Item Edit and Imputation: Evaluation Report

June 2012

This is one of a series of reports published to support the release of results from the 2011 Census. This series of methods and quality reports provides information on the different methods used to collect, process, clean, adjust and protect the census results. The series also reports on the quality assurance of the results and provides quality indicators.

Terms used in the series are explained in the 2011 Census glossary.

Contents

2011 Census item edit and imputation process

1 Executive summary

This report evaluates the 2011 Census Item Edit and Imputation project. It focuses on the processing of the England and Wales data; however the methodology was also applied to the Northern Ireland and Scotland census data, with small modifications to account for differences in the questionnaires. The report covers two aspects: firstly, a statistical evaluation of the imputation process with an assessment for each question treated. This is an assessment of the distributional performance of the imputation system, sources of possible bias and how the process addressed these. Secondly, the report aims to give an operational evaluation of the effectiveness of the project in implementing the agreed imputation strategy and meeting the overall objectives.

The project commenced in 2005 with an evaluation of the 2001 Census imputation, continued through development and live operation phases and completed with this evaluation report. The project was successful in meeting the overall aims and objectives set out, and the quality of item imputation in the 2011 Census was of a high standard with very few issues identified. These issues have been posed as areas for future research which may be used to build on the success of this project and form the basis of improvement for future large-scale imputation projects.

A new tool was adopted to implement edit and imputation in 2011 and this proved to be effective for large-scale statistical processing. The 2011 system was both faster and more effective at meeting the aims of the imputation strategy than its predecessors (for example achieving a higher rate of joint imputation). The use of separate algorithms to edit the relationship matrix question was a new approach and proved effective in addressing the response errors in the relationship matrix and improving the ratio of donors to failed records.

Another major change to the method was the move to simultaneous consistency editing and statistical imputation, rather than having a dedicated deterministic editing process prior to applying imputation, and subsequent consistency editing after imputation. Overall this was an efficient and effective processing strategy. However, three deterministic edits were added during processing for items associated with systematic non-response or response errors in order to improve the overall effectiveness of the imputation. A further possible edit to set activity last week to ‘no code required’ for those under 16 years old was identified during the evaluation. It is recommended that missing observations which can only have one outcome are deterministically edited prior to imputation. This would avoid imputing a different outcome that would also require another observation to be changed, and help to maximise the donor pool. An evaluation of known response errors could provide a list of deterministic edits to consider when analysing the live data and the capability to run a selection of deterministic edits should be built into the automated process for all data-driven statistical processes.

Similarly, 2011 was the first time that soft edits, which exclude certain characteristics from the donor pool, were omitted from the imputation process. This worked less well than

expected. Two soft edits were subsequently added to the system, and there was a proliferation of rare characteristics in the data, for example there was a disproportionate increase in the number of full-time students below the age of four. This was partly due to a shortcoming of the version of the system being used, and partly due to a failure in the diagnostic checks to detect the movements at the national level. Further testing of soft edits is required for future projects, in particular, whether deterministic edits might be required and whether the problem could be prevented by using reordering of persons within donor households.

Because imputation examines the multivariate relationships in the data in a very detailed way, it is a powerful tool for assessing the overall coherence of the data. The new 2011 system proved to be very useful in terms of identifying systematic coding errors and response errors in the data. Issues with earlier processes were also detected by the imputation process, which was data driven and therefore sensitive to unexpected values or distributional properties. It would therefore be beneficial in future projects to allow time in the processing schedule for further data cleaning after the first application of the imputation system. Separate time should also be allowed for parameterisation and tuning of the methods and systems

2 Project purpose and strategy

The purpose of this project was to develop and implement item-level edit and imputation for the 2011 Census. The process can be divided into two parts: deterministic editing and statistical imputation. Deterministic editing uses rules to deduce a valid value based on other information provided by the respondent; statistical imputation uses explicit or implicit models to estimate a valid value based on the characteristics of the respondent and the distributions in the data. Edit and imputation aimed to resolve non-responses and inconsistencies within the data obtained from completed or partially completed questionnaires. It did not address complete non-response where no questionnaire was returned. There were two main reasons for applying item editing and imputation. Firstly, it was desirable to provide a complete and consistent database for future research and analysis. Secondly, it provided an opportunity to adjust for non-response and inconsistencies, which can lead to bias or inconsistency in resulting estimates. For example, bias occurs where the non-responses would not have the same distribution as the observed values for an item (Durrent, 2005[1]).

Item editing and imputation addressed:

- **item non-response** - all responses that were missing or not valid, including, multi-ticks, out-of-range values and partially answered responses (for example in occupation and industry which were collected in multiple fields).
- **inconsistencies -** complete and valid responses that did not make sense in relation to other responses on the questionnaire, auxiliary information, or definitions. These were detected with pre-defined edit rule checks that compared values in different items and are also referred to as **edit failures**.

The population base for item imputation in 2011 was defined as:

- all responding individuals and households, including:
    - ‘dummy’ household returns completed by census collectors where there was no household return;
    - short term residents who were staying in the UK for three or more but less than 12 months;
    - students who lived elsewhere during term-time (for age, sex, marital status, student, term time indicator and second address).
- excluding:
    - visitors;
    - residents and households added during the relevant census coverage adjustment process (England and Wales, Northern Ireland or Scotland).

2.1 2011 Census edit and imputation strategy

As in previous censuses, the primary objectives of the 2011 item editing and imputation strategy were to:

- produce a complete and consistent database; and
- adjust for non-response bias by estimating the unobserved observations (non-responses).

The following three key principles were adopted from the 2001 strategy:

- Impute all missing data (except the voluntary question religion) to provide a complete and consistent database.
- Minimise the number of changes to inconsistent data.
- Ensure all changes made to observed data maintain the quality of the data.

In adhering to these principles the following key aims were defined:

- Imputation should not introduce bias or distortion in the data.
- Imputation facilitates the production of output data that are fit for purpose.
- Imputation methods help to ensure that pre-determined levels of data quality are met, with the highest priority given to those variables which define the population bases.
- Imputation supports the production of the population estimates by ensuring that the basic population estimates are not distorted by non-response or edit and imputation.
- Imputation should minimise changes to observed data by avoiding the use of deterministic editing.
- Imputation should draw from the complete observed distributions by keeping records that would fail soft edit rules in the donor pool.

3 Project overview

The project consisted of four phases: development, implementation, live processing and evaluation. Each is briefly described here and outlined in more detail in the following sections.

- **Methodological development** began in 2005 with an evaluation of the 2001 method and consideration of new or alternative methods. In 2006, an alternative imputation tool, the Canadian Census Edit and Imputation System (CANCEIS), was assessed and endorsed to replace the 2001 census edit and imputation system. Several years of research followed before a working desk-top prototype was established.
- **Implementation** for the automated census processing system occurred concurrently with development from early 2009.
- **Live processing** began in 2011. There was an initial iterative period of adjustment and tuning before the method was finalised. This phase was completed early in 2012.

- **Evaluation** began several months later once all the census processes had completed. Evaluation focused on the statistical outcomes of edit and imputation, providing an assessment for each question treated, however this report also aims to assess the operational aspects of the process and provide recommendations for future statistical projects.

4 Governance

An internal working group was set up with members from the England and Wales, Northern Ireland and Scotland census edit and imputation teams, census statistical design, census quality assurance, and subject matter experts from ONS population demography and the University of Southampton. All aspects of the imputation design and quality monitoring were agreed through this UK-level group and reported through relevant management boards and the UK Census Design and Methodology Advisory Committee.

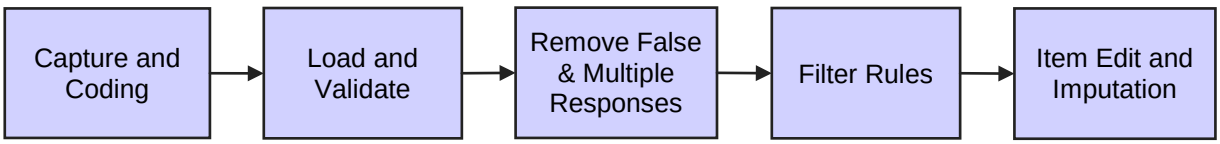
For complex concepts such as ethnicity and relationships, separate stakeholder engagement was conducted through attendance of census topic user groups and communication with external experts from academia and internal experts in population and demography subject matter areas. The working group commissioned an independent assessment of the relationships imputation methods from the ESRC Centre of Population Change at the University of Southampton (*UK Census Edit and Imputation Working Group paper (11)03*[2]).

Harmonisation and sharing with the devolved administrations was achieved through the working group, frequent ongoing communication and work-sharing.

5 Census processing overview

The edit and imputation process was automated and executed from within the automated census processing environment. Several earlier processes in this system helped prepare the data for imputation; these are shown in Figure 1.

Figure 1: Data processes for Census 2011



Firstly, at capture, questionnaires were scanned and complex coding was used to assign numerical values to written text and ticked boxes. This involved applying coding rules and standardised national coding frames, such as SIC07 (Standard Industrial Classification 2007) and SOC2010 (Standard Occupational Classification 2010), which allow data from different sources to be easily compared. The data were loaded into a database and validated to ensure that the values for each question were within the range specified in the relevant coding frame.

Next, false persons and multiple responses were removed. Multiple responses occurred when a household completed more than one questionnaire or recorded the same person more than once, while a false response was where there was not enough information recorded to identify a person. Following this, filter rules addressed inconsistencies in terms of the routing on the questionnaire, for example, setting a response to 'no code required' where the routing directed the respondent past the question. The data were then ready for item edit and imputation. Further information on the earlier processes can be found on the census pages of the ONS website.

6 Methodological development

It was desirable to base editing on that used in 2001 in order to maximise efficiencies where possible and to allow for comparability between the two censuses. With this in mind, an assessment of alternative methods was used to agree the 2011 strategy, objectives and detailed methodology. The pre-processing steps of capture, range checks and filter rules were all based on those from 2001, and the imputation framework remained a donor-based method which replaces missing or inconsistent values with valid values taken from a 'donor pool' of statistically similar records.

Donor imputation methods are ideal for census data because they can impute several variables simultaneously, including categorical and continuous, and the nearest neighbour donor method adopted in 2011 has been shown to be a robust method for estimating the distributional properties of the data (Chen and Shao, in Durant 2005[3])

The 2001 System

In 2001, a bespoke Edit and Donor Imputation System (EDIS) was created which first ran a series of deterministic edits on the data and then used donor imputation to replace any remaining invalid values. There were 13.7 million deterministic edits applied to 11.8 million people; the most frequently applied edits were to amend qualifications, provision of care, travel to work, supervisor status, general health and country of birth. In addition, the marital status of those under the age of 16 was set to single (never married) if missing. The system first attempted a joint (household level) donor imputation to simultaneously impute all members of the household from a single donor household of the same size. Joint imputation is desirable because it maintains the between-person distributional properties of the data. However, a single donor could not always be found and alternative methods included individual person imputation, the replacement of whole failed households with clean households, and manual imputation.

A full evaluation of the 2001 imputation method is available in the 2001 Census Edit and Imputation Evaluation Report[4]. While the results of the 2001 imputation method were of a suitable quality, a number of areas for improvement were identified. In 2001 only 34% of the population was imputed jointly with their household, losing some of the between-person information in the data, and over one million people were imputed using more than one method. In addition, each data block of around half a million people took 48 hours to be processed. With more data expected to be processed in 2011, a faster donor-based edit and imputation system was required that could deliver a higher rate of joint imputation.

The 2011 system

The Canadian Census Edit and Imputation System (CANCEIS) was identified as a possible option to replace EDIS. This system was developed by Statistics Canada specifically to impute census data and is executable from a range of platforms, making it appropriate for use in a large-scale statistical production environment. However, it should be noted that CANCEIS operates from a command line and must be manually integrated into a production system.

CANCEIS not only applies joint imputation, using a single-donor household for all members of a failed household, but also simultaneously imputes for both non-response and inconsistencies. The consistency editing, based on edit rules, identifies combinations of values that are not allowed and marks these for imputation as well as ensuring that all imputed values satisfy the edit constraints. Another major advantage of CANCEIS is that it seeks to minimise the number of changes required to repair a record when edit constraints are in place, thus minimising changes to observed data. This is achieved by ranking the suitable donors by their distance from the failed record, and the number of changes that would be required to repair the record, and giving the closest and minimum change donors the highest probability of selection. An evaluation of CANCEIS for the UK census was conducted by Wagstaff and Rogers (2006[5]) based on 2001 census data. This showed promising results and CANCEIS was later endorsed by the census project board for use in the 2011 Census.

Further information on CANCEIS is available in the Edit and Imputation Process[6] report; Bankier[7] (1999), Bankier et al[8]. (2000) and de Waal et al[9]. (2011).

6.1 Imputation process design

Once the tool for imputation was selected, development focused on designing a processing strategy. This included deciding how the edit rules would be implemented, and which variables should be imputed together. The edit and imputation process is fully described in the Edit and Imputation Process[6] report. The following provides an overview of the method.

6.1.1 2011 Census edit constraints

Pre-imputation deterministic edits

The 2001 imputation system included two separate editing processes as well as a process to impute for missing values. A deterministic editing process was run prior to imputation, and a consistency editing process followed the missing value imputation. The consistency editing process was not required in 2011 because CANCEIS edits inconsistencies and imputes non-response simultaneously. The results of early tests with 2001 census data were very positive and indicated that simultaneous imputation and editing would be effective. The CANCEIS algorithm minimises changes to observed data while reflecting the observable properties of the data within the imputed values. This is more favourable than hard programming edit rules which use a single predetermined value with no variation for differing characteristics. Based on this, a decision was taken by the working group not to apply a deterministic editing process prior to imputation.

Consistency edit rules for simultaneous editing

In 2011 there were 30 edit rules used within the imputation. These were largely based on the 2001 edit rules, but updated through consultation with users and subject matter experts. For example, the 2001 rule that did not allow same-sex couples was removed and replaced with rules that said married couples had to be of opposite sex and civil partners had to be the same sex. The edit rules implemented during imputation are provided in Annex A. There were two types of edit rule used in CANCEIS:

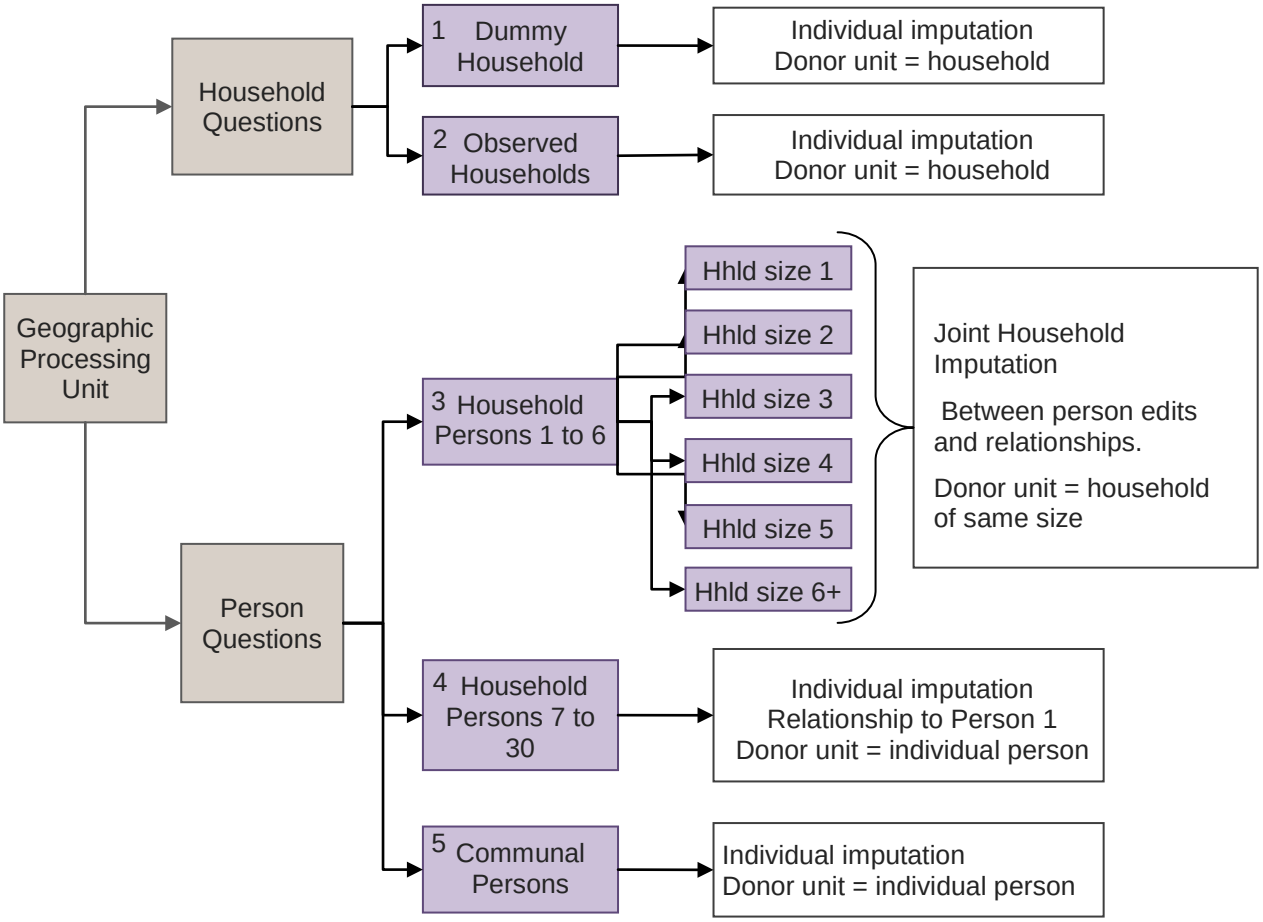
- **Hard edit rules** checked the plausibility of data and led to imputation if the record failed one of the rules. For example, 'a parent must be at least 12 years older than their child'. In addition, the questionnaire routing filters were implemented as hard edit rules to ensure that only persons required to give a response were imputed to have one.
- **Soft edit rules** were used to exclude records that failed the rule from being used as a donor, for example the rule 'it is unlikely that a person aged under 16 is a parent or partner' was used to exclude persons with this characteristic from the donor pool.

6.1.2 Imputation processing groups

As in 2001, the data were split into 101 geographical regions, or processing units (PUs), for delivery from the questionnaire processing supplier. In 2011 these contained on average 241,000 households with 530,000 person records and were processed separately through imputation. However, the data within each PU were still too complex to impute in one run. For example, there were household data and person data that required a different type of imputation. Furthermore, it would have been difficult to find an imputation model that could accurately estimate all the variables simultaneously. So the variables within each PU were further divided into separate modules for imputation.

