Household Analysis Review Group (HARG)

2012-Based Household Projections Methodology – Choosing Survey Data for Weighting Projections Based on Different Censuses

Contents

1. Purpose 2
2. Background 2
3. The Surveys Investigated 2
4. Scottish Household Survey (SHS) 3
   - 4.1 Sample Design 3
   - 4.2 Weighting 4
   - 4.3 National Records of Scotland (NRS) weighting of SHS household type data 4
5. The Scottish Government Pooled Dataset 7
   - 5.1 Sample Design 8
   - 5.2 Weighting 8
   - 5.3 Release of the Pooled Data 9
6. The Annual Population Survey (APS) 9
   - 6.1 Sample Design 9
   - 6.2 Weighting of the APS 11
   - 6.3 Issues with using the APS 12
7. Conclusion 13

1. Purpose

In this paper we discuss our investigation of a range of surveys. The aim was to identify the most suitable survey to use in production of 2012-based household projections. We recommend using the Scottish Household Survey (SHS) and this paper outlines the rationale behind that recommendation.

2. Background

National Records of Scotland (NRS) produces household projections for Scotland every two years. The proportion of adults classified as the head of a household in each Council area, age group and household type is projected forward. These are the 'headship rates'. They are applied to a projection of the population living in households to give the projected number of households.

Prior to the 2012-based household projections, a two point exponential model was used to project forward headship rates from the 1991 and 2001 Censuses. More detail on how the projections were produced can be found in the 'Methodology' section of the 'Household Projections for Scotland – 2010-based' publication on the NRS website.

The release of data from the 2011 Census prompted us to review the household projections methodology. In particular we needed to decide how data from the 2011 Census should be incorporated into the projections. Various methods were tested and compared and the results were presented at our last HARG meeting.

Our preferred option was to use a weighted modified two-point exponential model. Survey data on household type was used to assign weights to modified two-point exponential models based on the 1991 and 2001 Censuses and the 2001 and 2011 Censuses. This was done by minimising the difference, for Scotland as a whole, between the proportions of households of each type produced by the projection, and those estimated by the survey in the base year i.e. 2012.

HARG members agreed with NRS's preferred option. However they noted that we must be able to justify our choice of the survey data used to assign the weights. The Scottish Household Survey (SHS) was used in testing, prior to the last HARG meeting. We have now investigated the other major population surveys carried out in Scotland which contain information on household composition. We have also looked in more detail at the weighting methodology of the SHS. This paper gives an overview of this work and our rationale for choosing to continue using the Scottish Household Survey.

3. The Surveys Investigated

There are three major Scottish Government population surveys – the Scottish Household Survey (SHS), the Scottish Health Survey (SHeS) and the Scottish Crime and Justice Survey (SCJS).

We used the SHS in testing as it was specifically designed to provide information on (amongst other things) the composition of households in Scotland and it has a reasonably large sample size, around 10,000 households each year.

Both the SHeS and the SCJS contain information on household composition. The sample size of the SCJS is around 12,000 households but unlike the SHS it is carried out biennially and on a financial year basis. The SHeS is carried out annually

but has a much smaller sample size, around 4,000 households. The two surveys have been investigated because they form part of the Scottish Government's pooled survey dataset. Together with the SHS this provides a much larger sample size for a set of core questions included in all three surveys – potentially more than 20,000 households for the first year of pooled data. Household composition forms part of the set of core questions so the 'pooled dataset' was explored as a potentially more precise source of household type information.

We have also investigated the Annual Population Survey (APS) for Scotland. The APS is carried out by the Office for National Statistics. It combines information from the quarterly Labour Force Survey (LFS) with a 'boost' sample. The APS has the largest sample size of all population surveys in Scotland, around 20,000 households annually, so it offers the potential of more precise household type information.

The investigation of the surveys has involved exploring their sample designs and the methodologies used to weight their data.

Sample designs should reflect, as far as possible, the census definition of a household. No types of household should be missed out or incorrectly included (e.g. we don't want the communal establishment population included in a survey).

Weighting methodologies involve adjustments to account for sample design (i.e. to account for non-equal probabilities of selection) and to account for bias (i.e. a group or groups being under-represented or over-represented). An example of bias is that young men living alone are often under-represented in surveys. Most weighting methodologies use information on the structure of the population and/or households. Our main criterion here was that the weighting methodology did not make use of our household projections.

