Chapter 5
Data processing

Overview

5.1 This chapter outlines English Housing Survey (EHS) data processing procedures and gives information about the main derived variables and data outputs. The EHS has a number of quality assurance measures in place which are undertaken throughout the annual survey process, beginning at the point of data collection, both through the computer-aided personal interviewing (CAPI) system and through surveyors validating their forms using the online system developed by BRE (details below). As the data are collated, processed and modelled, additional validation procedures are undertaken.

Editing

Interview data

5.2 The CAPI program has numerous built-in checks for identifying obvious discrepancies so that they can be resolved by the interviewer during the interview. The discrepancies are resolved by either correcting a data entry error or by clarifying a response directly with the respondent. The CAPI checks include:

- range checks – e.g. if an unusually high/low weekly rent is entered
- conflicting answers to different questions – e.g. if the number of years living in the current accommodation is greater than the respondent's age

5.3 There are two types of checks:

- hard checks – where the interviewer cannot continue with the interview until they have changed the data entered in some way to remove the inconsistency. Hard checks are used when the inconsistency is impossible as with the example of the number of years living in current accommodation being greater than the respondent's age.
- soft checks (signals) – where the interviewer is told about the error but they can ignore it and move on to the next question. Soft checks are used when an answer is unlikely but not impossible, e.g. if a respondent says they have more than 5 bathrooms. These checks are used to get the

interviewer to confirm that the answer is correct and is not a data entry error, checking the answer with the respondent if appropriate.

Physical survey data

5.4 For the physical survey, a system of automatic data validation was introduced in 2008 as part of the move to using digital pens to collect the data. The process is subject to continuous development and operates in three stages.

5.5 First, a large number of checks are built into the EHS surveyors' website as surveys are uploaded. These include:

- range checks – where the entered answer falls outside a valid range of responses
- logic checks – where a combination of responses to certain questions are not logically consistent (e.g. to check that the sum of 'tenths of area' across rows added up to ten)
- consistency checks – to determine whether linked responses in different parts of the form are consistent with each other (e.g. that detailed room data is only entered where a room coded as existing), and
- plausibility checks – to determine whether a response is reasonable given that there is not a well-defined range of possible answers (e.g. ceiling height of a room entered as 24 metres instead of 2.4 metres)

5.6 Surveyors also visually check all pages to ensure that the digital pen entries mirror those on the paper form i.e. that handwritten numbers have not been misinterpreted by the software.

5.7 Second, the CADS Housing Surveys Regional Managers check the data and where necessary discuss with surveyors to agree on a final set of responses.

5.8 Once all EHS physical surveys have been submitted by the surveyors for the survey year, BRE undertakes further consistency and plausibility checks on the raw physical survey data. The purpose of these checks is, firstly, to detect and eliminate certain logical inconsistencies that would cause problems for modelling and, secondly, to identify highly implausible answers, which if deemed necessary after investigation, are corrected. In some cases the raw EHS physical survey data are altered following these consistency and plausibility checks as outlined below.

- Levels checks – data may be inconsistent with regard to the number of storeys in the building, and the floor occupied by the dwelling. The BRE checks test for the following possible errors:

    - a room on a floor that does not exist (e.g. 3rd floor of a three storey block, the 3 floors being recorded as ground, first and second)
    - a room on a level that is not part of the flat (e.g. room on the 3rd floor but flat on the 2nd floor)
    - a measured floor that is not part of the block (e.g. dimensions for 3rd floor when the dwelling only has three storeys)
    - a flat on a level that does not exist (e.g. flat on the 3rd floor when the module only has three storeys)
    - presence of a habitable attic/basement is inconsistent with the number of floors

- Plausible dimensions – checks are carried out on the dimensions, to identify any floor area that seems too large or too small. Where a reliable measurement is missing, BRE will attempt to work out the data from any measurements thought to be correct, or failing this by estimating the dimensions as best as possible from the photographs.

- Non-permissible values – on rare occasions a surveyor response may happen to be equal to a value that is reserved for special purposes. The numbers 77, 88 and 99 are reserved to indicate the section not applicable, question not applicable, or unknown. When these figures occur as real measurements or counts, they are reduced by one.

