Department for Communities & Local Government

ENGLISH HOUSING SURVEY

English Housing Survey

Technical Advice Note

Sampling and weighting: 2011-12 Update

© Crown copyright, 2013

Copyright in the typographical arrangement rests with the Crown.

You may re-use this information (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence. To view this licence, visit http://www.nationalarchives.gov.uk/doc/open-government-licence/ or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or e-mail: psi@nationalarchives.gsi.gov.uk.

This document/publication is also available on our website at www.gov.uk/dclg

Any enquiries regarding this document/publication should be sent to us at:

Department for Communities and Local Government
Eland House
Bressenden Place
London
SW1E 5DU
Telephone: 030 3444 0000

July, 2013

ISBN: 978-1-4098-3924-8

Contents

Introduction

Overview of sampling and weighting

Survey samples
- Interview survey sample
- Physical survey sample
- Size of physical survey sample to issue
- Sub-sampling cases for physical survey

Weighting methodology
- Overview
- Adjusting for the relationship between addresses and dwellings
- Weighting the full household sample
- Weighting the dwelling sample
- Calculating the two-year dwelling sample weights

Introduction

1. This is one of a series of Technical Advice Notes about the English Housing Survey (EHS) to give users further information about how EHS data is collected and quality assured and how some of the key derived measures are created.

2. This note provides information on how the samples of households and dwellings were selected for inclusion in the interview and physical survey elements of the EHS in 2011-12. It also provides information on the weighting methodology used to enable national-level estimates to be obtained from the survey results for both the full household sample and the dwelling sample.

Overview of sampling and weighting

3. The EHS consists of two main elements: an initial interview survey of approximately 13,300 households and a follow up physical inspection of a sub-sample of around 6,200 of these dwellings, including vacant dwellings. For more information on EHS methodology, see the survey overview and methodology technical advice note.

4. For the 2011-12 EHS the sampling and weighting for both of these survey components were carried out by the Office for National Statistics (ONS) who managed the survey on behalf of the Department for Communities and Local Government (DCLG). ONS also undertook the interviews with householders as well as the interview data validation processes. Physical surveys are carried out by a large field force of professional surveyors employed by MMBL-CADS, who worked in close co-operation with the interviewers from ONS to maximise response to the survey and deliver high quality data.

5. DCLG require the sample issued for interview to be a random sample of households in England. The cases for which an interview is obtained are known as the full household sample. For the physical survey, they require that a disproportionate number of the achieved sample of dwellings should come from properties rented from local authorities and housing associations, in order to meet set tenure targets and provide sufficient information on rarer tenures. All occupied dwellings in the physical survey are properties where an interview survey has been conducted previously. Vacant properties are also drawn from those addresses found to be vacant at the interview stage. The subset of cases for which a physical survey is achieved from the dwelling sample are also sometimes referred to as ‘paired cases’ because data are available on them from both the interview and physical surveys.

6. In 2011-12, 24,299 addresses were issued to interviewers. A small proportion of these were found to be ineligible, including addresses that

were found to be commercial premises, second and holiday homes or demolished properties. Of the remainder, interviews were achieved at 13,829 addresses (the full household sample). A further 983 addresses were found to relate to vacant dwellings. After sub-sampling, 9,508 dwelling addresses were passed to surveyors, and physical surveys were achieved at 6,459 of these (the 2011-12 component of the dwelling sample).

7. For analysis purposes, weights are calculated for the full household survey cases achieved in one year of the survey, but due to the smaller sample for which physical surveys are achieved, two years' dwelling sample cases are weighted together.

Survey samples

Interview survey sample

8. In 2011-12, DCLG reduced the number of household interviews undertaken as a result of a reduction in funding. This has resulted in some loss of precision. The calculation of the number of required interviews took into account a need to maintain a similar level of precision for social and private rented households, with a reduced level of precision for owner occupiers.

9. As the number of physical surveys conducted in the rented sectors was comparatively low in 2010-11, it was decided to draw a larger sample and to sub-sample properties likely to be owner occupied. Taking account of the sub-sample rates to be applied when identifying which properties are eligible for a physical survey, DCLG wanted to achieve approximately 13,300 household interviews and 6,200 physical surveys in 2011-12 divided across the tenures, as shown in Table 1.