4. Scottish Household Survey

The Scottish Household Survey (SHS) is carried out annually on a calendar year basis. SHS publications and background to the survey can be found on the Scottish Government (SG) website. The following information has been gathered from the SHS technical report – 'Scottish Household Survey – Methodology and Fieldwork Outcomes 2012' (SG website).

4.1 Sample Design

The SHS is a survey of private households. Communal establishments are not generally included. Addresses to be included in the sample are selected from the Royal Mail's small user postal address file (PAF) using systematic random sampling. The small user postal address file contains all addresses defined by the Royal Mail as receiving small volumes of mail. These will in the main be private households. However mobile homes, nurses' homes, student halls of residence, other communal establishments and sites for travellers are included in the SHS sample if they appear on the small user PAF and are the sole or main residence of the people living there. No SHS sampling is done on military bases so private households located in bases are not included.

The sample is disproportionately stratified by Council area. This ensures a minimum sample size is achieved in smaller Council area which allows data for councils to be published annually. The proportion of addresses in the sample from smaller cas is higher than their share of households in Scotland and the reverse is true for larger Council areas.

The interviewee for the household section of the questionnaire must be a householder and will usually be the highest income householder or their spouse/partner. Students' term-time addresses are taken as their main residence. In general, students in halls of residence should not be included as such communal establishment should not appear in the small user PAF (although some may).

In general, the sample design of the SHS doesn't have any major inconsistencies with the census definition of a household used for the projections. There are some issues with the inclusion of some communal establishments in the small user postcode file and the exclusion of private households on military bases but the impact of these is likely to be small.

4.2 Weighting

The first stage of the SHS weighting methodology involves calculating selection weights. Selection weights adjust for the fact that some households are more likely to be selected when the sample is drawn than others (i.e. they adjust for non-equal probabilities of selection). This happens because the proportion of households sampled in small cas is higher than their share of Scotland's households and vice versa for large Council areas. NRS's household estimates are used to calculate the proportion of households in Scotland that are in each Council area. This is then divided by the proportion of sample households in the Council area.

The second stage calculates calibration weights. Calibration weights adjust for non-response bias (i.e. some types of people being more or less likely to respond to the survey than others) and mean that weighted survey totals are equal to population totals. The data is weighted using the selection weights described above. The calibration process adjusts these weights to ensure that the weighted total of all members of responding households matches population totals for age-bands and gender within each Council area. NRS's mid-year population estimates are used for this. A scaling factor is then applied to ensure the weighted household total matches the NRS household estimates for each Council area. This is not necessary for the SHS data used in the projections as we make our own adjustment (as described in section 4.3).

The weighting process for the SHS does not involve household projections. Therefore there is no issue with circularity if we use SHS data to assign weights to the projections. Population estimates are used indirectly in our projections as they are the base for the population projections. Household estimates are used to constrain the projected number of households at Council area level for the base year and the following year. Neither contains any information on household type. It is worth noting that the weighting procedure uses the entire population from the NRS estimates, not the private household population (i.e. the population excluding those living in communal establishments).

4.3 NRS weighting of SHS household type data

Unweighted SHS data over-represents females and under-represents males. This is caused by young males, in particular, being difficult to contact or refusing to participate in the survey. Calibration weighting attempts to account for such biases. However this requires an assumption that non-respondents do not differ to respondents in characteristics not included in the weighting process, such as household type. Therefore people living in certain household types may be under or over-represented compared to people of the same age/gender living in another

household type. In the case of young males, this would mean that those living alone would be under-represented compared to those in other household types.

In general in the SHS (and other sample surveys) single adult households are under-represented[1]. This can be seen by comparing SHS data on household type with census data. Figure 1 compares 2001 SHS weighted data on household types with data from the 2001 Census. Figure 2 shows the same comparison but with 2011 data.

Figure 1: Comparing percentage of households of each type in the 2001 SHS with 2001 Census