- Incorrect number of flats – the dimensions of the surveyed flat are checked against the total floor area of the survey module to identify if the number of flats per module seems realistic.

- Incorrect roof type – certain roof types (chalet and mansard) can only occur where the dwelling has an attic. On occasions surveyors may mistake steep pitched roofs for chalet roofs. In this situation, the data for pitched and chalet roofs are swapped over.

- Implausible wall and window areas/fenestration ratio – where a dwelling seems to have a wall or window area/fenestration ratio that is either too high or too low the data are checked. The surveyor’s judgement is deemed correct unless there is clear evidence (e.g. from photographs) to amend the data.

- Wall thickness – cases are identified where the wall thickness as measured by the EHS surveyor is not typical of the wall selected i.e. cases where the EHS surveyors’ website has triggered a wall thickness range check. Each case is checked by viewing the EHS surveyors’ website and looking at the details recorded on the physical survey form in conjunction with the photos/EHS surveyor comments. Based upon the information gained, the action is decided upon for each case. This could be no action required or it could be that the physical survey data looks incorrect, either the wall thickness value or the way the surveyor has coded something as wall that should not be counted as wall. Where required, the appropriate modifications are applied to the physical survey data.

- Heating system consistency checks – cases which contain inconsistent heating system data on the physical survey form are flagged in the validation process at BRE. Each case is checked by returning to the raw data; in cases where alterations can confidently be made, the data are modified accordingly.

Comparison edits

5.9 A further important quality check involves comparing interview survey data with the corresponding physical survey data for each case. The first step is a series of global edits to resolve particular discrepancies in the data, e.g.

- If tenure in the interview survey (IS) = owner occupied AND tenure in the physical survey (PS) = another tenure then the PS tenure was changed to owner occupied.
- If tenure in the interview survey (IS) = renting from local authority AND tenure in the physical survey (PS) = another tenure then the PS tenure was changed to renting from local authority.

5.10 The remaining discrepancies between the two parts of the survey are flagged, investigated and recoded where applicable. This process is carried out in order to:

- check that the correct sampled dwelling was visited at both the interview survey and the physical survey, and
- correct any inconsistencies in key variables (e.g. tenure or property type) between the two different parts of the survey. Where possible other information from the survey (e.g. other variables, interviewer's and surveyor's comments, photo of the property) is checked to help decide what information is correct.

Houses in multiple occupation (HMO) edits

5.11 An HMO is a property rented by more than one person who are not from one 'household' (e.g. a family) but share facilities like the bathroom or kitchen. These differ from a shared house in that the residents generally have separate tenancy agreements and usually have begun their tenancies independently of each other. The identification of HMOs is critical in order to help ensure the accuracy of the weighting for the sample dwelling. The procedure for monitoring, reconciliation and validation of cases which have been flagged as HMOs by NatCen Social Research interviewers and/or CADS Housing Surveys surveyors is described below.

5.12 Cases are flagged as HMOs depending upon responses to certain key questions in the household questionnaire. Interviewers are trained in applying the EHS household definition and assessing the type of occupancy in complex situations, particularly in making the distinction between a group of sharers forming one household and separate households sharing facilities. Where necessary, reference is made to a check list of supplementary questions on the HMO Rules Card issued to interviewers (Annex 5.1) to help determine whether an address should be classified as an HMO.

5.13 Where the responses to the interview questions lead to the dwelling being flagged as an HMO or possible HMO, and the dwelling is eligible for a physical survey, the CADS Housing Surveys Regional Manager is notified. The Regional Manager will contact the interviewer to discuss the layout and occupation of the premises. The purpose of this contact is twofold:

- to confirm, as far as possible, that the address is an HMO for EHS purposes, and
- to determine whether the case is one that should be visited by the Regional Manager personally, as a complex HMO, or whether it should be allocated to a surveyor.

5.14 There will be occasions when a physical surveyor considers that a referred address appears to be an HMO despite not being flagged as such by the NatCen Social Research interviewer. In such cases, the surveyor will treat the case as an HMO, and a reconciliation process is applied to the interview and physical data during the final data validation stage.

