**An Initial Proposal for the Quality Assurance of the 2021 Census Data**

**1. Introduction**

The 2021 census results will be of value when they are sufficiently accurate and where users have adequate information on uncertainty to use them appropriately.

The Census and Data Collection Transformation (CDCTP) Roadmap specifies the level of accuracy required in order to meet the high-level goals of the census. This includes the requirements that the census population estimates have a 95% confidence interval of +/- 0.2 percentage points nationally and +/- 3 percentage points for all local authorities, with a bias of less than 0.5 per cent in the national estimate. More generally, we need to ensure that the estimates for each topic within the census are accurate enough to meet user requirements.

Ensuring the accuracy of the census results requires the meeting of quality standards for all aspects of the census operation – from easily understood questions and convenient ways of responding to the census to accurate identification of addresses which should be sent an invitation to respond, to application of statistical disclosure methods which protect the confidentiality of responses with minimal impact on the final results. These requirements are captured in a set of design goals and principles. These specify the standards each part of the census operation must meet in order to meet the high-level quality standards related to each dimension of statistical quality as set out by the UNECE. Progress against these design goals and principles will be monitored during the further development of the census.

If the design goals and principles are satisfied we will expect that the required standards for accuracy will be met. However, once the census data is collected, there remains the task of ensuring that the census results are indeed sufficiently accurate. This work will be led by the Quality Assurance of Census Data team. This paper outlines an initial proposal for this work.

We expect our plans to evolve in the light of stakeholder comments on this, and subsequent papers; as a result of experience from the Rehearsal in 2019/2020; and as a result of developments in the availability of tools and alternative data sources which can be used in the work.

**2. Quality Assurance Strategy**

The QA of the 2021 Census data needs to achieve the following:

* Ensure that the Census results provide a reliable basis for decision-making;
* Give data-users confidence that the Census results are fit for purpose;
* Minimise the risk that users of the Local Authority estimates of population would want to challenge the results when they are first released;
* Minimise the risk of obvious errors in the Census estimates – however minor – which might undermine the overall credibility of the statistics;
* Leave a legacy of methods, tools and skills for QA of post-2021 population statistics.

To achieve this, we will adopt the following strategic principles.

* Quantify accuracy wherever possible – being able to provide quantitative estimates of various aspects of error will allow us to prioritise quality assurance activities; help in making evidence-based decisions on any interventions required; allow us to assess the success of the Census against the agreed success criteria; and subsequently help users understand the reliability of the results;
* Ensure every QA check has a purpose – to inform a recommendation on an intervention or to help people use the data sensibly;
* Use corporate standards to produce flexible methods and tools – every census is different and we will need to be able to react quickly to unexpected anomalies in the census data, and this will develop skills and tools which will be useful for post-2021 population statistics;
* Work closely with teams developing transformed population statistics to develop, as far as possible, standard approaches to quality assuring these;
* Be outward looking – sharing experiences and learning from other census-taking organisations; and working in partnership with data-users to understand what would give them confidence in the Census estimates, and to provide that;
* Conduct quality assurance in parallel to processing – starting QA activities as soon as the first data comes into ONS so we can identify issues as soon as possible; and we can get as much (LAs or topics) as possible provisionally accepted before estimation and adjustment takes place;
* Take explicit account of all lessons learnt from 2011;
* Quality Assurance will be split into two themes: QA of processes and validation of Census estimates:

For QA of processes:

* Process Owners are responsible for QA of their process - the QA team is responsible for ensuring that an appropriate QA plan, including the development of meaningful quality indicators, is in place and is being followed; and for bringing quality information on all processes together to form a holistic picture of data quality.

For validation of estimates

* National and LA estimates by age and sex are the priority - these will be included in the first release, and are the fundamental outputs from the Census.
* Validation is user-centred - we approach QA with the ‘eyes of a user’ – prioritising statistics with their importance to a data-user, and using checks and comparator data sources which a user would be likely to use.
* Use expertise across ONS – making sure that topic experts on subjects such as demography, housing, labour market and health are assessing the stories shown by the provisional census results in the context of other evidence and trends;
* The team responsible for validation is not responsible for recommendations on whether an adjustment is required – though any recommendation would be informed by the analysis underpinning the validation.