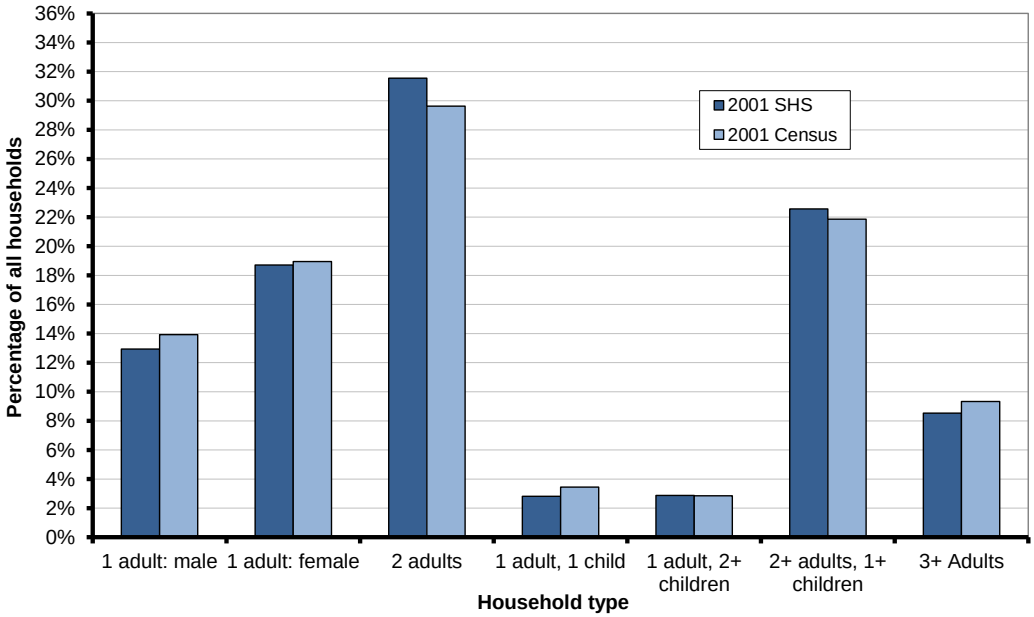
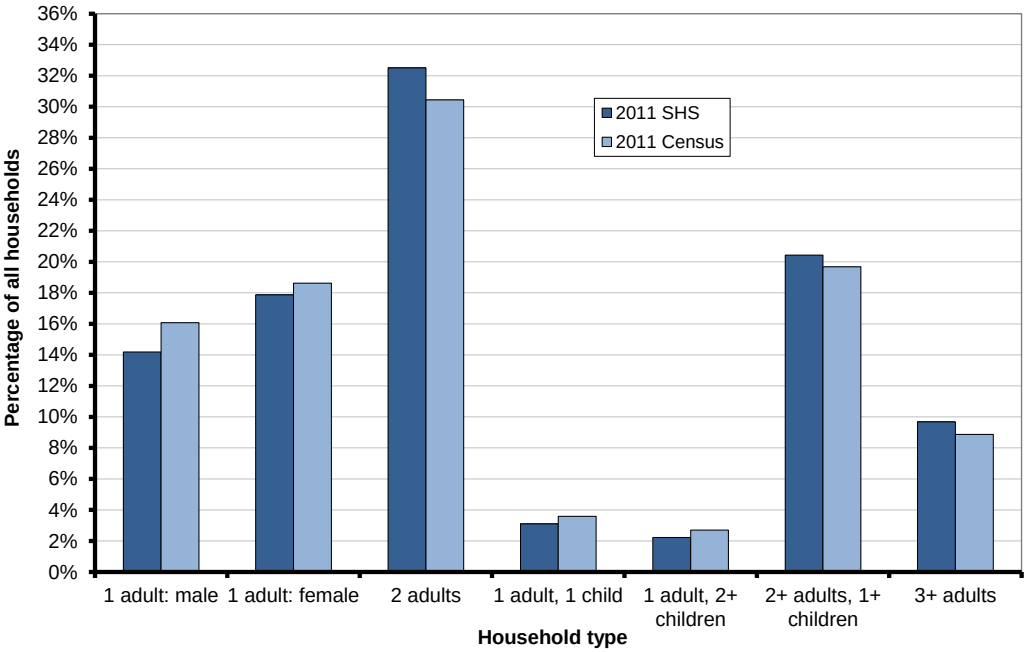


Figure 2: Comparing percentage of households of each type in the 2011 SHS with 2011 Census



Footnote

1) Refer to '[The Scottish Household Survey – Report of the 2001-Census linked study of non-response]', available on the Scottish Government website.