The modules were based on several factors: the order of the questionnaire, the questionnaire routing, the priority of each variable, the inter-relatedness of the variables, and the edit rules. The advantage of imputing in smaller modules is that it is possible to only include questions that help to predict each other, and the number of available donors is maximised for a given group by allowing records to be donors even if they have a missing value in unrelated questions. The trade-off is that it takes longer to process the data and it presents challenges for maintaining the edit rules when the applicable questions are imputed in separate modules. For the 2011 census the data were divided into the five groups shown in figure 2, based on the type of imputation required.

Figure 2: Data groups and imputation types within each processing unit

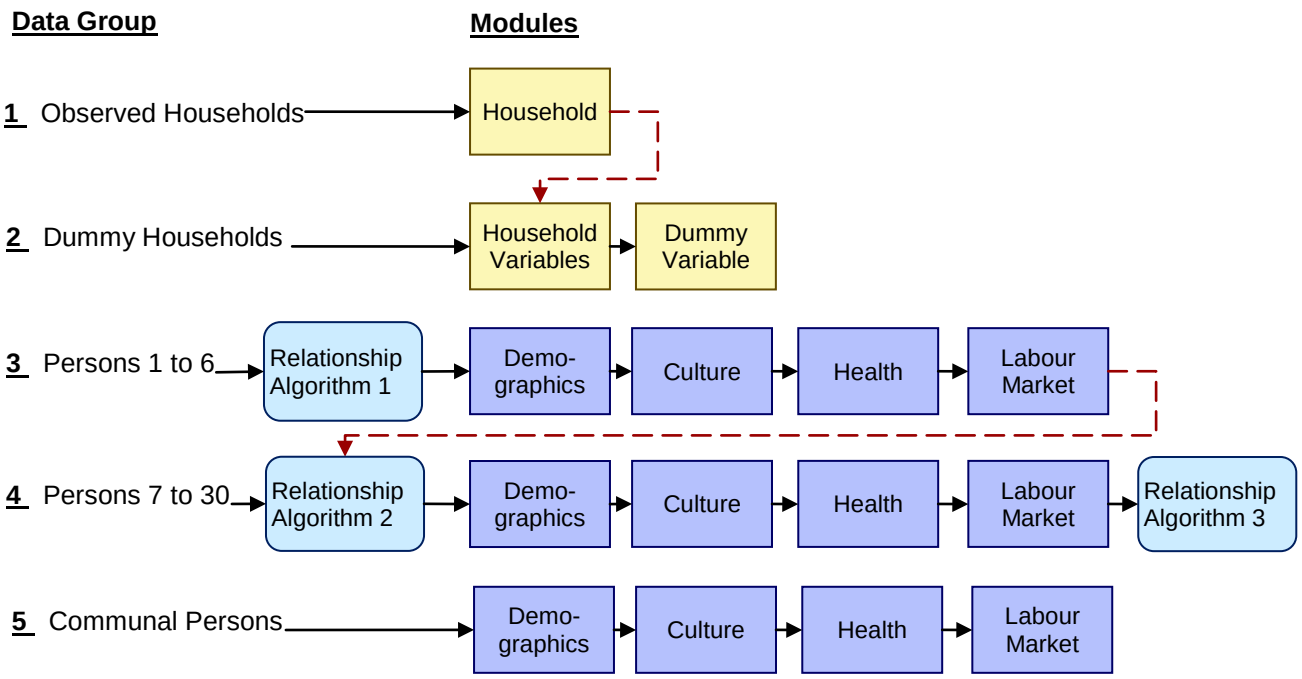


Groups 1 and 2 contained the household questions, where one response was required for the whole household. All of the dummy households (non-returned households where a census collector completed limited questions through observation) were imputed in group 1 and all of the observed households from returned questionnaires were imputed in group 2.

The next three groups contained the person data where a separate response was required for each person. In group 3, persons 1 to 6 collected on the main household questionnaire were imputed jointly, including their relationship to each other, using other households of the same size. In group 4, persons 7 to 30 collected on household continuation questionnaires were imputed as individuals, with only their relationship to person 1, and those in group 5 were imputed as individuals, using other persons from continuation questionnaires. Finally, persons from communal establishments collected on individual questionnaires were imputed as individuals, without any relationships, using persons within their own or a similar communal establishment.

CANCEIS was executed 15 times for each PU, in the sequence shown in figure 3. Each data group had one or more module, and where applicable the imputed data from the previous module were carried into the next. This allowed questions imputed in previous modules to be used in consistency editing or to find suitable donors in the subsequent modules. The modules used for each group are shown in figure 3.

Figure 3: Implementation of imputation groups in CANCEIS



Note that Figure 3 also shows the integration of three algorithms that were used for editing the relationship question. These were not modules in CANCEIS, but were applied to the data between instances. The relationship question was the most complex on the questionnaire and had the highest level of response errors. The purpose of algorithm 1 was to improve the consistency and completeness of the relationship question before imputation in order to improve the quality of the demographics module imputation. Algorithm 2 checked for consistency between persons 7 to 30 and any changes imputed for person one while algorithm 3 completed the imputation of the relationships between persons 7 and 30 because these were not included in CANCEIS. Algorithm 3 used triangulation rules to deduct a valid response. These rules triangulated the valid relationships for a missing value, based on what was observed in adjacent relationships. For example, if two persons were both the child of a third person, then they must be siblings or step-siblings. A scoring system, based on other observed properties of the relationships, was used to decide which relationship was most likely correct. For example, whether other step relationships had been recorded or there was more than one parent. Further information on the algorithms is given in sections 8.3, 8.4 and the Edit and Imputation Process[5] report.

The questions imputed within each of the modules are listed in Annex B. Observed households required only one module which contained all the questions about the accommodation. Dummy households required two modules: the questions in common with observed households were imputed first using additional donors from the observed household module and the dummy specific question, 'reason for dummy return', was then imputed using only dummy households as donors in the second module.

The persons data in groups 3 to 5 were divided into four modules for imputation. One of the aims of the imputation strategy was to give the highest priority to the questions defining

the population base, and to support the production of the population estimates. The key questions - age, sex and marital status - were treated first in the 'demographics' module with several variables that helped to predict them. These were student, relationship to person 1 and activity last week. These five variables were also connected by the edit rules, for example controlling the age at which you can work or be married, and the age difference between a parent and a child. Other variables included in this module were term-time indicator (important for defining the resident population base) and second address which was included here in order to treat all the variables that came before the term-time questionnaire filter. The next module, 'culture', contained variables describing the respondents' background, for example; ethnicity, country of birth, language, intention to stay, and passports. The third module contained the health-related questions and the final module 'labour market' contained all of the questions on employment and qualifications.

The dotted line between the modules shows where records, or information, were passed from one data group into the next. For example, information about the first six persons in the household was passed into 'Persons 7 to 30' in order to maintain consistency in the relationship data. The changes made by the relationship algorithms were also carried forward in a similar way.

Within each module a number of matching variables was selected to match the failed record with a donor. CANCEIS implements a complex distance function to give a statistical distance between each possible donor and the record which requires imputation. Those donors with the smallest distance are given the highest chance of being selected. There is also a maximum distance so that the system does not choose a donor which does not have anything in common with the failed record. The questions within each module were related and helped to predict each other. In addition, other key characteristics such as age, sex and country of birth group were used to identify appropriate donors. The matching criteria for each module was based on both priority outputs (like age and sex), and the planned categories for the 2011 Census outputs for each question within the module. This helped to estimate the unobserved distributions for key outputs. Each module also included geographic information and the data were sorted by this prior to imputation to give priority to donors that were geographically and statistically close to the recipient record.

6.1.3 Fallback methods

The imputation system was incorporated in a production environment where each PU automatically passed through all of the statistical processes. It was anticipated that around 5 to10 PUs would fail to impute automatically on the standard system settings and testing had shown that a small number of records would fail to impute in each PU (around 10 to 15 households). To ensure a complete and consistent database, fallback methods were required. These included:

- **Unique settings** – some PUs were run on individually specified settings, usually adjusting how many donors were considered for each failed record and how far away (geographically) the system would search for donors.
- **Manual imputation** - Records which failed to be imputed in successful PUs were treated outside of the automated environment by putting them back through CANCEIS with the observed and automatically imputed records from their PU. This meant that there was a small chance of using an imputed record as a donor. Occasionally

CANCEIS was executed with fallback settings which allowed the system to search further, and consider more donors. If this did not identify a suitable donor, persons in a household were imputed as a one-person household or expert judgements were made based on other information in the record, for example copying values from another household member or setting age based on the ages of a spouse or children.

7 Implementation and live processing

Due to the complexity of integrating the imputation system into the census processing environment, implementation needed to start before development had completed. Implementation included the following activities:

- specifying and acceptance testing for the derivation of the matching variables;
- specifying the input data requirements for CANCEIS, including formatting variables for CANCEIS and reformatting for the data base;
- delivering a prototype of the CANCEIS input files, and testing the implementation in the main system;
- negotiating requirements for fall back imputation methods, some of which needed to take place outside of the automated environment;
- designing and delivering a standard diagnostic tool for assessing the imputation of each PU in live processing;
- tuning the system once real data was available.

There was some overlap between implementation and live processing as the system was fine-tuned on the real data. Once the system was tuned, the data were automatically imputed in the live environment. The edit and imputation team only needed to address records that had failed to complete automatically. These were manually imputed in an offline environment using the fallback methods described in section 6.1.3. After any manual imputation each PU was signed off using the standardised diagnostic tool. Diagnostic checking was based on examining the pre- and post-imputation distributions in a similar way to the statistical evaluation given in the following section and annex C to G. This allowed the imputation process to be monitored to ensure that it was not causing any significant shifts in the proportional distributions of the output categories for any question. Increases or decreases of 0.05% (or more) were analysed to verify whether the change was related to the non-response mechanism. For example, there was a systematic increase in the proportion of persons with an activity last week of ‘retired’ because a disproportionate number of the missing responses were for those aged 65 or over. Changes of this type were accepted, while unwarranted changes were addressed through tuning or manual intervention. For example, during imputation students who lived elsewhere during term-time were sometimes changed to living at that address in order to meet the edit rules and these were put back to the original response with a deterministic edit.

The diagnostic process also provided an opportunity to check the quality of the data. For example by examining the response rates, edit failure rates and ratio of donors to failed records, a few systematic data issues were discovered which resulted in adjustments to earlier processes. More time was taken analysing the first few PUs and these formed a baseline for comparison when signing off subsequent PUs. The diagnostics also confirmed whether all of the hard edit rules were met in the final data, and monitored any change in the rate of soft edit rule failures.

8 Statistical evaluation

The evaluation phase of the project began once live processing was complete for all statistical processes and covered two aspects; a statistical evaluation of the imputation method and an operational evaluation of the execution of the project (an additional piece of work has looked at the possible variance introduced by the imputation process, which will be published separately). This section provides a statistical evaluation of:

- each question imputed;
- the relationship algorithms; and
- the implementation of the edit rules.

Overall the project was successful in meeting the main objectives and aims outlined in the strategy: a complete and consistent database was achieved with very few issues identified. One of the aims in 2011 was to improve on the performance of the EDIS system, for example by reducing the overall processing time and achieving a higher rate of joint imputation.

Table 1 compares operational results from each system. Despite a larger base population of 4.1 million people and a higher proportion of people requiring imputation in 2011, the average processing time for a PU in 2011 was up to four times faster than in 2001 indicating an overall gain in processing efficiency.

Of the 53.5 million persons who responded to the 2011 Census, 18.6 million (35%) required at least one question to be imputed. The increase in the rate of imputation from 28% (2001) to 35% (2011) is small considering that:

- there was a separate deterministic process in 2001 which applied millions of edits prior to imputation, and
- the system in 2011 ran editing and imputation simultaneously.

Although it is difficult to get a combined rate for 2001, it is likely that there was less editing in 2011 as imputation occurred in a single process rather than three separate processes.

In 2011 the majority of records (82%) were imputed in the automated environment jointly with their household, with the remaining 18% of people imputed individually being mainly single-person households, continuation questionnaire persons (7 to 30), and communal people for whom an individual imputation was planned. This is an improvement on the 2001 imputation which only imputed 34% with a joint imputation and thus captured less of

the implicit between-person distributional properties of the data. In 2011 these implicit distributions were accounted for completely in households of six or less persons, and between the first six persons of larger households.

Table 1: Operational comparison of CANCEIS and EDIS

	EDIS: 2001[a]	CANCEIS:2011[b]
Persons		
Person records processed	49.4 million	53.5 million
Average number of records in a processing unit	500,000	530,000
Average time to impute a processing unit	48 hours	12 hours
Persons needing at least one question imputed	13.8 million (28%)[c]	18.6 million (35%)
-Percent imputed as a household, taking into account multivariate joint distributions between persons and between questions	34%	82%
-Percent imputed as individuals	72%	18%
-Percent imputed using alternative methods to that implemented in primary imputation system	3%	0.10%
Persons imputed by more than one method	Over 1 million	Under 300
Households		
Household records processed	22.3 million	24.3 million
Households requiring at least one item imputed	2.5 million (11%)	2.8 million (9.5%)
Percent imputed taking into account multivariate joint distributions between questions	97%	100%

a. Census 2001 Review and Evaluation (ONS, 2003)
b. Data derived through the Census 2011 CANCEIS system diagnostics
c. Excludes overlap with deterministic applied to 11.8 million persons

The fallback imputation methods in 2011 were used for around 1/30th of the number of records treated with fallback methods in 2001. Manual imputation was limited to around 200 records and only 16 people were imputed separately from their household. Since the majority of records that were imputed using the fallback methods in 2011 were imputed using CANCEIS, only the 200 people imputed manually had more than one method applied, compared to over one million people in 2001.

There were almost 24.3 million household returns in 2011 and 2.8 million (11.5%) of these required one or more household questions to be imputed. All households were imputed by joint imputation with a single donor. This is an improvement on the 2001 household imputation which had a similar rate of imputation (2.5 million, 11%) and relied on fallback methods for around 3% of those households.

8.1 Assessing the performance of the imputation method

Imputation aims to estimate the distribution of non-response by taking into account differences between the characteristics of responders and non-responders (non-response bias). To be effective, the imputation model must reflect the underlying relationships between the different characteristics in the data. For donor-based imputation this translates to using appropriate characteristics in the donor selection model.

One way to assess the performance of the imputation is to compare the distribution of the observed values with the distribution of the imputed values. If the non-response was missing completely at random - that is, there were no differences between those who responded and those who did not - the imputed values would follow the same distribution as those observed. However, if there are differences between the people who responded and those who did not, then the imputed values should not follow the same distribution as the observed data. For example, in the 2011 Census the non-responders for economic activity were disproportionately of retirement age and the proportion of values imputed as 'retired' was higher than the observed proportion of 'retired' values. Although differences can be expected due to non-response bias, they should always be thoroughly checked because it is also possible for differences to arise due to misspecification of the matching variables or distance model. This was completed both during the tuning and diagnostic phase, and the evaluation for each question.

The other main diagnostic for assessing imputation is comparing the distribution of observed values to the complete distribution including the imputed values. Generally, imputed values should have a minimal effect on the distribution of the complete data unless the non-response rate is particularly high or the bias is very strong. In most cases donors were selected at random from the final donor pool of minimum change nearest neighbours, so there is some random variation in the distribution of the imputed responses and subsequently the complete data. However, early work on the imputation variance (to be published later) indicates that this variance is very small.

The following sections give an overview of the non-response, edit failure (inconsistency) and imputation rates. An assessment for each question imputed is provided in Annex C for the household questions and Annex D-G for the person questions. The tables for each question contain the observed, imputed and total distributions; highlighting where the imputed values had a different distribution to the observed values and whether this influenced the total distributions. In most cases, non-response was sufficiently small to prevent imputed values from causing a change to the observed distributions and investigation showed that the differences in the imputed distributions were due to the characteristics of the non-responders.

8.2 Household questions

This evaluation excludes the dummy returns, which were imputed separately. The household questions were completed to a high standard by the responding population, with low non-response and inconsistency (edit failure) rates. One of the main impacts on the quality of imputation is the rate of available donors compared to failed records. The rates for the household module, given in Table 2, are favourable; on average over 89% of records were available to use as donors and less than 1% had inconsistencies. Item non-response and imputation (Table 3) ranged from 2.3% (number of cars or vans and tenure)

to 3.6% (central heating). There was one edit rule for the household questions which stated that number of rooms must be greater than number of bedrooms, and there was a questionnaire filter between tenure and landlord. Less than 0.1% of observations for number of rooms, tenure and landlord were changed, and 0.2% of values for bedrooms were changed, due to failing the rule or the filter.