**3. Assurance of Processes**

The diagram provides a very simplified picture of the census process. Error in the census data can be introduced at any stage of the data journey.



Our approach to the assurance of the results of each process is to:

* Understand each process and the nature of error which might be introduced within it;
* Estimate the expected size and impact of the error within each process in order to prioritise QA work effectively;
* Define metrics which allow us to understand the size and nature of the error relating to each process during the 2021 Census.

For each process we are working closely with the owner of the process, drawing on their expertise to ensure that the defined metrics are meaningful and can be easily produced, avoiding duplication of work, and ensuring methods of intervention are available where the results of a process are causing concern. Our early involvement in the development of the diagnostics ensures that the impact of process quality on the accuracy of the data is taken into account.

(We acknowledge a complication in that error in one process might be corrected in a subsequent process and will seek to ensure that this possibility is appropriately reflected in the assessment of error).

See ***Annex A – Processes*** for further detail.

**4. Validation of Estimates**

This strand of work considers the likely accuracy of the census results themselves.

For this work we will use a suite of tools and methods that will provide standard automated checks but that will allow us to focus on areas, population groups or topics where there are inconsistencies or need for further analysis. This evidence will be reviewed by Assurance Panels who will make decisions on whether estimates are fit for purpose or require further work or adjustment (see Section 5).

With the time constraint of publishing the first census outputs by the end of March 2022, we will need to maximise the amount of QA work which is conducted in parallel with processing. Each variable collected on the Census will be subject to assurance. However, we will prioritise quality assurance activities relating to estimates, based on whether variables are:

* Used in population coverage estimation or imputation processes
* Used to produce LA population estimates by age and sex
* New variables not previously asked in a census
* Used as part of a key census output

The tools that we develop for quality assurance will be designed for flexibility, enabling analysts to interrogate data and adapt analyses as needed, rather than being constrained to a previously specified set of analyses. We will also seek to maximise commonalities with quality assurance developments on the admin-based population statistics – wherever possible, using common tools and approaches, and drawing on expertise elsewhere in ONS on alternative data sources and methods of quality assurance. Annex B summarises the planned systems and standard checks described in the following sections.

4.1 Internal Plausibility and Consistency

We are investigating a range of data science methods, modelling techniques and analysis of 2021 Census data itself to develop methods, tools and processes that will enable data-driven prioritisation of checks and analysis through identification of outliers and anomalies. These tools will help us to efficiently check the plausibility and consistency of estimates to identify areas requiring further analysis.

The Census is a highly complex and extensive operation and there are many ways that errors could be introduced. These include respondent error, through misunderstanding a question or entering information incorrectly, or errors in the processing of census data, for example scanning, coding or imputation errors. These errors can result in local under or over coverage and it is important that we are able to identify and reduce any systematic errors where possible.

*4.1.1 Outlier detection*

Building on initial research commissioned from the University of Southampton to evaluate different outlier detection methodologies, Methodology Data Scientists are researching and aiming to develop a set of tools using data science methods to identify

* Highly unusual records or outliers e.g. travel-to-work anomalies, groups of records with unlikely commuting patterns
* Groups of records with unusual characteristics e.g. a high number of people with the same occupation within a small geography
* Spurious distributions within areas or population groups within the data e.g. unexpected highly skewed demographic distributions in an area
* Unexpected issues within the data

These prototype tools are being developed using 2011 Census data and will be evaluated using data collected in the Rehearsal. The outcomes of this also have potential uses in other business areas across ONS. Simpler approaches to outlier detection will also be developed by the QA team as described in Annex B.

*4.1.2 Modelling approaches*

With the increasing role of integrated data within the transformation of the population statistics system, we need to explore other alternatives to comparison with administrative data. We also need to find ways to focus and prioritise our activities to quality issues where there could be the most impact on final census estimates.