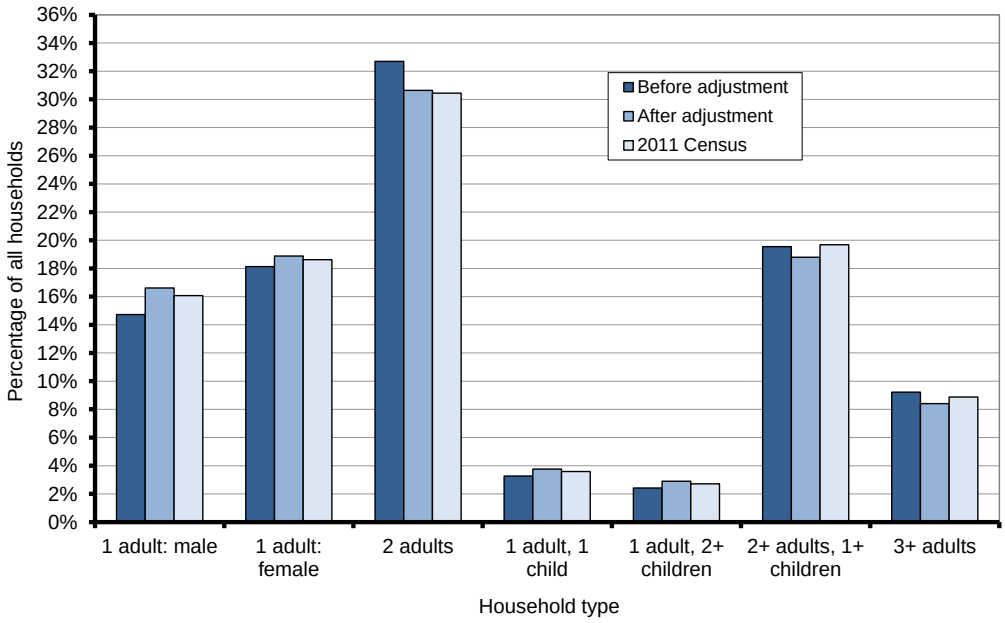
As expected, in both figures there are higher proportions of single adult households in the census than the SHS (although in 2001 the proportions of females living alone are very close). The SHS gives higher proportions of households containing two adults only and two or more adults with children, and lower proportions of households consisting of a single adult and one child (and two or more children in 2011). In 2001 there was a higher proportion of households containing 3 or more adults in the census than in the SHS. The opposite was seen in 2011.

In an attempt to account for biases in SHS data we make an adjustment based on the difference between household type information in the 2011 Census and the 2011 SHS. The adjustment made to 2012 data was as follows:

- Household type proportions are calculated using SHS weighted sample data.
- Differences in census and SHS household type proportions from 2011 are calculated.
- The differences are added to household type proportions from the 2012 SHS.
- Finally, the adjusted proportion is applied to NRS's 2012 household estimates total to give the number of households of each type.

Figure 3 shows that the adjustment brings the 2012 SHS proportions closer to 2011 Census values for the smaller household types but the opposite is true for household with two or more adults and children and three or more adults.

Figure 3: Comparing 2012 SHS data before and after adjustment with 2011 Census data

