Proceedings of the 21st GSS Methodology Symposium 2016

Methodology: The key to the door of innovation

**Improvements in Methodology for Matching the 2021 Census to the Census Coverage Survey (CCS)**

Sarah Cummins, Peter Jones, and Shelley Gammon[[1]](#footnote-1)

**Abstract**

Since 2001, the Census has been matched to a Census Coverage Survey (CCS) to facilitate the estimation of non-response. The quality requirements for this matching exercise are very high, since errors in matching will impact the estimates of the population. For the 2011 Census to CCS matching exercise, 70% of person matching was done via automated methods and the rest were matched clerically, resulting in exceptionally high quality matching.

This research investigates whether the cost and processing time of the 2021 Census to CCS matching exercise can be reduced by finding more matches using automated methods, without incurring unacceptable levels of error. Matching methods were tested using matched 2011 Census and CCS data as a ‘gold standard’, taking the links made in 2011 as the ‘true’ match status.

A hierarchical matching strategy was adopted using deterministic ‘match keys’ and probabilistic matching. Overall the matching left just under 4% of matches yet to be found at a 0.25% false positive error rate. This would substantially reduce the clerical resource needed for the 2021 Census to CCS matching exercise, but further work needs to be done to investigate the impact of a higher level of error and potential biases in that error rate.

Key Words: Matching; Census; Error; Estimation.

**1. Introduction**

Since 2001, the Census has been matched to a Census Coverage Survey (CCS) to facilitate the estimation of non-response [1]. The CCS is a complete enumeration of a 1% sample of postcodes undertaken around six weeks after the Census. The quality requirements for this matching exercise are very high, since errors in matching such as false positives (false links) and false negatives (missed matches) will impact the estimates of the population which arise from the coverage adjustment.

For the 2011 Census to CCS matching exercise, 60% of household matching and 70% of person matching was done via automated methods [2], including the use of exact matching and standard Fellegi-Sunter probabilistic methods. The rest were clerically matched (estimated at 30 full time staff over 30 weeks). Overall, the matched dataset obtained an estimated false positive rate of less than 0.1% and an estimated false negative rate of less than 0.25% [3]; good enough to be a ‘gold standard’ linked dataset.

Since the 2011 Census to CCS matching exercise, progress has been made in the field of automated matching methods due to the requirement to match large admin datasets in a secure environment [4,5]. This includes the use of sequential matching methods, including deterministic (exact and rule-based matching), probabilistic (Fellegi-Sunter) and associative matching methods.

This research aims to apply these methods to match the 2011 Census to CCS again. Since we know that the matched Census and CCS in 2011 was of a very high quality standard, we can compare our results to the ‘true’ matched status as obtained in 2011. The objective of this research is to investigate whether we can find more matches via the automated methods without affecting the quality of the links made. This is to show that there is potential to reduce the amount of clerical resource needed in 2021 and therefore reduce the cost of the Census to CCS matching, without risking the quality of the results.

**2. Methods**

**2.1 2011 Census-CCS matching**

A hierarchical strategy was used for linking the Census to CCS in 2011. First, matching was done at the household level and then individuals were matched within the household. A degree of flexibility was also required to enable potential individual matches to be made even if the household was not matched.

At both the household and the individual level, exact matching first identified the simplest records to match. This was followed by probabilistic matching to match the remaining records. The Fellegi-Sunter model was the probabilistic method chosen, with only very high scoring candidate pairs being automatically matched. Lower scoring record pairs were sent for clerical review. For individuals who had not been matched in a household, no automated matches were made by the probabilistic methods; instead any potential candidates were sent for clerical review [4].

**2.2 Census-CCS matching research**

The matching methods used in this paper are based on person matching only. We have not attempted to replicate the hierarchical matching methods (matching at household level and then individual level) used in 2011; however, should the methods be viable, they could be tested on household level variables as well.