Table 2: Rate of donors and failed records for the household module

	Mean percentage per PU[1]		
	Donors	**Invalid[2]**	**Inconsistent[3]**
Household modules	89.4	10.0	0.6

1: Processing unit (geographical area, n=105)
2: At least one question with a blank, out of range, multi-tick or irresolvable value
3. Inconsistent, or inconsistent and invalid; failed at least one edit rule

Table 3: Household question item non-response, edit failure and imputation rates

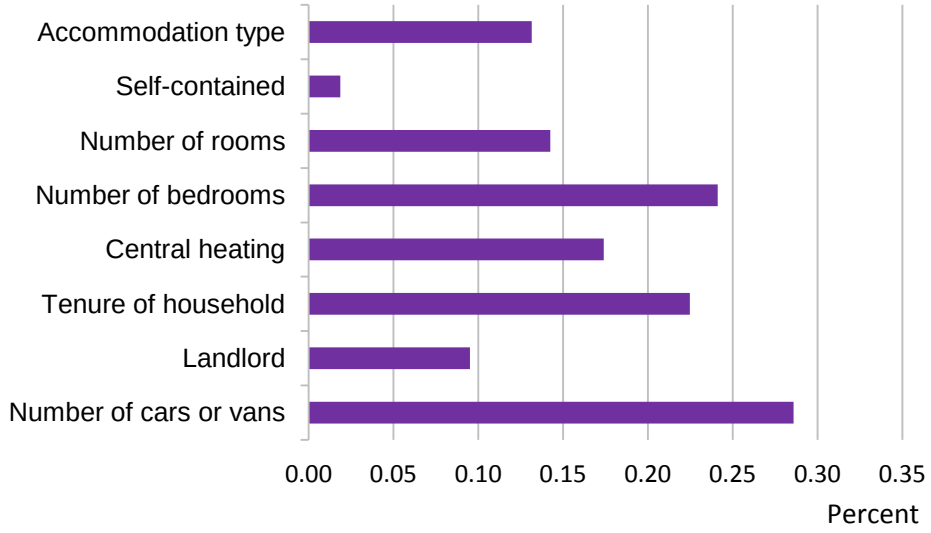
All eligible responding households England and Wales, 2011 — Thousands

	All N	**Non-response N**	**Edit failure N**	**Total imputed N**	***Non-response %***	**Edit failure *%***	***Total imputed %***
Type of accommodation[1]	22,877	583	<1	**583**	*2.5*	*<0.1*	***2.5***
Self-contained[1]	22,877	638	<1	**638**	*2.8*	*<0.1*	***2.8***
Number of rooms	22,877	710	12	**722**	*3.1*	*<0.1*	***3.2***
Number of bedrooms	22,877	600	45	**645**	*2.6*	*0.2*	***2.8***
Central heating	22,877	821	<1	**821**	*3.6*	*<0.1*	***3.6***
Tenure of household[3]	22,191	508	1	**509**	*2.3*	*<0.1*	***2.3***
Landlord[4]	7,718	215	<1	**215**	*2.8*	*<0.1*	***2.8***
Number of cars or vans[3]	22,191	501	<1	**501**	*2.3*	*<0.1*	***2.3***

1. Excluding dummy returns
2. Households with at least one resident
3. Households with at least one usual resident and tenure is renting, part renting or rent free

Figure 4 shows the maximum absolute percent change between the observed and final (imputed) distributions for all categories within each of the household questions. An individual breakdown of each category is given in tables 16 to 23 in Annex C. For the household questions, there were only minor changes in the proportional distributions of the reported categories ranging from <0.05% (self contained) to 0.28% (number of cars or vans). An individual assessment of each question is given in Annex C.

Figure 4: Maximum absolute percent change for any category after imputation - household questions



8.3 Person questions

One of the challenges in designing a large-scale statistical process is that it needs to be data driven, but with a census the data is often not available until processing begins, and then only in smaller processing units over time. As a result of real census data not being available during development, some aspects of the planned methodology had to be adjusted and additional steps added to overcome unforeseen properties in the person questions data. The following changes were made during tuning and parameterisation:

Change 1: Deterministic editing

With the exception of the relationship question which was known to have systematic response errors, the editing and imputation strategy addressed non-response and inconsistencies simultaneously using statistical imputation. However, during tuning higher than expected levels of non-response and inconsistency were detected in the demographics modules. The resulting reduction in the rate of available donors was a concern in terms of the quality of the imputation and the number of households which were failing to impute automatically. Analysis identified two further sources of systematic non-response / inconsistency in the marital and civil partner status and position in establishment questions and further scope for extending the edits to the relationship question. As a result, the first relationship algorithm was extended and

deterministic edits were specified for the other two questions. Tests using these changes demonstrated an improvement in the underlying coherence of the data (less inconsistency) and the ratio of donors to recipient records.

1. **Marital and civil partner status edit**

   The majority of values that were blank for this question belonged to persons who had not recorded a spouse or civil partnership on the relationship question (but had recorded all their relationships). An edit was applied to set these persons to 'Never married and never in a civil partnership'. Most of these respondents were aged less than 16 years old.

2. **Position in communal establishments edit**

   In communal establishments, particularly those for the elderly or children, it was common for a staff member to complete the questionnaire on behalf of residents. Sometimes they erroneously answered the position in communal establishment question as 'staff' for one or more residents. To correct for this any communal establishment persons reported as 'staff' that had not subsequently ticked 'working' in activity last week were changed from 'staff' to 'resident' (including where activity last week was missing).

3. **Additions to relationship algorithm one**

   Deterministic triangulation rules which treated missing values in the third relationship algorithm were implemented in the first relationship algorithm and applied to the first six persons in the household. For example, if two persons in the household shared a parent, they were set to either siblings or step-siblings depending on the information in the adjacent cells. This treated all of the missing and inconsistent values where there was observed information in adjacent cells that could be used to deduce a valid response, and meant that statistical imputation only needed to treat relationships where there was no information available or the rules could not resolve an inconsistency.

The counts and rates for deterministic editing are shown in Table 4. Values changed in this process could subsequently be changed by the imputation process in order to resolve inconsistencies with other imputed values or other persons in the house. Therefore, adding the editing and the imputation counts would slightly overstate the total number of records with an edited/imputed value.

The rate of editing for the relationship question appears quite small (0.8%) because only relationship to person one is presented. However the relationship edits were applied to all of the relationships in the demographics module and relationships between persons seven to 30 were treated entirely with the edit process. The evaluation of the relationship algorithms is provided in section 8.4.

Table 4: Person variable deterministic editing rates

All responding persons*, England and Wales, 2011 — Thousands

	All** persons N	Edited N	Edited %
Marital and civil partner status[1]	53,483	1,359	*2.5*
Relationship to person one[2]	30,335	247	*0.8*
Position in communal establishment[3]	958	67	*7.0*

* Includes short term residents in the UK for less than twelve months and students not at their term-time address where applicable

** Equals the number of persons in scope for the question:

All persons (including students at home address)

Persons two to thirty in households, including students at home address

All persons living in a communal establishment on census night

Change 2: Restricting the demographics module donor pool for households with five or more persons

Despite the additions to relationship algorithm 1 and deterministic editing, the system continued to encounter problems when attempting to impute the demographics and relationships in households of five or more persons. The issue was a shortfall in the memory allocated to hold all the possible imputation actions for the identified nearest neighbours. In more complex households, the system ran out of memory when evaluating the donors, which caused the system to terminate. The strategy employed to prevent this was to limit the donor pool to the absolute nearest neighbour, and take the minimum change imputation action from this donor for all five- and six-person households. This enabled the majority of PUs to be automatically processed, but removed the part of the process where donors were randomly selected from a pool of the best possible donors within the five-and six-person household demographics imputation. The trade-off was losing a small amount of the variation in the imputed values and possibly using a donor more times than if selecting from a pool of donors. However, a comparison of the two approaches in a sample area found little difference in the resulting distributions of age, sex, marital status and economic activity.

Change 3: Manual imputation before running in the automated environment

In six of the PUs, using a deterministic imputation for the demographics module was not enough to ensure the system did not run out of memory. In order to be able to process these areas, the household causing the failure had to be identified and manually imputed prior to automatic imputation.

The evaluation of the person questions is presented in the four modules used for imputation; demographics, culture, health and labour market.

8.3.1 Demographics module

The demographics module included personal characteristics such as age and sex; second address; activity last week and relationship. All the questions up to the student term-time question were answered by every person; consequently students with another address during term-time were counted at both their home and term-time addresses for all questions except activity last week.

The rates for the demographics imputation in each data group is shown in Table 5. On average, all data groups had at least 72% of records available as donors, with a rate close to 4:1 in the main population (persons 1 to 6). While there were more persons-7-to-30 donors available, this was primarily because each record contained only one individual and did not include all of the between-person edit constraints.

Table 5: Rate of donors and failed records for the demographics module

	Mean % of records per PU[1]		
Data groups	**Donors**	**Invalid[2]**	**Inconsistant[3]**
Persons 1 to 6	76.9	20.7	2.5
Persons 7 to 30[4]	89.2	9.9	1.0
Persons in a communal	71.8	27.3	0.9

1: Processing unit (geographical area, n=101)
2: At least one question with a blank, out of range, multi-tick or irresolvable value
3: Inconsistent, or inconsistent and invalid; failed at least one edit rule
4: Individual level imputation - did not contain all the between-person relationships

The non-response and edit failure rates for each question are given in Table 6. The highest level of item non-response was for second address postcode (9%) followed by second address type (6.7%) while the key variables age and sex were very low at 0.6% and 0.4% respectively. This module had the most edit rules (25), as well as three routing filters. However, imputation due to inconsistency/edit failure was fairly small; on average, only 2.5% of the person-1-to-6 households had inconsistencies and even fewer had inconsistencies in the other data groups. The highest rate of edit rule failure was for second address postcode (1%), followed by term-time address indicator (0.5%).

Table 6: Demographics module: non-response, edit failure and imputation

All eligible responding persons, England and Wales, 2011 — Thousands

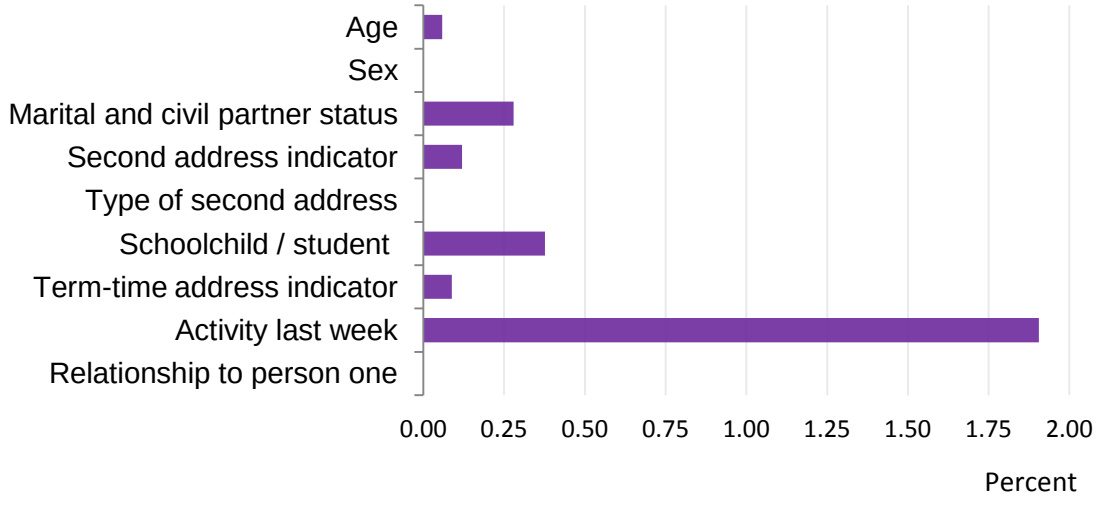
	All** persons N	Non-response N	Edit Failure N	Total imputed N	*Non-response %*	Edit Failure %	*Total imputed %*
Age[1]	53,483	319	81	**400**	*0.6*	*0.2*	***0.8***
Sex[1]	53,483	225	16	**240**	*0.4*	*<0.1*	***0.5***
Marital / civil partner status[1]	53,483	2,052	64	**757***	*3.8*	*0.1*	***1.4***
Second address indicator[1]	53,483	1,846	126	**1,972**	*3.5*	*0.2*	***3.7***
Second address country[1]	837	23	2	**25**	*2.7*	*0.2*	***3.0***
Second address postcode[1]	2,437	220	24	**243**	*9.0*	*1.0*	***10.0***
Type of second address[1,2]	3,274	219	9	**227**	*6.7*	*0.3*	***6.9***
Schoolchild / student[1]	53,483	1,745	62	**1,807**	*3.3*	*0.1*	***3.4***
Term-time address indicator[1,3]	11,607	159	59	**218**	*1.4*	*0.5*	***1.9***
Activity last week[4]	43,041	2,172	34	**2,206**	*5.1*	*0.1*	***5.1***
Relationship to person one[1,5]	30,335	1,203	125	**1,328**	*4.0*	*0.4*	***4.4***

* The shortfall between total imputation and non-response + edit failures was addressed by deterministic editing
** All persons eligible to respond:

1. All persons, including students at their home address
2. With a second address in the UK or abroad
3. In full-time education
4. Where aged 16 or over, excluding students at their home address
5. Where person two to thirty living in a household

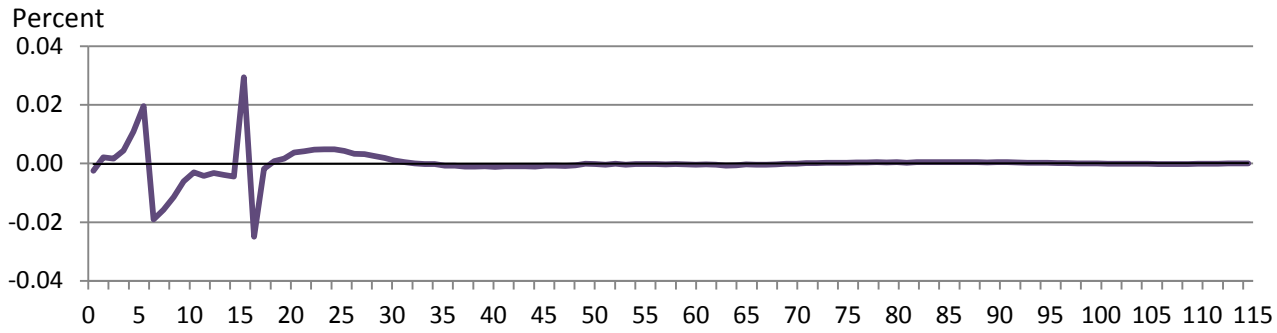
Figure 5 shows the maximum absolute percent change between the observed and final (imputed) distributions across all categories within each of the demographic questions. A breakdown for each category is given in tables 24 to 32 in Annex D. The amount of change was generally low; sex, type of second address and relationship to person 1 had the least change (<0.25%) while the largest change was for activity last week at around 1.85%. A full evaluation for each question in the module is provided in Annex D.

Figure 5: Maximum absolute percent change in any category post imputation – demographic questions



The majority of these questions had no - or just minor - issues, however there was an issue identified in the imputation of age. Figure 6 shows the difference in the proportional distributions for each year of age before and after imputation. While all of the differences were 0.3% or less, the majority of movement concentrated around the school age and working age boundaries.

Figure 6: Difference in the proportional distributions of single year of age before and after imputation



These ages were particularly affected by edit rules between student-age, working-age, marital/civil partner-age and parent-age. For example, age may have been changed in order to make the person old enough to work (> 15) or if they were under 16 they may have been made older because of a reported or imputed spouse in the house. For respondents aged between 10 and 14, just over half (52%) of inconsistencies were resolved by changing age, most often making the respondent older than 16. This resulted in a net decrease of 2,489 10 to 14–year-olds excluding values imputed from non-response (and subsequent adjustments).

Almost 34,000 ages were changed in the 15 to 19-year-old band due to edit failures. Around 17,500 of these were 16 to 19-year-olds with a missing value for activity last week who were matched to a 15-year-old donor. Because 15-year-olds did not answer the working questions, these records received ‘no code required’ for activity last week and the value of age also had to be changed in order to keep the record consistent with the edit rules. The non-response for activity last week was particularly high for 16-years-olds, who accounted for 80% of these cases. This resulted in a net increase of 14,061 15-year-olds and net decrease of 13,858 16-year-olds excluding imputation of non-response (and subsequent adjustments).

This movement had already been partly reduced by adjusting the weights in the distance model during tuning, however, there was a trade-off between heavily penalising changes to age, and maintaining small sub-populations, such as those with second address. The movement may have been further mitigated by applying deterministic editing for persons around the working-age boundary. For example, setting activity last week to ‘student’ if no employment information had been provided would have prevented the need to impute activity last week in the joint household imputation. It is also possible that being able to use a ‘reordering’ function in newer versions of CANCEIS would minimise this problem. Reordering allows the persons in the donor household to occur in any position in the record and makes it easier to find very close statistical matches. These are areas for future research for large-scale imputation projects.

8.3.2 Culture module

The culture module included questions about cultural characteristics such as country of birth, national identity, ethnicity, language and religion. It was answered by all persons except students who had another address during term-time. The mean rates of available donors for the culture module are in Table 7; on average, the household person imputation groups had more than 81% of records available as donors while communal persons had on average around 75%. For the main population (persons 1 to 6) there were approximately 4 donors for each failed record.

Most questions in culture had a non-response rate of less than 4% (Table 8), the exceptions were date of arrival to UK and intention to stay at 4.8% and 14.5% respectively. Religion also had a higher rate of 7.1%, but it was voluntary and could validly be left blank.

There were two edit rules and five questionnaire filters in this module. Inconsistency was generally low; on average less than 1.5% of records had inconsistent values. The highest rate of imputation due to edit rule failure was for arrival in the UK (1.9%) followed by address one year ago country (0.7%).

