Scotland's Census 2022
Statistical Disclosure Control & Outputs

Household Record Swapping Methodology

October 2020

Contents

1. Plain English Abstract

Some people in Scotland may have characteristics which make them rare or unique within their local area. In such cases, these people might be easy to identify in tables published from Scotland's Census 2022. National Records of Scotland (NRS) will therefore take steps to protect the privacy of people and households in Scotland and ensure that they cannot be identified in any published outputs (tables, reports etc.) from the census.

One of the methods that will be used will be to swap a small number of households with other households in nearby areas. Although all households in Scotland will have a chance to be selected for swapping, this method focuses on swapping households containing individuals with rare or unique characteristics who might be easy to identify in census tables.

Swapping households in this way means that data users cannot be certain whether a household or individual that appears in outputs from Scotland's Census 2022 is real, or if it has been swapped with another household in a nearby area.

2. Abstract

All households in Scotland are required to complete a census return for all usually resident persons. Some individuals and households will possess characteristics that make them unusual or even unique within their local area. This poses a potential disclosure risk as it may be possible to identify an individual or household in tables created from the census data without measures to protect their confidentiality.

In order to protect the confidentiality of Scotland's people and households in outputs from Scotland's Census 2022, National Records of Scotland (NRS) will employ a number of Statistical Disclosure Control (SDC) techniques.

Household record swapping formed a key component of the SDC strategy for Scotland's Census 2011. Internal and external evaluation of 2011 SDC indicated that an appropriate level of confidentiality protection had been introduced into census tables as a result of the application of the record swapping process. This protection had been achieved without excessively impacting data quality such that the utility of 2011 Census data was not compromised.

As such, NRS plans to use household record swapping as part of the SDC strategy for Scotland's Census 2022 in an updated form. A key innovation for the dissemination of outputs from Scotland's Census 2022 will be the availability of the flexible table builder tool. This will provide census data users with greater flexibility in

specifying their desired outputs. However it also creates unique challenges from a disclosure protection perspective.

NRS are therefore proposing a minor amendment to the 2011 record swapping method to include an additional first stage of swapping which targets households and individuals who are rare or unique at the Output Area level (OA), the lowest level of geography that will be available to users of the flexible table builder. A proportion of OA rare or unique households (specified by NRS) will be swapped prior to the application of the main 2011 record swapping method.