Table 1: Required number of interviews and physical surveys, 2011-12

	Interviews	Physical surveys
Owner occupiers	8,533	2,602
Private rented	1,691	1,176
Local authority	1,487	1,158
Registered social landlord	1,588	1,255
Total	13,299	6,191

10. In order to achieve the number of interviews and surveys described above, an initial sample of 38,416 addresses was drawn from the Postcode Address File (PAF). These addresses were drawn as a systematic random sample from the Royal Mail's Small User PAF. For each address, the predominant tenure within the postcode that contained the sampled address was identified and attached to the record.

11. As a cost cutting measure, the number of addresses issued to interviewers on the EHS was reduced from 32,100 in 2010-11 to 24,299. However, because reducing the sample size uniformly across the full sample would reduce the numbers within certain tenures to a level that would not facilitate reliable analysis, owner occupied addresses were sub-sampled. To achieve reliable numbers, the originally drawn sample of 38,416 addresses was sub-sampled in postcodes that were predominantly owner occupied, while all sampled addresses from other postcodes were retained.

12. Predominant tenure was identified using Experian's Residata[1] classifications; addresses were then grouped into strata and sub-sampled at the rates of 54.5% for owner occupied and 100% for other tenures, see Table 2. As a result, 24,299 addresses were issued to interviewers.

Table 2: Sub-sampling of PAF addresses, 2011-12

Assessed tenure from Residata (predominant tenure)	PAF sample	Sub-sampling rate	Issued EHS sample
Owner occupied	31,027	54.5%	16,910
Private rented	731	100%	731
Social rented	6,474	100%	6,474
Mixed tenure	122	100%	122
Unknown tenure	62	100%	62
Total	**38,416**		**24,299**

Physical survey sample

13. DCLG also require the delivery of sufficient paired cases with physical surveys to deliver the same precision as achieved by the English House Condition Survey (8,000 cases a year through a clustered sample design). The EHS sample is stratified to ensure disproportionate numbers of the rarer tenures (social renters). The alternative to this approach would be to issue a very much larger random sample and carry out unnecessary owner occupier and private renter surveys in order to accumulate sufficient social renting cases.

[1] Experian possess a database that contains information obtained from a number of sources including insurance companies, Census, etc. referred to as Residata. It is from this that information is taken on predominant tenure within a postcode as well as other information. The matching of the EHS sample to Residata is carried out by BRE.

14. Table 3 compares the target paired case annual tenure distribution with the national stock updated to show the 2011-12 totals.

Table 3: Tenure distribution of target achieved physical survey sample compared with the national stock

Tenure	Target sample		National stock [a,c]
	N	%	%
Owner occupied	2,602	42	68
Private rented	1,176	19	14
Local authority	1,158	19	18
Registered social landlord	1,255	20	
Total [b]	**6,191**	**100**	**100**

[a] Taken from Table 1.01 of the English Housing Survey Headline Report 2011-12
[b] Percentages may not add up to 100 per cent because of rounding
[c] The split between local authority and registered social landlord is not available

Size of physical survey sample to issue

15. The issued sample for the physical survey is drawn as a stratified sub-sample of the dwellings of those households who respond to the interview survey, together with a stratified sub-sample of dwellings found to be vacant at that stage. Calculation of the size of the sample to be issued takes account of the expected physical survey response rates by tenure.

Sub-sampling cases for physical survey

16. The sub-sampling of interview survey cases for the physical survey is carried out in the field during the interview, using the tenure of the dwelling established during the interview. Interviewers are advised via the CAPI (computer-aided personal interviewing) instrument if the household they are interviewing is eligible for the physical survey, and they then attempt to gain agreement from respondents to take part before passing on the address details to MMBL-CADS.

17. Different sub-sampling rates are applied to each tenure group to identify cases eligible for the physical survey. These rates can be varied quarterly if required in order to achieve the required annual total sample.