Table 7: Rate of donors and failed records for the culture module

	Mean % of records per PU[1]		
Data groups	**Donors**	**Invalid[2]**	**Inconsistent[3]**
Persons 1 to 6	81.5	17.1	1.4
Persons 7 to 30	96.7	3.1	0.2
Persons in a communal	74.7	24.4	0.9

1: Processing unit (geographical area, n=101)
2: At least one question with a blank, out of range, multi-tick or irresolvable value
3: Inconsistent, or inconsistent and invalid; failed at least one edit rule

Table 8: Culture module: non-response, edit failure and imputation

All eligible** responding persons, England and Wales, 2011 — Thousands

	All** persons	Non-response	Edit Failure	Total imputed	*Non-response*	Edit Failure	*Total imputed*
	N	N	N	N	%	%	%
Country of birth[1]	52,791	800	7	**806**	*1.5*	*<0.1*	***1.5***
Arrival in the UK[1]	6,858	326	128	**453**	*4.8*	*1.9*	***6.6***
Intention to stay[1,2]	594	86	6	**92**	*14.5*	*0.9*	***15.5***
National identity[1]	52,791	1,023	5	**1,027**	*1.9*	*<0.1*	***2.0***
Ethnic group[1]	52,791	1,595	21	**1,616**	*3.0*	*<0.1*	***3.1***
Welsh language[1,3]	2,861	96	1	**97**	*3.4*	*<0.1*	***3.4***
Main language[1]	52,791	1,328	41	**1,369**	*2.5*	*0.1*	***2.6***
Proficiency in English[1]	3,929	142	<1	**143**	*3.6*	*<0.1*	***3.6***
Religion[1]	53,068	3,759	<1	**<1**	*7.1*	*<0.1*	***0.0***
Address one year ago[1,4]	52,150	2,004	97	**2,101**	*3.8*	*0.2*	***4.0***
Address one year ago country[1,5]	605	22	4	**26**	*3.6*	*0.7*	***4.4***
Address one year ago postcode[1,6]	4,074	235	1	**236**	*5.8*	*<0.1*	***5.8***
Passports held (UK)[1]	52,791	1,222	33	**1,254**	*2.3*	*0.1*	***2.4***
Passports held (other)[1]	3,963	93	1	**94**	*2.4*	*<0.1*	***2.4***

** All persons eligible to respond:

1. All persons at, or without, another address during term time
2. Where arrive to live in UK was less than 12 months ago
3. Where country of residence was Wales
4. Where aged at least one year old
5. Where address one year ago was within UK
6. Where address one year ago was outside UK

Figure 7: Maximum absolute percent change in any category after imputation – culture questions

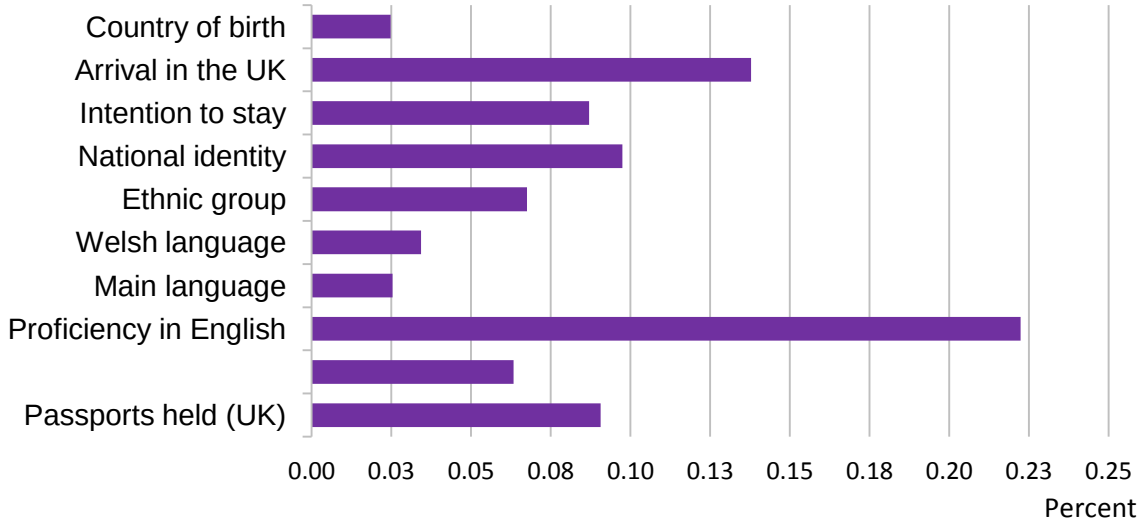


Figure 7 shows the maximum absolute change between the observed and final (imputed) distributions for any category within the culture questions. A breakdown of all categories is provided in tables 33 to 42 in Annex E. The adjustments to the distributions were all minor; country of birth and main language had less than 0.03% change while ethnicity had the largest change of 0.23%. A full evaluation for each question is provided in Annex E.

8.3.3 Health module

The health module treated the questions related to health and caring responsibilities. These were required for all persons except students who had another address during term-time. The mean rate of available donors for the health module was high; on average 88% to 90% of records were available as donors (Table 9) with moderate non-response ranging from 1.6% to 3.5% (Table 10). There was one edit rule and one questionnaire filter governing this module and inconsistency with these was low; less than 1% of households failed due to inconsistency and the edit failure rate was less than 0.1% for all questions.

Table 9: Rate of donors and failed records for the health module

	Mean % of records per PU[1]		
Data groups	**Passed**	**Invalid[2]**	**Inconsistent[3]**
Persons 1 to 6	88.0	11.2	0.9
Persons 7 to 30	96.7	3.1	0.2
Persons in a communal	90.1	9.5	0.4

1: Processing unit (geographical area, n=101)
2: At least one question with a blank, out of range, multi-tick or irresolvable value
3: Inconsistent, or inconsistent and invalid; failed at least one edit rule

Table 10: Health module non-response, edit failure and imputation

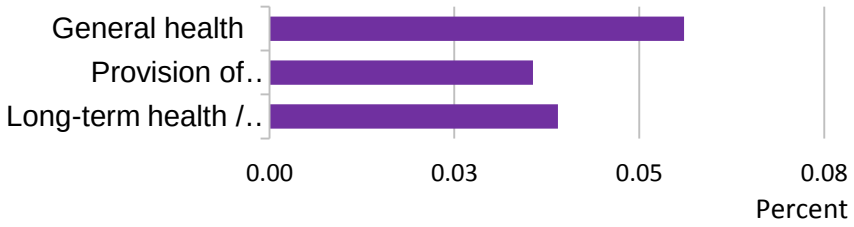
All eligible responding persons, England and Wales, 2011 — Thousands

	All* persons N	**Non-response n**	**Edit Failure n**	**Total imputed n**	***Non-response %***	**Edit Failure %**	***Total imputed %***
General health	52,791	853	5	**857**	*1.6*	*<0.1*	***1.6***
Provision of unpaid care	52,791	1,855	8	**1,863**	*3.5*	*<0.1*	***3.5***
Long-term health / disability	52,791	1,675	5	**1,680**	*3.2*	*<0.1*	***3.2***

*Excludes students living somewhere else during term-time

Figure 8 shows the maximum absolute percent change between the observed and final (imputed) distributions for any category in each of the health questions. A full breakdown of the categories is in tables 43 to 45 in Annex F. There were only very minor changes of between 0.03%-0.05% for the health questions; the full assessments are in Annex F.

Figure 8: Maximum absolute percent change in any category after imputation - Health questions



8.3.4 Labour market module

The labour market module contained all the questions about qualifications and working. It was answered by all persons over the age of 15, except students who had another address during term-time. There were three edit rules and three routing filters within the module. Generally, the rate of available donors for labour market module was lower; on average 64.1% of records were available as donors for persons 1 to 6 and persons in communal establishments (Table 11). Persons 7 to 30 had a high pass rate, but were more likely to be children who did not answer the labour module. The lower pass rates were due to some higher non-response rates, for example, workplace address (8.9% to 12.5%), last year worked (10.9%) and industry ever worked (17.2%). The latter two questions may have had a lower response rate because respondents had difficulty recalling the correct answer, while workplace address had a lot of missing postcodes because they were not validated prior to imputation.

The persons-1-to-6 labour modules also had the highest average rate of inconsistency failures (3.5%) which was primarily attributable to edit failures in the workplace address, transport and hours of work questions (Table 12). The inconsistency in workplace address was mainly due to the high number of postcodes set to missing before imputation, which disrupted the way that the three parts of the question were coded for imputation. The remainder of the questions had low to moderate non-response (1.8% to 7.2%) and low edit failure (0.0% to 0.8%).

Table 11: Rate of donors and failed records for the labour market module

	Mean % of records per PU[1]		
Data groups	**Passed**	**Invalid[2]**	**Inconsistent[2]**
Persons_1to6	64.1	32.5	3.5
Persons_7to30	94.9	4.4	0.7
Persons_Communal	64.1	34.0	1.9

1: Processing unit (geographical area, n=101)
2: At least one question with a blank, out of range, multi-tick or irresolvable value
3: Inconsistent, or inconsistent and invalid; failed at least one edit rule

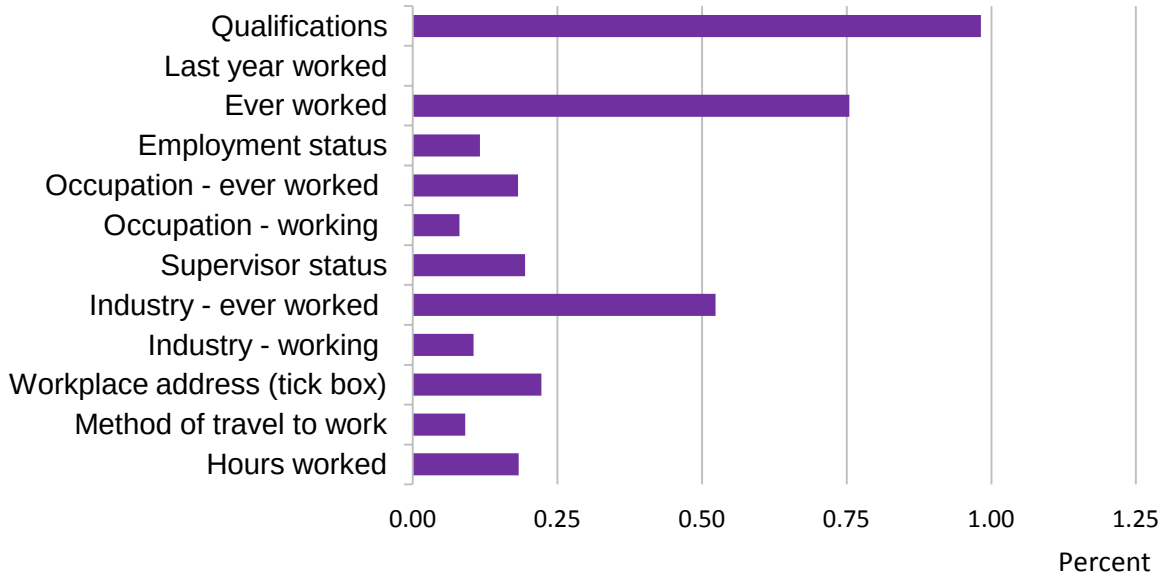
Table 12: Labour market module non-response, edit failure and imputation

All eligible responding persons, England and Wales, 2011 — Thousands

	All persons	Non-response	Edit Failure	Total imputed	*Non-response*	Edit Failure	*Total imputed*
	N	n	n	N	%	%	%
Qualifications	43,041	2,433	14	**2,447**	*5.7*	*<0.1*	***5.7***
Ever worked	17,787	316	143	**459**	*1.8*	*0.8*	***2.6***
Last year worked	14,433	1,569	45	**1,614**	*10.9*	*0.3*	***11.2***
Employment status	39,687	1,582	85	**1,668**	*4.0*	*0.2*	***4.2***
Occupation - ever worked	14,433	940	<1	**940**	*6.5*	*<0.1*	***6.5***
Occupation - working	25,255	578	85	**663**	*2.3*	*0.3*	***2.6***
Supervisor status	39,687	1,711	85	**1,796**	*4.3*	*0.2*	***4.5***
Industry - ever worked	14,433	2,481	<1	**2,481**	*17.2*	*<0.1*	***17.2***
Industry - working	25,255	1,813	85	**1,899**	*7.2*	*0.3*	***7.5***
Workplace address (tick box)	4,845	481	155	**635**	*9.9*	*3.2*	***13.1***
Workplace country	39	4	1	**5**	*9.8*	*3.1*	***12.9***
Workplace postcode	20,371	2,548	371	**2,919**	*12.5*	*1.8*	***14.3***
Method of travel to work	25,255	796	460	**1,256**	*3.2*	*1.8*	***5.0***
Hours worked	25,255	854	457	**1,312**	*3.4*	*1.8*	***5.2***

Figure 9 shows the maximum absolute percent change between the observed and final (imputed) distributions for all categories within each labour question. A full breakdown is given in tables 46 to 56 in Annex G. There were some slightly larger changes in the module: qualifications and industry for persons in work had a 0.5% change; while ever worked had a 0.8% change. These changes are explained in the full assessment of the questions in Annex G.

Figure 9: Maximum absolute percent change after imputation - labour module



8.4 Relationship algorithms

Combined with the statistical imputation of relationships to persons one to five, the relationship algorithms were successful in ensuring complete and consistent relationships data. The method for imputing relationships was endorsed by an independent review by The University of Southampton[2]. As explained in section 8.3 the triangulation rule based imputation from algorithm three was moved in to the first algorithm. This meant that only invalid values or inconsistencies that the deterministic editing could not resolve needed to be imputed in the automated method.

The change was primarily in response to a memory resource issue which occurred when running the demographics module. It was thought that improving the relationships data would help to free up memory resource by minimising the need to search widely for unique or complex household structures. Using the triangulation rules up front had the added benefit of increasing the number of available donors, however it also meant the imputation of relationships was predominantly deterministic. There were a few minor issues with the triangulation rules which were detected and re-coded during early processing.

This strategy was successful in improving the number of available donors and with other tuning activities ensured the successful imputation of each demographics module in the automated system. Because almost all of the imputation occurred prior to CANCEIS, the second algorithm - which checked for changes to person 1 during imputation - became redundant. Although it was no longer required, it was more efficient to let it run than remove it from the integrated system. This is why the figures for Algorithm 2 in Table 13 are low compared to the other two algorithms. Algorithm 1 did most of the work, editing 1.7 million people and 2.1 million values within those records. The third algorithm primarily edited persons 7 to 30, although some checks were made on all persons; just under 675,000 values were edited in almost 535,000 person records.

Table 13: Relationship algorithm edits

	Total persons edited	% of all persons	Total number of values edited
Algorithm 1	1,671,501	3.1%	2,133,902
Algorithm 2	70	<0.1%	74
Algorithm 3	534,975	1.0%	674,046

8.5 Consistency checks

The strategy for editing in 2011 was to edit for inconsistencies and impute missing values in one step where persons who failed the soft edit checks were kept in the donor pool. This allowed the donor pool to be maximised and for the process to be more data driven in terms of basing imputation on the full observed distributions. One disadvantage of the change was that the comparability to the 2001 Census was somewhat reduced. For example the imputation rates appeared higher in 2011 because they included the majority of the editing when in fact imputation due to non-response only was very similar between censuses. This has made it more difficult to draw comparisons with the 2001 results, and to explain the results to users.

During tuning some issues were detected with the imputed values where the model was occasionally selecting donors that resulted in unlikely combinations of responses.

This was resolved with the addition of several new edit rules, the last two of which were exclusions from the donor pool:

- A mother must be less than 67 years old at the birth of her child (the oldest recorded mother at the time).
- A person born outside of the UK must be at least 10 years old to be married/separated /divorced/widowed (based on expert knowledge of the minimum age for marriage around the world).
- A person not currently working cannot have a position of 'staff' in a communal establishment (kept consistency with the deterministic edit added rule for position).
- Households containing a person aged less than 16 with a partner or step-child cannot be donors
- Households containing fathers who are more than 65 years older than their child cannot be donors

Overall 347,274 (1.56%) persons had at least one hard edit rule failure. The editing rates have been presented for each question throughout this report and the rate of failures in the observed data for each rule taken into CANCEIS is provided in Annex A. The hard edit failure rates were all very low (<0.5%) except for the rules that you cannot be a parent under 12 years (0.96%) or be less than 12 years older than your child (2.39%).These relate to a common response error on the relationship matrix question where the relationship is recorded the wrong way around.

CANCEIS ensures that no hard edit rules are broken in imputed data as long as each rule is programmed correctly. The hard edit rules were very effective in CANCEIS and were also maintained in relationship algorithm 3 for persons 7 to 30. However, one of the changes to the specification of the edit rules was overlooked and not included in the CANCEIS system; this meant that persons were allowed to have more than one partner in the household, although they could not have a spouse/civil partner and a partner.

Many of the soft edits were more relaxed versions of the hard edits, and were equally uncommon with the majority being present in less than 0.5% of households. As with the hard edits, the conditions around age of parents were somewhat higher at 1.14% and 2.43%. Tables 14 and 15 show the change in the prevalence of the between-person and within-person soft edits after imputation respectively.

Table 14: Between-person soft edit condition rates

	Observed		Imputed		*Change*		
	Total	%	Total	%	*Total change*	*Percent change*	*Percent of condition*
Number of households	**22,231,013**	**100**	**22,231,013**	**100**	-	-	-
At least one condition	301,507	1.36	43,373	0.20	*-258,134*	*-1.16*	-
More than two parents/step-parents	99,494	0.45	4,910	0.02	*-94,584*	*-0.43*	*-95.07*
Aged <16 years; partner	15,404	0.07	2,300	0.01	*-13,104*	*-0.06*	*-85.07*
Aged <16 years; step-parent	40,318	0.18	1,239	0.01	*-39,079*	*-0.18*	*-96.93*
Aged <14 years; mother/father	252,655	1.14	312	<0.01	*-252,343*	*-1.14*	*-99.88*
Parent <14 years older than child	539,919	2.43	14,605	0.07	*-525,314*	*-2.36*	*-97.29*
Mother >49 years older than child	11,774	0.05	13,883	0.06	*2,109*	*0.01*	*17.91*
Siblings >37 years apart in age	12,942	0.06	6,944	0.03	*-5,998*	*-0.03*	*-46.35*
Stepchild older than step-parent	48,207	0.22	9,549	0.04	*-38,658*	*-0.17*	*-80.19*
Has parent aged <30; not 'never married / civil partnership'	208,481	0.94	66	<0.01	*-208,415*	*-0.94*	*-99.97*
Has grandparent aged <40; not 'never married / civil partnership'	33,423	0.15	1	<0.01	*-33,422*	*-0.15*	*-100.00*
Parent aged <28; COB "elsewhere"; married/separated/divorced/ widowed	99,494	0.45	19	<0.01	*-99,475*	*-0.45*	*-99.98*
Aged < 16 years; COB "elsewhere"; husband/wife	2,957	0.01	223	<0.01	*-2,734*	*-0.01*	*-92.46*
At least one usual resident not 16 years or older	8,213	0.04	2,239	0.01	*-5,974*	*-0.03*	*-72.74*
Father; > 65 years older than child	4,897	0.02	6,957	0.03	*2,060*	*0.01*	*42.07*

Table 15: Within-person soft edit condition rates

	Observed		Imputed		Change		
	Total	%	Total	%	Total change	Percent change	Percent of condition
Number of persons	**53,483,440**	**100**	**53,483,440**	**100**	**-**	**-**	**-**
Persons with a condition	772,555	1.44	860,326	1.61	*87,771*	*0.16*	-
Divorced/ dissolved same-sex civil partnership; <18 yrs old	4,904	0.01	1,142	<0.01	*-3,762*	*-0.01*	*-76.71*
COB "elsewhere"; married/ separated/ divorced/ widowed <16 yrs old	2,997	0.01	1,538	<0.01	*-1,459*	*<0.01*	*-48.68*
Second address type 'student's home address'; <4 yrs or >64 yrs old	1,908	<0.01	1,807	<0.01	*-101*	*<0.01*	*-5.29*
Reason for second address "another address when working away from home"; <16 yrs old	1,276	<0.01	2,119	<0.01	*843*	*<0.01*	*66.07*
In full-time education; <4 years old	513,087	0.96	552,087	1.03	*39,000*	*0.07*	*7.60*
In full-time education; >64 years old	13,156	0.02	15,302	0.03	*2,146*	*<0.01*	*16.31*
Welsh national identity; no Welsh language; cannot speak English well or very well; not aged <4 or speak BSL*	1,099	<0.01	335	<0.01	*-764*	*<0.01*	*-69.52*
English national identity; cannot speak English well or very well; not aged <4 or speak BSL*	30,441	0.06	27,104	0.05	*-3,337*	*-0.01*	*-10.96*
Read/Writes Welsh; < 3 years old	642	<0.01	938	<0.01	*296*	*<0.01*	*46.11*
“Student term-time/boarding school address" one year ago; <5 or >65 years old	2,801	0.01	1,462	<0.01	*-1,339*	*<0.01*	*-47.80*
Aged < 35 years and retired	3,460	0.01	6,768	0.01	*3,308*	*<0.01*	*95.61*
Aged > 64 and student	10,197	0.02	12,137	0.02	*1,940*	*<0.01*	*19.03*
Last worked < 16 years after DOB.	71,346	0.13	96,888	0.18	*25,542*	*0.05*	*35.80*
Address one year ago UK; date of arrival in UK after March 2010	35,518	0.07	36,063	0.07	*545*	*<0.01*	*1.53*
UK qualification (except apprenticeship/ professional/ foreign/ vocational) and cannot speak English well/very well or speak BSL /other*	80,525	0.15	105,632	0.20	*25,107*	*0.05*	*31.18*
UK qualification (except apprenticeship/ professional / vocational); cannot read / write Welsh or speak English well/very well or speak BSL / other*	35	<0.01	17	<0.01	*-18*	*<0.01*	*-51.43*
Apprenticeship; never worked	14,356	0.03	17,818	0.03	*3,462*	*<0.01*	*24.12*

* British sign language / other UK language

It was possible for the soft edit conditions to be propagated during imputation because (with the two exceptions added during processing) soft edits were not used in CANCEIS. There were several reasons for not including them. Firstly, a decision was taken in the edit and imputation working group to allow rare characteristics to be propagated in proportion to their observed distribution because they may exist in the non-responding population. In addition, the data model in the demographics module was very complex due to the relationships question and adding additional edit constraints would have risked higher levels of non-imputed records or system failures due to memory resource errors.