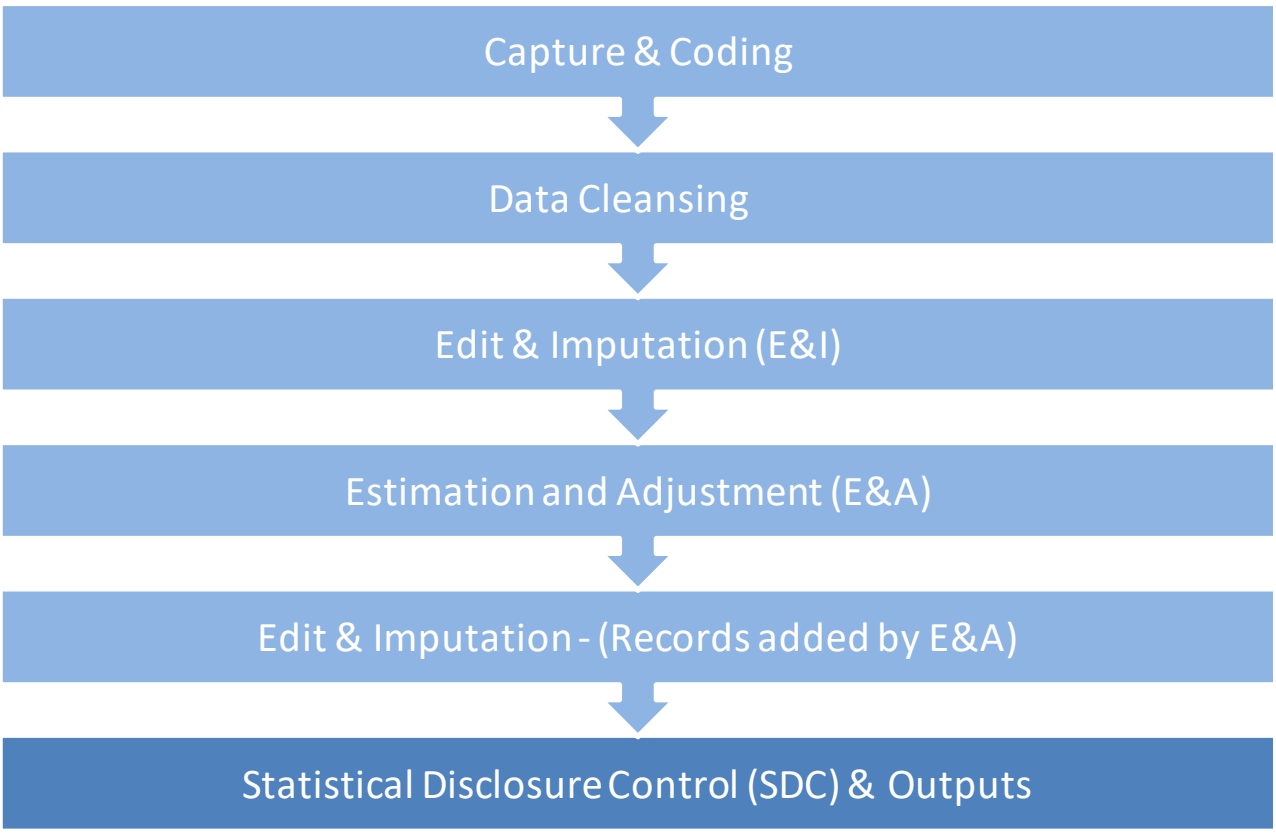
Testing of this method using data from Scotland's Census 2011 has indicated that the inclusion of the first stage of swapping introduces additional uncertainty into tables created at the lowest levels of geography. Moreover, incorporating the additional swapping stage affords NRS greater flexibility to achieve an appropriate balance between protecting the privacy of individuals and households in tables created by users in the flexible table builder without impacting on data utility.

3. Introduction & Background

NRS has a legal obligation to ensure that the privacy of individuals and households in Scotland is protected in all outputs from Scotland's Census 2022. For more information on confidentiality protection in the context of the census and relevant legislation, please see the Confidentiality page on the Scotland's Census website.

In order to prevent the release of confidential information, NRS will utilise a number of Statistical Disclosure Control (SDC) techniques. SDC refers to the range of methods which can be used to prevent the release of confidential information about an individual or household in census outputs.

The application of statistical disclosure control techniques is the final step in the census statistical methodology prior to the creation and dissemination of outputs. For more information on the full statistical methodology for Scotland's Census 2022, please refer to the Statistical Methodology page on the Scotland's Census website.



Household record swapping was a key component of our SDC strategy for Scotland's Census in 2011. Record swapping involves swapping the geographic information of two demographically similar households in nearby areas in order to introduce uncertainty into published census outputs. Although all households have a

chance to be selected for swapping, households containing individuals with rare or unique characteristics within their local area are much more likely to be swapped. A more detailed explanation of the record swapping process is given in section 4 of this paper.

Internal and independent evaluations of the Scotland's Census 2011 SDC strategy have provided assurance that the record swapping process achieved its intended purpose of protecting the confidentiality of Scotland's people and households. Hence, NRS intends to use household record swapping again for the 2022 Census, incorporating some minor amendments to account for innovations in the way outputs from Scotland's Census 2022 will be disseminated.

This paper proceeds by outlining how the household record swapping was applied in 2011 before detailing the update for its application for Scotland's Census 2022. A summary is also provided on testing of the updated 2022 record swapping method using historical Census data and data collected during the 2019 Census rehearsal.

It is important to note that detailed information on the household record swapping is restricted, as knowledge of parameters or specifics of the swapping process would compromise the protection afforded to the Census data. As such, detailed information on swapping parameters and the post-swapping analysis metrics have not been included in this paper.

