For example, the difference between two population counts exceeds 15 per cent in the city of Edinburgh.

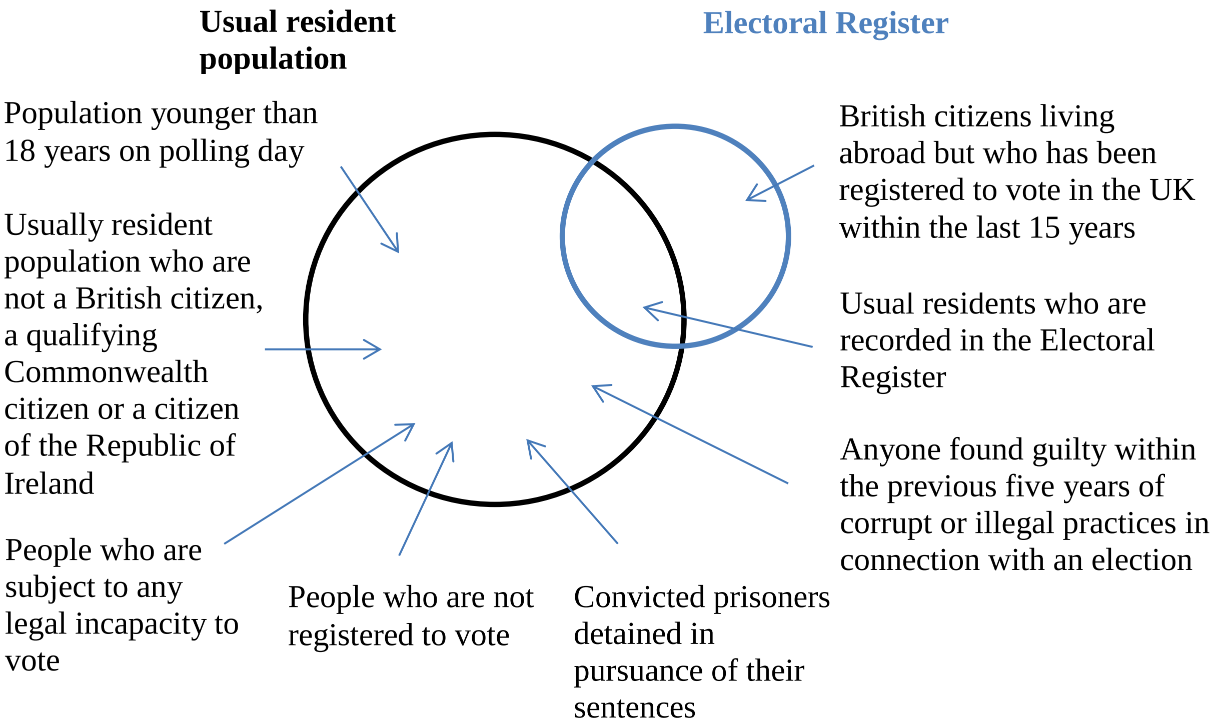
Usual resident population
Electoral Register
Population younger than 18 years on polling day
Usually resident population who are not a British citizen, a qualifying Commonwealth citizen or a citizen of the Republic of Ireland
People who are subject to any legal incapacity to vote
People who are not registered to vote
Convicted prisoners detained in pursuance of their sentences
British citizens living abroad but who has been registered to vote in the UK within the last 15 years
Usual residents who are recorded in the Electoral Register
Anyone found guilty within the previous five years of corrupt or illegal practices in connection with an election

*Figure 6. The relation between usually resident population and Electoral Register Data*

### *3.3 Summary*

The assessment of the major administrative data sources showed that none of these administrative data sources is sufficient to estimate Scotland’s population alone. It is important to recognise the difference between the population statistic required (concept) and the population captured by the particular administrative data source. Data adjustments, estimation and combination may be used to augment the deficient data. In theory, it should be possible to estimate the usual resident population. Making better use of administrative data sources has the potential to save money and resources. In the next section, we introduce some methods that can be used to estimate population statistics based on administrative data.

## 4. Methods for combining administrative data

In this section, we present several alternative approaches that can be used to estimate local populations by age and sex in Scotland. We focus on the basic population statistics with the idea that these could then be combined with surveys or other data sources to provide more detailed characteristics. We distinguish three approaches: the development of a population register, estimation by combining existing administrative data at the aggregate level and estimation by combining existing administrative data at the individual level. These approaches may overlap in practice.