Generally the between-person conditions were not increased during imputation because they were matched with a corresponding hard edit rule. For example, the only persons who could have the 'mother less than 14 years old' soft edit condition were girls aged 12 or 13 years old, because being a mother under the age of 12 was removed by the hard edit rules. There was a slight increase of 0.01% in the number of households that contained a mother who was more than 49 years older than her child, which equated to an 18% increase on the observed total.

There were a lot more increases of the within-person soft edits. For example, the number of persons aged over 64 years who were in full-time education increased by almost 100%, although this was only an increase of 0.01% across all of England and Wales. These increases mainly occurred when values taken from a donor record made a soft edit condition with a value already in the recipient record. It was less common for the condition to come from a donor. Making rare conditions was more frequent because under between-person edit constraints - such as the rules governing age differences in parents and children - it was not possible to use reordering of persons within households (in the version of CANCEIS implemented). For example, older persons, such as a grandparent or lodger, were sometimes matched with a child because they were recorded in the same position; similarly, children either side of the school-age working-age boundaries were matched. This was because at the household level, the households were a close match despite this disparity. This led to an increase in rare traits such as being retired under 35 years old, or being in school under the age of four. However, across the entire population these increases were very small, the largest being a 0.7% increase in the number of children under four who are in full-time education.

Newer versions of CANCEIS allow reordering under between-person edit constraints and it is expected that this would resolve much of this issue. Testing of this assumption is required for future statistical projects, and the use of selected pre-imputation editing to improve the coherence of the data should be evaluated during tuning. The diagnostics for soft edits focused on changes at the population level, which meant that the large shifts in the actual occurrences of the conditions were undetected. Future diagnostics should also include a '% of condition' calculation to highlight shifts during processing.

9 Operational evaluation

This section gives a brief outline of the execution of the project, key challenges and recommendations for the future. Generally the project ran smoothly, with development and implementation occurring simultaneously in the last two years before census. The working group formed an excellent base for addressing methodological questions about, and agreeing, the proposed methods. It was also the mechanism for agreeing what aspects of

quality were to be monitored by edit and imputation, and what aspects would be monitored by census quality assurance.

Implementation

Integrating CANCEIS into the automated production environment required significant additional IT infrastructure because it is executed using a command line and does not have a programming interface with which to connect to other platforms. The IT strategy stipulated that the imputation sequence be hard programmed into the system and IT development time was required to accommodate this. Using a hard programming strategy meant that changes to the sequencing of the imputation would have been time consuming and costly to implement. It is possible to program a more generic solution and although this would take significantly longer to develop it may be worthwhile if the imputation method is not fully developed at the time of integration.

One of the key challenges during development was obtaining data on which to base the research. It took around 6 months to build a data set which augmented 2001 census data with more recent social survey data. This did cause some delay in delivering the first prototype and made the task of requirements specification more difficult. Being able to communicate the statistical requirements in an appropriate IT language was another challenge, and having a good relationship with the IT developers/analysts proved crucial to achieving efficient communication between the two teams.

The general strategy and methodology for imputation was designed in advance based on an understanding of previous censuses and other social surveys. However, because imputation is an estimation process based on observed data, the detailed system settings and tuning of the method needed to be completed when real census data were available. The initial prototype delivered for the census processing environment required more tuning than had been anticipated and the need for tuning had not been sufficiently recognised during the planning of the census processing timetable. It was also assumed in planning that there would be one generic set of parameter files for running all PUs, and this was not achieved with the live data. As a result, the system had not been set up to easily receive updates, and changes made through tuning were difficult and time consuming. Once the need for data-driven changes was recognised, delivering unique input files - or updates to the relationship algorithms - became quicker and more automated.

For large-scale statistical projects it is important to recognise that data-driven processes are very difficult to fully specify up front, and peculiarities in the live data may require last minute changes to the methods. The processing timetable needs to allow several weeks for analysing the live data and finding solutions to any issues arising. Particularly for imputation which looks across the entire data for the first time and often finds errors in earlier processes. The production system also needs to be flexible in terms of making frequent updates or indeed large changes to the method.

Processing

The automated system worked very well, taking on average 12 hours to complete a PU, and parallel processing meant several PUs could be delivered each day. The offline environment was also very effective as an interface between the automated and manual processes. Four to six PUs could be manually imputed and signed off each day (given two staff working).

Timetable and resource

Processing was completed by two full-time staff, which was one less than originally planned. On average, each person could complete two PUs per day, so an additional person would allow processing time to be cut down by around 30%. This would help to create some contingency in the timetable for additional analysis when issues arise.

10 Conclusion

The imputation of the 2011 Census data was successful in meeting the main aims and objectives of the project:

- The 2011 method produced a higher rate of joint imputation in a much shorter space of time than achieved in previous censuses. This meant that the hierarchical (between-person) distributions were better accounted for than in previous methods.
- After edit and imputation the database of responding persons was complete and consistent.
- Changes to observed data were minimised by imputing inconsistencies and non-response simultaneously with CANCEIS, which minimises the changes made to failed records by selecting from a list of nearest minimum-change donors.
- Priority was given to the key variables that defined the population estimates and bases. These were treated in the first module with variables that helped to predict them.
- The distributions of non-response were estimated for all questions, with the majority having no issues. For age, there was some movement in the distributions of single years of age around the working and student age boundaries. This has identified areas for future research. However, with a maximum 0.3% difference in the proportional distribution for any single year of age, the accuracy was sufficient (Section 8.3.1).
- Non-response bias was evident in some of the questions where the imputed values followed a different distribution to that observed (Annex C to G). This resulted in appropriate minor adjustments to the final distributions. The most notable of these was for activity last week, where the final distributions were adjusted by 1.91%. These changes maintained the quality of the data.
- There were some post-imputation edits required for type of second address and term-time indicator in order to correct for cases where an appropriate donor was not found during imputation (See Annex D). It is thought that using reordering may mitigate this type of issue, and it is an area for future research.
- The relationship algorithms were very effective in addressing the response errors in the relationship matrix and improving the ratio of donors to failed records. The second algorithm would not be required in the future. (Section 8.4)
- The proliferation of rare characteristics (soft edit conditions) was higher than expected (Section 8.5). This has been identified as an area of future research, with

the possibility of using reordering within the donor household, or including additional edits before or during imputation.

- Additional deterministic edits were applied prior to imputation (Section 8.3). Although imputing non-response and editing inconsistencies worked well for the majority of questions, deterministic editing is likely to be required to address any systematic response errors in the data. Some of these may be anticipated based on past census data, but new errors are also likely to arise. Flexibility for adding new deterministic edits prior to imputation would therefore be an advantage.

Processing was generally smooth and efficient; however there were some challenges in implementing the process in an automated production environment:

- The implementation of the method was more iterative than originally planned. Future statistical projects such as this would benefit from an iterative design and testing approach with strong feedback loops and flexibility to make changes to the methods during live operations.
- The period for tuning and parameterisation was also longer than expected. This is an important phase in the implementation and should be adequately timetabled.
- Edit and imputation is the first step that validates consistency between different questions. It would be beneficial to allow time for analysis and modification to the underlying database when edit and imputation is first run on the live data.

11 Further information

The item non-response, editing and imputation rates and quality statements for each 2011 Census question are available from the census website.

The information about variables and classifications gives detailed information about each question and provides the national rates of non-response and item editing and imputation.

The item non-response, editing and imputation rates are also available to download for regions, counties, unitary authorities and local authorities.

12 References

1 Durrent, G.B. (2005) Imputation Methods for Handling Item-Non-response in the Social Sciences: A Methodological Review, *NCRM Methods Review Papers /002,* University of Southampton

2 Falkingham, J. and Gowan, T., February 2011, Independent report on the imputation method for the relationship matrix by Southampton University ESRC Centre for Population Change, *UK Census Edit and Imputation Working Group paper (11)03*, (available on request)

3 Chen, J. and Shao, J. (2001) Jackknife Variance Estimation for Nearest Neighbour Imputation, *Journal of the American Statistical Association*, 96, 453, 260-269.

4 Edit and Imputation Evaluation Report, February 2001, *Census 2001 Review and Evaluation*, http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-conduct/review-and-evaluation/evaluation-reports/edit-and-imputation/index.html

5 Wagstaff, H.F. and Rogers, S. (2006) Application of CANCEIS to 2001 Census Data Technical Report. *ONS Internal technical report* (available on request).

6 2011 Census Item Edit and Imputation Report, December 2012, *2011 Census*: *Methods and Quality Report,* http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/index.html

7 Bankier, M. (1999) Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses. *Working Paper No. 24, UN/ECE Work Sessions on Statistical Data Editing*, Rome

8 Bankier, M., Poirier, P., Lachance, M. and Mason, P. (2000) A generic Implementation of the Nearest Neighbour Imputation Methodology (NIM). *Proceedings of the second International Conference on Establishment Surveys*, Buffalo, pp 571-578

9 De Waal, T., Pannekoek, J. and Scholtus, S. (2011) Handbook of Statistical Editing an Imputation, Wiley, New Jersey

Annex A: Edit constraints for the 2011 Census

Edit rule required	Type of check	Module	% records failing before imputation
A household cannot have more bedrooms than rooms.	Hard	Household	0.18%
A person aged less than 16 cannot have a marital status of being: in a registered same-sex civil partnership/separated but still in a registered same-sex civil partnership/legally dissolved same-sex civil partnership/surviving partner from same-sex civil partner	Hard	Demographics	0.02%
A person aged less than 16 cannot have a marital status of married/separated but still legally married/divorced/widowed (unless country of birth (COB) elsewhere and age >9).	Hard	Demographics	0.01%
A person aged between 6 and 15 must be a student in full-time education unless limited a lot by a health problem/disability.	Hard	Demographics	0.13%
A person cannot have arrived to live in the UK earlier than their date of birth (DOB).	Hard	Culture	0.07%
A person aged less than 5 years cannot be a carer.	Hard	Culture	0.01%
A person's year last worked cannot be before their DOB.	Hard	Labour Market	<0.01%
If second address is the same as workplace address, second address reason should include 'another address when working away from home' or 'armed forces base address'.	Hard	Demographics	<0.01%
A person cannot be aged under 17 and usually travel to work by driving a car or van.	Hard	Labour Market	0.02%
A person who has never worked cannot have the second address reason of 'another address when working away from home'.	Within Person Hard Check	Demographics Labour Market	<0.01%
A person intending to stay less than 6 months in the UK cannot have arrived in the UK before September 2010.	Hard	Culture	0.01%
A person cannot have more than one husband or wife/civil partner.	Hard	Demographics	0.14%

Edit rule required	Type of check	Module	% records failing before imputation
A person cannot have a husband or wife/same-sex civil partner and a partner.	Hard	Demographics	0.04%
Two people with at least one parent in common cannot be married to or in a same-sex civil partnership with/partners of each other.	Hard	Demographics	0.44%
If two people are in a same-sex civil partnership, both must be of the same sex.	Hard	Demographics	0.03%
A person aged less than 16 years cannot be a same-sex civil partner.	Hard	Demographics	0.02%
A person aged less than 12 years cannot be a mother/father.	Hard	Demographics	0.96%
A parent cannot be less than 12 years older than their child.	Hard	Demographics	2.39%
A mother cannot be more than 66 years older than their child.	Hard	Demographics	0.02%
A person aged less than 24 years cannot be a grandparent.	Hard	Demographics	0.17%
A grandparent cannot be less than 24 years older than their grandchild.	Hard	Demographics	0.21%
A person aged less than 16 years cannot be a husband/wife (unless COB 'elsewhere').	Hard	Demographics	0.08%
A person aged less than 12 years cannot be a partner/step-parent (unless COB 'elsewhere').	Hard	Demographics	0.04%
A person with a parent aged under 28 in the household cannot have a marital status of married/separated but still legally married/divorced/widowed (unless COB elsewhere).	Hard	Demographics	0.43%
A person with a parent aged under 28 in the household cannot have a marital status of being: in a registered same-sex civil partnership/separated but still in a registered same-sex civil partnership/legally dissolved same-sex civil partnership/surviving partner.	Hard	Demographics	0.01%
A person with a spouse in the household cannot have a marital status other than married or separated, but still legally married.	Hard	Demographics	0.33%
A person with a same-sex civil partner in the household cannot have a marital status other than in a registered same-sex civil partnership or separated, but still legally in a same-sex civil partnership.	Hard	Demographics	0.11%

Edit rule required	Type of check	Module	% records failing before imputation
At least one usual resident must be 12 years old or above.	Hard	Demographics	0.02%
A father is unlikely to be more than 65 years older than his child.	Soft	Demographics	0.02%
A person aged less than 16 years is unlikely to be a partner/step-parent.	Soft	Demographics	0.07%

Annex B: Variables in each imputation group

Observed households
Accommodation type
Self-contained
Number of rooms
Number of bedrooms
Central heating
Tenure of household
Landlord
Number of cars or vans

Dummy households - household
Accommodation type
Person count
Self-contained

Dummy households - dummy
Reason for dummy

Demographic variables
Relationships (Person 1)
Sex
Age
Marital and civil partner status
Second address
Type of second address
Schoolchild or full-time student indicator
Term-time address indicator
Country of birth indicator
Activity last week

Culture variables
Country of birth
Arrive in UK
Intention to stay in UK
National identity
Ethnicity
Welsh language (Wales only)
Main language
Proficiency in English
Address one year ago
Passport held (UK)
Passport held (other)

Health variables
General health
Provision of unpaid care
Long-term health problem or disability

Labour market variables
Qualifications
Ever worked
Last year worked
Employment status
Occupation
Supervisor status
Industry
Workplace address
Method of travel to work
Hours worked
Workplace address
Method of travel to work
Hours worked

Communal persons - additional variables
Position in establishment (demographics)

Annex C: Household questions evaluation

Type of accommodation

For type of accommodation 2.4% of responses were missing and 0.6% were multi-ticked. While being detached or semi-detached was imputed at a slightly lower rate than observed (-6.8%, -5.2%) and living in a flat or tenement was imputed at a higher rate (+7%), there was 0.2% or less change between the observed and total distributions (Table 16).

Table 16: Distribution of type of accommodation

All eligible responding households, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		*Difference (imputed - observed)*	Total including imputed		*Change (total – observed)*
	N	%	N	%	*%*	N	%	*%*
Detached	5,256	23.6	98	16.8	***-6.8***	5,354	23.4	***-0.2***
Semi-detached	7,057	31.7	155	26.5	***-5.2***	7,212	31.5	***-0.1***
Terraced	5,483	24.6	151	25.9	***1.3***	5,634	24.6	***<0.1***
Flats or tenements	3,381	15.2	129	22.1	***7.0***	3,510	15.3	***0.2***
Converted or shared	824	3.7	33	5.7	***2.0***	857	3.7	***0.1***
Commercial	211	0.9	12	2.0	***1.1***	223	1.0	***<0.1***
Caravan or mobile	82	0.4	6	1.0	***0.7***	88	0.4	***<0.1***
Total	**22,294**	**100**	**583**	**100**	**0.0**	**22,877**	**100**	**0.0**

Self-contained accommodation

There was 2.8% non-response in self-contained, which included less than 0.01% multi-ticks. The observed and imputed distributions were about the same (± 0.6%) and there was less than 0.1% change between the observed and total data (Table 17).

Table 17: Distribution of self-contained

All eligible responding households, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Self -contained	21,969	98.8	628	98.2	**-0.6**	22,597	98.8	***<0.1***
Not self-contained	270	1.2	10	1.8	**0.6**	280	1.2	***<0.1***
Total	**22,239**	**100**	**638**	**100**	**0.0**	**22,877**	**100**	**0.0**

Number of rooms

Non-response in number of rooms included 3.04% missing values and a further 0.06% which could not be resolved. There was less than 5% difference between observed and imputed distributions for all categories and the distributions for the totals were within 0.1% of the observed for all categories.

Table 18: Distribution of rooms

All eligible responding households, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
1	152	0.7	8	1.2	**0.5**	161	0.7	**<0.1**
2	588	2.7	28	3.9	**1.3**	617	2.7	**<0.1**
3	2,113	9.5	101	14.1	**4.5**	2,215	9.7	**0.1**
4	4,155	18.8	150	20.7	**2.0**	4,305	18.8	**0.1**
5	5,511	24.9	170	23.6	**-1.3**	5,681	24.8	**<0.1**
6	4,405	19.9	127	17.7	**-2.2**	4,532	19.8	**-0.1**
7	2,307	10.4	61	8.4	**-2.0**	2,368	10.4	**-0.1**
8	1,457	6.6	37	5.1	**-1.5**	1,494	6.5	**<0.1**
9	783	3.5	20	2.7	**-0.8**	803	3.5	**<0.1**
10	374	1.7	10	1.3	**-0.3**	384	1.7	**<0.1**
11 or more	309	1.4	10	1.3	**-0.1**	318	1.4	**<0.1**
Total	**22,155**	**100**	**722**	**100**	**0.0**	**22,877**	**100**	**0.0**

Number of bedrooms

There were 2.6% missing or invalid responses for number of bedrooms, with less than 0.01% unresolved responses. Zero, one and two bedrooms were imputed at a higher rate than observed and three, four and five bedrooms were imputed at a lower rate (Table 19). This is consistent with imputing a higher rate of compact accommodation such as flats, tenements, converted and shared accommodation. The biggest differences were for one bedroom (+8.5%) and three bedrooms (-6.8%). However, these differences did not lead to a substantial change in the total distribution when including the imputed values (≤0.2%).