4. Household Record Swapping during the 2011 Census

4.1 Statistical Disclosure Control for Scotland's Census 2011

During preparations for Scotland's Census 2011, a number of proposed SDC methods were investigated and their strengths and weaknesses considered. This work was carried out in collaboration with colleagues at the UK Census offices conducting the censuses in England & Wales and Northern Ireland, ONS and NISRA respectively. One of the SDC methods considered during this investigation stage was household record swapping. For more detail on the SDC measures considered in preparation for Census 2011, please see the ONS report on Evaluating an SDC strategy for 2011 outputs.

The central trade-off around the application of SDC measures is achieving an appropriate balance between the protection of the privacy of data subjects and preserving sufficient utility for data users by minimising the impact on data quality. Based on these considerations, the pre-2011 evaluation recommended household record swapping as the preferred SDC method from those assessed. Record swapping was found to be the optimal method for achieving the necessary balance between protection against the disclosure of confidential information and preserving data utility.

This section details how the record swapping was applied during Scotland's Census 2011.

4.2 2011 Household Record Swapping Method

The 2011 household record swapping method firstly involved assessing every individual and household for uniqueness or rarity on a number of demographic characteristics at different levels of geography. Following this assessment, every household in Scotland was assigned a household "risk score", a metric providing an indication as to whether the household or its residents possess any rare or unique characteristics that could make them easy to identify in published Census outputs, posing a statistical disclosure risk.

Once household risk scores had been assigned, a sample of households was selected for swapping. All households had a chance to be selected or "flagged" for swapping, however those with higher household risk scores were much more likely to be chosen to ensure that those individuals and households with rare or unique characteristics were afforded the greatest protection.

The next step was to match each household in the swapping sample with another demographically similar household in a nearby geographic location so that these households could be swapped. For example, if Household A in Area 1 is swapped with Household B in Area 2, then in all census outputs, Household A will now appear to be in Area 2 and vice versa.

Households selected for swapping were matched using a number of demographic characteristics to ensure that the impact on data quality was minimised. The characteristics used to match households included household size to preserve the counts of individuals and households in each geographic area.

Furthermore, the restriction that households could only be swapped over short distances ensured that, for example a household in the Scottish Borders could not be swapped with another in Orkney. In 2011, the furthest that a household could be swapped was between adjacent Council Areas, for example a household with very unusual characteristics in Glasgow City might be swapped with a demographically similar household in Renfrewshire. However, swapping at this level was kept to a minimum and almost all swapping occurred over short distances within council areas. Only swapping households over short distances helps to minimise the impact of the record swapping process on overall data quality.

This process of swapping households introduces a level of uncertainty into all cell counts in published census outputs such that a data user cannot be certain whether a cell count is real or the result of swapping. For example, a data user may believe that they would be able to identify themselves, or indeed someone else, in a 2011

census table. However, the application of record swapping means that there is a chance that any given household has been selected for swapping, particularly if they possess rare or unique characteristics.

Imputation, although not part of the 2011 SDC strategy, also introduced a measure of uncertainty into census outputs, as data users could not be certain that a count within a census table was not the result of imputed records. For more information on imputation, including plans for the 2022 imputation process, please see the Edit & Imputation page on the Scotland's Census website.

The household record swapping method is highly configurable and parameters such as the size of the swapping sample, characteristics used to assess households for their risk score and the demographic profiles used to match households for swapping can be specified by the user.

As such, for Scotland's Census 2011, swapping parameters were specified by NRS and swapping commenced via an automated process created using SAS code. Once the swapping had been completed, a comprehensive catalogue of diagnostic checks were used to confirm that the process had been completed as expected without anomalies, such as poor quality matching of households or excessive level of swapping in a particular geographic location.

Once diagnostic checks had been completed, the post-swapped data was analysed to determine whether or not the swapping process had achieved its intended aim of balancing disclosure protection whilst preserving data utility.

In order to determine whether or not sufficient protection against disclosure had been applied, the NRS SDC and Outputs team assessed the level of doubt that had been introduced to the post-swapped dataset. "Doubt" metrics give an indication into the level of certainty that a data user could have around whether a cell value in an output table is real or has been altered, either because of the household record swapping or imputation.

