
# Assessing Administrative Data: A Comparison of Population Counts from Aggregated Administrative Data and the Mid-Year Population Estimates

**Published on 13 July 2012**

**Contents**

**List of Figures**

# 1. Introduction

National Records of Scotland (NRS) is investigating new approaches to producing census-type statistics without the physical requirement for traditional census enumeration and the high cost this involves. Data from administrative sources is one of the key inputs that is being considered. This paper presents some results from the work undertaken to assess the suitability of aggregated administrative data to contribute to such approaches. It provides an overview of the key sources of administrative data available to NRS and reports comparisons between the population counts derived from them and the Mid-Year Estimates (MYEs – available in the Population Estimates section of the NRS website). Table 1 lists these data sets and the corresponding MYE against which a comparison was undertaken.

**Table 1: Overview of the data sets featured in this report**

| Administrative data set | Comparator | Reference year |
|---|---|---|
| 1 NHSCR Demographic extract | MYEs | 2010 |
| 2 Community Health Index | MYEs | 2010 |
| 3 DWP Customer Information System | MYEs | 2010 |
| 4 School Census 2009/10 | MYEs | 2010 |
| 5 Child benefit | MYEs | 2009 |
| 6 Super Older Persons' Database | MYEs | 2010 |
| 7 Electoral Register | MYEs | 2010 |

As all administrative systems, the data sources used to extract population counts for this exercise are set up and maintained in order to support specific administrative processes. They are not intended to measure population stocks and cannot be expected to fully reflect population definitions used for Census and MYEs. From the perspective of the population estimation process departures in the definition of the target population constitute errors of representation (Bakker 2010). There may also be other, measurement errors arising from the way target populations are captured and population counts are derived. That is, the differences in population counts observed in the comparisons reported here may reflect any of several problems including coverage, data collection and processing issues. This report does not engage in details with such possible 'errors' in using administrative data for population estimation and no attempt is made to make adjustments for them (as for example in Office for National Statistics, 2012). In some cases, NRS has had little input in specifying the rules for deriving population counts, and while there may be scope for some marginal improvements in processing rules, this is a subject for future work.

This report should not therefore be read as an assessment of the quality of administrative data or the accuracy of population estimates. Its objective is to provide a first look at the broad patterns of agreement or disagreement between sources of data that are relevant to the estimation of population stocks in order to inform further research. It focuses on comparisons by gender and age at the level of council areas.

The paper is organised as follows. There is a section devoted to each administrative source, giving some background and reporting results from the comparisons on the dimensions described above.

Differences are expressed as simple percentage difference:

$D_i = (A_i - P_i)/P_i \times 100$

where A is the administrative data count, and P is the population estimate for the ith age and sex group. This is followed by some longitudinal data showing trends over the last ten years, where such information is available. The paper concludes with a section summarising key observations about the patterns which are displayed by the data.

# 2. NHS Central Register (NHSCR) Demographic Extract

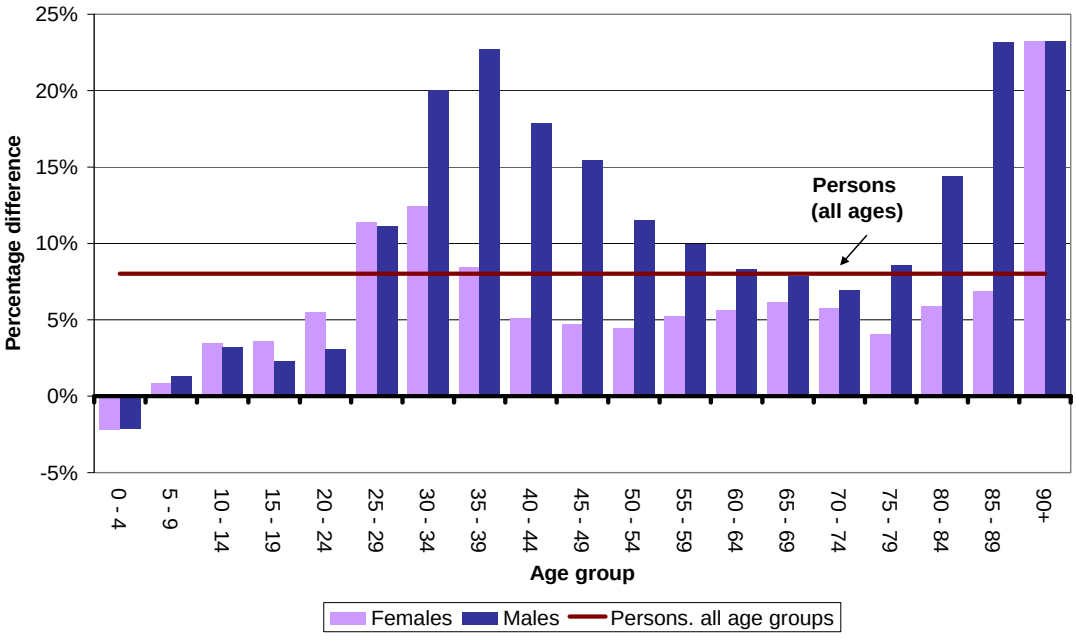
## 2.1 Background

The NHSCR is a database of people born in Scotland or registered with general practitioners (GPs) in Scotland. An edited extract is supplied to the National Records of Scotland (NRS) at specified intervals for the purpose of producing demographic statistics. The extract includes information on gender, dates of birth, health board registration and post code of residence. Information on council area location can be derived from the post code. When people die or leave Scotland their records are not removed from the source register but the fact of death or migration is recorded. However information on such events does not always reach the register (for example when deaths occur abroad or emigrants do not inform their GP) and this is thought to be the main explanation for the excess of records found on the NHSCR in comparison to Census or the Mid-Year Estimates (MYEs).

## 2.2 Gender and age

Figure 1a shows the excess of NHSCR records expressed as a percentage of the 2010 MYEs for males and females by five-year age group. The average difference across all age groups is 5.7 per cent for females, 10.3 per cent for males and 8.0 per cent for all people, but there is considerable variation across age with NHSCR population counts for some groups exceeding the MYEs by up to 20 percentage points. The high relative differences observed at very old ages are largely driven by small population sizes in these age groups.

**Figure 1a: Comparison between NHSCR and Mid-Year Estimates by Age and Gender, 2010**

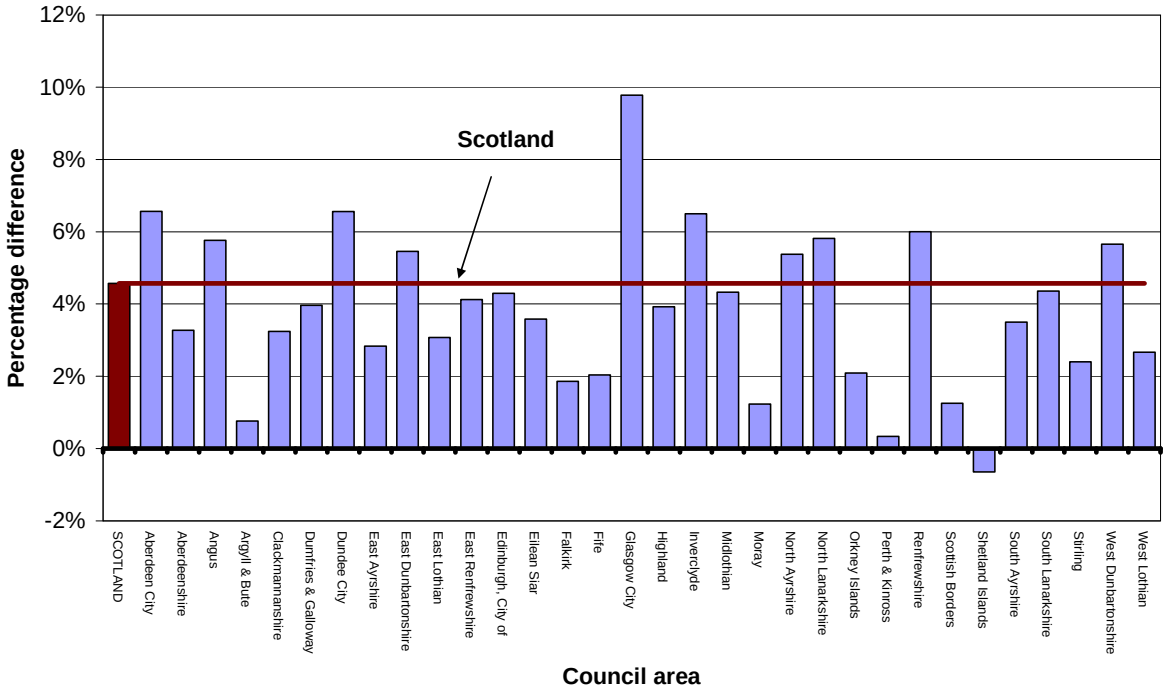



**Note**: Values for age group 90+ have been trimmed to the value of the next highest category. Because of small population size in this age group measures based on per cent difference tend to become inflated and are not particularly informative