We plan to explore the potential of different modelling approaches to contribute to the assurance of estimates in a number of ways, including

* Data reduction – understanding which variable relationships to prioritise in our assurance activities
* Model comparison – comparing model parameters based on 2021 Census data to those generated from previous census or alternative sources as an indicator of similar/expected relationships that could be used to highlight potential quality issues
* Produce modelled estimates – use models to generate expected estimates that we could use as comparators

We initially plan to use log-linear models and the development of a 2021 comparator dataset (fundamentally based on rolling forward 2011 Census data) for the above and welcome additional suggestions for approaches.

4.2 Comparison with other sources

Quality assurance of 2011 Census estimates used a range of independent sources of population information as comparators, including GP Patient Registers, Council Tax and School Census information, as well as information from the 2001 Census. These comparators were used to assess the numbers and characteristics of key population groups.

In 2021, we will aim to make best use of all sources available, including (rolled-forward) 2011 Census, mid-year population estimates, research population outputs (including Statistical Population Datasets (SPDs)), administrative, survey or other sources. The use of alternative sources will be based on assessment of their statistical quality e.g. coverage, definitional differences and timeliness. We will ideally condense as much of the information as possible into a single comparator data source allowing efficient comparisons against the census data, but recognise the need to be able to understand and explain differences between the census results and individual alteranative sources.

We are especially interested in data sources available to local authorities, such as Council Tax data, that could provide additional local intelligence and comparators to support our assurance processes at small and local area levels.

To determine how we will use these sources, we are working with teams across ONS, and, in particular, elsewhere in Statistical Design and Research, to understand the use of alternative sources across the Census design and how the population statistics system will have evolved by 2021, so we can ensure we are using the most appropriate comparators. Our approach to comparator analysis will be to automate checks to be applied to all areas and population groups where possible, to highlight, and direct resource to where further investigation and/or more complex analysis may be needed. We will prioritise checks on key information and, in particular, checks related to the census LA population estimates by age and sex.

Uses of alternative data could include

* Automated comparison between aggregated sources/local information and census estimates to highlight areas for further investigation
* Comparison with single sources, aggregated from national to OA/postcode level, to better understand any quality issues at small area level
* Researching the potential to use sources available in generation of modelled comparator estimates
* Linking 2021 Census data to other sources where further investigation/triangulation is required to inform QA outcome decisions

Evaluation of 2011 Census Quality Assurance recommended comparisons with the previous census, particularly at lower levels to highlight any issues affecting small areas, for example, placement of communal establishments. It was also identified, where applicable, as the most reliable source for checking distributions in topic analysis, in conjunction with topic lead expertise. Other potential uses include the use of rolled forward characteristics estimates as additional comparators and use in further analysis of change over time. With these recommendations in mind, 2011 Census data will be a key comparator within our processes.

We will also use information generated from the 2021 census collection operation to understand the quality of estimates. We will use census field operation indicators, including overall return rates, local variation and other field intelligence in reviewing our analysis and checks.

4.3 Demographic and Topic Analysis

In addition to the use of comparator sources discussed above, we will also undertake more specialised analyses to understand the quality of our estimates.

*4.3.1 Demographic analysis*

We will work with the ONS Population Statistics Division to develop demographic analysis and tools to assess the quality of our population estimates. Demographic indicators will be produced based on 2021 Census data and compared with indicators produced using other sources, including ONS mid-year population estimates and administrative data. These, and other analyses could include

* Sex ratios – the ratio of the number of men per 100 women in the census population estimates could be checked by single year of age to identify any implausible values
* Fertility rates – calculated based on census population estimates of women aged 15-44. Unusually high or low fertility rates could be compared to rates calculated from mid-year population estimates and other sources, with further analysis of variation within Local Authorities
* Mortality rates – age-standardised mortality rates could be calculated using census population estimates and comparing to equivalent rates calculated using mid-year population estimates and further research carried out on age-specific mortality rates based on census population estimates
* Comparison of census-based estimates with official migration estimates
* Analysis of the ONS Longitudinal Study to understand coverage quality and plausibility of results

We are working with the ONS Longitudinal Study (LS) team to identify opportunities for its use in the assurance process. Analysis of the LS was used to identify people who were missed in the 2011 Census (including legitimate reasons) and those who had been included more than once at different locations.