The doubt metrics used to evaluate output tables from 2011 were developed in collaboration with UK colleagues at ONS and NISRA, and were agreed with the National Statistician. These measures assessed the number of attribute disclosures in each census output table. Attribute disclosure refers to situations where it is possible to learn something new and potentially disclosive about an individual or household from a published table.

Specifically, these doubt metrics took the form of thresholds for acceptable counts of:

- **real** attribute disclosure cases that have been protected either by the swapping process or by imputation
- **apparent** attribute disclosure cases that are not real (i.e. appear as a result of imputation or the record swapping process)

Each 2011 census output table was evaluated using these doubt criteria. This provided assurance that a sufficient level of uncertainty had been introduced by the swapping process, such that a user of the data could not be certain that any apparent attribute disclosures they observed were in fact real.

As noted previously, the Census SDC methodology must balance the protection of respondent confidentiality with maintaining an acceptable level of data utility for data users. Hence in 2011, once post-swapping doubt had been analysed, the data was assessed again to ensure that the swapping had not excessively damaged the census data.

Data utility measures assessed the post-swapped data to determine the extent to which the original census data had been altered during the swapping process. As with the doubt metrics, these data measures were developed in collaboration with colleagues at ONS and NISRA. Evaluation of the 2011 post-swapped data indicated that data utility had been preserved.

The 2011 household record swapping method involved an iterative process of specifying and refining swapping parameters before evaluating the post-swapped data to ensure that an appropriate balance between statistical disclosure protection and the preservations of data quality has been achieved.

As noted, detailed information on the swapping parameters, diagnostic checks and the doubt and data utility metrics used to assess the post-swapped data cannot be provided in this paper to ensure that the SDC protection in 2011 census outputs is not compromised.

4.3 Communal Establishment Swapping

A slightly different method of record swapping was used in 2011 to protect the confidentiality of residents in Scotland's Communal Establishments. In the context of the census, Communal Establishments refers to establishments providing managed residential accommodation (for example, prisons, large hospitals and hotels).

The Communal Establishment swapping method was broadly similar to the household swapping approach, except that rather than households being swapped a sample of individuals is selected for swapping between establishments of the same type.

The Communal Establishment swapping followed the same principles of identifying risky data subjects and swapping a sample weighted towards those at highest risk of disclosure. As with the household swapping, data quality is preserved through a set of demographic matching criteria for swapped individuals and swaps only occur between Communal Establishments of the same type. The key difference is that in

this case, individuals are swapped between establishments of the same type, rather than the whole establishment being swapped with another in a nearby geographical area.

4.4 Independent reviews of security and confidentiality

During internal review of the swapping process, thresholds for acceptable levels of "doubt", developed in collaboration with UK colleagues at ONS and NISRA and approved by the National Statistician were used to determine that an acceptable level of uncertainty had been introduced into the census data once swapping had been applied. This provided reassurance that a user of the data could not be certain that any apparent attribute disclosures they observed were real.

In addition to internal evaluation, an Independent Review Team scrutinised the privacy and confidentiality measures employed by all three UK Census Offices during 2011 Census operations. This included an evaluation of the respective SDC approaches, and their findings were published in a report concluding that:

"the public can be assured that the information that they have provided to the 2011 Census has been well protected"

4.5 UK Census Data Users

Feedback gathered during the phase 3 UK Statistics Authority Assessment of the 2011 Censuses indicated that data users understood the need for a statistical disclosure control strategy to protect the privacy of people and households in published census outputs. Users did, however, express a need for more information on the specific aspects of the NRS SDC methodology.

During engagement events for Scotland's Census 2022, stakeholders in Scotland indicated that they understood the requirement for the application of disclosure control measures and expressed no concerns around the household record swapping method in 2011 or as a proposed method for 2022. Presentation materials and an event summary from these events is available on the June 2017 Statistical Disclosure Control & Outputs event page of the Scotland's Census website.