### *4.1 Population register*

In this section, two types of population registers are introduced. The first represents a European-style population register based on residence. This type of register requires every member of the population to notify the local authorities when changes of address are made with access to services dependent on registration. The second represents the adaptation of an existing administrative data source to be used for population estimation. In Scotland, there is only one potential candidate: the National Health Service Central Register. Note, the development of the first option would involve substantial legal and operational changes and is therefore is only presented for completeness.

#### 4.1.1 Register based on residence

The Netherlands and the Nordic countries have developed highly advanced and integrated registers. These registers are based on place of residence and can be used to continually estimate the populations by age, sex and local areas at any point in time. Since the register is continuous over time, usual residence populations can be estimated by measuring how long each person has lived at their current address, i.e., those that have lived for most of a year in a particular place.

Population registers based on residence are very flexible and can be used to estimate populations for different geographical levels including small area populations. Statistics Finland (2004) describes some key preconditions that should be met before replacing traditional censuses with register-based censuses. These preconditions are legal basis, public approval, unified identification code systems, comprehensive and reliable register systems developed for administrative needs and cooperation among administrative authorities. Once all preconditions are met the register-based census provides frequent up-to-date information at lower costs and reduces the response burden on population.

### 4.1.2 Using an existing administrative data source as a basis for population estimation

The NHSCR collects information on everyone who is registered with a General Practitioner in Scotland. While most of the other administrative data sources collect information for specific age groups, the NHSCR collects information for all age groups. In theory, everyone who is resident in Scotland should be registered with a local General Practitioner. Therefore, it is one of the most suitable administrative data sources which can be used as a base register to estimate the population of Scotland. As mentioned in Section 3.1.1, the NHSCR does not have information about those who are not registered with a General Practitioner, such as people working in armed forces and their families, unreported births and prisoners. It also includes people who are not usual residents, such as short term migrants, people who have left the country but not informed the register, people who died abroad and people who have duplicate National Health Service numbers. Thus, in order to use the NHSCR to estimate the population totals, the coverage would need to be adjusted.

One way to adjust the NHSCR to match a usual resident population would be to carry out a coverage survey. A census coverage survey is a follow up survey conducted a few weeks after the census which aims to measure the coverage of the census at local authority level by age and sex. Census coverage surveys were carried out in the United Kingdom (Scotland, Northern Ireland, England and Wales) after 2001 and 2011 censuses (Brown et al. 2011). The linked census coverage survey and census records are used to check for any duplicated records in the census and to provide an estimate of the overcount in the census. Then the undercount is estimated by employing dual system estimation (see Section 4.3) and ratio estimation. The Office for National Statistics (2012b) describes the 2011 Census coverage assessment and adjustment process.

## *4.2 Combining two or more administrative sources at aggregate level*

### 4.2.1 Rule-based approaches

To combine administrative data, one option is to identify which sources best capture segments of the population by age, sex and geography, possibly after adjustments, and then uses those sources to represent them. An example of such an exercise is provided in Figure 7, where NHSCR is used for estimating the population between 0 and 4 years, School Census is used for estimating population between 5 and 14 years, the Department of Work and Pensions' CIS is used to estimate the population between 15 and 64 years and either the Super Old Persons Database or NHSCR is used to estimate population aged 65 years and over. A more sophisticated frame can also be employed to combine data sources according to sexes or council areas in addition to age groups.