18. Because the rented sector is smaller than the owner occupied sector, the rented sector is over-sampled to ensure sufficient numbers for analysis. Sub-sampling rates are reviewed at the start of each year and are kept under review throughout the year in order to ensure a sufficiently large sample is achieved. The sub-sampling rates for 2011-12 were as follows:

Table 4: Sub-sampling rates for the EHS 2011-12 physical survey

	Sub-sample rates			
	Quarter 1	Quarter 2	Quarter 3	Quarter 4
Owner occupiers	54.5%	45%	40%	40%
Private renters	100%	100%	90%	90%
Local authority	100%	100%	85%	90%
Registered social landlords	100%	100%	85%	90%

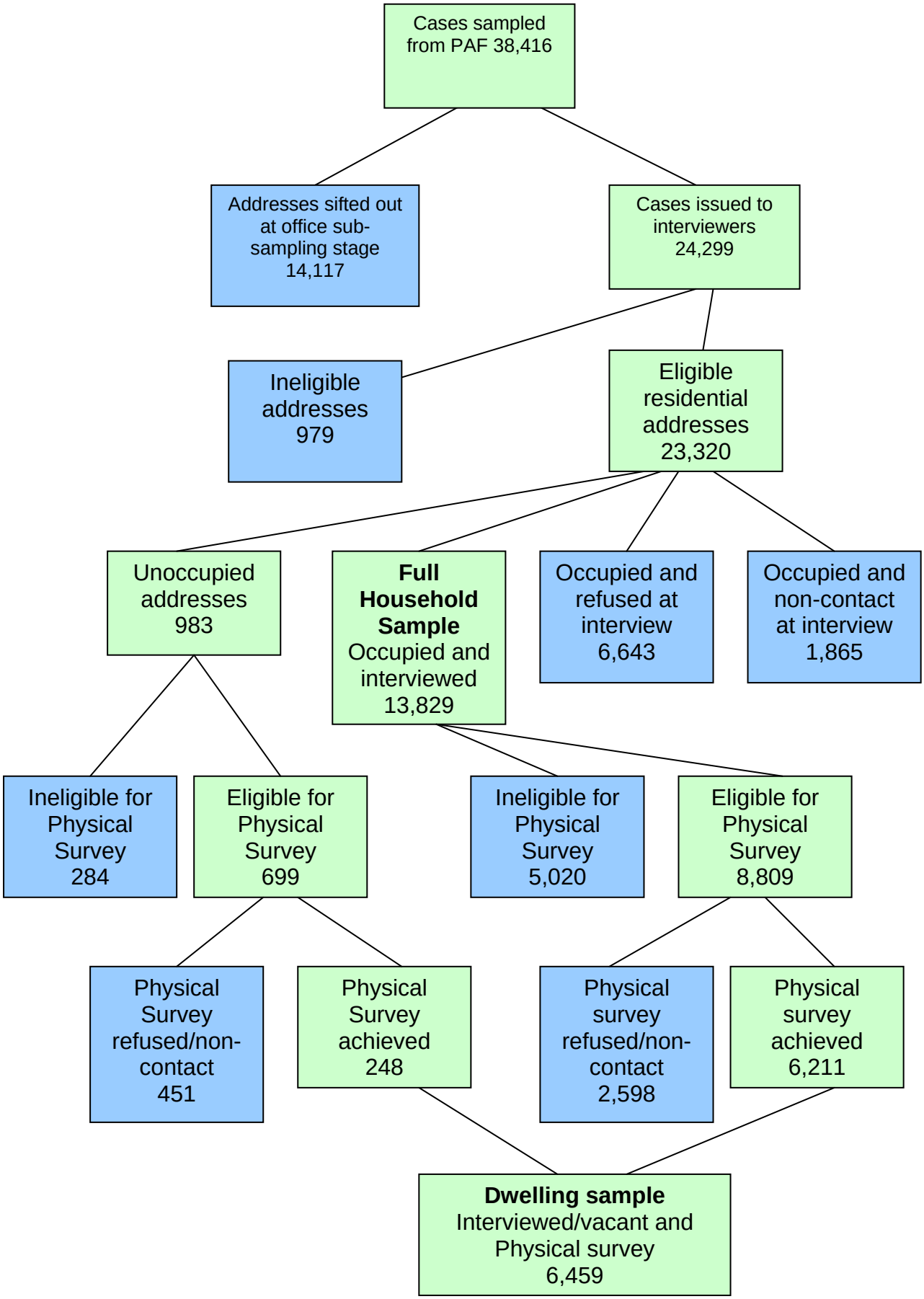
19. The sub-sample rate used in 2011/12 varied throughout the year. The original plan was to use a sub-sample rate of 50% for owner occupiers throughout the year, to achieve the 6,200 physical surveys required. However the process was complicated by an incorrect sub-sample rate being applied in quarter 1, and also better than expected response rates achieved throughout the year. As a result, the sub-sample rate was revised throughout the year to achieve the required number of interviews and physical surveys. As the surveyor field force is split into two, with half working in quarters 1 and 4 and the other half working in quarters 2 and 3, it was also necessary to adjust sub-sampling rates to ensure that roughly equal numbers of surveys were conducted by each group of surveyors.