We can also link census data to itself and other sources to compare distributions and assess quality issues arising from questions on address one year ago, second residences and visitors. We will also identify areas of population change from e.g. migration, or population cohorts and undertake further research in these areas.

*4.3.2 Topic analysis*

The purpose of topic analysis is to examine the census results relating to each topic covered in the census (for example, housing or health) to check for any evidence of systematic errors in the results for that or related topics, and to understand, as far as possible, the reasons for any differences between the census results and other sources of information on that topic.

We will prioritise these checks based on whether the information is derived from a question not previously asked on a census, a key output, contribution to the production of population estimates or coverage estimation and imputation, or where any potential errors could have the greatest impact and potentially negatively affect the public trust in census estimates.

We will work with topic experts within and outside ONS to devise and review the prioritised topic analysis checks. They will be asked to review the development of our checks, in particular by reflecting their knowledge of the ways in which the data will be ultimately used, and, where appropriate, invited to contribute to the quality assurance work on the data itself.

This work will also encompass any assurance required for Census outputs using alternative data sources (‘enhanced Census outputs’), where we will be particularly reliant on experts on those data sources outside the Quality Assurance of Census Data team.

**5. Development and Management of the Quality Assurance Process**

5.1 Development

We will seek stakeholder views in the development of our proposals for the Quality Assurance process to ensure that the QA carried out will reflect good practice and include, where possible, checks that other stakeholders would plan to conduct, and that it will provide the evidence required by stakeholders to understand the reliability of the Census results. Particular ways in which we will seek stakeholder views are:

* Developing plans in conjunction with suppliers (e.g. Census Processing and Data as a Service) to ensure that the plans are feasible given the data and infrastructure likely to be available;
* Seeking comment on early proposals from the Census Research Assurance Group (a group of methodological and demographic specialists, from inside and outside ONS, and staff who have previously worked on Census QA) (this paper reflects comments made at that Group in March 2019);
* Seeking comment on proposals (in particular identifying any additional assurance that should be carried out) from Census Advisory Groups covering key users of Census results;
* Seeking endorsement of the proposed approach from the External Assurance Review Panel;
* Participating as appropriate in stakeholder activities led by the Statistical Design and Research Strategy and Communications team, the Census Stakeholder and Communications team, and the Census Outputs and Dissemination team;
* Collaborating with a number of LAs to understand their requirements and issues around the quality of census data.
* Inviting comments from all stakeholders on a published proposal for quality assurance.

Proposals developed in the light of the above, experience from the 2019/20 Rehearsal, and other developments in ONS, will be formally signed off at appropriate level Boards within the Census and Data Collection Transformation Programme (CDCTP).

5.2 Management

The Quality Assurance process will be managed as follows.

The Joint Heads of Census Quality Assurance will be responsible for agreeing and co-ordinating the work of the various teams involved in QA of the data. These teams will be:

* The Process QA team - ensuring that the work described in Section 3 is completed, working closely with Census teams responsible for each process.
* Topic QA groups (Housing; Labour Market; Education; Health; Income; Ethnicity/Identity/Language/Religion; Gender Identity/Sexual Orientation) – ensuring that the work described in Section 4.3 is completed. These Topic QA groups will be informal groups bringing expertise from across ONS
* Population Estimates QA team – ensuring that the main demographic statistics are plausible for England and Wales and its constituent LAs (and as a lower priority – other geographies used for statistical outputs). This team will be closely linked with the Methodology Census Coverage Estimation and Adjustment teams and draw on expertise from Population Statistics Division and teams working on transformed population statistics.

Each of these teams will be responsible for producing a stream of evidence on the reliability of the Census figures. We currently propose that this evidence will be considered by a set of QA Panels, consisting of a small number of ONS experts on that subject who were not involved in the QA process. The recommendations of these panels – whether to accept the data or whether an intervention is required – would be passed to a High-Level QA Panel, containing ONS staff and external experts, for agreement. The final conclusions of this High-Level Panel would be provided to the National Statistician as the evidence needed to obtain final sign-off of the data.