Table 19: Distribution of bedrooms

All eligible responding households, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
0	62	0.3	5	0.7	**0.5**	67	0.3	**<0.1**
1	2,378	10.7	124	19.2	**8.5**	2,502	10.9	**0.2**
2	6,097	27.4	192	29.8	**2.4**	6,289	27.5	**0.1**
3	9,375	42.2	228	35.3	**-6.8**	9,603	42.0	**-0.2**
4	3,279	14.7	70	10.9	**-3.9**	3,349	14.6	**-0.1**
5	809	3.6	19	2.9	**-0.8**	827	3.6	**<0.1**
6	166	0.7	5	0.7	**<0.1**	171	0.7	**<0.1**
7	39	0.2	1	0.2	**<0.1**	40	0.2	**<0.1**
8 or more	27	0.1	1	0.2	**0.1**	28	0.1	**<0.1**
Total	**22,232**	**100**	**645**	**100**	**0.0**	**22,877**	**100**	**0.0**

Central heating

Type(s) of central heating was a multi-tick question. Non-response was 3.6%, including 0.8% which had 'none' as well as one or more types of central heating. The most common type, gas, was imputed at a lower rate than observed (-7.7%) and the majority of this difference went into electric (+4.8%) and none (+1.7%). This is consistent with imputing a higher rate of flats and tenements which had the highest observed rates of electric storage and no central heating respectively.

Table 20: Distribution of central heating

All eligible responding households, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
None	590	2.7	36	4.4	**1.7**	626	2.7	**0.1**
Gas	17,395	78.9	584	71.2	**-7.7**	17,979	78.6	**-0.3**
Electric	1,727	7.8	104	12.7	**4.8**	1,832	8.0	**0.2**
Oil	947	4.3	38	4.7	**0.4**	986	4.3	**<0.1**
Solid fuel	167	0.8	7	0.9	**0.1**	174	0.8	**<0.1**
Other	334	1.5	18	2.2	**0.7**	352	1.5	**<0.1**
Two or more types	895	4.1	33	4.1	**<0.1**	928	4.1	**<0.1**
Total	**22,056**	**100**	**821**	**100**	**0.0**	**22,877**	**100**	**0.0**

Tenure

Tenure was required for all households with at least one usual resident; non-responses totalling 2.3% included 0.02% multi-ticks. Ownership with a mortgage or loan was imputed at a lower rate than observed (-9.6%) while renting was imputed at a correspondingly higher rate (+9.8%). A smaller difference occured between owning outright and living rent free.

This is consistent with tenure being missing less often for those living in detached homes which have the highest rate of ownership, and the propensity for imputing compact accommodation which has a lower rate of ownership. These differences had almost no impact on the distribution of the total data which remained within 0.2% of the observed distributions.

Table 21: Distribution of tenure

All eligible responding households, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total – observed
	N	%	N	%	%	N	%	%
Owns outright	6,857	31.6	158	31.1	**-0.5**	7,015	31.6	**<0.1**
Mortgage or loan	7,335	33.8	123	24.2	**-9.6**	7,458	33.6	**-0.2**
Part owns and rents	167	0.8	4	0.8	**<0.1**	171	0.8	**<0.1**
Rents	7,041	32.5	215	42.3	**9.8**	7,256	32.7	**0.2**
Rent free	283	1.3	9	1.7	**0.4**	292	1.3	**<0.1**
Total	**22,191**	**100**	**509**	**100**	**0.0**	**21,682**	**100**	**0.0**

Landlord

Landlord was only required if there was at least one usual resident, and tenure was rent, part rent or rent free. Non-response consisted of 2.6% missing and 0.2% multi-ticked responses. There were less than 1,000 values imputed because of inconsistency with tenure. Private landlord and letting agencies were imputed at a lower rate than observed (-6.9%) while housing associations, co-operatives, trusts and social landlords, councils and relatives or friends were imputed at a higher rate. This is consistent with imputing a higher rate of (and landlord being missing more often for) terraces and flats or tenements which have a higher rate of social and council landlords than other accommodation types. These differences resulted in 0.2% or less change between the observed and total distributions.

Table 22: Distribution of landlord

All eligible responding households, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total – observed
	N	%	N	%	%	N	%	%
Housing association, co-operative, trust or social landlord	1,861	24.8	61	28.2	**3.4**	1,922	24.9	**0.1**
Council	2,041	27.2	61	28.6	**1.4**	2,102	27.2	**<0.1**
Private landlord / letting agency	3,113	41.5	74	34.5	**-6.9**	3,187	41.3	**-0.2**
Employer	94	1.3	2	1.1	**-0.1**	96	1.3	**<0.1**
Relative or friend	301	4.0	12	5.7	**1.7**	314	4.1	**<0.1**
Other	93	1.2	4	1.8	**0.5**	96	1.2	**<0.1**
Total	**7,503**	**100**	**215**	**100**	**0.0**	**7,718**	**100**	**0.0**

Number of cars or vans

A response to the number of cars or vans in the household question was required where at least one usual resident lived in the property. Non-response was 2.3% which included only 2,171 (<0.01%) multi-ticks and unresolved write-ins. Having no cars or vans was favoured in the imputation (+12.7%) while having two vehicles was less favoured compared to the observed (-10.2%). This is consistent with imputing a higher rate of compact accommodation than was observed. The imputed distribution caused little change when combined with the observed values and the distribution of the total data was within 0.3% of the observed.

Table 23: Distribution of number of cars

All eligible responding households, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	n	%	n	%	%	N	%	%
None	5,254	24.2	185	36.9	**12.7**	5,439	24.5	**0.3**
1	9,189	42.4	218	43.6	**1.2**	9,407	42.4	**<0.1**
2	5,568	25.7	77	15.4	**-10.2**	5,645	25.4	**-0.2**
3	1,241	5.7	15	3.1	**-2.6**	1,257	5.7	**-0.1**
4	323	1.5	4	0.7	**-0.7**	326	1.5	**<0.1**
5	75	0.3	1	0.2	**-0.2**	76	0.3	**<0.1**
6 or more	41	0.2	1	0.1	**-0.1**	41	0.2	**<0.1**
Total	**21,690**	**100**	**501**	**100**	**0.0**	**22,191**	**100**	**0.0**

Annex D: Demographics question evaluation

Age

Age was derived from date of birth and is presented in Table 24 in five-year age bands. Date of birth was double-coded for quality (by computer and by person) and had a low non-response rate of 0.6%, the majority of which were blank values. There were 14 edit rules in this module (which governed age, relationship, marital and civil partner status, and schoolchild / student status) and a working-age filter which decided whether activity last week should be completed. Resolution of the edits and filter resulted in around 81,000 ages (0.02%) being changed during imputation.

Most five-year age bands were imputed in similar proportions to the observed data, except 15 to 19-year-olds (+7.8%) and 10 to 14-year-olds (-2.5%). These movements were discussed in section 8.3.1. The differences between the observed and imputed proportions for the remainder of the population were less than 2%. Because non-response was low, the total distributions for each band, including imputed values, remained within 0.1% of the observed distributions.

Table 24: Distribution of five-year age band

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		*Change (total – observed*
Age band	N	%	N	%	%	N	%	%
0	636	1.2	6	1.5	**0.3**	642	1.2	**<0.1**
1 – 4	2,510	4.7	27	6.7	**1.9**	2,536	4.7	**<0.1**
5- 9	2,893	5.5	24	6.0	**0.6**	2,917	5.5	**<0.1**
10 – 14	3,080	5.8	13	3.3	**-2.5**	3,093	5.8	**<0.1**
15 – 19	3,459	6.5	57	14.3	**7.8**	3,516	6.6	**0.1**
20 – 24	3,771	7.1	36	9.0	**1.9**	3,807	7.1	**<0.1**
25 – 29	3,428	6.5	30	7.4	**0.9**	3,457	6.5	**<0.1**
30 – 34	3,340	6.3	22	5.5	**-0.8**	3,362	6.3	**<0.1**
35 – 39	3,458	6.5	20	4.9	**-1.6**	3,477	6.5	**<0.1**
40 – 44	3,858	7.3	22	5.6	**-1.7**	3,880	7.3	**<0.1**
45 – 49	3,903	7.4	24	5.9	**-1.4**	3,927	7.3	**<0.1**
50 – 54	3,463	6.5	21	5.3	**-1.2**	3,484	6.5	**<0.1**
55 – 59	3,078	5.8	19	4.7	**-1.1**	3,096	5.8	**<0.1**
60 – 64	3,281	6.2	19	4.7	**-1.4**	3,300	6.2	**<0.1**
65 – 69	2,598	4.9	15	3.8	**-1.1**	2,613	4.9	**<0.1**
70 – 74	2,113	4.0	14	3.4	**-0.6**	2,127	4.0	**<0.1**
75 – 79	1,719	3.2	12	2.9	**-0.3**	1,731	3.2	**<0.1**
80 – 115	2,496	4.7	20	5.1	**0.4**	2,516	4.7	**<0.1**
Total	**53,083**	**100**	**400**	**100**	**0.0**	**53,483**	**100**	**0.0**

Sex

Non-response for sex was 0.4%, including 4,689 multi-ticks (0.1%). The edit rules stated that marital spouses had to be of opposite sex and civil partners had to be the same sex. Around 16,000 sexes (<0.01%) were changed due to edit failures. There was 0.1% difference between the imputed and observed distributions, and less than 0.1% in the final distribution of total values (Table 25).

Table 25: Distribution of sex

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		*Change (total – observed*
	N	%	N	%	%	N	%	%
male	25,897	48.6	117	48.5	**-0.1**	26,014	48.6	**<0.1**
female	27,346	51.4	124	51.5	**0.1**	27,470	51.4	**<0.1**
Total	**53,243**	**100**	**240**	**100**	**0.0**	**53,483**	**100**	**0.0**

Marital and civil partner status

Non-response for marital and civil partner status was 3.8%. The deterministic edit rule discussed in section 8.3 set 2.5% of these to 'never married or in a civil partnership' before the remaining 1.4% were imputed. There were several edit rules governing marital/civil partner status which ensured consistency with age and recorded relationships. This resulted in around 64,000 marital and civil partner statuses (0.1%) being amended by imputation.

There were some large differences between the observed and imputed distributions: 'never married' or 'civil partner' was 19.7% more frequent in the imputed values, while 'married' was 21.1% less frequent. This is largely due to the characteristics of the non-responders. For example, over 65% of the non-responders were under 16 years old and 63% were the child or step child of person 1, making them more likely to be never married or in a civil partnership. As a result, there was a minor adjustment of 0.3% in the total distributions for being never married/civil partner and being married.

Table 26: Distribution of marital and civil partnership status

All eligible responding persons, England and Wales, 2011 Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	n	%	n	%	%	N	%	%
Never married / civil partner	24,260	46.0	498	65.7	**19.7**	24,758	46.3	**0.3**
Married	20,477	38.8	134	17.8	**-21.1**	20,612	38.5	**-0.3**
Separated	1,060	2.0	18	2.3	**0.3**	1,078	2.0	**<0.1**
Divorced	3,816	7.2	52	6.9	**-0.3**	3,868	7.2	**<0.1**
Widowed	2,982	5.7	53	7.1	**1.4**	3,035	5.7	**<0.1**
Civil partner	91	0.2	1	0.1	**<0.1**	92	0.2	**<0.1**
Separated civil partner	11	<0.1	<1	<0.1	**<0.1**	11	<0.1	**<0.1**
Dissolved civil partner	10	<0.1	<1	<0.1	**<0.1**	10	<0.1	**<0.1**
Widowed civil partner	20	<0.1	<1	<0.1	**<0.1**	20	<0.1	**<0.1**
Total	**52,726**	**100**	**757**	**100**	**0.0**	**53,483**	**100**	**0.0**

Second address

Second address was asked in three parts: a ticked indicator, a UK address field and a country field. Non-response for the indicator was 3.5%, and where applicable 9% of postcodes and 2.7% of countries were missing or otherwise invalid. However, a separate process was used to clerically match postcodes from the written address and 40,138 (1.65%) of the imputed second address postcodes were later updated with a clerically matched postcode.

Not having a second address was slightly more frequent (and having one, less frequent) in the imputed values (Table 27). Around 9% of respondents who indicated having another address were changed to 'no second address' by imputation where inconsistencies were present or a corresponding postcode/country had not been supplied, and the best donor selected did not have a second address. However, when including the imputing values there was negligible change (≤ 0.1%) between the observed and total distributions.

Table 27: Distribution of second address indicator

All eligible responding persons, England and Wales, 2011 Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	n	%	n	%	%	N	%	%
No second address	48,296	93.8	1,914	97.0	**3.3**	50,209	93.9	**0.1**
UK second address	2,395	4.6	42	2.1	**-2.5**	2,437	4.6	**-0.1**
Other second address	821	1.6	17	0.9	**-0.7**	837	1.6	**<0.1**
Total	**51,511**	**100**	**1,972**	**100**	**0.0**	**53,483**	**100**	**0.0**

Type of second address

Type of second address was asked to all persons who had another UK or international address where they spent more than 30 days a year (including students who lived somewhere else during term-time). As this question was a multi-tick, the non-response rate of 6.7% consisted entirely of blank responses. Table 28 shows the number of persons eligible to respond and those who had a response imputed. The counts and proportions for each type of address indicate whether or not that option was selected (or imputed) either on its own or in combination with other responses.

There were some large differences between the observed and imputed distributions; other and working addresses were 15% and 8% more frequent respectively in the imputed values, while students' home and term-time addresses were 15% and 13% less frequent respectively. These differences were mainly due to resolving the edit rules. For example, working was imputed to satisfy the rule that second address must contain working where workplace and second address were the same.

There were also post-imputation deterministic edits for second address type which influenced the distribution of imputed values. These were used where a valid second address had been changed to 'no second address' because second address type was blank and the donor did not have a second address (or second address type). The valid address was returned and second address type was set according to the characteristics of the non-responder: 'students term-time address' if term-time indicator was 'address in question 5'; else 'other parent/guardian' if the person was aged under 16 or 'other' if the person was aged 16 years or over.

Table 28: Distribution of type of second address

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	n	%*	n	%*	%	N	%*	%
Eligible persons	**3,047**	**100**	**227**	**100**	**.**	**3,274**	**100**	**.**
Armed forces	62	2.0	8	3.4	**1.4**	70	2.1	**0.1**
Working	208	6.8	33	14.7	**7.9**	241	7.4	**0.5**
Students home	625	20.5	18	7.8	**-12.7**	643	19.6	**-0.9**
Students term-time	649	21.3	14	6.3	**-15.0**	663	20.2	**-1.0**
Parent / guardian	644	21.1	59	25.9	**4.8**	703	21.5	**0.3**
Holiday	391	12.8	27	11.8	**-1.0**	418	12.8	**-0.1**
Other	583	19.2	78	34.1	**15.0**	661	20.2	**1.0**

* Percent of eligible persons who have that type of second address

Schoolchild or student – full-time education indicator

Non-response for schoolchild/student was 3.3% and consisted primarily of blanks with only 1,161 multi-ticks. There was one edit rule which stated that persons aged between 6 and 15 had to be in full-time education, unless they were limited by a long-term disability or illness. Around 62,000 (0.1%) of the values were amended due to failing this edit rule. In the imputed values, being in full-time education was 11% less frequent than was observed. This is consistent with the ages of the non-responders; only 6% were aged between 5 and 15 years old, and a further 6% were aged up to 25 years old. There was a minor adjustment of 0.4% in the distribution of the total including imputed values.

Table 29: Distribution of schoolchild or student

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
In full-time education	11,409	22.1	198	10.9	**-11.1**	11,607	21.7	**-0.4**
Not full-time education	40,267	77.9	1,609	89.1	**11.1**	41,877	78.3	**0.4**
Total	**51,677**	**100**	**1,807**	**100**	**0.0**	**53,483**	**100**	**0.0**

Term-time indicator

The term-time address question was asked to everyone who said they were in full-time education. This was where students who lived somewhere else during term-time were routed out of the questionnaire. Because students who have a different term-time address were a relatively small subset of the population, it was harder to match donors that had another term-time address. Sometimes a donor was selected that lived at the same address during term-time, which moved the respondent into the resident population for that area. These persons were manually put back to their observed value after imputation.

The non-response rate was low at 1.4% and there were only 1,253 multi-ticks. Around 59,000 values (0.5%) were amended because they were inconsistent with the answer to the schoolchild / student question or subsequent questions that should have been routed out. The imputed values (Table 30) favoured living at the same address by 4.7%, while living at your second address (question 5) was 4.4% less frequent than in the observed values. As the difference was small and the non-response rate low, there was little difference (≤ 0.1%) on the distribution for the total including imputing values.

Table 30: Distribution of term-time indicator

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Same address	10,700	93.9	215	98.6	**4.7**	10,915	94.0	**0.1**
Address in question 5	650	5.7	3	1.3	**-4.4**	653	5.6	**-0.1**
Another address	39	0.3	<1	0.1	**-0.2**	39	0.3	**<0.1**
Total	**11,389**	**100**	**218**	**100**	**0.0**	**11,607**	**100**	**0.0**

Activity last week

Activity last week was derived from responses to questions 26 to 30 which asked which activities a person was undertaking in the last working week, and whether they were available for work, looking for work or waiting to start work. All persons over the age of 15 who did not live somewhere else during term-time were asked to respond. Around 5.1% of responses were incomplete or blank and 0.1% (34,000) were edited due to being inconsistent with the age and term-time routing filters.

There were two large differences between the observed and imputed distributions (Table 31); working was 33.1% less frequent while being retired was 37.2% more frequent. This was consistent with non-responders tending to be over 65 years of age. Sixteen-year-olds were also less likely to respond, which helps explain why being a student was the only other more frequent category. These differences lead to small adjustments in the total distributions; being retired increased by 1.9% and working fell by 1.7%.