## 2.3 Council areas

Figure 1b gives the percentage difference between the NHSCR-based population counts and the MYEs for males and females combined for each of Scotland's 32 council areas. The overall figure is 2.6 per cent for females, 6.7 per cent for males and 4.6 per cent for all people. This figure of 4.6 is smaller than the 8 percentage point difference reported in Figure 1a since in some cases it was not possible to identify the council area of residence because of missing or invalid postcode information and this artificially depressed the population count derived from the NHSCR. Data for the very old age groups are affected disproportionately: the difference between those aged 90 and over in the NHSCR database and the MYEs has dropped from an excess of 72 per cent if Figure 1a to 2.3 per cent. For those aged between 60 and 90, differences between the NHSCR-based population counts and the MYEs are now all within 2 percentage points.

**Figure 1b: Comparison between NHSCR and Mid-Year Estimates by Council Area, 2010**



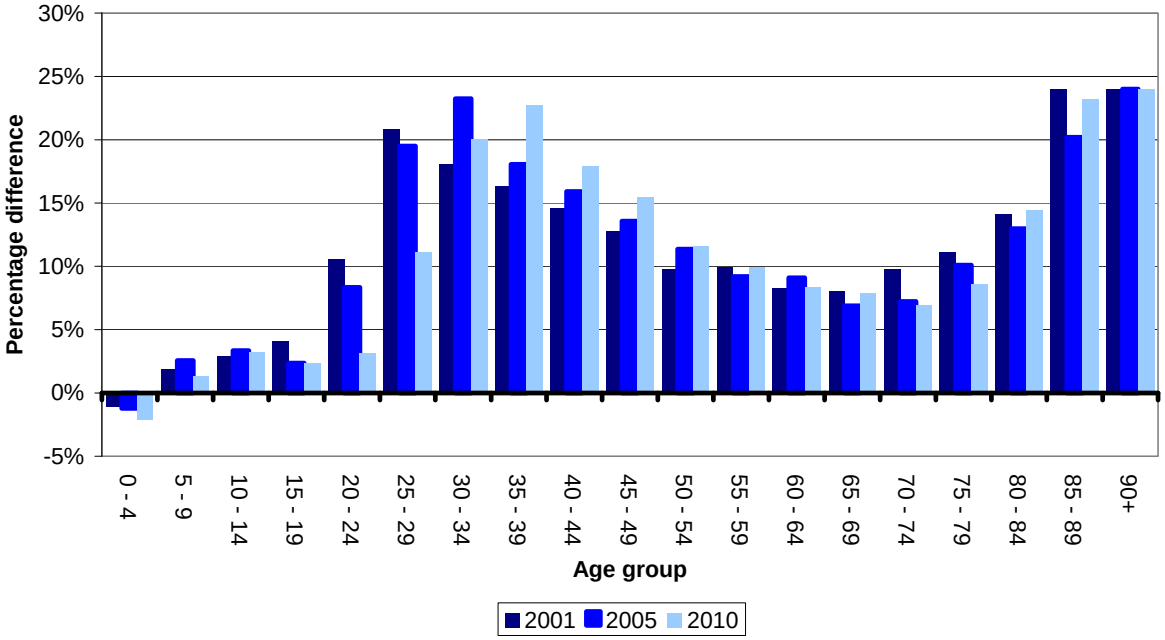

There are several council areas where the difference between the population count derived from the NHSCR extract and the 2010 MYEs is reasonably small (within 2 percentage points). In many more however this difference is substantially higher, with that for Glasgow City being particularly marked.

## 2.4 Longitudinal trends

The NHSCR contains historical information which allows the creation, from current extracts, of population counts relating to earlier years. As historical information on postcode of residence is only available since 2008, these population counts for earlier years can only be derived at Scotland level. Figures 1c and 1d give per cent difference for males and females by five-year age group in 2001, 2005 and 2010 as compared to the Census and the MYEs in those years. Although age patterns remain broadly similar over this period, there is some evidence of a cohort effect for young adults, especially among men. The highest relative differences in 2001 were observed among men in their late 20s. In 2010 this cohort, now aged 35-39, is

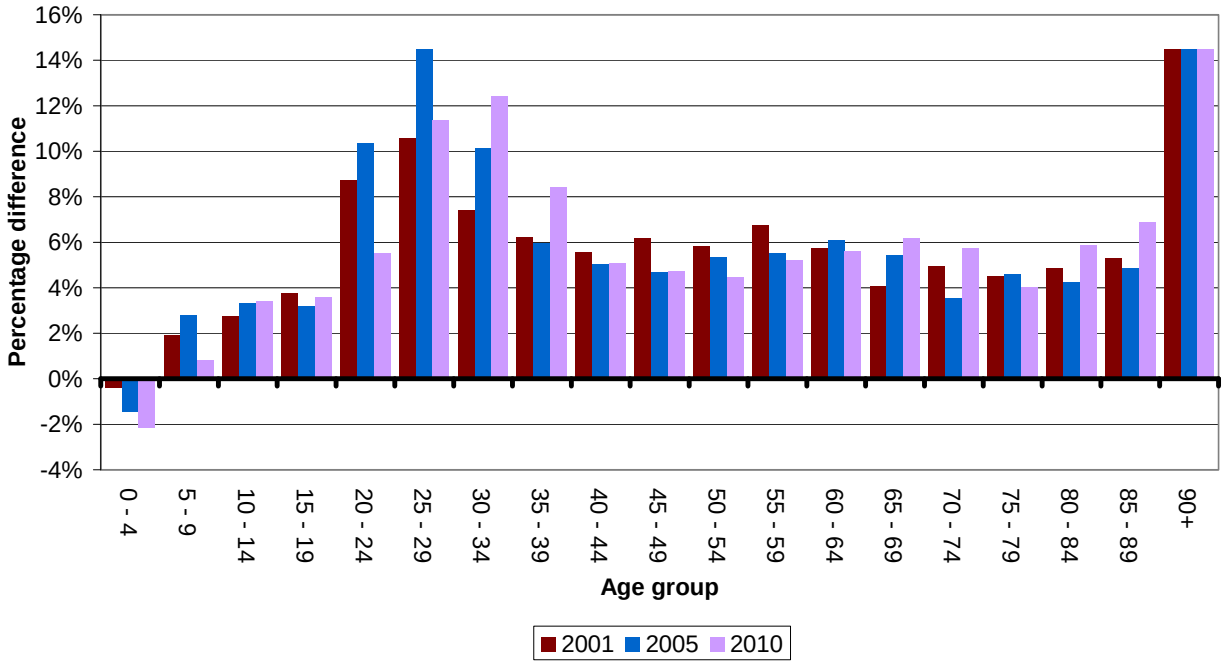
again responsible for the highest percentage difference between the NHSCR-based count and the MYEs. This pattern points to potential methodological issues that need further investigation when 2011 census data become available.