5.15 CADS Housing Surveys Regional Managers compile and maintain a database of all cases they know to be HMOs. These cases, along with cases flagged as HMOs at the interview survey but which did not have a subsequent physical survey, are reviewed by BRE for data validation as part of the comparison edits process. The HMO checking process also includes cases that were not identified as HMOs at interview survey but which the data suggest could potentially be HMOs. BRE checks relevant interview and physical survey data such as number of households (NumHhld) and number of accommodation units (AcNumber). Where there are inconsistencies further investigation is undertaken and the data altered to the correct values.

5.16 During the HMO comparison process BRE also derives the ratios of addresses to dwellings and dwellings to households. This information is required to ensure the correct numbers of dwellings and households are used in the production of weights. As part of the QA process, DCLG conducts spot checks on these ratios as well as the HMO edits resulting from the process above.

5.17 A record of all address changes are kept by interviewers and/or Regional Managers for HMO cases as part of a comprehensive system for recording address changes for all issued cases. This feeds into the address file supplied to DCLG at the end of fieldwork.

Coding

5.18 After the interview, the data are coded and edited by trained coders and editors at NatCen Social Research. An edit programme is utilised by these staff to code open answers and back-code responses as appropriate. For example, at the interview, respondents are asked how they pay for their electricity (question HmpyElec2), and the respondent is shown eight possible answers (e.g. direct debit) on a card. If their payment method is not on the list the interviewer will code ‘other’ and is asked to enter the details of the payment method at a follow up question (Hmelothr). After the interview, the coder will look at the details given at Hmelothr and check it against the eight answer codes to see whether it could be classified as one of these payment methods and if it can they will change the answer as appropriate (i.e. backcode the answer). Job details are coded to the Standard Occupation Classification (SOC) and the Standard Industry Classification (SIC).

5.19 Errors detected by the edit programme are resolved by referring back to the original questionnaire documents by experienced editors. Individual corrections are made to the data and the corrected data are rerun through the edit programme until it confirms that the data are clean. Queries arising from the coding and editing process are recorded in a standardised way and these are examined by the supervision team on completion of each batch of work to ensure that they have been carried out correctly.

5.20 After the coding and editing stage further internal consistency checks on the data are carried out by a data manager and the data are corrected where appropriate.

Derived variables

5.21 Derived variables are created either by simply recoding a particular survey question or by combining the information collected from a number of questions, which can involve complex modelling. Examples of basic derived variables include dwelling age and dwelling type and examples of complex derived variables include basic repair costs, usable floor area and energy efficiency rating. The derived variables and geo-demographic variables, such as region, rurality and Index of Multiple Deprivation, included in the key EHS derived datasets interview.sav, physical.sav and general.sav can be found in Annex 5.2.

5.22 In addition to the three key EHS derived datasets, further detailed derived files such as actual costs.sav, energy performance.sav and HHSRS.sav are available on the EHS database, as listed in Table 5.1.

5.23 Further details on the derivation of these derived and detailed variables are available in the EHS Data Dictionary, made publicly available via the UK Data Archive (http://ukdataservice.ac.uk/).

5.24 The EHS derived variables are included in the datasets deposited at the UK Data Archive. To comply with the data disclosure control guidance issued by the Government Statistical Service, some of the variables are released under the more restricted Special Licence rather than through the End User Licence. The further detailed derived files are also available only via the Special Licence. In addition, the very disclosive geo-demographic variables (local authority and postcode) are available only through the Archive's Secure Data Service.

Modelling

5.25 The derivation of some of the derived variables involves complex modelling. A detailed description of how the more complex derived variables are defined and modelled is covered in Annexes to this chapter:

- Annex 5.3: Accessibility indicators
- Annex 5.4: Household derived indicators
- Annex 5.5: Housing conditions
- Annex 5.6: Energy efficiency
- Annex 5.7: Dimensions
- Annex 5.8: Poor quality environments

Imputation

5.26 As part of the modelling processes, it is sometimes necessary for any missing data to be substituted with imputed values. The imputation of missing data is more prevalent with the interview survey data than the physical survey data. This is because the interview survey data are based on information provided by the householder who can choose to refuse questions or who may not know the answer to particular questions leading to missing values. The physical survey data are based upon a physical inspection of the property and there are only a few sections of the physical survey form where the trained surveyor can select 'information unknown' as an option; the most notable is the loft inspection, where surveyors cannot always obtain access.