A consideration when matching the 2011 Census to the CCS is that whilst the CCS has approximately 700,000 person records within 350,000 households [1], the Census has 49 million counted people. Matching to the full Census would be computationally demanding, especially for probabilistic matching where every record is compared to every other record. Since 98% of matches made in 2011 agreed on postcode, the approach was taken to block the linkage by postcode (i.e. only record pairs that agreed on postcode would be considered). Postcode was also the main blocking strategy in 2011 for matching at the household level.

Blocking by postcode substantially reduces the time taken to match, since the 2011 Census can be subset to the postcode sample for the CCS. This approach also improves the quality of the links made, since records from a smaller subset are more likely to be unique and therefore less likely to be linked incorrectly. However, this makes the assumption that the postcode data is accurate for both sources, which it may not be.

**2.3 Linkage quality**

The term ‘link’ is used to denote the record pairs that we think belong to the same entity. The term ‘match’ is used to denote the record pairs that truly belong to the same entity (unknown). Since the Census was matched to the CCS in 2011 to such a high quality standard, for this research it is assumed that the links made in 2011 are true matches.

In data linkage, there are two types of error: False positives (FP) are record pairs declared as links incorrectly and false negatives (FN) are record pairs that are not linked but should have been (missed matches). In this paper, the FP rate and FN rate will be reported to demonstrate the quality of the matching method.

The quality requirements for the 2021 Census to CCS matching exercise are provisionally to obtain both FP and FN rates of less than 0.25%.

**3. Results**

**3.1 Deterministic matching**

In deterministic matching, there must be an exact agreement on each of the matching variables. Where all variables are used, this is called ‘exact matching’. This should always be done first to get the easy matches. ONS have also developed a set of ‘match keys’ which can be applied sequentially after the exact matching [5]. Each of the match keys are a concatenation of matching variables, some of which have been relaxed or omitted to allow for error in some variables. These match keys are also used in a deterministic way so that there must be exact agreement on each match key.

Exact matching was used (counted as match key 1) followed by six further match keys (match keys 2-7). Match keys 2-7 account for error in different variables; although it was a requirement for each to match exactly on postcode. Table 1 shows the results of matching on the seven match keys. A total of 539,160 links were made. Each match key had a FP rate below 0.25%, except for match key 7 which had a FP rate of 0.64. This indicates that this match key should possibly be omitted.

Overall, there was a FP rate of 0.07%. This is within the FP quality requirements of less than 0.25%. Match keys have found a total of 83.54% of all true matches (a FN rate of 16.46%). This is already better than the automatched records in 2011 which found around 60% of household level matches and 70% of individual level matches and can already demonstrate a potential cost saving for the 2021 Census to CCS clerical matching.

**Table 1: Deterministic matching results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Match key** | **Match key variables** | **Links** | **% FP rate** | **% FP rate\*** | **% FN rate\*** |
| 1 | Forename, surname, DoB, sex, postcode | 245356 | 0.03 | 0.03 | 61.97 |
| 2 | Forename initial, surname initial, DoB, sex, postcode | 161765 | 0.05 | 0.04 | 36.90 |
| 3 | Forename initial, DoB, sex, postcode | 51909 | 0.12 | 0.05 | 28.86 |
| 4 | Surname initial, DoB, sex, postcode | 31065 | 0.16 | 0.05 | 24.06 |
| 5 | Forename, surname, DoB, postcode | 2784 | 0.04 | 0.05 | 23.62 |
| 6 | Forename, surname, sex, postcode | 45968 | 0.18 | 0.07 | 16.51 |
| 7 | Transposed forename-surname, DoB, sex, postcode | 313 | 0.64 | 0.07 | 16.46 |

\*Cumulative

**3.2 Probabilistic matching**

Probabilistic matching is performed on the residuals from the deterministic matching in order to find more matches. Probabilistic matching is where all records on one dataset are compared to all records on another dataset, computing a score for each candidate pair. It is more computationally demanding than deterministic methods; therefore blocking[[2]](#footnote-2) is used to reduce the number of comparisons. Residuals from the Census and CCS datasets from the deterministic stage were blocked by postcode since it was known that 98% of known matches from 2011 had agreement on postcode.