**Annex A Processes**

**A.1 Processes**

This Annex describes the approach to understanding where error in the data could be introduced within the census process.

It is important to remember that the workpackage for processes looks at the stages that the data travels through to understand if error can be introduced which will either directly, or indirectly, impact the census results. It does not cover operational quality or progress towards programme quality goals.

**A.2 Data touchpoints**

The initial approach required that we understand where the touchpoints for a process and the data are. The Business Architecture and Requirements team (BAR), responsible for translating the business needs into a cohesive set of requirements, have developed an artefact that maps every process to the corresponding service. This formed the basis of our initial research.

**A.3 Service leads**

By working with the process owner and their teams, we encourage discussion about the process and potential concerns. As the business areas are in the early stages of defining their QA, getting involved at the beginning of these activities ensures that quality is considered.

**A.4 Approach**

Based on our current understanding, we prioritised the list:

* H - there is a **high**likelihood that the service may introduce material error
* M - there is a **medium**likelihood that the service may introduce material error
* L - there is a **low**likelihood that the service may introduce material error
* C - the service has little impact on the data or is controlled via KPIs etc

**A.5 Prioritised list**

Each category was then prioritised based on how difficult it was to understand and the perceived complexity. Timing, as to when the service is required was also considered.

|  |  |  |  |
| --- | --- | --- | --- |
| **Priority** | **Process theme** | **Priority** | **Process theme** |
| 1 | Address | 2 | Frame |
| 3 | Processing | 4 | Matching |
| 5 | E&I | 6 | Estimation |
| 7 | Adjustment | 8 | SDC |
| 9 | Capture | 10 | Transfer of data |
| 11 | Authenticate | 12 | RCA |
| 13 | FWMT | 14 | Outputs |
| 15 | UAAs | 16 | Field |
| 17 | Community events | 18 | processing access |
| 19 | Contact/support | 20 | Recruit |
| 21 | CCS\* |  |  |

\*it is recognised that CCS is key to the census but uses a lot of the processes covered in the other priorities.

**A.6 Impact Maps**

By mapping the concerns and asking, ‘so what?’ we are able to understand actions and/or mitigation required.

Working with the service lead, each process, or sub-process, is assessed to understand, in effect, what can go wrong and what preventative measures are in place. If the event does occur, the impact on the data can be understood, and the appropriate actions can be taken.

An example Impact Map is shown below.



**A.7 Initial Proposal**

An initial proposal for the quality assurance of each process is outlined in the remainder of this Annex.

**A.7.1 Address Frame**

There are two aspects of the Address Frame which are of particular importance to the accuracy of the census results even after the data is collected. The first is that census returns are accurately georeferenced (that is, an address being correctly assigned to a geographical point). The second is that the Address Frame’s coverage and classification of household and Communal Establishment (CE) addresses is sufficiently accurate to avoid systematic inaccuracies in the type and location of imputed records.

The quality assurance of this geographical information in advance of the census will be led by the Address Register and Geography teams in the context of the quality standards agreed with Census. This assurance is likely to include:

* A field address check covering a large number of addresses in each local authority.
* Desk research for larger communal establishments.
* Evaluation of the accuracy of the geo-referencing, and, if possible, identification of any large CEs which cross LA (or other area) boundaries.

Quality assurance following the census data collection might include:

* Checks of census Output Area (OA) population estimates against those implied by the Address Frame.
* Checks of census Communal Establishment OA population estimates against those for 2011 for that and neighbouring OAs (to identify whether a large CE might have been missed or assigned to the wrong OA).
* Evaluation of information (e.g. dummy forms or requests for forms) from the census collection operation.

**A.7.2 Data Collection**

**A.7.2.1 Completeness of Enumeration**

Whilst the Census process is designed to adjust for non-response, the quality of results at a local level may be compromised if there is a failure to conduct the enumeration effectively in that area – for example, resulting in a very poor response rate within an area or a large communal establishment being missed.