5.27 Imputation of data also takes place in the modelling of derived variables where a value provided in the raw data falls outside of consistency/plausibility checks. Such values are interrogated and only changed when it is almost certain that the data are incorrect. See Annexes 5.3, 5.4, 5.5, 5.6, 5.7 and 5.8 for further details.

5.28 Examples of imputation that occurred in the modelling of EHS 2014-15 derived variables are as follows (figures are based upon weighted data):

- In the modelling of the derived variables from the EHS 2014-15 interview survey, 40% of the full sample had some form of income imputation (the highest imputation rate of all of the derived variables due to the sensitive nature of the questions), 13% of renters had weekly rents imputed and 14% of households with a mortgage had their weekly mortgage payments imputed. These imputations were due to a combination of missing raw data and implausible values. The 40% figure for imputation of income includes any change to any component of household income. This may only be to change the amount received from a particular benefit by a very small amount, which would not significantly impact the total household income.
- In the modelling of derived variables from the EHS 2014-15 physical survey on the dimensions of the property e.g. derivation of floor area, external wall area etc., a total of 140 cases in the paired single year dataset had some form of alteration to the raw physical survey data following consistency and plausibility checks on the raw physical survey data.
- In the derivation of loft insulation (which also feeds into SAP12 energy modelling and the modelling of Decent Homes), for the EHS 2014-15 single year paired sample, 9% of dwellings with a loft had a value for loft insulation imputed, due to either the property having a flat roof or because no access to the loft space was possible during the physical inspection of the property. This is the largest imputation rate in the derivation of the energy efficiency rating.

5.29 Where appropriate, the EHS Annual Reports contain details on the approach used to handle the cases that are missing from the raw physical and interview data during analysis.

Data outputs

5.30 A range of EHS datasets are produced annually and released via the UK Data Archive under the End User Licence or the Special Licence, Table 5.1.

Table 5.1 List of annual datasets

Physical datasets	**Interview datasets**	**Detailed derived datasets**	**Derived datasets (paired sample)**	**Derived datasets (full household sample)**
Available via the Special Licence only			**Available under both End User Licence and Special Licence**	
Amenity.sav	Adapt_hhld.sav*	Actual costs.sav	general.sav	generalfs.sav
Around.sav	Adapt_person.sav*	Dimensions.sav	physical.sav	interviewfs.sav
Chimney.sav	Adaptation_hhld.sav*	Energy performance.sav	interview.sav	
Commac.sav	Adaptation_person.sav*	Full and paired sample equivalised income.sav		
Common.sav	Attitudes.sav	HHSRS.sav		
Damppc.sav	Contact.sav	Standardised costs.sav		
Doors.sav	Disability.sav			
Dormers.sav	Dwelling.sav			
Elevate.sav	Employment.sav			
Firstimp_PS.sav	Energy.sav			
Flatdets.sav	Fire.sav			
Hhsrs.sav	Firstimp_IS.sav			
HQ.sav	HhldType.sav			
Interior.sav	Identity.sav			
Introoms.sav	Income.sav			
Numflats.sav	Owner.sav			
Plotlvl.sav	People.sav			
Roofcov.sav	Renter.sav			
Rooffeat.sav	Rooms.sav			
Roofstru.sav				
Services.sav	Vacant.sav			
Shape.sav	WaitList.sav			
Shared.sav				
Structure.sav				
Wallfin.sav				
Wallstru.sav				
Windows.sav				

* The 2014 adapt and adaption interview datasets are located on the full sample files only.

5.31 The data, user guides and supporting documentation are publicly available from the UK Data Archive (http://ukdataservice.ac.uk/). Datasets can be downloaded in SPSS and SAS format.

5.32 Prior to releasing the data in the UK Data Archive, all disclosive variables are removed to maintain the confidentiality of respondents. Some response categories are also condensed, several variables are top coded, and, in a few rare situations, data swapping between cases takes place for disclosure control reasons.