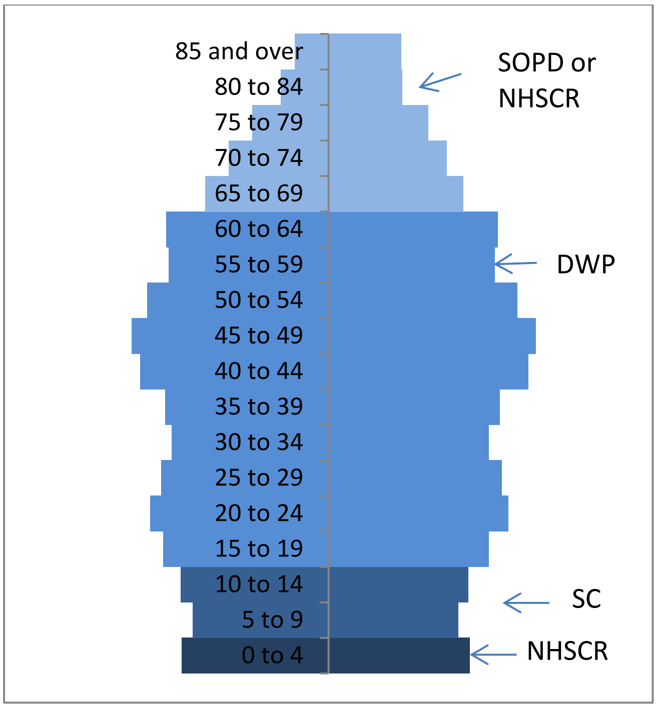

*Figure 7. Population pyramid based on 2011 Census day usual resident population, Scotland* (Source: National Records of Scotland, 2013)

### 4.2.2 Model-based approaches

Another way to combine data is to use statistical modelling. For example, the NHSCR provide information by age, sex and geography but no attribute information. Raymer et al. (2007, 2011) and Smith et al. (2010) showed how these data can be extended to include attribute information from other sources by using iterative proportional fitting and log-linear modelling. For England and Wales, they combined NHSCR data and census migration data to estimate elderly migration flows in England and Wales, NHSCR data and 1991 and 2001 Censuses to estimate ethnic migration patterns; and data from the National Health Service Patient Register Data System, Labour Force Survey and 2001 Census to estimate economic activity flows. It is possible to extend their work by combining attribute information from different administrative sources, registers or surveys to provide census type statistics.

The modelling approach assumes that one data source is incomplete, or inaccurate in certain aspects, and the other data source is providing auxiliary information. For example we can add ethnicity information to incomplete National Health Service data. A similar model to the one presented in Smith et al (2010), used to combine census, patient register and survey data for estimating internal migration in England and Wales,  can be applied to estimate cell counts by age, sex and ethnicity:

$$\log\left(\mu_{xyz}^{ASE}\right) = \lambda_x^A + \lambda_y^S + \log(m_{xyz}^{ASE}),$$

where $\mu_{xyz}^{ASE}$ represents the expected population by age, sex and ethnicity, $\lambda_x^A, \lambda_y^S$ represent the age and sex characteristics of National Health Service data, and $m_{xyz}^{ASE}$ represents the auxiliary ethnicity information in the other data source (obtained from a survey or recent census). The above formula can be extended to combine National Health Service data with one or more other data sources in order to produce detailed census type statistics. This approach could also be used to distribute age and sex across geography if an accurate estimate of the age-sex-geography association structure could be obtained from an alternative source.

### *4.3 Combining two or more administrative sources at individual level*

#### 4.3.1 Record linkage

The aim of record linkage is to bring two or more data sources which have information about the same population together by using identifiers. Record linkage is also useful for removing duplicates and assuring the general quality of administrative data sets. There are two types of record linkage. Rule-based record linkage is employed when a unique identifier which is universally available for all records in both data sources is available or a concatenated unique identifier can be generated by using a set of partial identifiers. Probabilistic record linkage is employed when a unique identifier is not available in data sources and a set of variables act as potential identifiers. Currently, ONS is conducting research on matching anonymised data with rule-based linkage using match-keys and score based (probabilistic) matching using logistic regression (ONS 2013a).

Rule based linkage is also called *exact matching*. In the absence of a unique identifier, a combination of partial identifiers such as forename, surname (or forename initial, surname initial), date of birth, sex and post code can be used to generate match-keys. ONS (2013a) constructed a series of match-keys to solve different inconsistencies in the records. According to the research on patient register, 99.55 per cent of people in the UK are assumed to have a unique match-key which consists of forename initial, surname initial, sex, date of birth and postcode district(ONS 2013a). A set of match keys is used together to increase the match rate. Consider the illustration presented in Figure 8. Here, the linked data source includes both records which have the same identifier for both data sources and records whose identifier does not match with any identifier in the other data source.