**Figure 1c: Comparison between NHSCR and Mid-Year Estimates for Men by Age**




**Note**: Values for age group 90+ have been trimmed to the value of the next highest category.

**Figure 1d: Comparison between NHSCR and Mid-Year Estimates for Women by Age**




**Note**: Values for age group 90+ have been trimmed to the value of the next highest category

# 3. Community Health Index (CHI)

## 3.1 Background

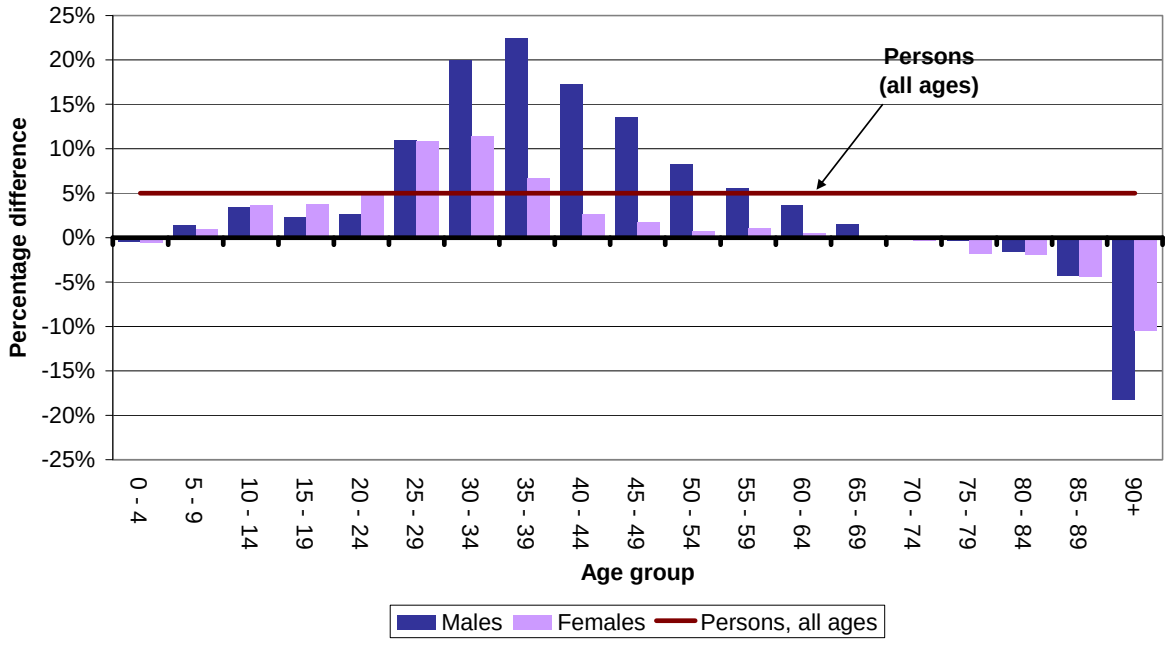
The Community Health Index (CHI) is a database intended to provide a common identifier for health care provision purposes and covers everyone registered with a GP in Scotland or in receipt of screening or other health services. The CHI system uses the NHS Central Register (NHSCR) to validate GP registrations and maintain the quality of the index and while the two administrative systems largely coincide in terms of the population they cover, there may also be some differences. For example CHI contains records for persons receiving screening services even where they are not registered with GPs, while the NHSCR does not. The CHI and the NHSCR systems are continuously being synchronised and differences in coverage are most likely to be short term and affecting predominantly migrant populations.

The analysis reported in this paper used two sets of population counts obtained from the CHI which are not entirely consistent. The first one was derived from an extract taken in 2010 and covers all persons registered on CHI at that time. It is compared to the Mid-Year Estimates (MYEs) for the same year. The second is based on an extract of the CHI taken in 2008 which uses historical information covering the preceding 10 years to derive population counts relating to 2001. These counts were compared to 2001 census figures. They benefit from retrospective corrections to CHI records based on information about health service use and place of residence that has become available after 2001.

## 3.2 Gender and age

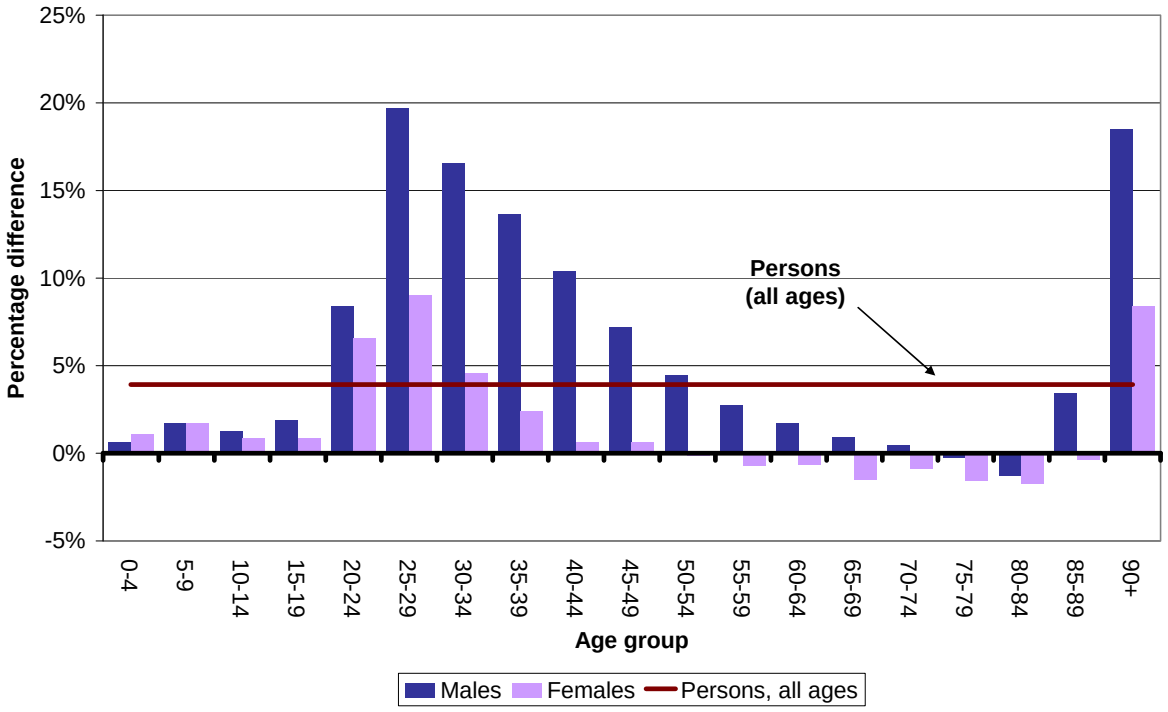
Figure 2a shows the difference between CHI-based population counts and the 2010 MYEs, expressed as a percentage of the 2010 MYEs, for men and women by five-year age groups. The excess records on the CHI across all age groups is 7.4 per cent for men, 2.7 per cent for women and 5.0 per cent for all people.

**Figure 2a: Comparison between CHI and Mid-Year Estimates by Gender and Age, 2010**



The equivalent figures for 2001 in comparison to the 2001 Census are shown in Figure 2b. The percentage difference across all age groups is 1.6 per cent for women, 6.4 per cent for men and 3.9 per cent for all people. The age profile remains broadly similar over this period with some indication of a cohort effect similar to that observed for the NHSCR. The highest relative difference in population counts among men is observed for those in their late 20s in 2001 and is still present in 2010 for the same cohort. Figures for the older age groups are less stable across years, possibly because of the smaller population size which makes results for these groups more vulnerable to processing inconsistencies.