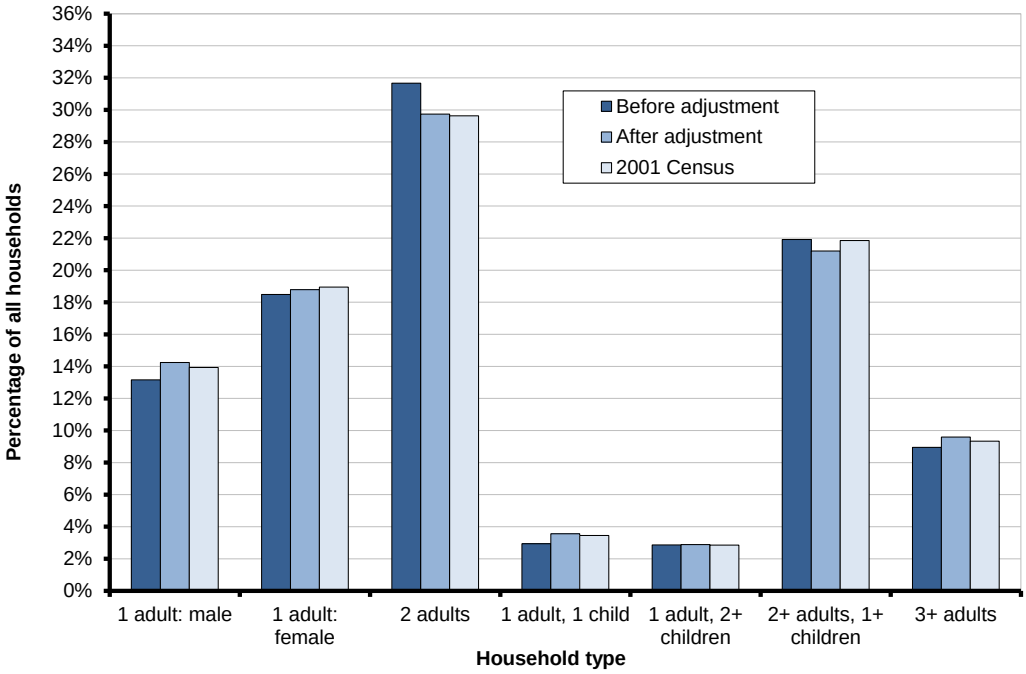


For 2002 to 2010, SHS figures are adjusted using a linear interpolation of the differences between the Census and SHS in 2001 and 2011. Figure 4 shows the effect of this on 2002 SHS data. Overall the effect is similar to that seen in Figure 3 for 2012 SHS data, with some small differences as expected. The adjusted numbers and proportions for each household type are published annually alongside NRS's household estimates in 'Estimates of Households and Dwellings in Scotland', available on the NRS website.

We have decided to continue to use this method to adjust the SHS data as it produces estimates of smaller household types which are generally closer to census estimates than the unadjusted data.

The UK Department for Communities and Local Government's (DCLG's) household projections make use of household representative rates from the Labour Force Survey (LFS). Before the data is used it is adjusted, based on the differences between the census and LFS in the census year, to account for potential biases[2]. This further justifies our decision to adjust the SHS to counter any bias.

Figure 4: Comparing 2002 SHS data before and after adjustment with 2001 Census data



5. The Scottish Government Pooled Dataset

Since January 2012 the three main Scottish Government population surveys Scottish Household Survey (SHS), the Scottish Health Survey (SHeS) and the Scottish Crime and Justice Survey (SCJS). have included a core set of questions. All ask identical questions on household composition. Sampling is coordinated so no address is sampled by more than one survey in the same year. This allows the creation of a 'pooled dataset' containing information on the core questions from each survey, which has a much larger sample size than the individual surveys. A larger sample size should mean more precise estimates therefore we have investigated the pooled dataset as an alternative to the SHS.

The following information (available on the Scottish Government website) comes from papers on the centralised sample design and weighting for Scottish Government surveys, dissemination options for the pooled dataset and the most recent technical reports for the SHS, the SHeS and the SCJS.

Footnote

2) Refer to 'Updating Department for Communities and Local Government's (DCLG) household projections to a 2008 base' (legislation.gov.uk/ website). 2008-based projections are the latest full set produced by DCLG. Interim 2011 household projections were produced but no mention is made of whether the LFS data used was adjusted in a similar manner.

5.1 Sample Design

The sample frame for all three surveys is the Royal Mail's small user Postal Address File (PAF). This mainly includes private residential addresses however some ineligible addresses (e.g. some communal establishments and small businesses) will also appear. More detail on the small user PAF is given in section 4.1. No address is sampled by more than one survey in each year.

Scottish Household Survey (SHS) sample design has been discussed in section 4.1 and it was considered suitable for use in the projections. To investigate the pooled dataset we also need to look at the sample designs of the SHeS and the SCJS.

The SHeS uses a two-stage clustered sample design. Intermediate geographies (data zones for the islands) are randomly selected at the first stage. Addresses in these intermediate zones are then selected using systematic random sampling. The sample is disproportionately stratified by council area. The interviewee for the household questionnaire is usually the highest income householder or their spouse/partner. Around 4,000 households are surveyed in each calendar year.

The SCJS surveys around 12,000 households biennially on a financial year basis. Systematic random sampling is used to select addresses from the small user PAF. The sample is disproportionately stratified by council area. The interviewee is randomly selected from all those aged 16 or over in the household.

There are no aspects of the sample designs of the SHeS and SCJS that would make them unsuitable for use in the household projections. We came to the same conclusion about the SHS. Therefore based on sample design alone, the pooled data could be used.

5.2 Weighting

The project established by the Scottish Government on weighting the pooled sample recommended that it should involve combining the weights from the component surveys (although the exact method wasn't detailed). This means that we must look at the weighting strategies of the three surveys. The method used by the SHS has been discussed in section 4.1 and our adjustments to it in section 4.3.

The first stage of the SHeS weighting procedure calculates a combined selection weight which corrects for non-equal probabilities of selection caused, in part, by the disproportionate stratification of the sample. This combined selection weight is used as the input in the calculation of calibration weights. Calibration weighting attempts to correct for non-response bias. The combined selection weights are modified so that the totals of the weighted survey data equal population totals for each health board, age/gender totals for Scotland as a whole and population totals for the 15 per cent most deprived areas. The population living in private households is used. This is calculated by applying communal establishment rates (taken from our communal establishments data collection) to NRS's mid-year population estimates.