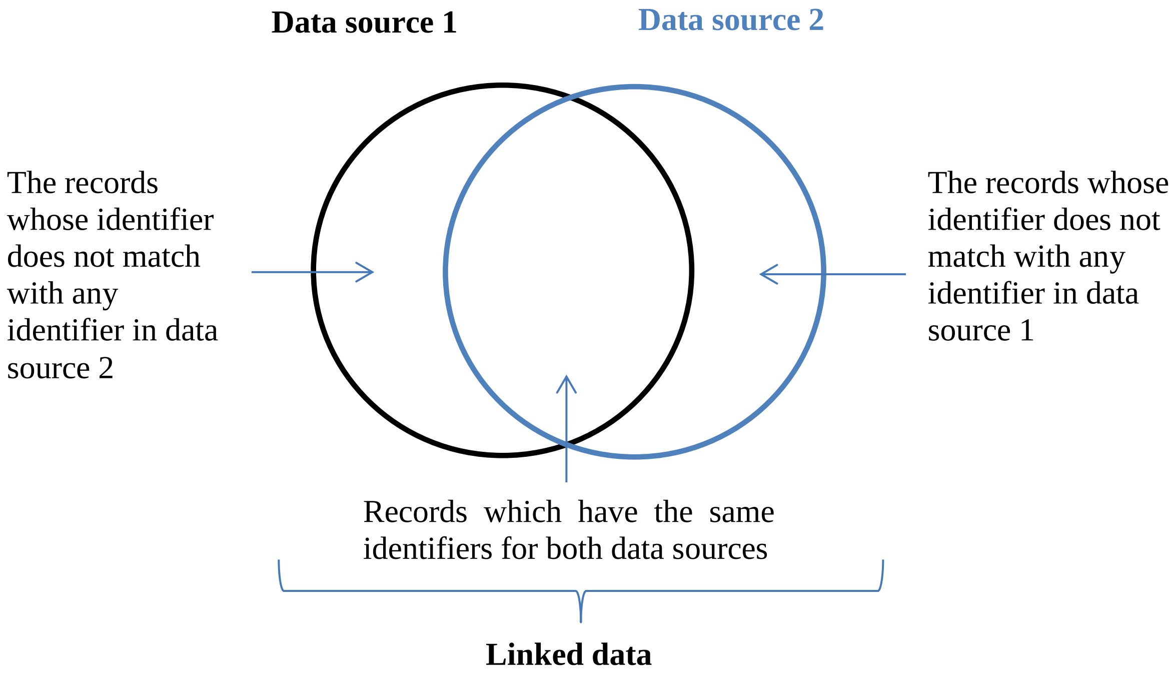
Data source 1
Data source 2
The records whose identifier does not match with any identifier in data source 2
The records whose identifier does not match with any identifier in data source 1
Records which have the same identifiers for both data sources
Linked data

*Figure 8. Combining two data sources by exact matching*

If more datasets will be combined a more sophisticated method is needed. For example, Harper and Mayhew (2012) applied a set of rules to combine locally available administrative datasets. They use truth tables which employ Boolean algebra to test whether the person in one dataset is the current resident of the address. Each address is represented by a unique property reference number. Figure 9 illustrates the possible categories when three dataset is combined, whereby

- 0 = people not in any dataset,
- 1 = people only in GP Register,
- 2 = empty properties,
- 3 = people only in the other administrative dataset,
- 4 = people in GP Register and in address register,
- 5 = people in GP Register and in other administrative dataset,
- 6 = people in address register and in other administrative dataset, and
- 7 = people in GP Register, in address register and in other administrative dataset.

Basically, a person is listed as usual resident if he or she has an assigned Unique Property Reference Number (UPRN) and is on the GP Register or at least one of the data sources. People in Categories 4, 6 and 7 are listed as current residents in their addresses which have an assigned UPRN. The remaining categories represent persons who are not confirmed as residents. This method misses people who are not included in GP Register or in any datasets. They propose to combine additional datasets to lower the proportion of missing people in category 0.

Data sources are linked by using a combination of potential identifiers such as name, date of birth and post code if a unique and universally available identifier is not available. According to Gill (2001), a probabilistic linkage score is calculated for each potential identifier based on previous knowledge or preliminary exercise with the current data. This score is measuring the probabilities of matching by chance for current pair if they were correctly linked and if they were not correctly linked. The scores for potential identifiers are represented as S1 (score for the first potential identifier), S2, S3, S4, S5 and S6 in Figure 10.