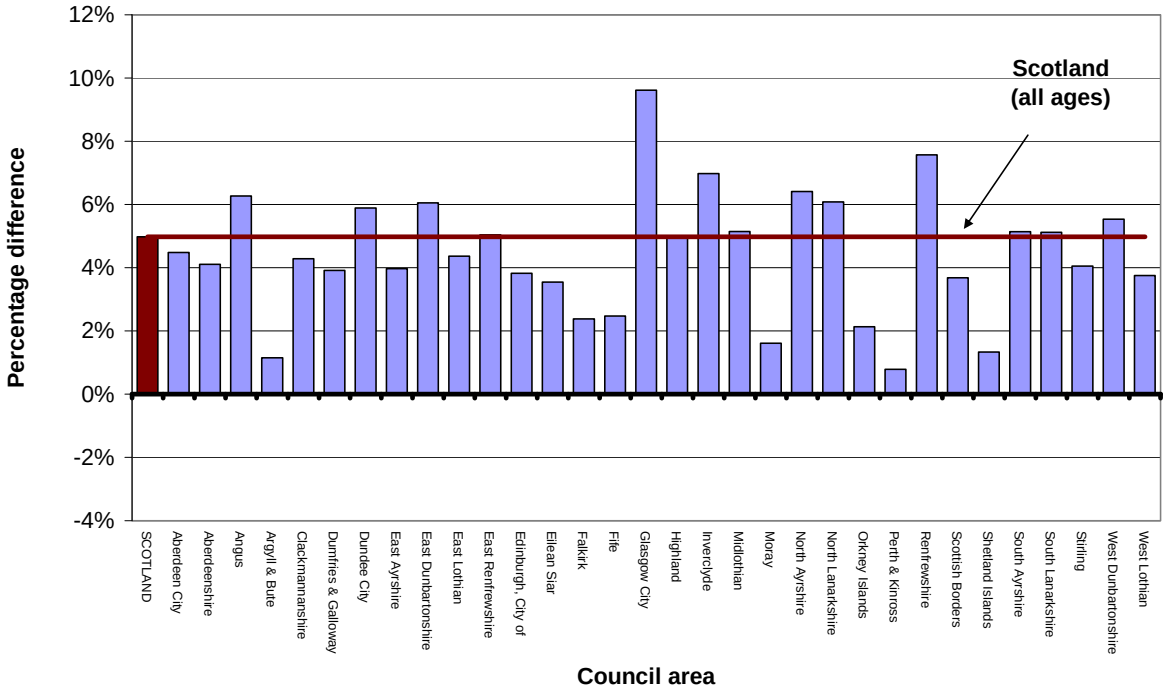
**Figure 2b: Comparison between CHI and Census by Gender and Age, 2001**



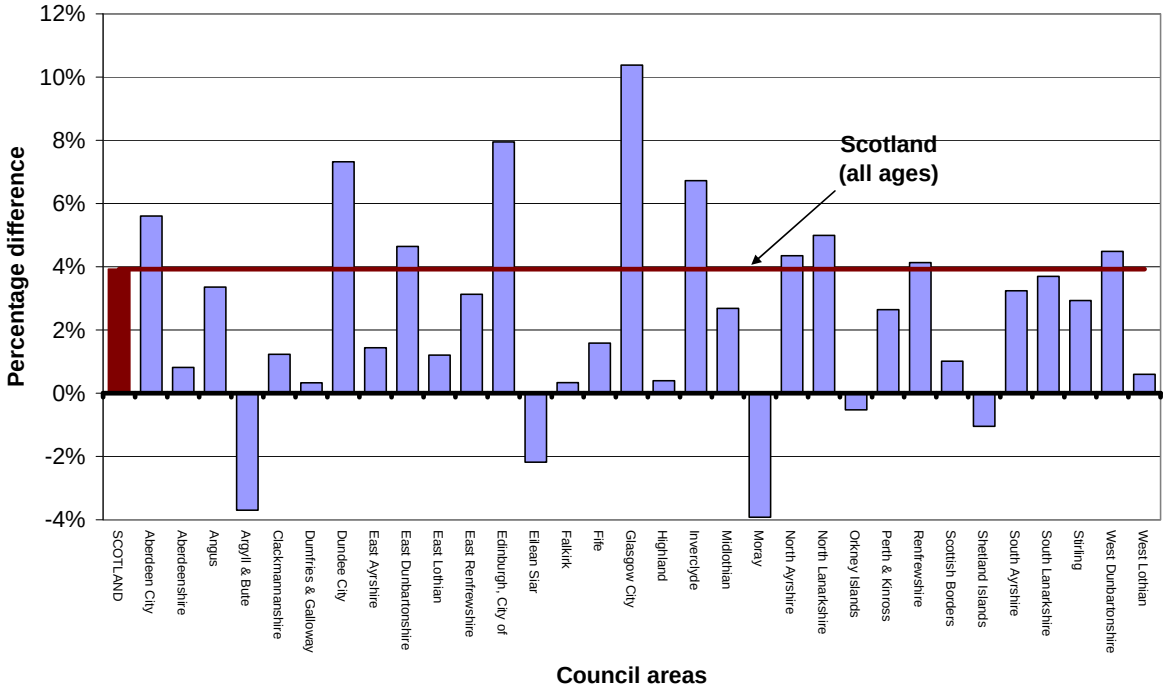

## 3.3 Council areas

Figures relating to 2010 are reported in Figure 2c, and those for 2001 are shown in Figure 2d. The overall pattern of the relationship of CHI to the comparator data is similar to that of the NHSCR demographic extract. Population counts for Glasgow City stand out with the highest levels of excess records both proportionally and in absolute terms. The 2001 comparison for CHI appears closer to the respective population estimate than the 2010 figures, which can have a number of explanations, not least the potentially better quality of population counts that can be achieved using retrospective information.

## Figure 2c: Comparison between CHI and MYEs by Council Area, 2010




## Figure 2d: Comparison between CHI and Census by Council Area, 2001



# 4. Department for Work and Pensions Customer Information System

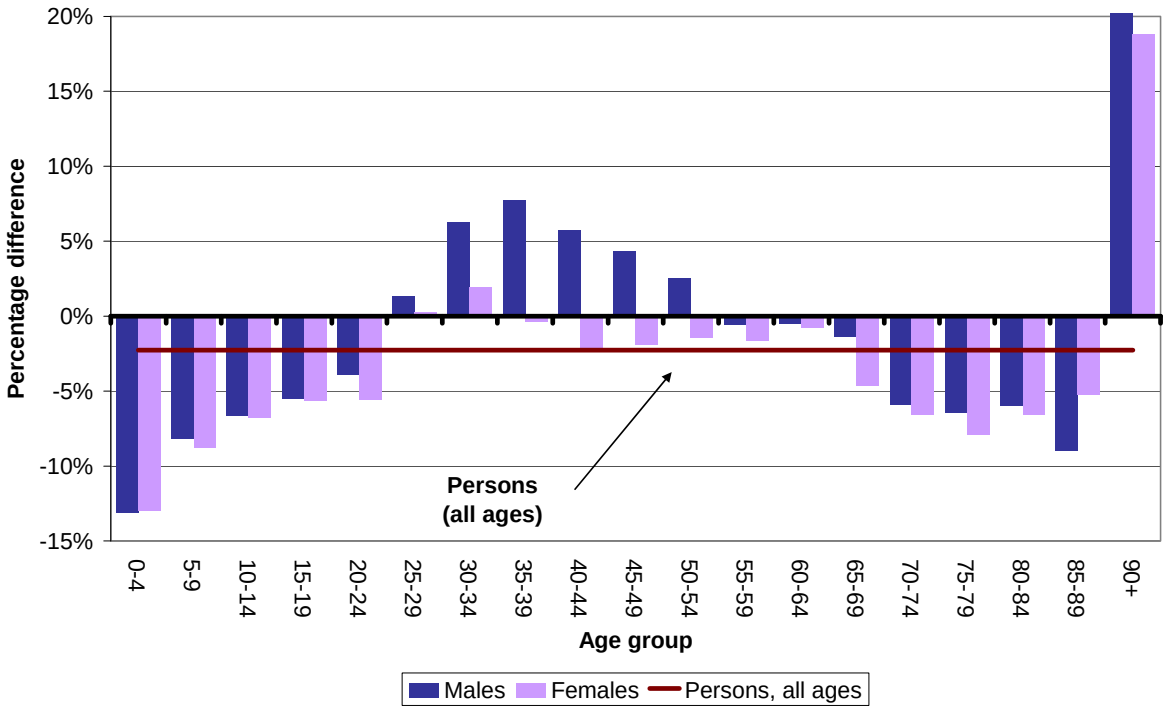
## 4.1 Background

The Customer Information System (CIS) is a central repository of personal details for the Department for Work and Pensions (DWP) and parts of Her Majesty's Revenue and Customs (HMRC). It covers all individuals who have been a client or customer of the DWP since 6 April 1999 in Great Britain (GB) and all individuals who have been a client or customer of HMRC since 2005 in the United Kingdom (UK). An extract containing population counts by age, gender and postcode sector relating to 2010 was used for the comparisons reported here. This extract was subject to statistical disclosure control prior to release by DWP.