Any such problems may be identified as part of the response management work conducted within the census operations and the Quality Assurance team will be part of a data quality management forum at which these issues would be discussed, ensuring that the team is aware of such issues which might be relevant for later quality assurance work. Further assurance of the quality of the enumeration will be provided by work described in Section 4 (in particular 4.2) of this paper.

**A.7.2.2 Accuracy of Collected Data**

We know that the information contained is not always correct – for example, the respondent may have misunderstood the question, accidentally selected the wrong option, or have provided incorrect information as a proxy response for someone else in the household. The obvious approach to understanding the scale of these errors is to compare the information from a particular census form with that collected for the same household or person in another source.

To estimate the expected scale of this error we can use the results of the 2011 Census Quality Survey (CQS), the analysis of the linkage of the 2011 Census data with the Longitudinal Study, and information from testing of the 2021 questions and collection tools.

Possible approaches in 2021 include:

* Linking Census returns to a 2021 Census Quality Survey – this would allow coverage of all the Census questions and the collection of additional information such as whether the individual completed their census return in person or whether it was a proxy response.
* Linking Census returns to the Census Coverage Survey (CCS) – this will provide a much larger sample than the CQS but will only cover a subset of the Census questions. Though the CCS returns were linked to the Census returns in 2011 the information was not used as part of the quality assurance.
* Linking Census returns to alternative (administrative or survey) data sources, including the 2011 Census and the Longitudinal Study – this would be likely to provide a large number of matched records, but a limited set of variables which could be compared.
* Analysis of 2021 data by mode of collection, focusing on possible quality effects related to mode – for example, if responses to drop-down menu questions on the online form are problematically different to those on paper.
* Outlier detection methods described in Section 4.1 and the identification of spurious (i.e. clearly fictional) responses.

Once we have estimated the likely scale and nature of this error we can estimate the likely impact of the error – if, for example, a question is answered incorrectly by 1% of respondents this may have a large impact if this substantially distorts the estimates for a small group, or a very small impact if the individual errors largely ‘cancel each other out’ when aggregating results.

**A.7.3 Data Load and Coding**

Returned paper questionnaires need to be scanned. Both paper and online returns need to have responses to some questions coded. Both of these processes risk introducing error.

To estimate the expected scale of this error we can use the evaluation of the 2011 capture and coding, the quality standards agreed with the Supplier for this work, and (as part of the total error in responses) the results of the 2011 Census Quality Survey.

Possible approaches in 2021 include:

* Checking a sample of returns to understand the rate of error in capture and coding and the ‘transition matrix’ of errors (this will be important in understanding any potential bias in the results).
* Evaluation of the completeness of the coding lists (e.g. how many write-in answers could be automatically assigned to an existing code).
* Using the results of the dummy imputation exercise described under Item Imputation to identify possible systematic errors.
* Checks on specific identified risks (e.g. ‘major’ in job title being interpreted as military officer when not appropriate).

**A.7.4 Resolution of Multiple Records/Removal of False Persons**

This process seeks to identify and remove incorrect census returns which relate to a person already correctly covered on another return from the same or very close location, or which do not contain sufficient information to be treated as a response.

To estimate the expected scale of this error we can use evidence from the linkage of the 2011 census data with the Longitudinal Study together with assumptions on changes in response patterns as a result of greater use of online/individual response.

Possible approaches in 2021 include:

* Analysis of the linkages of the census returns with data sources as described under Completion of the Census Form
* Analysis of names and other variables to identify false people not identified in the RFP process.

**A.7.5 Item Imputation**

Item imputation is the process of imputing for a missing or clearly incorrect value on a Census return, and is conducted in the Census using a donor-imputation methodology. Quality issues around item imputation include predictive accuracy (how accurate is the method in imputing the correct value for an individual); distributional accuracy (how accurate is the method in reflecting the true multivariate distributions of variables) and ‘spikes’ - where the same donor is selected for multiple cases of missing values when that donor has rare characteristics.