The Fellegi-Sunter method was chosen for the probabilistic matching phase. This model for record linkage was developed by Fellegi and Sunter [6] and its use has been well documented in literature. It was also used in the 2011 Census to CCS automated matching.

The Fellegi-Sunter model gives us statistical theory by which to set weights for each variable. Each matching variable has an ‘m’ probability and a ‘u’ probability: The m probability is the probability of agreement on a variable given the record pair is a match (i.e. measuring freedom from error) and the u probability is the probability of agreement on a variable given the pair is a non-match (i.e. measuring agreement by chance). The m-probability and u-probability are used to calculate an agreement weight and a disagreement weight for each variable. An agreement weight (always positive) is applied where there is agreement on a variable and a disagreement weight (always negative) is applied where there is disagreement on a variable. All the weights are added up across each variable to calculate a total score for a record pair.

To estimate the m and u-probabilities, an unsupervised iterative method called the EM algorithm was used [7]. The advantage of this method is that it doesn’t require clerically resolved training data. The deterministic links made in the previous stage were included in the EM algorithm to ensure that the proportion of true matches in the blocked records remains above 0.05 (which is one of the pre-conditions of the EM algorithm to get accurate weight estimates).

The standard Fellegi-Sunter model does not account for partial agreement. In order to account for some partial agreement, an agreement score was calculated for forename and surname using the Levenshtein Edit Distance and this was used to interpolate between the disagreement and agreement weight. Similarly for the date of birth variable, partial agreement was accounted for by calculating different agreement scores depending on how many of the components (i.e. day, month or year) there is agreement on.

Only record pairs which were the highest scoring for both the Census and CCS records were taken, to avoid multiple links. Chart 1 shows the cumulative FP and FN rates for the Fellegi-Sunter method by score threshold. The score threshold 0.7 gives a very low FP rate (around 0.1%) and reduces the matches left to be found to around 8.3%. The score threshold 0.55 gives a relatively low FP rate (just over the accepted error rate of 0.25%) whilst reducing the matches left to be found to 3.8%. Below 0.3, not many more links are being made and no more true matches are found, hence the FN rate does not change.

This demonstrates that whilst the use of probabilistic matching will increase the FP error rate, there is a substantial reduction in the matches left to be found. This would largely reduce the clerical resource needed in 2021.

**Chart 1: Cumulative FP and FN rates for the Fellegi-Sunter method by score threshold**

**4. Discussion**

**4.1 Impact on clerical review**

In the 2011 Census to CCS matching there were 195,000 matches left for clerical review after the automated process. If we were to use match keys in 2021, this would reduce the number of matches left for clerical review by almost half to 100,000. If in addition to this, if probabilistic matching was used at an overall FP rate of 0.25%, this would further reduce the matches left for clerical review by a quarter to 25,000.

However, reducing the number of matches left does not necessarily proportionally reduce the time taken for clerical review. This is because the matches left are likely to be of very low data quality and therefore be the hardest ones to match. Clerical matching may use clerical comparisons and clerical searching. Clerical comparisons are where the clerical matcher is presented with two records and they have to decide whether they belong to the same person. Clerical searching on the other hand involves a clerical matcher taking a residual record and querying the other dataset for its pair. Clerical searching is far more time consuming and it is likely that a higher proportion of the 25,000 matches remaining will require this technique. However, further research needs to go into efficient blocking strategies to pool potential candidate pairs to ensure that clerical comparison is maximised.

**4.2 Impact of FP error**

Any error (FPs and FNs) will affect the estimate of the population. FPs will make the population estimate lower than it should be because if the Census matches to the CCS incorrectly, it will mean that an adjustment is not made where it should have been. FNs will make the population estimate higher than it should be because if the Census fails to match to the CCS when it should have done, an adjustment will be made where it shouldn’t have been made. The level of error has a direct impact, so 0.01% FPs will mean a decrease in the population by 0.01%.