## 4.2 Gender and age

Figure 3a shows the difference between the CIS and the Mid-Year Estimates (MYEs), expressed as a percentage of the 2010 MYEs, in the number of males and females by five-year age groups. The CIS appears to hold more records for the population in working age and fewer records for those in the younger or older age groups in comparison to the MYEs. On average the CIS population count across all age groups is lower by 3.6 per cent for females, 1.0 per cent for males and 2.3 per cent for all people. There appear to be distinct differences by gender among the working age population with male counts on the CIS consistently exceeding the respective MYE, while those for women are remarkably close to the MYE figure.

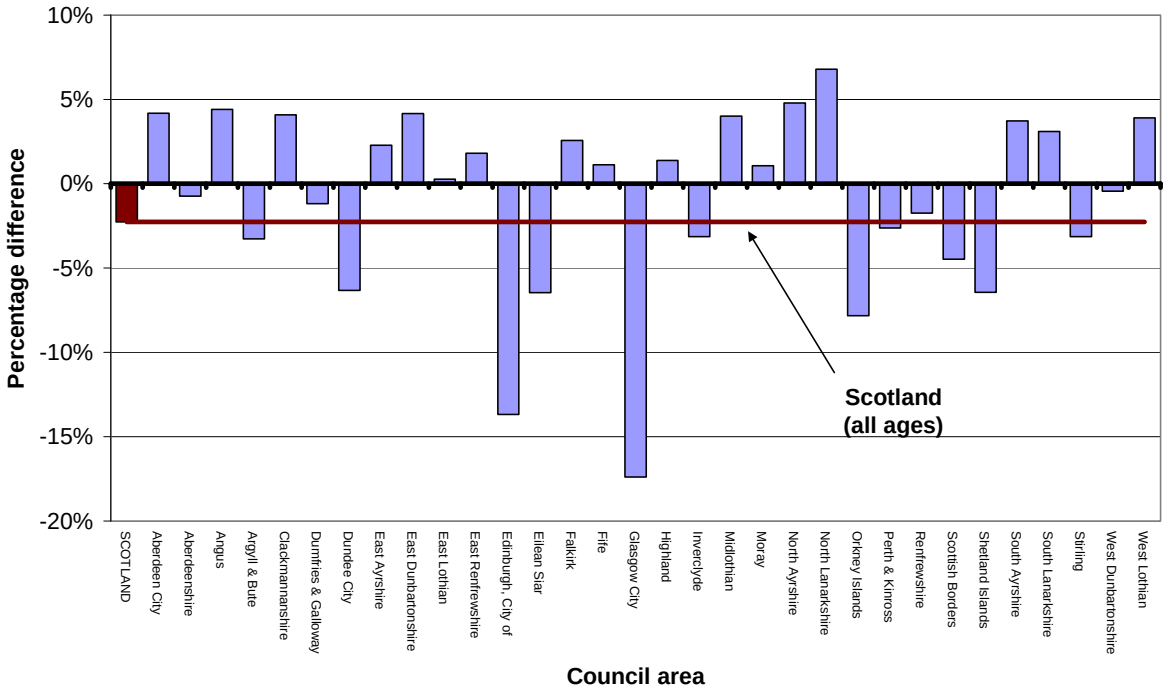
**Figure 3a: Comparison between CIS and MYEs by Age and Gender, 2010**



## 4.3 Council areas

Figure 3b gives the average CIS percentage difference for each council area in Scotland. There is considerable variation across council areas. Glasgow and Edinburgh City stand out with the highest levels of CIS undercount compared to the population estimate, which is in stark contrast to the results from comparing health service-based data with the MYEs.

**Figure 3b: Comparison between CIS and MYEs by Council Area, 2010**



# 5. School Census

## 5.1 Background

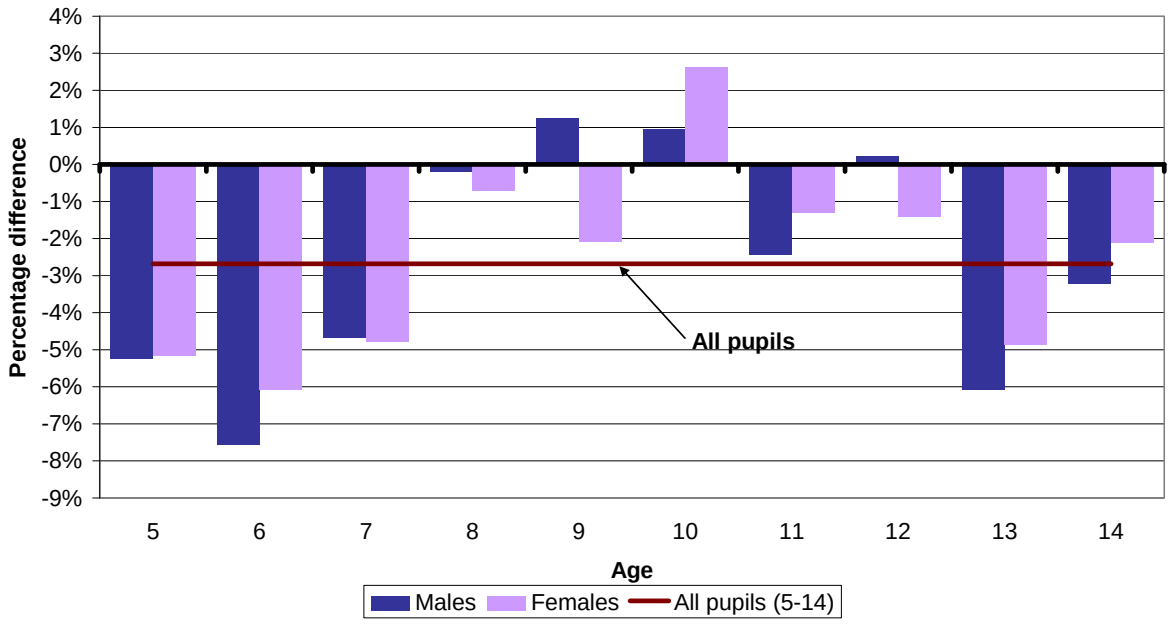
Data on pupils in the publicly funded school sector are collected in September each year through school management systems. This data collection is well established, comprehensive and allows longitudinal analysis which can contribute to the study of migration. Because of some flexibility in school entry and school leaving age, the underlying children population is best captured between the ages of 5 and 14. This is the age group we use in the comparison reported here. Data for the school year 2009/10 are compared to the Mid-Year Estimates (MYEs) for 2010.

The information collected in the School Census does not include pupils in the independent school sector or pupils who are educated at home. Whilst at the national level around 96 per cent of children are in the publicly funded school system, for some council areas this proportion falls as low as 82 per cent.

## 5.2 Gender and age

Figure 4a compares the 2010 male and female population counts at each single year of age from 5 to 14 between the 2009/10 School Census and the corresponding MYE. Counts from the School Census are on average lower than the MYEs by 2.6 per cent for females, 2.8 per cent for males and 2.7 per cent for all children.