To estimate the expected scale of this error we can use the evaluation of the 2001 and 2011 imputation processes, looking at both the rate of imputation for each item and the accuracy of that imputation.

Possible approaches in 2021 include:

* Comparing imputed values against those collected in the CQS/CCS/other data sources.
* Dummy imputation on 2021 data – applying the observed pattern of missingness to records without missing data; applying imputation to those records, and comparing the results with the collected values.

For each of these approaches it will be important to derive the ‘transition matrix’ for imputation (as described under Coding and Capture), and to consider impacts on multivariate distributions.

**A.7.6 Coverage Matching, Estimation and Adjustment**

Quality assurance of these processes will be largely conducted separately by the Coverage and Estimation team. The adjusted estimates resulting from this process will be covered by the QA processes described in Section 4.

**A.7.7 Disclosure Control**

The statistically adjusted census data is subject to two statistical disclosure control processes before publication. These processes are targeted record swapping, where the geographical location of some records is changed, and post-tabulation cell key perturbation, covered in Section A.7.9.1.

 We can estimate the expected scale of the error introduced by record swapping by looking at the evidence from 2011 in conjunction with any change in the parameters and rules being applied in this process.

The approach in 2021 is expected to be a combination of the above and a simple comparison of the adjusted figures with the unadjusted.

**A.7.8 Create/Amend/Revise Geographies**

Whilst provisional 2021 Output Area boundaries will be derived before the census data is collected, the final boundaries will be derived by the Methods, Data, Research (MDR) Geography team using the Census database following coverage adjustment. That team will be responsible for the main quality assurance of that work, with the Census Quality Assurance team contributing where required.

**A.7.9 Outputs**

The main responsibility for the Census Quality Assurance team is the quality assurance of the data itself rather than outputs derived from the data. However, there are some aspects of the Outputs process which the team will work on.

**A.7.9.1 Cell Key Perturbation**

Cell key perturbation is a statistical disclosure control method which introduces small adjustments to cells within output tables. We can confidently estimate the expected scale of error introduced by the post-tabulation adjustment from first principles – that is, derived mathematically from the properties of the perturbation algorithm. As with Record Swapping, the quality assurance approach in 2021 is expected to be a combination of the above and a simple comparison of the adjusted figures for a particular set of tables with the unadjusted figures.

**A.7.9.2** **‘Pre-canned’ Tables**

Where outputs are tables specified as standard outputs, cell counts produced using the Outputs team’s tabulation tool will be compared (for all or selected areas) with counts produced using an independent system. The primary purpose of this is not to check that the tabulation tool can aggregate data correctly, but to help identify any errors or ambiguities in the coding of tables (for example, in presenting non-standard classifications).

**A.7.9.3 Origin-Destination Outputs**

In addition to the planned QA of origin-destination related variables covered within the work described in Section 4, the Quality Assurance team will conduct additional assurance of the complex origin-destination outputs to help ensure the tabulations are correct.

**ANNEX B: Validation Tools and Checks**

B.1 The tools described below for use by the Quality Assurance of Census Data team will be developed in Python and implemented in DAP, drawing on views of the census database which will allow analysis on the data as it stands, or stood, at any point in time or processing cycle. Standard checks are identified for each system, but these will be a starting point for further analysis as required.