20. Vacant properties are sub-sampled at the same rates as occupied cases based on their last known tenure gathered by interviewers as part of their initial visit. For vacant properties, cases are selected for physical survey based on the interviewer's best estimate of tenure using local enquiries and using the same sub-sampling rates as for occupied cases. Permission and access for the survey is then sought by the MMBL-CADS surveyors.

21. A lower proportion of full physical surveys are obtained in unoccupied dwellings, compared with occupied dwellings, because of the difficulty in gaining access to a property that is currently unoccupied. Surveyors managed to gain access and obtain full physical surveys in 35% of dwellings that were unoccupied at interview in 2011-12. This is an increase from 32% in 2010-11.

22. Figure 1 below shows details of the numbers of cases going forward at each successive stage of the survey process in 2011-12.

Figure 1: Sample structure of the EHS, 2011-12



Weighting methodology

Overview

23. The weighting methodology involves a series of steps, each of which is designed to take account of the selection and response processes involved in successive steps in the sampling and interview/physical survey response process. These steps in the weighting should give us unbiased estimates for the populations of households and dwellings in England.

24. The method is very similar to that of the former Survey of English Housing, the main difference being that much more detailed bias adjustment is carried out in the EHS. The weighting process for a single year's data is broken down into 12 stages.

Stages 1 and 2 apply to both the full household sample and the dwelling sample and adjust for:

- Stage 1: Address selection probabilities
- Stage 2: Address to dwelling relationship

Stages 3-6 apply to the full household sample and adjust for:

- Stage 3: Dwelling to household relationship
- Stage 4a: Accounting for office refusals (refusal to cooperate prior to the interview)
- Stage 4b: Probability of contact at the interview survey
- Stage 5: Probability of response at the interview survey
- Stage 6: Calibration weighting for the interview survey – matching to known population totals as at 1st October of the survey year

Stages 7a-12 apply to the dwelling sample and adjust for:

- Stage 7a: Office refusals – refusal to cooperate prior to contact at the dwelling
- Stage 7b: Probability of contact at the dwelling
- Stage 8: Probability of cooperation at the dwelling
- Stage 9: Sub-sampling by tenure for the physical survey
- Stage 10: Co-operation at the physical survey
- Stage 11: Calibration weighting to dwelling totals, including an adjustment for new builds
- Stage 12: Calibrating the household weights for the paired sample to those from the full household sample
- Stage 12a: Recalibrating the dwelling weights from Stage 11 to be consistent with the household weights from Stage 12

Stages 1 to 5 and 7a to 10 are implemented using a set of SPSS syntax scripts in combination with the specialist SPSS module AnswerTree.
Stages 6, 11, 12 and 12a are referred to as calibration stages because they scale the weighted data to predetermined totals. These stages use the Generalised Estimation System (GES) macro implemented in SAS.

Once the results obtained at Stage 12a for a single year's dwelling sample have been checked and signed off, a further weighting phase repeats the methodology of stages 11 – 12a on the combined data from two years' dwelling sample data, to create the final dataset for analysis.

These processes are described in more detail below.