The first stage of weighting the SCJS is also to calculate selection weights to account for non-equal probabilities of selection. These act as entry weights to the calibration weighting process. The calibration weighting stage corrects for non-equal probabilities of selection in different geographic areas and for response bias. The entry weights are modified so that the weighted household totals match estimates of the following:

- Households of each type in each Police Force Area (PFA)/Community Justice Authority Area (CJAA).
- Age of head of household in each PFA/CJAA.
- Households in urban and rural areas in each Council area.

The information on household type and age of head of household comes from our household projections.

The weighting methods of the SHeS and the SHS are compatible with use of their data in the household projections. However the SCJS uses our household projections in its weighting, which introduces a circularity we would rather avoid. The weights for the pooled dataset are based on those of the component surveys. Therefore our concerns about the SCJS means we are reluctant to use the pooled data. Some information from the pooled dataset, including household composition, was published on 21 May 2014[3]. However the release doesn't contain detailed information on the weighting procedure used.

5.3 Release of the Pooled Data

The pooled data has been released as 'Data Being Developed'. This means it shouldn't be the preferred source for various measures, its publication is to allow for testing and comment from expert users. Their feedback will assist in further quality assurance and development of the data. This designation is another reason we wouldn't use pooled data in the current projections. It may not be of the standard required of National Statistics and the dataset may be subject to change following testing and feedback.

6. The Annual Population Survey

The Annual Population Survey (APS) has a sample size of around 20,000 households per calendar year, around double that of the SHS. Estimates from the APS should therefore be more precise so we have investigated whether it can be used instead of the SHS.

The APS is run by the Office for National Statistics (ONS). It is a combination of information taken from ONS's quarterly Labour Force Survey (LFS) and an annual survey 'boost' paid for by the Scottish Government. ONS publishes a range of user guides and papers covering different aspects of the LFS and the APS. The following information has been gathered from the most up-to-date versions of the LFS User Guide Volume 1 – Background and Methodology, LFS User Guide Volume 6 – Local Area Data and LFS User Guide 8 – Household and Family Data. Links to these guides can be found on the LFS User Guidance web page of the ONS website.

6.1 Sample Design

The APS sample design, and indeed that of the LFS, is more complicated than the other surveys discussed. As mentioned above, it combines data from the LFS with an annual boost.

Footnote

3) Refer to 'Pooled Sample from Population Surveys in Scotland' on the Scottish Government website.

6.1.1 Labour Force Survey data in the Annual Population Survey

The Labour Force Survey (LFS) is carried out on a quarterly basis. Once a household is selected it is part of the survey for five successive quarters. In each quarter there are five 'waves'. Those in wave 1 are being interviewed for the first time, those in wave 2 for the second time and so on. Households in wave 5 are being interviewed for the last time (one year after they joined). In the following quarter wave 1 becomes wave 2 etc. and a new set of wave 1 households is added (the same number of new addresses is added each quarter). It is the address that stays in the sample, not its occupants. Therefore if they move away and are replaced, the new residents will be interviewed.

Data from waves 1 and 5 from 4 consecutive quarters of the LFS is pooled to make up part of the rolling 4-quarterly Annual Population Survey (APS) dataset. Only waves 1 and 5 are used to ensure that no household appears more than once in each APS dataset.

The target population is all those resident in private households. One exception to this is students living in halls of residence (or at boarding schools). They are included as residents at their parental home. In our household projections such students are part of the communal establishment population, not the private household population. This inconsistency may result in the LFS overestimating average household size. Another small difference is that the LFS includes a sample of NHS accommodation. This appears to be housing for NHS staff such as nurses. A different sampling frame is used for NHS accommodation and any impact it has on the APS data on household types is likely to be minimal.

The main sampling frame for the LFS is the Royal Mail's small user postal address file (PAF). Systematic random sampling is used to select the new (wave 1) addresses for each quarter. More details on PAF can be found in section 4.1.

PAF is used for sampling south of the Caledonian Canal. North of the Caledonian Canal the sampling frame consists of telephone directories and interviews are by telephone rather than face-to-face. Systematic random sampling is used to select addresses from the relevant directory. Using telephone directories as a sampling frame has some disadvantages, mainly the potential for bias caused by excluding those not listed, e.g. because they are ex-directory, have a mobile phone only or no phone at all. It may take some time to include residents of newly built homes. For ONS the cost savings of sampling this way outweighed the disadvantages.