5. Household Record Swapping Method for 2022

Internal and external evaluation of the household record swapping, provided assurance that this method had achieved the intended aim of protecting the privacy of Scotland's people and households whilst preserving acceptable levels of data utility in outputs from Scotland's Census 2011. Hence, NRS intends to use household record swapping again, as part of the Scotland's Census 2022 SDC

strategy. As in 2011, the NRS SDC strategy has been developed in close collaboration with UK census colleagues at ONS and NISRA.

The next section outlines plans for the dissemination of outputs from Scotland's Census 2022 that necessitate a change in the NRS SDC approach.

5.1 Dissemination of Outputs from Scotland's Census 2022

A key innovation for the dissemination of outputs from Scotland's Census 2022 will be the availability of a flexible table builder tool. This will allow users to specify their desired variables and output geographies to create their own census tables. The table builder will drastically reduce the need for NRS to create standard output tables, of which there were more than 400 in 2011, each of which needed to be individually SDC checked.

For more information on our plans for the dissemination of outputs from Scotland's Census 2022, please see the Outputs page of the Scotland's Census website. For information on our current plans for output geographies to be available in the flexible table builder, please refer to the slides and event summary report from our June 2019 Scotland's Census Outputs stakeholder engagement events.

Though the flexible table builder will provide easier access to Census data, it also potentially increases the risk of attribute disclosures by allowing users to create tables at low levels of geography containing very small counts. As such, NRS have developed an SDC strategy for Scotland's Census 2022 focused on protecting tables created by users in the flexible table builder, particularly at the lowest levels of geography.

5.2 Statistical Disclosure Control Strategy for Scotland's Census 2022

In addition to household record swapping, NRS intend to incorporate two other SDC methods specifically targeted at preventing disclosure in tables generated by the flexible table builder. Cell-key perturbation will be used to introduce a small amount of statistical noise to some of the cells in a table and is applied automatically at the point where a user requests a table from the table builder.

Furthermore, SDC rules will be in place within the flexible table builder that will prevent users from creating tables with very high levels of detail or which are excessively "sparse" (containing a large number of counts of 1 or 0) such that the table may be disclosive.

As an additional layer of disclosure protection, NRS are also proposing a minor amendment to the household record swapping process, specifically targeting individuals and households at risk of disclosure at Output Area, the lowest level of geography that will be available to users in the flexible table builder.

For more information on our SDC strategy for Scotland's Census 2022, including additional material provided during stakeholder engagement events, please refer to the Statistical Disclosure Control page on the Scotland's Census website.

5.3 Household Record Swapping in 2022

NRS has investigated the application of an additional stage of record swapping targeted at individuals and households possessing characteristics that make them a disclosure risk at the Output Area (OA) level, the lowest geography level that will be available to users of the Scotland's Census 2022 table builder. The development of this additional swapping stage has been carried out in close collaboration with colleagues at the other UK Census offices, ONS and NISRA.

During this additional stage of swapping, in a similar manner to the 2011 record swapping, all individuals will be assessed on a number of demographic characteristics to determine whether they possess any attributes that make them rare or unique at the OA level. Households containing these rare or unique individuals will be flagged and the NRS SDC & Outputs team will then specify a proportion of these OA rare or unique households to be swapped. This proportion can be manipulated by NRS as required.

The process of finding a match for households during the first stage will follow the same approach as in the 2011 method, whereby households selected for swapping are matched with demographically similar households in nearby geographic areas to minimise the impact on data quality.

This first stage of swapping will be in addition to the 2011 record swapping method rather than replacing it. However, this first stage of swapping takes place separately to the method detailed in section 4 of this paper (i.e. the 2011 method), and any households swapped during this first stage of swapping will be excluded from the 2011 record swapping process (i.e. households will not be swapped multiple times).

Directly targeting rare or unique records at the lowest geography levels will provide greater assurance that the record swapping process is protecting those individuals and households who are at greatest risk of disclosure in tables created by the flexible table builder.