A final score is calculated by using the scores of each potential identifier. Two records are linked if the final score exceeds the pre-set threshold and they are not linked if it does not exceed the pre-set threshold. The example in Table 1 illustrates a simple linking of two different datasets.

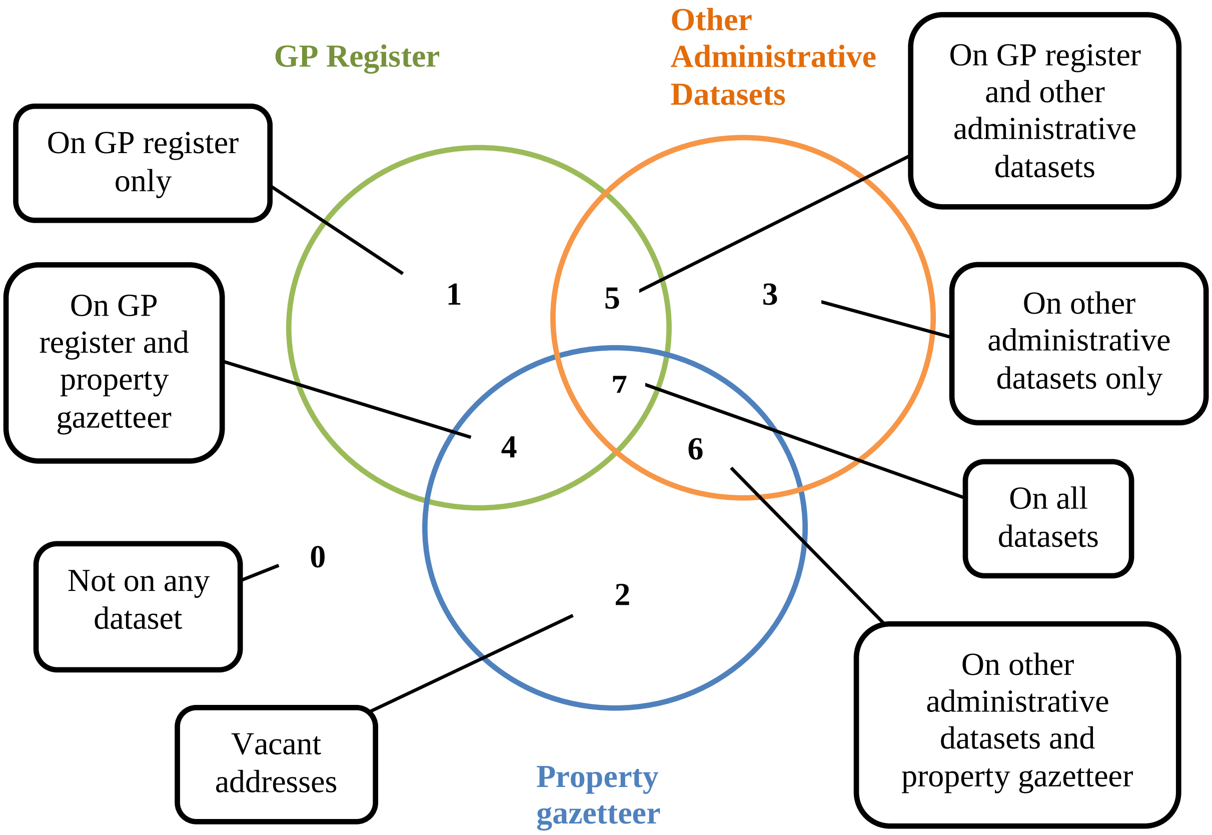
GP Register
Other Administrative Datasets
Property gazetteer
On GP register only
On GP register and other administrative datasets
On GP register and property gazetteer
On other administrative datasets only
On all datasets
Not on any dataset
Vacant addresses
On other administrative datasets and property gazetteer
0
1
2
3
4
5
6
7

*Figure 9. Simple Venn diagram partitioning different categories of administrative data with and without addresses* (Source: Figure 1 in Harper and Mayhew, 2012)

*Table 1. Probabilistic linkage for two data sources*

| | Fore-name | 2nd Fore-name | Surname | Gender | Date of birth | Post code | Final Score |
|---|---|---|---|---|---|---|---|
| Data 1 | John | W. | Thompson | Male | 20/03/1973 | SE 15 | |
| Data 2 | Jon | - | Thomson | Male | 20/03/1973 | SE 15 | |
| Scores | S1 | S2 | S3 | S4 | S5 | S6 | |

### 4.3.2 Dual system estimation

With record linkage, there is a chance that some of the population will not be included in any of the data sets. Linking as many datasets as possible decreases the probability of missing individuals. However, it may not always be possible to link many datasets. In this case, dual system estimation (or more general capture-recapture methods) may be used to estimate the number of missing individuals in the linked datasets.