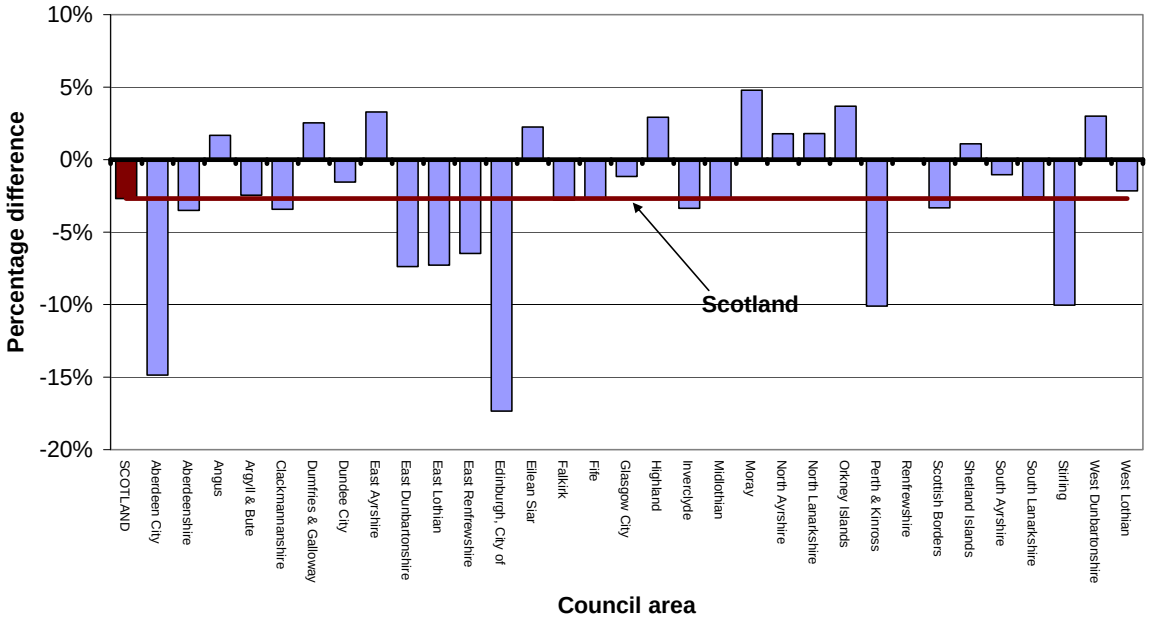
**Figure 4a: Comparison between School Census and MYEs by Age and Gender, 2010**



## 5.3 Council area

Figure 4b breaks down this difference by council area for all ages combined. The figures show that while the differences are broadly in line with the distribution of independent sector education, there are a few areas where school census counts exceed the size of the MYE population.

**Figure 4b: Comparison between School Census and MYEs by Council Area for children aged 5-14, 2010**



# 6. Child Benefit Data

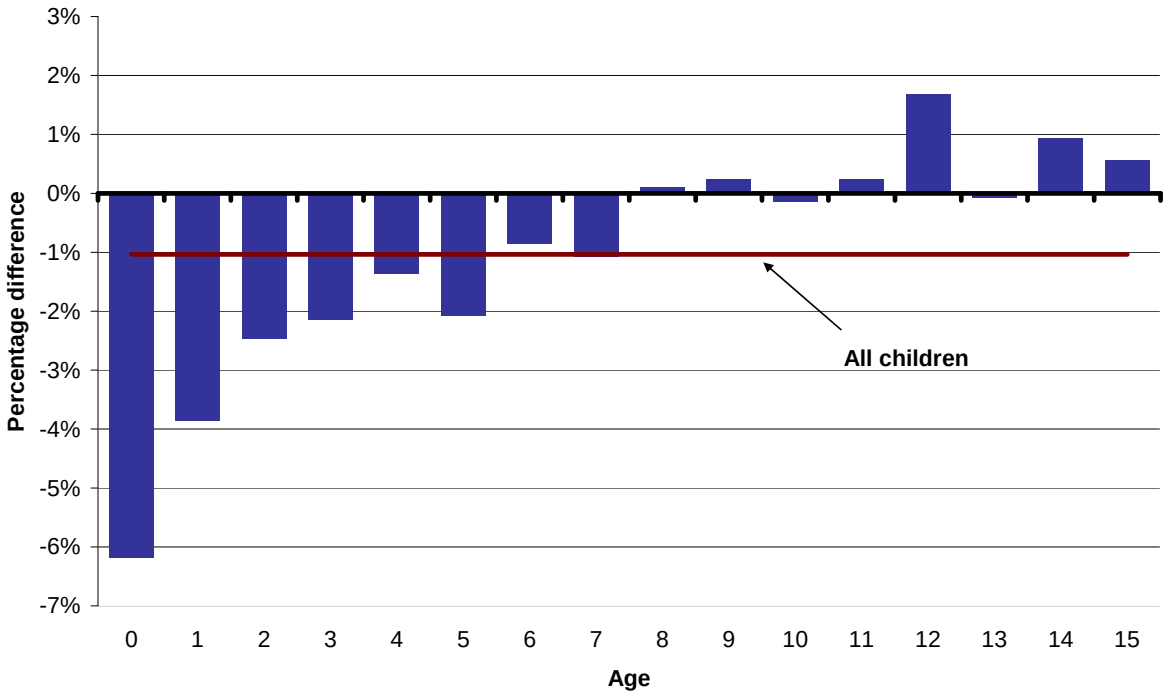
## 6.1 Background

Child benefit is currently a universal benefit administered by Her Majesty's Revenue and Customs (HMRC) (and Department for Work and Pensions (DWP) prior to 2003). The data set used for this assessment is extracted from the management systems used by HMRC to administer the benefit. It contains information on number of claimants by age of child and data zone of residence and has been subject of statistical disclosure control to protect the confidentiality of the data. Child benefit is paid to the carers of children, usually into bank, building society or post office accounts, and although people are expected to inform HMRC when they move home, address information may not always be up to date. It is estimated that child benefit is claimed for about 98 per cent of eligible children. From January 2013, child benefit will become subject to means testing which will reduce the value of the data in estimating the size of the child population.

## 6.2 Gender and age

No gender information was available from the child benefit data for this comparison. Figure 5a shows the total population at each single year of age from 0 to 15 estimated from the child benefit data in comparison to the corresponding Mid-Year Estimates (MYE). On average the child benefit population count is 1 percentage point below the MYE for this age group. However, there seems to be a distinct age-related pattern, with child benefit data recording fewer younger children and more older ones relative to the MYEs.

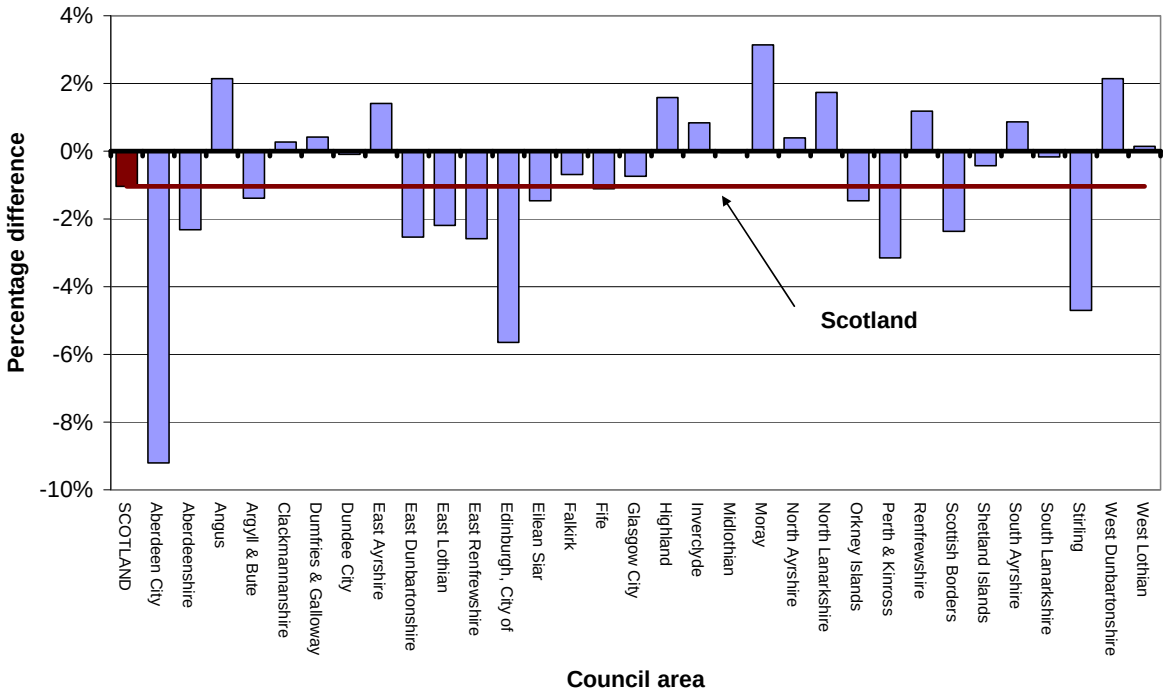
**Figure 5a: Comparison between Child Benefit data and MYEs by Age, 2009**



## 6.3 Council areas

Figure 5b gives the average percentage difference for all children aged 0-15 for each council area. For most council areas this difference is within 3 percentage points of the MYE. There are a few, such as Aberdeen and Edinburgh City, where child benefit counts are substantially lower than the MYE.