This additional stage of swapping also provides NRS with greater flexibility in specifying swapping parameters that achieve the appropriate balance between disclosure protection and the preservation of data quality. Specifically, changing the proportion of OA unique records to swap will ensure that appropriate protection is applied without compromising data utility.

5.4 Testing of the proposed 2022 Household Record Swapping method using 2011 Census Data

This section summarises testing carried out by NRS on the application of the additional swapping stage using 2011 census data and data from the 2019 Census Rehearsal.

NRS have tested the proposed 2022 household record swapping method incorporating the additional first stage targeted at OA rare or unique households. This testing was carried out on Scotland's Census 2011 data. The amended 2022 method was applied to the 2011 data as a first step before the application of the main 2011 record swapping method. Once the swapping process had been completed, the full range of diagnostic, doubt and data utility checks were carried out to determine the level of protection that had been applied with the amended 2022 method.

This testing revealed that the additional first stage of swapping had indeed created additional uncertainty in output tables, particularly those at the OA level. These results suggest that the additional stage of record swapping will be useful for protecting records which pose a disclosure risk at the lowest levels of geography.

Following these results, NRS are proposing to include the additional first stage to swap a proportion of OA rare or unique households prior to the application of the 2011 record swapping method for Scotland's Census 2022.

5.5 Testing proposed 2022 Household Record Swapping method using 2019 Census Rehearsal Data

In 2019, NRS carried out a Census rehearsal to test some of the systems and services that will be integral to the success of Scotland's Census 2022. Data was collected from approximately 72,000 households in areas of Glasgow City, Dumfries & Galloway and Na h-Eileanan Siar. These areas were selected to allow NRS to test approaches in rural, urban and diverse communities. Participation in the 2019 Rehearsal was voluntary. For more information on the 2019 Census Rehearsal, please refer to the Census Rehearsal Evaluation Report.

Data collected from the 2019 Census rehearsal has been used to test several aspects of the Statistical Methodology for Scotland's Census 2022. This included the proposed 2022 household record swapping methodology. Testing with the 2019 rehearsal data provided a useful opportunity to assess functionality of the record swapping process with new data in preparation for Scotland's Census 2022.

This testing indicated that the record swapping had been successfully applied to the 2019 rehearsal data and provided assurance of the processes that take place prior to and during the application of SDC.

6. Conclusion

Testing of the amended 2022 record swapping method, incorporating an additional first stage of OA rare or unique swapping, indicated an increased level of disclosure protection in census output tables at the lowest geography levels.

Given these findings, NRS are recommending that a proportion of OA special unique households should be swapped prior to the application of the main sampled record swapping method for Scotland's Census 2022. This provides assurance that the data for individuals and households at greatest risk of disclosure are protected in output tables created by users in the flexible table builder. This additional stage of swapping also affords NRS with greater flexibility in achieving the appropriate balance between disclosure protection and maintaining acceptable levels of data utility.

7. Annex

7.1 Glossary

Attribute Disclosure	The ability to learn something new about a respondent (or group of respondents) from a table
Communal Establishment	Establishments providing managed residential accommodation. Managed means full-time or part-time supervision of the accommodation (for example prisons, large hospitals and hotels).
Data Utility metrics	Measures used by NRS to determine whether the swapping process has caused an excessive level of damage to data such that it is no longer useful for data. These take the form of thresholds on acceptable levels of difference between the un-swapped and post-swapped datasets
"Doubt" metrics	Measures used by NRS to determine whether the swapping process has created sufficient uncertainty within the post-swapped data such that the confidentiality of people and households are protected.
Household record swapping	Household record swapping involves assessing all households in Scotland for rarity based on a set of demographic characteristics before selecting a sample of households to swap. Households selected for swapping are matched with demographically similar households in nearby areas and the geographic information of these two households is then swapped. Although all households have a chance to be selected for swapping, households containing individuals with rare or unique characteristics within their local area are much more likely to be chosen.

Statistical Disclosure Control (SDC)	Statistical Disclosure Control refers to the methods that we use to protect the privacy of personal information. This includes making small changes to data, controlling access to data, and controlling the level of detail that is available to census data users.