Table 31: Distribution of activity last week

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Working	24,653	60.4	602	27.3	**-33.1**	25,255	58.7	**-1.7**
Unemployed	1,880	4.6	65	2.9	**-1.7**	1,945	4.5	**-0.1**
Student	1,987	4.9	113	5.1	**0.3**	2,100	4.9	**0.0**
Retired	8,208	20.1	1,264	57.3	**37.2**	9,472	22.0	**1.9**
Sick / disabled	1,580	3.9	72	3.3	**-0.6**	1,652	3.8	**<0.1**
Home / family	1,653	4.0	48	2.2	**-1.9**	1,701	4.0	**-0.1**
Other	875	2.1	43	1.9	**-0.2**	917	2.1	**<0.1**
Total	**40,836**	**100**	**2,206**	**100**	**0.0**	**43,041**	**100**	**0.0**

Relationship to person one

Every household was asked to complete the relationship matrix which included how each person was related to person 1 and to each person listed previously to them on the questionnaire. The automated (nearest neighbour) imputation only treated the relationships to persons 1 to 5. The remainder of the relationships were treated by the relationship algorithms. Only relationship to person 1 is considered here, further analysis of the relationship algorithms is provided in section 5.1.2. Non-response for relationship to person 1 was moderate including 3.3% blanks and 0.7% multi-ticks. Before imputation, the first relationship algorithm edited 247,431 (0.8%) of the relationship-to-person-1 values. This primarily corrected parents and children who were recorded the wrong way around and set in-law relationships to ‘other relation’. A further 0.4% of values were changed during imputation due to failing the edit rules.

The main difference between the distributions of the observed and imputed values (Table 31) was a 14.8% lower frequency of spouses and 7.3% higher frequency of unrelated persons in the imputed values. Relationship had the most applicable edit rules, and the distribution of imputed values was influenced by both complex edit constraints and the characteristics of the non-responders.

The differences in the imputed responses caused minor adjustments to the total distributions: spouses decreased by 0.6% while being unrelated increased by 0.3%.

Table 32: Distribution of relationship to person one

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		**Imputed responses**		**Difference (imputed - observed)**	**Total including imputed**		***Change (total - observed***
	N	%	N	%	%	N	%	%
Spouse	9,550	32.9	241	18.1	**-14.8**	9,790	32.3	**-0.6**
Civil partner	33	0.1	<1	<0.1	**-0.1**	34	0.1	**<0.1**
Partner	2,196	7.6	131	9.9	**2.3**	2,327	7.7	**0.1**
Child	14,228	49.1	639	48.1	**-1.0**	14,867	49.0	**<0.1**
Stepchild	384	1.3	29	2.2	**0.8**	413	1.4	**<0.1**
Sibling	283	1.0	33	2.5	**1.5**	317	1.0	**0.1**
Stepsibling	6	<0.1	3	0.2	**0.2**	9	0.0	**<0.1**
Parent	300	1.0	19	1.4	**0.4**	319	1.1	**<0.1**
Stepparent	7	<0.1	2	0.1	**0.1**	9	0.0	**<0.1**
Grandchild	331	1.1	21	1.6	**0.4**	352	1.2	**<0.1**
Grandparent	6	<0.1	<1	<0.1	**<0.1**	6	0.0	**<0.1**
Other relation	435	1.5	56	4.2	**2.7**	491	1.6	**0.1**
Unrelated	1,246	4.3	154	11.6	**7.3**	1,400	4.6	**0.3**
Total	**29,006**	**100**	**1,328**	**100**	**0.0**	**30,335**	**100**	**0.0**

Annex E: Culture questions evaluation

Country of birth

Country of birth had low non-response (1.5%), which included 14,025 (0.03%) multi-ticked or irresolvable responses. Although there were no edit rules, around 7,000 values (0.01%) were changed due to inconsistency with the student / term-time filter. Table 33 shows the high-level country of birth group used in imputation. There were only minor differences in the proportions of respondents belonging to each group in the observed and imputed data, and less than 0.1% change in the total distributions after imputation.

Table 33: Distribution of country of birth group

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	n	%	n	%	%	N	%	%
UK	45,224	87.0	709	87.9	**0.9**	45,933	87.0	**<0.1**
Ireland	362	0.7	19	2.3	**1.6**	381	0.7	**<0.1**
Europe	2,089	4.0	26	3.2	**-0.8**	2,115	4.0	**<0.1**
Africa	1,170	2.3	16	2.0	**-0.2**	1,186	2.2	**<0.1**
Asia	2,366	4.6	23	2.9	**-1.7**	2,389	4.5	**<0.1**
North America	482	0.9	9	1.2	**0.2**	492	0.9	**<0.1**
South America	128	0.2	1	0.2	**-0.1**	129	0.2	**<0.1**
Oceana / other	163	0.3	2	0.3	**<0.1**	166	0.3	**<0.1**
Total	**51,985**	**100**	**806**	**100**	**0.0**	**52,791**	**100**	**0.0**

Arrival in the UK

Only persons born outside the UK were asked arrival in the UK. Month and year of arrival were converted to ‘months since arrival’ for imputation. Where only month was invalid, months since arrival was calculated using the year only, and non-response for the derived variable was 4.8%. There were two edit rules: arrival must be after date of birth, and if intending to stay less than six months arrival must be less than six months prior to census night. Around 128,000 months of arrival in the UK (1.9%) were changed due to inconsistency with the edit rules or the term-time / country of birth routing filters. Table 34 shows months since arrival aggregated to four groups. The observed and imputed distributions were quite similar (≤ 2.1% difference) with 0.1% or less change in the total distribution including imputed values.

Table 34: Distribution of arrival to the UK

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Less than 3 months	145	2.3	20	4.4	**2.1**	165	2.4	**0.1**
3 to 6 months	222	3.5	14	3.1	**-0.3**	236	3.4	**<0.1**
7 to 12 months	183	2.9	12	2.6	**-0.3**	194	2.8	**<0.1**
More than 12 months	5,855	91.4	408	89.9	**-1.5**	6,263	91.3	**-0.1**
Total	**6,405**	**100**	**453**	**100**	**0.0**	**6,858**	**100**	**0.0**

Intention to stay

Only persons born outside the UK who had arrived less than 12 months before census night were required to answer intention to stay. The non-response rate, comprised almost entirely of blanks, was much higher than for other questions at 14.5%, but is partly a reflection of the small size of this sub-population. Around 6,000 values (0.9%) were changed because of inconsistency with the routing filters or because they arrived more than six months ago but said they intended to stay less than six months. There was little difference (0.7% or less) between the distributions of the observed and imputed values; and similarly, 0.1% or less difference in the total distributions (Table 35).

Table 35: Distribution of intention to stay in UK

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Less than 6 months	43	8.6	7	7.9	**-0.7**	51	8.5	**-0.1**
6 to 12 months	86	17.1	16	17.6	**0.6**	102	17.2	**0.1**
12 months or more	373	74.3	68	74.5	**0.2**	441	74.3	**<0.1**
Total	**502**	**100**	**92**	**100**	**0.0**	**594**	**100**	**0.0**

National identity

National identity had low non-response (1.9%), including 97,085 (0.2%) irresolvable responses. Although there were no edit rules, around 5,000 values (0.01%) were changed due to inconsistency with the term-time filter. National identity was a multi-tick question with unlimited allowable combinations of responses. These were aggregated into the broad groups shown in Table 36 for analysis. There were some small differences between the observed and imputed distributions; UK identities were 5.4% less frequent in the imputed values and other identities were 5% more frequent. However**,** because non-response was small the change in the total of observed and imputed values was negligible (0.1%).

Table 36: Distribution of national identity group

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
Group	n	%	n	%	%	N	%	%
UK	47,412	91.6	885	86.2	**-5.4**	48,297	91.5	**-0.1**
UK and other	445	0.9	13	1.2	**0.4**	458	0.9	**<0.1**
Other	3,907	7.5	129	12.6	**5.0**	4,036	7.6	**0.1**
Total	**51,764**	**100**	**1,027**	**100**	**0.0**	**52,791**	**100**	**0.0**

Ethnic group

Ethnic group was collected with 18 tick boxes and five write-in boxes. Coding rules were used to allocate each response to a single code from the ethnicity index. These have been aggregated to the 18 categories provided in Table 37 for analysis. Non-response for the derived variable was 3%, including around 0.3% which were unresolved because they included more than three high-level ethnic groups (for example Other White, Bangladeshi and Caribbean). While there were no edit rules for ethnicity, around 21,000 values (0.04%) were changed because of inconsistency with the term-time routing filter. In the imputed values, English was 11.2% less frequent while the other categories tended to be more frequent than observed. This was mainly due to geographic effects.

There was one issue with the imputation, and this was for the Arab category. Due to an error in the coding of ethnicity, those that ticked ‘Arab’ on the questionnaire ended up with the code for ‘Other’. Although these were all corrected after imputation, it meant that the Arab ethnic group was underrepresented in the donor pool, and only 2,000 persons were imputed into the Arab ethnicity. However, this did not affect the final distribution which decreased by less than 0.01%. There was a minor downward adjustment of 0.3% in the total distribution for English, while the changes to other categories were negligible (≤ 0.1%).

Table 37: Distribution of ethnic group

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
English	41,832	81.7	1,140	70.6	**-11.2**	42,973	81.4	**-0.3**
Irish	487	1.0	16	1.0	**<0.1**	503	1.0	**<0.1**
Gypsy/traveller	51	0.1	1	0.1	**<0.1**	52	0.1	**<0.1**
Other white	2,245	4.4	105	6.5	**2.1**	2,350	4.5	**0.1**
Mixed Caribbean	348	0.7	10	0.6	**<0.1**	358	0.7	**<0.1**
Mixed African	135	0.3	6	0.4	**0.1**	141	0.3	**<0.1**
White Asian	312	0.6	10	0.6	**<0.1**	322	0.6	**<0.1**
Other Mixed	158	0.3	5	0.3	**<0.1**	163	0.3	**<0.1**
Indian	1,329	2.6	63	3.9	**1.3**	1,392	2.6	**<0.1**
Pakistani	1,049	2.0	53	3.3	**1.2**	1,102	2.1	**<0.1**
Bangladeshi	404	0.8	19	1.2	**0.4**	423	0.8	**<0.1**
Chinese	349	0.7	22	1.4	**0.7**	372	0.7	**<0.1**
Other Asian	687	1.3	55	3.4	**2.0**	742	1.4	**0.1**
African	887	1.7	64	3.9	**2.2**	951	1.8	**0.1**
Caribbean	536	1.0	24	1.5	**0.4**	560	1.1	**<0.1**
Other Black	146	0.3	7	0.4	**0.1**	153	0.3	**<0.1**
Arab	20	<0.1	2	0.1	**0.1**	21	<0.1	**<0.1**
Other ethnicity	200	0.4	15	0.9	**0.5**	215	0.4	**<0.1**
Total	**51,175**	**100**	**1,616**	**100**	**0.0**	**52,791**	**100**	**0.0**

Welsh language

Welsh language was only asked to persons living in Wales. Respondents were asked to tick all the Welsh skills that applied to them. Non-response was 3.4%, including 3,915 unresolved multi-ticks of least one skill and no Welsh. Around 1,000 values (<0.1%) were changed due to inconsistency with the term-time filter. Table 38 shows the distributions for each skill / no skills independently. The imputed and observed values followed a similar distribution, with negligible changes when including imputed values.

Table 38: Distribution of Welsh language proficiency

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Eligible	2,765	100	97	100	**.**	2,861	100	**.**
Understands spoken	619	22.4	20	20.9	**-1.5**	640	22.4	**-0.1**
Speaks	514	18.6	17	17.6	**-1.0**	531	18.6	**<0.1**
Reads	490	17.7	16	16.4	**-1.3**	506	17.7	**<0.1**
Writes	417	15.1	14	14.1	**-1.0**	431	15.1	**<0.1**
No Welsh	2,041	73.8	73	75.3	**1.5**	2,114	73.9	**0.1**

Main language

Main language had low non-response (2.5%), including 68,806 unresolved responses (0.1%). A further 41,000 values (0.08%) were edited because of inconsistency with the term-time filter. High-level main language group was evaluated. There were only minor differences between the observed and imputed distributions, with less than 0.1% change when including the imputed values (Table 39).

Table 39: Distribution of main language

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Arabic	134	0.3	1	0.1	**-0.2**	135	0.3	**<0.1**
African- northern	3	<0.1	<1	<0.1	**<0.1**	3	<0.1	**<0.1**
African-other	213	0.4	11	0.8	**0.4**	225	0.4	**<0.1**
Asian- eastern	363	0.7	9	0.6	**-0.1**	372	0.7	**<0.1**
Asian-south	1,221	2.4	46	3.4	**1.0**	1,267	2.4	**<0.1**
Asian-other	243	0.5	8	0.6	**0.1**	251	0.5	**<0.1**
English	47,614	92.6	1,248	91.1	**-1.5**	48,862	92.6	**<0.1**
European	1,606	3.1	46	3.3	**0.2**	1,652	3.1	**<0.1**
American	2	<0.1	<1	<0.1	**<0.1**	2	<0.1	**<0.1**
Oceanic	21	<0.1	1	<0.1	**<0.1**	22	<0.1	**<0.1**
Other	1	<0.1	<1	<0.1	**<0.1**	1	<0.1	**<0.1**
Total	**51,422**	**100**	**1,369**	**100**	**0.0**	**52,791**	**100**	**0.0**

Proficiency in English

Only persons who gave a main language other than English were required to give their proficiency in English. Non-response was 3.6%, including 304 multi-tick responses (0.01%). There were no edit rules for proficiency in English and less than 1,000 values were changed due to routing filters.

There were some differences between the observed and imputed distributions; not speaking English at all was 6.1% more frequent in the imputed values while speaking English very well was 5.2% less frequent. This was consistent with the majority of non-responders being aged under 4 years old. As non-response was small, there were only minor changes between the observed and total distribution.

Table 40: Distribution of proficiency in English

All responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Very well	1,522	40.2	50	35.0	**-5.2**	1,572	40.0	**-0.2**
Well	1,401	37.0	48	33.8	**-3.2**	1,450	36.9	**-0.1**
Not well	670	17.7	29	20.0	**2.3**	698	17.8	**0.1**
Not at all	193	5.1	16	11.2	**6.1**	209	5.3	**0.2**
Total	**3,787**	**100**	**143**	**100**	**0.0**	**3,929**	**100**	**0.0**

Religion

Religion was not imputed because it was a voluntary question.

Address one year ago

Address one year ago was asked for everyone (at or without a term-time address) however, persons aged zero on census night were later amended to be ineligible. The question consisted of a tick-box indicator and a written in UK address or country where applicable. Non-response for the indicator was 3.8%, and where applicable, 3.6% of UK addresses and 5.8% of countries were missing or invalid. However, almost 1% of the imputed postcodes were later updated with an expert clerical match. A further 0.2% of indicators, 0.03% of postcodes and 0.7% of countries were edited for consistency with the term-time filter.

The distribution of postcodes and country codes was not included in the evaluation; Table 41 shows the distribution for the second-address indicator. There were only small differences between the observed and imputed distributions (1.6% or less), and negligible changes to the total distribution when including the imputed values.

Table 41: Distribution of address one year ago indicator

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Same as person 1	17,895	35.8	757	36.0	**0.3**	18,652	35.8	**<0.1**
Address on front	27,627	55.2	1,193	56.8	**1.6**	28,820	55.3	**0.1**
Student UK term-time	282	0.6	6	0.3	**-0.3**	289	0.6	**<0.1**
Another UK address	3,663	7.3	122	5.8	**-1.5**	3,785	7.3	**-0.1**
Outside the UK	582	1.2	24	1.1	**<0.1**	605	1.2	**<0.1**
Total	**50,049**	**100**	**2,101**	**100**	**0.0**	**52,150**	**100**	**0.0**

Passports held

The response for passports held was collected in two parts; a multiple tick-box response for UK, Irish, other or none, and a write-in response for 'other' passports. There were 128,337 (0.2%) multi-ticks that included 'none' or irresolvable responses. The responses for each tick box are presented independently in Table 42. Non-response for UK passports was 2.3% and where applicable, 2.4% for non-UK passports. There were no edit rules for passports held but around 33,000 (0.1%) UK passports and 1,000 (0.02%) non-UK passports were changed due to questionnaire filters. There were some small differences between the observed and imputed distributions for having particular passports; no passport was 7.6% more frequent in the imputed values while having a UK passport was 6.5% less frequent. However, these differences had a negligible impact on the distribution of the totals when including the imputed values.

Table 42: Distribution for passports held

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Eligible	51,537	100.0	1,254	100.0	.	52,791	100.0	.
UK	39,302	76.3	875	69.7	**-6.5**	40,177	76.1	**-0.2**
Irish	370	0.7	9	0.7	**<0.1**	380	0.7	**<0.1**
Other	3,889	7.5	73	5.8	**-1.7**	3,963	7.5	**<0.1**
No passport	8,542	16.6	303	24.2	**7.6**	8,846	16.8	**0.2**

Annex F: Health questions evaluation

General Health

General health had a low non-response rate of 1.6%, including 4,483 (<0.01%) multi-ticks. Editing related solely to inconsistencies with the student/term-time filter. There were some small differences between the imputed and observed distributions which had little impact on the total distribution (Table 43).

Table 43: Distribution of general health

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		*Difference (imputed - observed)*	Total including imputed		*Change (total - observed*
	N	%	N	%	*%*	N	%	*%*
Very good	24,276	46.7	378	44.1	***-2.7***	24,654	46.7	***<0.1***
Good	17,801	34.3	268	31.3	***-3.0***	18,069	34.2	***<0.1***
Fair	6,930	13.3	144	16.8	***3.5***	7,074	13.4	***0.1***
Bad	2,262	4.4	51	6.0	***1.6***	2,314	4.4	***<0.1***
Very bad	664	1.3	16	1.8	***0.6***	680	1.3	***<0.1***
Total	51,934	100	857	100	***0.0***	52,791	100	***0.0***

Provision of unpaid care

The provision of unpaid care had a moderate non-response rate of 3.5% including 19,064 multi-ticks (0.03%). The imputed and observed distributions were similar and there was less than 0.01% change between the observed and total distributions (Table 44).