Adjusting for the relationship between addresses and dwellings

25. Stage 1 and 2 are common to the weighting of both the full household sample and the paired cases sample. Stage 1 takes account of the sampling fractions involved in drawing the equal-probability sample of addresses from the PAF. The weight for this stage is simply:

$$w_1 = \frac{\text{delivery points on PAF}}{\text{delivery points sampled}}$$

Subsequent stages involve multiplying these initial weights by successive adjustment factors.

26. The EHS analyses are concerned with dwellings and households rather than addresses, and there is not always a one-to-one relationship between an address, a dwelling, and a household. (For the purposes of the survey, a dwelling is defined as 'a self-contained unit of accommodation where the occupants of that accommodation have sole use of all the rooms and facilities'). Usually there is only one dwelling at each address sampled from the PAF, but addresses are occasionally found to cover more than one dwelling (for example if a house has been converted into self-contained flats) or only part of a dwelling (for example a bedsit which shares facilities with a household at a separate postal address). The weighting methodology must take account of this.

27. Where an address refers to more/less than one dwelling, each dwelling at that address would have a lower/higher chance of selection. Stage 2 adjusts the weight calculated at Stage 1 to take account of the address to dwelling relationship for such cases.

Weighting the full household sample

28. Stage1 of the weighting process in 2011-12 was revised to take into account the office based sift by predominant tenure. This resulted in

two design weights being produced, one for cases identified as in postcodes that were predominantly owner occupied, and another design weight for all other cases.

29. Stage 3 makes a similar adjustment where the dwelling contains more than one household, and each household therefore had a lower chance of selection.

30. Refusal to co-operate prior to interview, non-contact at interview and refusal to co-operate at interview do not happen completely at random and the factors associated with each of these three processes may differ. Stages 4a, 4, and 5 each use available information about each case as predictor variables in a model to partition the sample into groups in order to describe as much variation in the response variable as possible. These models are constructed using the CHAID algorithm in the SPSS AnswerTree software. Typical predictor variables for Stage 4a are geographical area; predominant tenure, dwelling age and dwelling type in the area; urban/rural classification. For subsequent stages, information collected by the interviewer is also used.

31. Response weights are calculated for each of the groups produced by AnswerTree at each of Stage 4a, 4 and 5.

32. The previous five stages account for the sampling and response probabilities. Applying the weight from Stage 5 to the household-level data would provide a survey estimate of the total number of households in England. However, this will differ from the true value because of sampling error, under-coverage of the frame and inadequacies in capturing the non-response mechanisms. Estimates for subgroups such as tenures will differ from their true values for the same reason. These differences in the survey estimates can be reduced by adjusting the weights so that the total final weights match certain control totals.

33. The control totals used for Stage 6 were based on ONS population projections by sex and age group by geographical area, and tenure (owner occupied, social sector and private rented sector) from the Labour Force Survey as at 1 October of the survey year.

Weighting the dwelling sample

34. The approach to weighting the paired cases which make up the dwelling sample is dwelling-based rather than household-based so, in contrast to the weighting of the full household sample described above, all dwellings identified as vacant at which a physical survey was achieved are included in the sample to be weighted.

35. This process is more complex than that for the full household sample, partly because of the need to create internally consistent household and dwelling weights, and also because of the need to combine two years' dwelling sample data to obtain a large enough sample for analysis.

36. In outline, the approach is to first weight up the dwellings data for the current year to estimated dwelling controls by tenure based on DCLG published data on dwelling stock, then to adjust this weighting so that the number of weighted households that result from it is consistent, within tenure and geographical area, with the weighted full household sample. The data for two consecutive years are combined and reweighted to the reference date of 1st April midway between the two survey periods: this process is described in the next section.

37. The weighting process starts by using Stages 1 and 2, described above, to take account of the initial sampling fractions and the address to dwelling relationship.

38. Stage 7a then compensates for office refusals at either interview or physical survey stage. This process uses AnswerTree as in Stage 4a but with the inclusion of dwellings identified as vacant.

39. Stages 7 and 8 adjust for interviewer contact at the dwelling and for co-operation with the interview survey. All vacant dwellings are deemed to have been 'contacted' provided they have been located by the interviewer, and to have 'cooperated' with the interview survey phase with the exception of those in respect of which a refusal was sent to the Field Office.