|  |  |
| --- | --- |
| **Tool** | **Description** |
|  |  |
| ***Outlier Identification System*** | A set of programs which identify individual records which seem unlikely to be correct (allowing investigation if these are concentrated by geography or topic) |
|  Spurious Response Identification | Simple and easily updated program to identify e.g. unlikely names or the effects of organised attempts to provide systematically inaccurate data. |
|  Data Science Outlier Identification | Application of Data Science methods to identify unlikely records |
|  Outlier Summariser (summarises outliers by topic and separately by OA) | Simple analysis tool which takes output from the previous Identification tools and summarises results by topic and geography (OA level), to inform discussions on whether action needs to be taken on any group of outliers. |
|  |  |
| ***Demographic Analysis Tool*** | This is the main tool used for checking the population estimates by age and sex for any area selected. It will draw on comparator data sources (to be identified).  |
| *Checks to be implemented* |  |
|  Population by age/sex: current and predicted final | Meaningful demographic analysis can start early in the processing timeline by estimating final population estimates based on data received to that point (e.g. using forecast response profiles)  |
|  Population by age/sex: comparator sources | The census estimates are compared with other sources (e.g. MYEs, SPDs, 2011 Census/SAPE for OAs, births for recent years) based on an understanding of the issues with each data source.Comparisons will be accompanied by a summary measure of how well the two sources correspond (e.g. the residual from a regression of the census estimates against comparator sources).This will include specific checks for ages above 90 against evidence from death registrations. |
|  Sex ratios | Ratio of males to females by SYOA: checked for spikes and single value summary for each area |
|  Fertility rates, mortality rates | These demographic rates are calculated using registrations data as needed. They are examined for spikes in the age distributions, and summarised as single values (TFR/Life Expectancy at Birth) to compare an area’s change since 2011 with those of similar areas. |
|  Types of address | Household addresses (split as occupied, vacant, second homes) compared to e.g. Council Tax data. |
|  Households | Numbers of households and the size distribution (number of usual residents, number of visitors) versus a comparator. This will also provide any evidence of the effects of limits of form size (people not requesting continuation forms) |
|  Communal Establishments | To include:Distribution of CEs by size and type compared to 2011.Check numbers of individuals recorded at each large CE (size to be decided) against ‘expected’ value.Location of large CEs consistent with 2011 (where appropriate) |
|  Communal Establishment populations | Size and sex/age distribution; change since 2011 compared to national pattern of change. |
|  Migration | Checks of international, cross-border and internal migration flows against comparator sources |
|  Armed Forces populations | Check against comparator source |
|  Qualitative information on LA population and issues | To be prepared ahead of census to help put census results in context and understand apparent anomalies |
|  For all above: automated report production for national; region; county; LA | For consideration by QA panels and others  |
|  |  |
| ***Topic Analysis Tool*** |  |
|  For any area/topic selected |  |
|  Distribution by category against comparator source |  |
|  Location quotient for selected categories | A measure of the relative concentration of the category compared to other areas: for comparison with corresponding values in 2011 |
|  Age standardised measures | Allowing meaningful comparisons of populations with different age structure |
|  Loglinear measures of association with other variables | Assessment of how relationships between variables (e.g. between ethnic group and religion) may have changed since 2011 and whether anomalies in estimated relationships for certain areas suggest need for further investigation. |
|  |  |
| ***Write-in Answer Analysis Tool*** |  |
| *Checks to be implemented* |  |
|  Distribution of write-in answers | Check for any issue in the coding of these write-in answers (for example if a large number of people write in the same response for religion it will be particularly important that this is coded to the correct standard category)  |
|  Evidence of truncated answers  | Check for issues caused by limited space on form. |
|  Mode effects  | After allowing for confounding factors - e.g. as a result of dynamic drop down menus.  |
|  Characteristics of non-coded answers |  |
|  |  |
| ***Travel to Work QA Tool*** | Identification of pairs of OAs/LSOAs with unlikely Travel to Work patterns (including consideration of method of travel to work). |
|  |  |
| ***CQS Reporting Tool*** | Automated production of agreement rates and transition matrices. Will only be developed if CQS is to take place in 2021. |
|  |  |
| ***Linkage Analysis Tool*** | For each linked dataset (individual and household): automated production of join proportions, agreement rates and transition matrices by area. |
|  |  |
| ***Other Validation Checks*** |  |
| Internal consistency of relationship matrix |  |
| Geography lookups are correct  | To avoid e.g. OA referenced to wrong LA |
| Sex consistent with names  | Automated check on sample of names that the reversal of ordering of sex tickbox does not lead to systematic error |
| CE classifications on address frame  | To check e.g. for risk that a Halls of Residence is classified as 400 households |
| Specific checks on Veterans question | Distribution by age and sex, apparent conflicts with ocupation or industry data |
| Date of birth analysis | Distribution of dates within year and evidence of ‘century errors’. |