Table 44: Distribution of provision of unpaid care

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
Hours per week	N	%	N	%	%	N	%	%
None	45,498	89.3	1,684	90.3	**1.0**	47,181	89.4	**<0.1**
1 to 19	3,434	6.7	103	5.5	**-1.2**	3,537	6.7	**<0.1**
20 to 49	722	1.4	25	1.3	**-0.1**	746	1.4	**<0.1**
50 or more	1,274	2.5	52	2.8	**0.3**	1,326	2.5	**<0.1**
Total	50,928	100	1,863	100	**0.0**	52,791	100	**0.0**

Long-term health problem or disability

Long-term health problem or disability was similarly responded with non-response of 3.2%, which included 3,806 (<0.01%) multi-ticks. The observed and imputed values followed a similar distribution and there was little change (≤0.1%) to the distribution when including the imputed values (Table 45).

Table 45: Long-term health problem or disability

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Limited a lot	4,389	8.6	165	9.8	**1.2**	4,554	8.6	**<0.1**
Limited a little	4,888	9.6	168	10.0	**0.4**	5,056	9.6	**<0.1**
Not limited	41,835	81.8	1,347	80.2	**-1.7**	43,182	81.8	**-0.1**
Total	51,112	100	1,680	100	**0.0**	52,791	100	**0.0**

Annex G: Labour Market questions evaluation

Qualifications

Only those aged 16 years or over were required to provide their qualifications, excluding students living somewhere else during term-time. Qualifications was a multi-tick question and was imputed as a binary string allowing any combination except having none and a qualification. A filter rule was applied to resolve this combination prior to imputation and the non-response of 5.7% consisted entirely of blanks. The editing rate was less than 0.1% and related to inconsistency with the working-age and term-time routing filters.

The imputed values favoured having no qualifications by 17.2%, while having a degree level qualification was 7.9% less frequent and having other qualifications was just over 5% less frequent. This was primarily driven by a higher rate of having no qualifications in the areas with the highest non-response and a slight over representation of persons who were more likely to have no qualifications: females and persons aged over 60 years. This resulted in a 1% adjustment to the total distribution in favour of having no qualifications, while degree level decreased by 0.5%, and other qualifications by 0.3%.

Table 46: Distribution of qualification group - highest level attained

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
None	8,876	21.9	958	39.1	**17.3**	9,834	22.8	**1.0**
Entry level[1]	4,477	11.0	241	9.9	**-1.2**	4,719	11.0	**-0.1**
Intermediate[2]	6,031	14.9	355	14.5	**-0.3**	6,386	14.8	**<0.1**
Advanced[3]	3,752	9.2	161	6.6	**-2.7**	3,913	9.1	**-0.2**
Degree / Higher[4]	8,148	20.1	297	12.2	**-7.9**	8,446	19.6	**-0.5**
Other[5]	9,310	22.9	434	17.7	**-5.2**	9,744	22.6	**-0.3**
Total	**40,594**	**100**	**2,447**	**100**	**0.0**	**43,041**	**100**	**0.0**

1. < 5 O Level/CSE/GSCEs, NVQ level 1, Foundation Diploma/GNVQ, Entry Level, Basic Skills
2. 5+ O Level/CSE/GSCEs, 1 A Level, 2-3 AS Levels, School Certificate, Higher Diploma, NVQ Level 2, Intermediate GNVQ, City and Guilds Craft, BTEC, RSA, Apprenticeship
3. 2+ A Level/VCEs, 4+ AS Levels, Higher School Certificate, Progression/Advanced Diploma, NVQ Level 3, Advanced GNVQ/City and Guilds/RSA, ONC, OND, BTEC National
4. Degree, Higher Degree, NVQ Level 4, HNC, HND, RSA/BTEC Higher, Professional
5. Other professional / work related, Foreign

Ever worked

Only persons not currently in employment were asked whether they had ever worked. Non-response was low at 1.8% with almost no multi-ticks (around 100) and only 0.8% failed the edit rules. Inconsistency was comprised of three routing filters into the question and two further filters based on Ever worked: if Ever worked was 'yes', then a valid Last year worked was required, and if it was 'no', the remaining questions were not required. Having never worked was 29.2% more frequent in the imputed values than in the observed. This was partly due to resolution of inconsistencies with Last year worked which resulted in Worked in past being changed to 'never worked'. Although Ever worked was changed in a minority of such inconsistencies, these edits represented 28% of the imputed values. There were also some geographic effects; the areas with the highest non-response had amongst the highest observed proportions of persons who had never worked. For example, 32 out of 35 of the largest contributors had a never-worked rate above the national average. This resulted in an adjustment of 0.8% into the Never worked category.

Table 47: Distributions of ever worked

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Worked in past	14,191	81.9	242	52.7	**-29.2**	14,433	81.1	**-0.8**
Never worked	3,137	18.1	217	47.3	**29.2**	3,354	18.9	**0.8**
Total	**17,327**	**100**	**459**	**100**	**0.0**	**17,787**	**100**	**0.0**

Last year worked

Last year worked was only required for persons not currently working who had worked in the past. There were three routing filters and one edit rule which stated you could not have last worked before you were born. Non-response was 10.9% including 0.12% irresolvable responses, and a further 0.3% failed the edit rule/routing. Last year worked was imputed as a discrete variable and is represented in Table 48 as percentiles; the distribution of Last year worked remained unchanged after imputation.

Table 48: Distribution of last year worked

	Percentile						
	1st	***5th***	***25th***	***50th***	***75th***	***95th***	***99th***
Observed	1958	1976	1992	2002	2008	2010	2011
Total (including imputed)	1958	1976	1992	2002	2008	2010	2011

Employment status

All persons who were working, or had worked in the past were required to record their employment status. There was moderate non-response of 4%, including 0.9% multi-ticks. Imputation due to edit failures was low at 0.2%, and related solely to inconsistency with the term-time, age and working/ever-worked routing filters. Being an employee was 2.8% more frequent in the imputed values while being self employed without employees was 2.4% less frequent (Table 49). This produced a negligible adjustment of 0.1% in the total distribution.

Table 49: Distribution of employment status

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Employee	32,806	86.3	1,485	89.0	**2.8**	34,291	86.4	**0.1**
Self employed no employees	3,895	10.2	131	7.9	**-2.4**	4,026	10.1	**-0.1**
Self employed with employees	1,319	3.5	51	3.1	**-0.4**	1,370	3.5	**<0.1**
Total	**38,020**	**100**	**1,668**	**100**	**0.0**	**39,687**	**100**	**0.0**

Occupation

Any person who was working or had worked in the past was asked for their occupation. This was captured in two parts: job title and job description. Automatic coding was applied with clerical coding attempted for unresolved responses. The results are presented for the two sub-populations: those currently working, and those who worked at some time in the past. For those currently working non-response was 2.3% with 0.8% being unresolved responses. Editing was low at 0.3% and consisted of values imputed due to a change in the routing criteria.

The non-response for those not currently in work was higher at 6.5% which included almost 1% unresolved written responses, however there were fewer than 1000 values edited. The imputed values for both groups followed a similar distribution to their observed values (tables 50 and 51), and there were only marginal changes between the observed and total distributions.

Table 50: Distribution of current occupations

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
Major occupation group	N	%	N	%	%	N	%	%
1 (managers / directors)	2,682	10.9	63	9.5	**-1.4**	2,745	10.9	**<0.1**
2 (professionals)	4,294	17.5	100	15.1	**-2.3**	4,394	17.4	**-0.1**
3 (technical roles)	3,100	12.6	76	11.5	**-1.1**	3,176	12.6	**<0.1**
4 (administrative roles)	2,840	11.5	77	11.6	**0.1**	2,917	11.6	**<0.1**
5 (skilled trades)	2,823	11.5	73	10.9	**-0.5**	2,896	11.5	**<0.1**
6 (care / leisure services)	2,322	9.4	60	9.1	**-0.4**	2,383	9.4	**<0.1**
7 (sales / customer services)	2,054	8.4	71	10.7	**2.4**	2,125	8.4	**0.1**
8 (process /plant operatives)	1,774	7.2	49	7.4	**0.2**	1,823	7.2	**<0.1**
9 (elementary roles)	2,703	11.0	93	14.0	**3.1**	2,796	11.1	**0.1**
Total	**24,591**	**100**	**663**	**100**	**0.0**	**25,255**	**100**	**0.0**

Table 51: Distribution of previous occupations

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
Major occupation group	N	%	N	%	%	N	%	%
1 (managers / directors)	1,001	7.4	61	6.5	**-0.9**	1,062	7.4	**-0.1**
2 (professionals)	1,578	11.7	86	9.2	**-2.5**	1,664	11.5	**-0.2**
3 (technical roles)	961	7.1	55	5.8	**-1.3**	1,016	7.0	**-0.1**
4 (administrative roles)	2,045	15.2	152	16.1	**1.0**	2,196	15.2	**0.1**
5 (skilled trades)	1,596	11.8	109	11.6	**-0.3**	1,705	11.8	**<0.1**
6 (care / leisure services)	1,098	8.1	71	7.5	**-0.6**	1,169	8.1	**<0.1**
7 (sales / customer services)	1,405	10.4	102	10.8	**0.4**	1,507	10.4	**<0.1**
8 (process /plant operatives)	1,371	10.2	110	11.7	**1.5**	1,480	10.3	**0.1**
9 (elementary roles)	2,438	18.1	196	20.9	**2.8**	2,634	18.3	**0.2**
Total	**13,493**	**100**	**940**	**100**	**0.0**	**14,433**	**100**	**0.0**

Supervisor status

Those working, or who had worked in the past, were also asked whether they supervised other employees. Non-response was moderate at 4.3% including 0.09% multi-ticks. This question was only affected by the term-time and working-age routing filters, and the 0.2% of values edited related to persons whose routing criteria were changed in an earlier module. Not supervising others was 4.3% more frequent in the imputed values than the observed and there was a minor adjustment of 0.2% in favour of doing no supervision when including the imputed values (Table 52).

Table 52: Distribution of supervisor status

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Supervises others	11,525	30.4	469	26.1	**-4.3**	11,994	30.2	**-0.2**
Does not supervise	26,366	69.6	1,327	73.9	**4.3**	27,693	69.8	**0.2**
Total	**37,891**	**100**	**1,796**	**100**	**0.0**	**39,687**	**100**	**0.0**

Industry

Industry was also required for all persons who had ever worked. Automatic coding was applied before attempting clerical coding for unresolved cases. The results are presented for the two sub-populations: those currently working, and those who worked at some time in the past. For those currently working, non-response was 7.2% with 1.9% being unresolved responses. Editing was low at 0.3% and consisted of values imputed for respondents whose routing criteria were changed by imputation, for example because their age had been imputed to be over 15. The imputed values for persons currently working (Table 53) followed a similar distribution to the observed values, and there were negligible changes between the observed and total distributions.

The non-response rate for those not currently working was high at 17.2 %, including 3.7% unresolved written responses, however there were fewer than 1,000 values edited due to edit rule failures. While most of the imputed values followed a similar distribution to that observed (Table 54); manufacturing was 3% higher in the imputed values leading to a 0.5% increase when including the imputed values.

Table 53: Distribution of industry where currently working

All responding persons, England and Wales, 2011 — Thousands

Major industry group	Observed responses N	Observed responses %	Imputed responses N	Imputed responses %	Difference (imputed - observed) %	Total including imputed N	Total including imputed %	Change (total - observed %
A, B	247	1.1	29	1.5	**0.4**	276	1.1	**<0.1**
C (manufacturing)	2,073	8.9	194	10.2	**1.3**	2,267	9.0	**0.1**
D, E	299	1.3	23	1.2	**-0.1**	322	1.3	**<0.1**
F (construction)	1,770	7.6	170	9.0	**1.4**	1,940	7.7	**0.1**
G (trade, vehicle repairs)	3,706	15.9	312	16.4	**0.6**	4,018	15.9	**<0.1**
H, J	2,068	8.9	171	9.0	**0.2**	2,240	8.9	**<0.1**
I (accommodation/food)	1,272	5.4	118	6.2	**0.8**	1,390	5.5	**0.1**
K (finance, insurance)	1,014	4.3	64	3.4	**-1.0**	1,078	4.3	**-0.1**
L, M, N	3,011	12.9	236	12.4	**-0.5**	3,247	12.9	**<0.1**
O (public, defence)	1,425	6.1	92	4.9	**-1.2**	1,517	6.0	**-0.1**
P (education)	2,357	10.1	169	8.9	**-1.2**	2,526	10.0	**-0.1**
Q (health/social work)	2,955	12.7	226	11.9	**-0.8**	3,181	12.6	**-0.1**
R,S,T, U	1,158	5.0	95	5.0	**0.1**	1,253	5.0	**<0.1**
Total	**23,356**	**100**	**1,899**	**100**	**<0.1**	**25,255**	**100**	**<0.1**

Table 54: Distribution of industry where not currently working

All eligible responding persons, England and Wales, 2011 — Thousands

Major industry group	Observed responses N	Observed responses %	Imputed responses N	Imputed responses %	Difference (imputed - observed) %	Total including imputed N	Total including imputed %	Change (total - observed %
A, B	177	1.5	49	2.0	**0.5**	226	1.6	**0.1**
C (manufacturing)	1,740	14.6	437	17.6	**3.0**	2,177	15.1	**0.5**
D, E	162	1.4	35	1.4	**<0.1**	196	1.4	**<0.1**
F (construction)	779	6.5	180	7.2	**0.7**	958	6.6	**0.1**
G (trade, vehicle repairs)	2,072	17.3	426	17.2	**-0.2**	2,497	17.3	**<0.1**
H, J	899	7.5	190	7.7	**0.2**	1,089	7.5	**<0.1**
I (accommodation/food)	819	6.9	162	6.5	**-0.3**	982	6.8	**-0.1**
K (finance, insurance)	420	3.5	63	2.5	**-1.0**	483	3.3	**-0.2**
L, M, N	1,219	10.2	235	9.5	**-0.7**	1,454	10.1	**-0.1**
O (public, defence)	722	6.0	125	5.0	**-1.0**	847	5.9	**-0.2**
P (education)	1,142	9.6	211	8.5	**-1.1**	1,352	9.4	**-0.2**
Q (health/social work)	1,248	10.4	264	10.6	**0.2**	1,513	10.5	**<0.1**
R,S,T, U	553	4.6	105	4.2	**-0.4**	658	4.6	**-0.1**
Total	**11,952**	**100**	**2,481**	**100**	**0.0**	**14,433**	**100**	**0.0**

A,B agriculture, forestry and fishing, mining quarrying

D,E electricity, gas, steam, air conditioning / water supply, sewerage, waste management

H, J transport and storage / information and communication

L, M, N real estate / professional, scientific and technical / administrative and support services

R,S, T,U arts, entertainment, recreation / other services / households as employers / extraterritorial bodies

Workplace address

The address of a person's workplace was only required if the person was working in the last week. The question was divided into three parts, tick boxes, a postcode or a country. The overall non-response rate for the question was 12%, however as a proportion of the values within each category,12.5% of postcodes, 9.8% of countries and 9.9% of tick responses were imputed for non-response and a further 1.8%, 3.1% and 3.2% of values were edited in each category respectively. Edit failures comprised inconsistencies with the term-time, age and working routing filters as well as responding for more than one address category. The distribution of the type of workplace was similar in the observed and imputed values, with a minor change of 0.2% between working from home and having a UK postcode (Table 55). A separate clerical matching process updated 3% of imputed postcodes and 1% of imputed countries after imputation.

Table 55: Distribution of workplace address

All eligible responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Mainly at/from home	2,390	11.0	330	9.3	**-1.7**	2,720	10.8	**-0.2**
Offshore	39	0.2	5	0.1	**<0.1**	44	0.2	**<0.1**
No fixed address	1,781	8.2	300	8.4	**0.2**	2,081	8.2	**<0.1**
UK (postcode)	17,452	80.4	2,919	82.0	**1.6**	20,371	80.7	**0.2**
International (country)	34	0.2	5	0.1	**<0.1**	39	0.2	**<0.1**
Total	**21,695**	**100**	**3,559**	**100**	**<0.1**	**25,255**	**100**	**<0.1**

Hours of work

Hours of work was answered by those currently working, non-response was low at 3.4%, including 3,326 (0.01%) multi-ticks. An additional 1.8% of values were imputed because their routing criteria had changed during imputation. The observed and imputed values followed a similar distribution with a slight tendency towards imputing fewer than 31 hours per week. There were minor changes to the total distribution when including the imputed values.

Table 56: Distribution of hours worked

All responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
Per week	N	%	N	%	%	N	%	%
Up to 15	2,301	9.6	172	13.1	**3.5**	2,474	9.8	**0.2**
16 to 30	4,684	19.6	287	21.9	**2.3**	4,971	19.7	**0.1**
31 to 48	13,778	57.5	699	53.3	**-4.2**	14,478	57.3	**-0.2**
49 or more	3,179	13.3	153	11.7	**-1.6**	3,332	13.2	**-0.1**
Total	**23,943**	**100**	**1,312**	**100**	**0.0**	**25,255**	**100**	**0.0**

Method of travel to work

Method of travel to work was only required for those currently working. Non-response was low at 3.2% with an additional 1.8% of values edited. Editing occurred due to inconsistency with the age, term-time and working routing filters, and where persons under the age of 17 stated that they were driving to work. The observed and imputed distributions (Table 57) were quite similar, except for driving to work which was 5.7% less frequent due to the edit rule for age and driving. In turn there was a minor adjustment of 0.3% in the total data when including the imputed values.

Table 57: Distribution of transport

All responding persons, England and Wales, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (total - observed
	N	%	N	%	%	N	%	%
Work at home	1,316	5.5	86	6.9	**1.4**	1,403	5.6	**0.1**
Underground / metro	870	3.6	53	4.2	**0.6**	923	3.7	**<0.1**
Train	1,231	5.1	55	4.4	**-0.7**	1,287	5.1	**<0.1**
Bus, coach	1,711	7.1	110	8.7	**1.6**	1,820	7.2	**0.1**
Taxi	123	0.5	8	0.6	**0.1**	131	0.5	**<0.1**
Motorcycle /scooter	193	0.8	9	0.7	**-0.1**	202	0.8	**<0.1**
Driving car or van	13,978	58.2	660	52.6	**-5.7**	14,639	58.0	**-0.3**
Passenger car / van	1,220	5.1	77	6.1	**1.0**	1,297	5.1	**0.1**
Bicycle	678	2.8	34	2.7	**-0.1**	712	2.8	**<0.1**
On foot	2,524	10.5	155	12.3	**1.8**	2,679	10.6	**0.1**
Other	155	0.6	9	0.7	**<0.1**	164	0.6	**<0.1**
Total	**23,999**	**100**	**1,256**	**100**	**0.0**	**25,255**	**100**	**0.0**