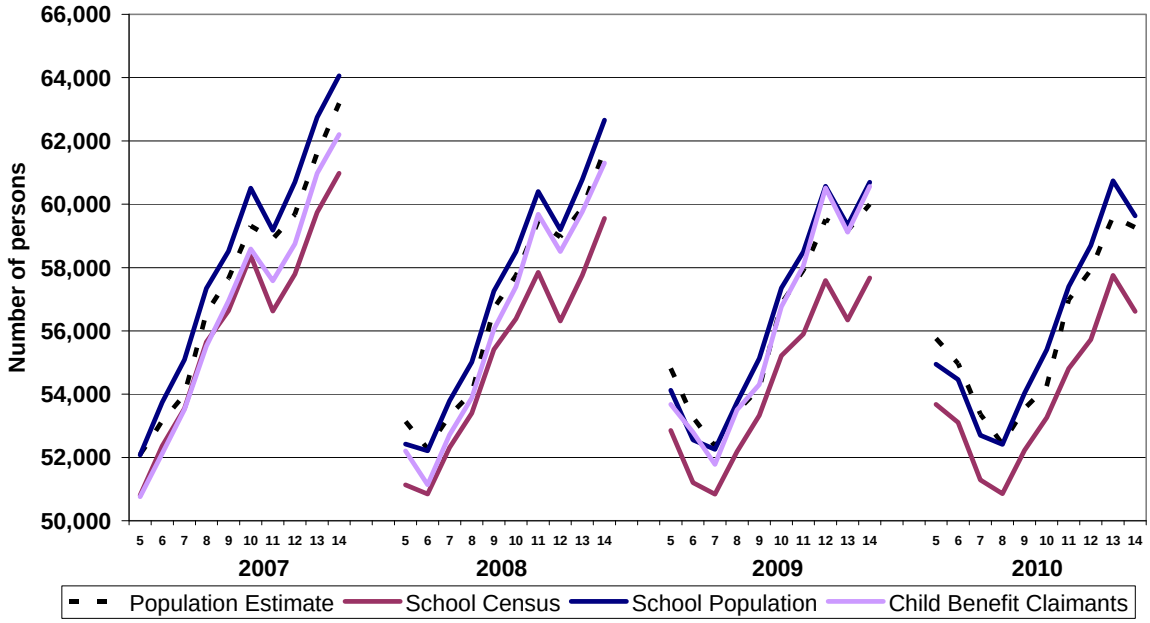
**Figure 5b: Comparison between Child Benefit Data and MYEs by Council Area for children aged 0-15, 2009**




## 6.4 Longitudinal trends

Figure 5c shows how different administrative sources represent the population of children aged 5 – 14 for each year between 2007 and 2010. The category 'school population' is formed by adding aggregate figures for children attending independent schools in Scotland to the count derived from the School Census. For children who have entered the school education system prior to 2007, the overall school population count exceeds the respective MYE, while the opposite is true for younger children.

**Figure 5c: Comparison of MYEs and alternative data sources for children aged 5-14**



# 7. Super Old Persons Database

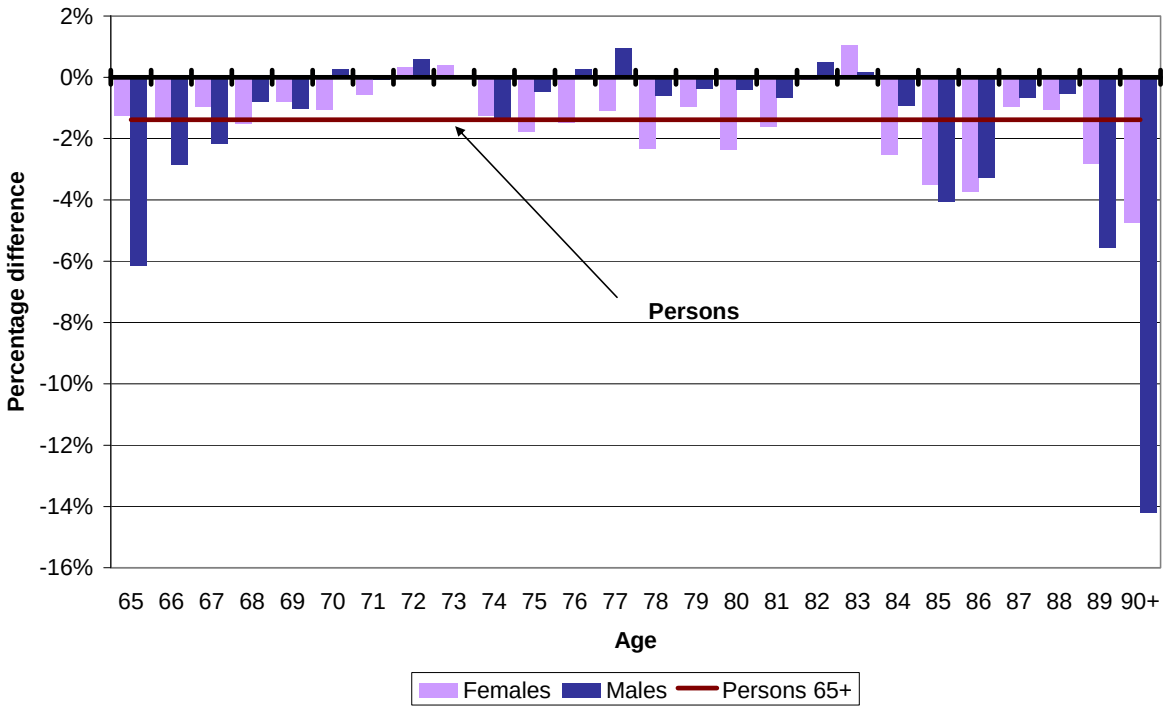
## 7.1 Background

The Super Old Persons Database (SOPD) is derived from individual Department for Work and Pensions (DWP) databases for state pension and other pension age benefits, such as pension credit and attendance allowance, and covers persons aged 65 years and over. Studies have generally found good agreement between these data and the mid-year population estimates (for example Office for National Statistics, 2009).

## 7.2 Gender and age

Figure 6a shows the percentage difference between SOPD and the Mid-Year Estimates (MYEs) in the number of males and females for each single year of age between 66 and 89 and for all those aged 90+ taken together. Across age on average there is just over 1 percentage point undercount on the SOPD compared to the MYEs, and this is similar for men as well as women. Except at the very old ages, most differences are reasonably small.