The dual system estimation uses a cross-classified table of counts representing those included in both data sets and those included in only one of the data sets. This information is used to estimate the persons not included in any of the data sets (see, for example, Swanson and Tayman 2012). Basically, the dual system estimation uses the observed values in the table to estimate the missing cell by assuming the values in the cells have the same relative frequency to their column and row totals. For example, the population total, $X_{++}$,in Table 2 is estimated by solving the following equation: $X_{++} = X_{1+}\frac{X_{+1}}{X_{11}} = (120 \times 250)/100 = 300$.

Thus, $X_{22}$ = 300 – 100 – 20 – 150 = 30. Note, $X_{22}$ can also be estimated as ($X_{12}$×$X_{21}$) / $X_{11}$ = (20 × 50) / 100 = 30.

*Table 2. Dual system estimation example*

| | Yes | No | Total |
|---|---|---|---|
| Yes | $X_{11}$ =100 | $X_{12}$ =20 | $X_{1+}$ =120 |
| No | $X_{21}$ =150 | $X_{22}$ | $X_{2+}$ |
| Total | $X_{+1}$ =250 | $X_{+2}$ | $X_{++}$ |

### *4.4 Demographic accounting system*

A demographic accounting model based on administrative data could be used to estimate populations by age, sex and local areas. Unlike the methods presented above, this method relies on estimating the demographic components of change (i.e., fertility, mortality, internal migration and international migration) and embedding them within a cohort component demographic model. This approach has been used by United Kingdom statistical offices to produce post-census population estimates. The main problem with this method is the estimation of internal and international migration, which are the weakest components in terms of data quality. However, methods could be developed to improve the measurement of these components. Furthermore, coverage surveys or other administrative data could be used periodically to adjust population estimates.

## 5. Recommendations for developing an administrative data base to estimate current population totals

Having undertaken this work, we first outline what we believe are feasible, and possibly overlapping options, for estimating populations by age, sex and council areas in Scotland. We then provide some recommendations for taking this work forward.

### *5.1 Options*

As described in Section 4, there are several options for making better use of administrative data to estimate population totals by age, sex and local areas. This includes developing a single and comprehensive administrative data source or developing an estimation system that relies on two or more existing administrative data sources.

For the single administrative data base, National Records of Scotland could consider (a) creating a brand new European-style population register, (b) modifying the current most appropriate administrative data source (NHSCR) with the aim of complete coverage, (c) use current administrative data sources with adjustments based on external sources and coverage surveys by using statistical methods. According to NRS, options (a) and (b) are not feasible at this point in time. If option (c) is adopted, it seems logical that there would also be incremental improvements over time in the main administrative data source.

For combining two or more administrative data sources to make optimal use of the best features of each source, the options are to (a) combine administrative sources at the aggregate level, (b) stratify the problem and use different aggregated administrative sources for particular ages, sexes, council areas,(c) use record linkage and capture-recapture methods to estimate the whole population and (d) use record linkage with the aim of producing a comprehensive database at the individual level. Note, only option (d) provides individual level data.

### *5.2 Initial recommendations for the Beyond 2011 project*

Before any estimation work, it is important to first carry out a thorough assessment of the available administrative data sources. This needs to be done in at least two ways. First, National Records of Scotland should perform analyses along the lines of recent work by the Office for National Statistics in comparing administrative data in England and Wales with 2011 Census population totals and sex ratios.[2] Second, there needs to be work in assessing the ability of various administrative data sources to be combined and the accuracy of the combined data with regard to the 2011 Census structures. This would be particularly useful for combining two or more administrative data sources at the aggregate level and also for adding detailed information from survey or other sources. There should also be work carried out at the record level. This includes exploring various record linkage approaches, as well as options for incorporating capture-recapture methods to estimates the usual residents not found in administrative data sources.