From the third quarter of 2010 onwards a change was made to the treatment of households containing only those aged 75 or over. Follow-up interviews were no longer conducted with such households i.e. they would only be interviewed once. The reasoning was that their economic status is likely to remain stable. Any such household in wave 1 will represent this household type in waves 2 to 5 through a specific weight. As the household type information for such households will still form part of wave 5, the impact on the data we're interested in is likely to be very small.

6.1.2 Annual boost data in the APS

The Scottish Government pays for an annual boost of the LFS, which increases the sample size from around 7,000 households to in excess of 20,000 allowing more precise estimates at council level to be produced. The boost sample has four annual waves. Households selected for wave 1 are interviewed four times with each interview a year apart. In any given year, a quarter of the sample will be newly

selected addresses (wave 1) and three quarters will have been interviewed at least once before.

The sample is chosen in a similar manner to the main LFS, as described in the previous section.

6.1.3 Suitability of APS sample design

There are no major issues with sample design which would prevent the APS from being used in our projections. However a couple of minor issues are worth bearing in mind. These are that students living in halls of residence are treated as living in their parental home, and the use of telephone directories to sample north of the Caledonian Canal. The former may mean the APS overestimates average household size and the latter may introduce biases. However it's likely that these issues would have a minimal effect on household type information at Scotland level. The relatively complex sample design of the APS makes it harder for users to understand how the data is collected and what the implications of the design may be.

6.2 Weighting of the APS

Weighting of the APS household dataset involves two stages. In the first stage, design weights are created to correct for non-equal probabilities of selection. The next stage, calibration weighting, adjusts for non-response bias. The design weights are modified so that the survey data sums to population totals for different breakdowns (known as calibration groups). These breakdowns involve combinations of all or some from location (e.g. local authority, Government Office Region etc.), age and gender. This process also assigns the same weight to all members of the same household so that household level figures can be produced. The weighting procedure uses the whole APS dataset (i.e. data for the entire UK).

The population data used is based on mid-year population estimates. However APS data is generally published before the relevant year's population estimates. Therefore a projection of the mid-year population is used. Local authority population projections are produced by rolling forward five year average growth rates from mid-year population estimates and constraining the final figures to official national population projections. The communal establishment population is estimated by assuming the proportions of communal establishment residents in each five year age group, gender and region are the same as in the 2001 Census. The private household population is given by the difference between these two sets of figures. Reweighting of APS data takes place from time to time to incorporate more up-to-date population estimates. However there doesn't seem to be a fixed schedule for this. Total numbers of households from the APS are survey estimates, there is no constraining to household estimates or projections.

Weighting the APS household data does not, therefore, involve any household variables. Although a projection of the population is involved, the method used is different to that used in NRS's population projections. Therefore the APS weighting methodology does not create any major obstacles to its use in our household projections.

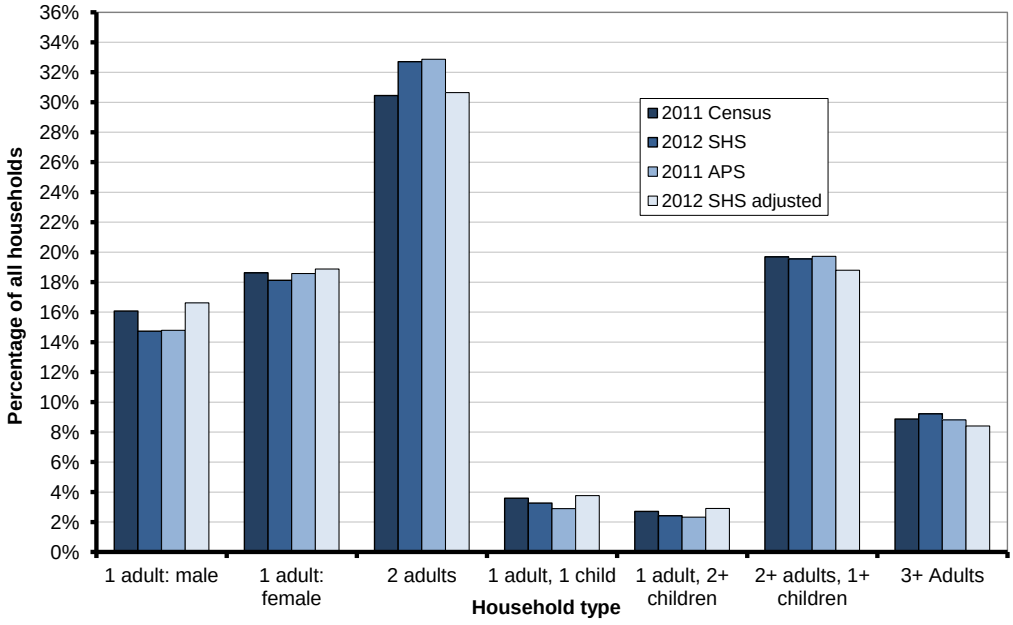
6.3 Issues with using the Annual Population Survey