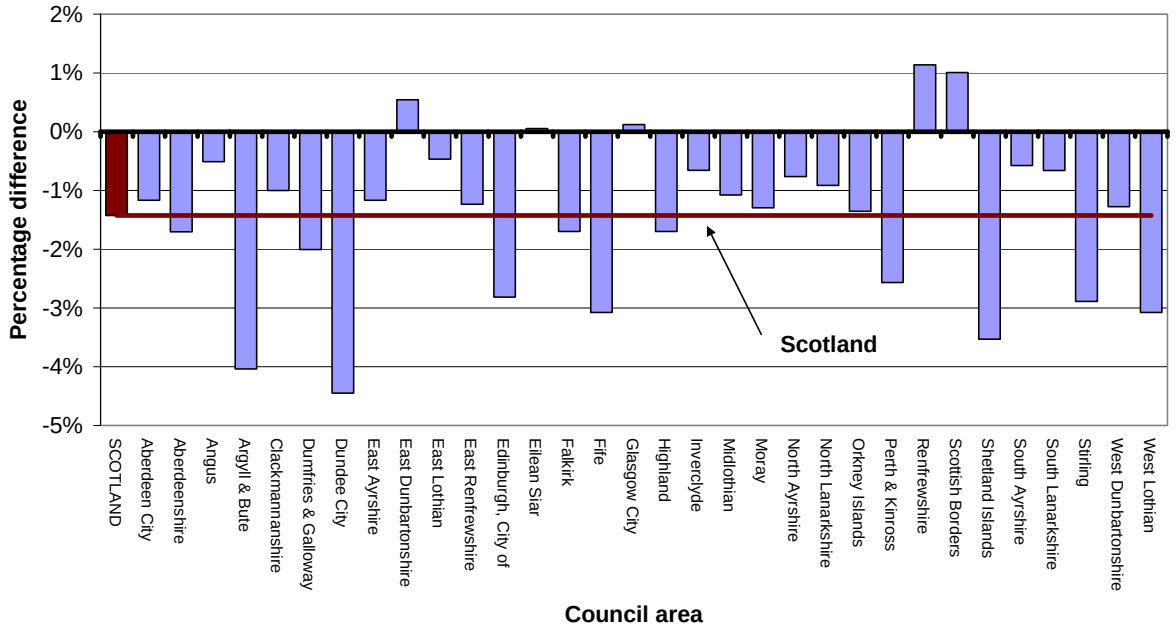
**Figure 6a: Comparison between SOPD and MYEs by Gender and Age, 2010**



## 7.3 Council Areas

Figure 6b shows how the difference between the SOPD and the MYEs breaks down by council area. For most council areas the difference falls within 2 percentage points of the MYE. There are a few unusual outliers which in other comparisons between the MYEs and administrative data sources have tended to conform to the general pattern.

**Figure 6b: Comparison between SOPD and MYEs by Council Area, 2010**



# 8. Electoral Register

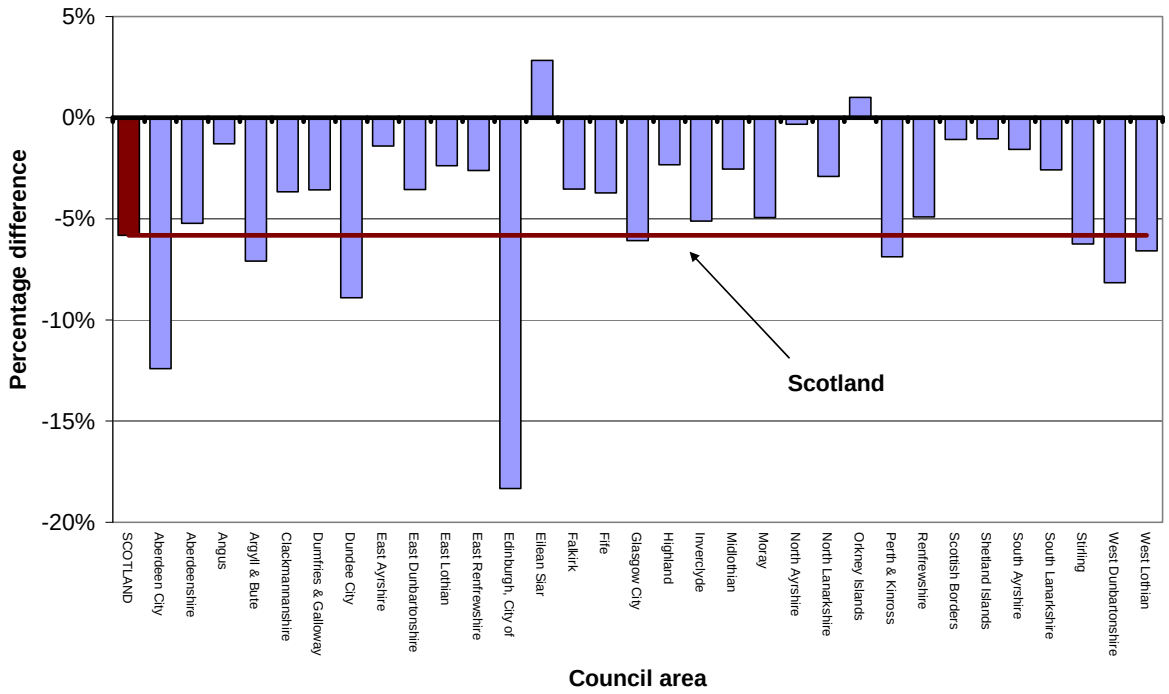
## 8.1 Background

The Electoral Register is maintained by local Electoral Registration Offices. Eligibility for registration is complex and does not fully conform to the definition of resident population used by the Mid Year Estimates (MYEs). Moreover, it is known that not all people eligible to vote register and some categories such as students, those who rent in the private sector, migrants and minority ethnic groups tend to be under-represented on the Electoral Register. It is also known that some duplication of records may exist, as certain types of voters are allowed to register at more than one address. The data used for this comparison was extracted from the Register as published on 1st December 2010. The Register does not hold information on the age or gender of registered voters (other than that they are 18 or older). The comparison is therefore limited to the population aged 18 and over.

## 8.2 Council areas

Figure 7a gives the percentage difference between the Electoral Register and the MYEs for each council area. On average across all council areas the Electoral Register count is lower than the MYE by 6 percentage points, but there is substantial variation around this figure. In Edinburgh City and Aberdeen differences are most marked and the Electoral Register count appears least well suited to represent the size of the resident population. There are a few areas where the number of records on the Electoral Register exceeds the size of the population as estimated by the MYEs.

**Figure 7a: Comparison between the Electoral Register and the MYEs by Council Area for People Aged 18 or Older, 2010**



## 9. Conclusion

From this broad overview it would appear that health service-based data tend to overestimate the size of the resident population in comparison to our usual population statistics. Counts derived from employment and benefit records display closer agreement with population estimates on average, but there are substantial variations which are not well understood. Discrepancies are most manifest with respect to the working age population, while alternative counts of children and people of pensionable age demonstrate closer agreement. From these results it is evident that the population of the bigger cities in Scotland would be harder to estimate from administrative data alone, as different sources sometimes provide conflicting information. Further research, not reported here, shows that moving to lower levels of geography introduces more variation as the significance of up-to-date address information increases.

It is of course important to remember that administrative systems are designed for specific purposes and target populations that are not as a rule consistent with the definition of resident population as measured by the census or the inter-censal population estimates. There are variations in the quality and timeliness of the individual items of data they hold depending on the purpose for which they operate: for example, it may not be necessary to update address information in a timely manner or remove records for people who no longer belong to the target population. While it is important to understand how administrative systems operate and what populations they target, this may not be sufficient to support the use of aggregated administrative data in population estimation. It is crucial to assess what population coverage administrative systems ultimately achieve, and methods such as individual level record linkage offer a way to do that. This is one line of inquiry that National Records of scotland will be pursuing at the next stage of its research into developing alternatives to the traditional census.

Our findings are consistent with research in the rest of the UK and other developed countries which do not operate population registers (for example Bye and Judson, 2004; Statistics New Zealand, 2012). Studies in such context find that while administrative data provide useful indication of the resident population, they have limited potential for use as direct counts and are best used in combination. Despite all limitations they hold sufficient promise to warrant further investigations into ways of using administrative data to improve the quality and efficiency of population statistics.

## References

Bakker, Bart, 2010, Micro-Integration: State of the Art, Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, May 2010, Working Paper 10

Bye, Barry and D.Judson, 2004, Census 2000 Testing, Experimentation and Evaluation Program Synthesis Report No.16, Results from the Administrative Records Experiment in 2000, US Census Bureau, Washington, DC

Office for National Statistics, 2009, Interim Report on the Potential Use of Department for Work and Pensions data to Improve Population and Migration Statistics

Office for National Statistics, 2012, Using administrative data to set plausibility ranges for population estimates, available on the Office for National Statistics website.

Statistics New Zealand, 2012, Transforming the New Zealand Census of Population and Dwellings: Issues, options and strategy. Wellington: Statistics New Zealand.