[2] The ONS Beyond 2011 Programme will make its final recommendation in 2014 in terms of producing their future census-type statistics. They have already narrowed their options to two: (a) an online census which will be a modernised census encouraging online completion of census forms and (b) the administrative data option which will re-use already collected administrative data and will be supplemented by a four per cent annual survey (ONS 2013b).

To summarise, we believe all of the options presented in this report could be used to obtain population estimates by age, sex and local area. However, a recommendation cannot be made until the feasibility, accuracy and costs of each has been assessed. This will be clearer once the initial investigation of the data sources described above has taken place.

## References

Brown J, Abbott O and Smith PA (2011) Design of the 2001 and 2011 Census Coverage Surveys for England and Wales. *Journal of the Royal Statistical Society*(*Statistics in Society*) 174 (4):881-906.

Bycroft C (2011) Social and population statistics architecture for New Zealand. Wellington: Statistics New Zealand.

Community Health Index (2013) NHS Greater Glasgow and Clyde. Available from: http://www.nhsggc.org.uk/content/default.asp?page=home_chi[Accessed 18 February 2013].

Electoral Commission (2012) Who is eligible to vote at a UK general election?The Electoral Commission. Available from: http://www.electoralcommission.org.uk/faq/voting-and-registration/who-is-eligible-to-vote-at-a-general-election[Accessed 25 February 2013]

General Register Office for Scotland (2013) National Health Service Central Register. Edinburgh: General Register Office for Scotland. Available from: www.nrscotland.gov.uk/statistics-and-data/nhs-central-register/about-the-register [Accessed 11 February 2013].

Gill L (2001) Methods for automatic record matching and linkage and their use in national statistics, The National Statistics Methodology Series, Office for National Statistics. Available at http://www.statistics.gov.uk/downloads/theme_other/GSSMethodology_No_25_v2.pdf[Accessed March 2013].

Harper G and Mayhew L (2012) Using administrative data to count local populations.*Applied Spatial Analysis and Policy* 5(2):97-122.

HMRC (2013) New arrivals to the UK and Child Benefit. Her Majesty's Revenue and Customs. Available from: http://www.hmrc.gov.uk/childbenefit/start/who-qualifies/new-arrivals-uk.htm [Accessed 25 February 2013]

National Records of Scotland (2012) Assessing administrative data: A comparison of population counts from aggregated administrative data and the mid-year population estimates. Edinburgh: National Records of Scotland.

National Records of Scotland, 2011 Population Pyramid available http://www.scotlandscensus.gov.uk/en/censusresults/visualisations/2011pyramid.html [Accessed 19/03/2013]

Office for National Statistics (2012a) Beyond 2011: Administrative data sources report: NHS Patient Register. Titchfield: Office for National Statistics.

Office for National Statistics (2012b)The 2011 Census coverage assessment and adjustment process, 2011 Census: Methods and Quality Report. Titchfield: Office for National Statistics.

Office for National Statistics (2013a) Beyond 2011: Matching anonymous data. Titchfield: Office for National Statistics.

Office for National Statistics (2013b) Beyond 2011: Newsletter- July 2013. Titchfield: Office for National Statistics.

Raymer J, Abel GJ and Smith PWF (2007) Combining census and registration data to estimate detailed elderly migration flows in England and Wales. *Journal of the Royal Statistical Society Series A*(*Statistics in Society*) 170(4):891-908.

Raymer J, Rees P and Blake A (2012) A conceptual framework for UK population and migration statistics. Titchfield: Office for National Statistics. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/imps/latest-news/conceptual-framework/index.html

Raymer J and Smith PWF (2010) Editorial: Modelling migration flows. *Journal of the Royal Statistical Society Series A* (*Statistics in Society*) 176(4):703-705.

Raymer J, Smith PWF and Guilietti C (2011) Combining census and registration data to analyse ethnic migration patterns in England from 1991 to 2007.*Population*, *Space and Place* 17:73-88.

Smith PWF, Raymer J and Guilietti C (2010) Combining available migration data in England to study economic activity flows over time. *Journal of the Royal Statistical Society Series A* (*Statistics in Society*) 173(4):733-753.

Statistics Finland (2004) Use of registers and administrative data sources for statistical purposes. Handbook, Statistics Finland.

Swanson DA and Tayman J (2012) *Subnational population estimates*. Dordrecht: Springer.