Although there are no major issues with the sample design or weighting procedure used for the Annual Population Survey (APS) data that prevent its use in our projections, there are a number of smaller issues to consider. There is the use of the telephone directory as a sample frame north of the Caledonian Canal and, more importantly, the inclusion of students in halls of residence (or boarding schools) at their parental homes. Also, the weighting procedure involves a form of population projection.

Compared with the other surveys discussed in this paper, in particular the SHS, the APS is relatively complex. Detailed documentation of the survey methodology is produced by ONS but it is spread across a number of documents all of which are updated at different times. To understand the method used to produce household data from the APS requires reading sections in Volumes 1, 6 and 8 of the LFS user guide. Providing information on household type is not the main purpose of the LFS or APS therefore it's more difficult to build up a clear picture of the methodology behind this data.

We have been unable, at this stage, to get 2012 household type information from the APS. We have managed to get data for earlier years but there are some issues with the figures from 2010. Therefore Figure 5 shows a comparison of 2011 APS data with unadjusted 2012 SHS data, 2011 Census data and adjusted 2012 SHS data[4]. In most cases, the adjusted 2012 SHS percentages are closest to the census values. Differences between the 2012 SHS data and the other figures are expected as they cover different years (also as a result of sampling variability in both surveys) but the absence of 2012 APS data means we can't compare data from the same year. This comparison doesn't show a clear advantage, in terms closeness to census values, if we use the APS instead of the adjusted SHS.

Figure 5: Comparing 2011 Census data with 2012 unadjusted SHS data, 2011 APS data and 2012 adjusted SHS data



Footnote

[4] As described in section 4.3 we adjust the SHS to account for bias using the differences between the survey data and census. This means that for 2011 the adjusted SHS values are simply Census values.

Taking into consideration the issues discussed above, most importantly not having 2012 APS data and the comparison shown in Figure 5, we recommend not using the APS in the 2012-based household projections, and to use the adjusted SHS figures instead. We can look again at using the APS when work begins on the 2014-based household projections.

7. Conclusion

This paper explores the three main Scottish Government (SG) population surveys Scottish Household Survey (SHS), the Scottish Health Survey (SHeS) and the Scottish Crime and Justice Survey (SCJS) the SG's pooled dataset and Office for National Statistics' Annual Population Survey (APS) to decide on the most appropriate source to use in our new projections methodology. Our main concern was that the survey design and weighting methodology should not be incompatible with our projections. Namely, that the sampling design matches our definition of households as closely as possible, and that the weighting method does not make use of our household projections (to avoid 'circularity').

There were no major issues with the designs of the surveys investigated although the treatment of students in halls of residence or boarding schools in the APS is a minor issue. The SCJS uses our household projections in its weighting procedure, therefore we would not use SCJS data. There were no major issues with the weighting methods of the other surveys. However as the intention is that the weights for the pooled dataset will be based on those of the component surveys, the issue with the SCJS means we would be reluctant to use the pooled dataset. Furthermore, the pooled data has been released as 'Data Being Developed'. It may not be of the standard expected of National Statistics and may be subject to changes after initial publication therefore isn't suitable for our purposes at the moment.

The SHeS sample size is less than half that of the SHS. So our choice was between the SHS and the APS. The sample size of the APS is around twice that of the SHS so estimates from it should be more precise. However it is more complex than the SHS making it difficult to understand the survey methodology and hard to gauge the impact of any issues with the data. We haven't been able to get APS data for 2012. Although it should be possible to obtain this data, the timescale for our projections means it is no longer possible to wait for it. Comparing the APS, SHS and 2011 Census did not give a clear indication of which survey gave the more accurate results. Therefore we recommend that we continue using the SHS. We have SHS data for 2012, we can adjust it (using census household type information) to account for any biases towards particular household types and the survey methodology is well documented and relatively simple to understand.

Household Analysis Review Group (HARG) members are asked to comment on our recommendation to use the SHS to weight the results of the household projections, at Scotland level.
Any other comments on this work or suggestions would be much appreciated.

NRS Household Estimates and Projections Branch
June 2014