40. The dwelling sample is required to contain a disproportionate number of social rented properties. This is achieved by sub-sampling the issued sample during the interview survey once information about tenure, including tenure of vacant properties, has been collected. Only dwellings selected at this stage are eligible for the physical survey. Sub-sampling rates for the physical survey can also be varied between quarters, as was done for 2011-12. Stage 9 calculates adjustments to the weights to take account of this sub-sampling.

41. Stage 10 adjusts for response to the physical survey, within weighting groups created by AnswerTree. For this stage, data collected during the interview survey are also used to help determine the weighting groups, and vacant cases are treated separately if the data show this to give a better result.

42. As with the interview survey weighting, the initial weighting stages of the physical survey (Stages 1, 2, and 7a - 10) attempt to account for sampling and response probabilities and so the total weight of the data gives a survey estimate of the total number of dwellings in the population. However, this will differ somewhat from the true value because of factors such as sampling error, under-coverage of the frame and inadequacies in capturing the non-response mechanisms. These differences in the survey estimate can be reduced by adjusting the weights to match chosen control totals.

43. After stage 10 of the weighting process, a small number of cases had weights that were identified as outliers because they were considerably higher than those for other cases in the same tenure and geographical

area. These outliers had their value adjusted to the value of the highest 'non-outlier' within the same group. The difference between the original and new values was then redistributed over all cases within the same tenure and geographical area.

44. The calibration of the single-year weights to control totals was carried out in Stages 11 – 12a. At Stage 11, the weights from Stage 10 were first calibrated to DCLG dwelling estimates split by tenure and geographical area, projected forward to the required 1 October reference date.

45. The achieved sample of dwellings does not include any new dwelling built since the creation dates of the two PAF files from which the sample was taken, so the weights are adjusted using the number of new dwellings built between the PAF date and the reference date for the weighting. This process involves separate adjustments, for each PAF sub-sample, for

- areas with a high/low rate of new build, and
- private/social sector housing (excluding local authority housing)

Because of the small number of recently-built dwellings in the survey, the weights of all cases with a construction date of 1990 onwards are adjusted in this process. The calibration of the revised weights to tenure by area totals were then rerun

46. In 2011-12, an error was noticed with the way the new build adjustment was applied to the EHS weighting. This error was corrected for 2011-12, however the 2009-10 and 2010-11 weights were slightly affected. This error primarily affected the number of post-1990 dwellings reported in the survey, with a lower number of these dwellings reported than would have been with the corrected dwelling weights. Investigation into the impact of this error concluded that the change in results was negligible, and there was no plan to reproduce these weights for 2009-10 and 2010-11 EHS data.

47. At Stage 12, paired cases household weights were derived from the dwelling weights produced at the end of Stage 11, using the dwelling to household relationships found during the interview and physical surveys.

48. These weights were then recalibrated to give the same distribution of population by age and sex; geographical area; and tenure, as the weights from the full household sample (at Stage 6).

49. At Stage 12a, the averaged adjustments made to each tenure by type by geographical area group of household weights were then applied to the corresponding dwelling weights, including those for vacant dwellings. This ensured that the dwelling and household weights resulting from this stage were consistent for each dwelling case while ensuring that the household weights produced the same weighted totals as those for the full household sample.

50. Dwelling totals split by tenure and geographical area are used to weight the EHS data. As data for the required time point and tenure split were unavailable at the time of the 2011-12 weighting process, earlier data were projected forwards to the control point of 1 October 2011.

51. The results were then used as control totals to weight up the survey data to provide national estimates of dwellings for the survey year.

Calculating two year weights

52. Because of the smaller annual sample sizes involved, analysis of the dwelling sample is normally carried out using 2 years' weighted data. This section sets out how combined weights were calculated for the '2011' dwelling sample.

53. The two-year dataset was produced by combining data from the 2010-11 and 2011-12 EHS. As in previous years, the two datasets were combined after the sampling and response adjustments, and the combined dataset was then calibrated to control totals as at 1 April 2011.

54. The 2010-11 and 2011-12 EHS datasets were individually reweighted to the 1 April 2011 reference point. These datasets were then merged together and the weights halved so that each dataset had equal influence.