This research has shown that over 96% of matches can be found at an overall FP rate of 0.25%. A higher overall level of FP error may be adjusted for if that error is random; however, it is not likely that the FP rate across different subgroups of the population will be the same. Since the non-response adjustment is done for age, sex and geography, if there are many subgroups above this target level of error, the adjustment for non-response might be poor in those groups.

Chart 2 shows a scatter plot of the FP rate by FN rate for each sex/quinary age group. Whilst most sex/quinary age groups have a FP rate below 0.25%, there a number of groups above this. The groups which have the highest FP rate are ones where either sex or age is missing. This is to be expected given that if age or sex are missing, there is less information to use in order to determine whether a record pair are a link and therefore more likely to link incorrectly. There are only three other quinary age/sex groups above 0.25% - males 15-24 and females 15-19. This indicates that generally there is little variation in FP rate other than where there is missing information.

**Chart 2: FP rate and FN rate by sex/quinary age groups at overall 0.25% FP error**

Chart 4 shows the equivalent chart for LAs at an overall FP rate of 0.25%. The FP rate ranged from 0% to 1.84%, with 67 out of 348 LAs having a false positive rate above 0.25%. The range in the FP rate across LAs indicates a lot of variation above the overall FP error rate, more so than for quinary age and sex. If there is a lot of variation in FP error above the defined target of FP error, the non-response for LAs with high FP error will not be adjusted for accurately. If this is the case, it might be more appropriate to use a lower overall target for FP error to ensure that all subgroups have a FP error rate below 0.25%.

**Chart 3: FP rate and FN rate by LA at overall 0.25% FP error**

**5. Conclusions and further work**

The linkage that was done on record pairs with agreement on postcode obtained around 96% of true matches within the quality requirements of false positives set at 0.25%. The match keys alone found 84% of true matches with a very low level of error (0.07%), reducing the matches left to 100,000 - half of what was clerically reviewed in 2011.

On the residuals of the match keys, the Fellegi-Sunter probabilistic model performed to a very good standard. This method could be used to allow more automated acceptance of links whilst still maintaining a FP rate below the target of 0.25%. This would also further quarter the number of matches left to 25,000 and therefore substantially reduce the amount of clerical review needed.

However, the variation of FP rates across subgroups also needs to be taken into account in terms of deciding on a suitable method and target FP error rate for 2021 – potentially a higher probabilistic score threshold should be used to ensure there is less variation in FP error across subgroups.

Priorities for further work in this research include:

* Investigating methods of blocking or prioritising residual record pairs for clerical review.
* Investigating the impact of the change to online data entry in 2021 – how will the quality of matching differ for online entry compared to hand-written forms (scanning software)?
* Continuing to investigate tolerance for FP error - if we know the error, can we adjust for it?

**References**

[1] Office for National Statistics (2012). The 2011 Census Coverage Assessment and Adjustment Process. Available at: <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/coverage-assessment-and-adjustment-process.pdf>

[2] Office for National Statistics (2012). Automatic Match Rates for the 2011 Census to the Census Coverage Survey. Available at: <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/automatic-match-rates-for-the-2011-census-to-the-census-coverage-survey.pdf>

[3] Office for National Statistics (2013). An Assessment of the Quality of the Matching between the 2011 Census and Census Coverage Survey. Available at: <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/coverage-assessment-and-adjustment-methods/census-coverage-survey--ccs-/matching-quality-report-ns.pdf>

[4] Abbott, O., Jones, P., Ralphs, M. (2015). Large scale linkage for total populations in official statistics. Chapter 8 in Methodological Developments in Data Linkage.

[5] Office for National Statistics (2013). Beyond 2011: Matching Anonymous Data (M9). Available at: <http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011-matching-anonymous-data.pdf>

[6] Fellegi, I., Sunter, A. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, Vol. 64, No. 328.

[7] Winkler, W. (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 667-671.

1. Sarah Cummins, ONS, sarah.cummins@ons.gov.uk; Peter Jones, ONS, peter.jones@ons.gov.uk; Shelley Gammon, ONS, shelley.gammon@ons.gov.uk [↑](#footnote-ref-1)
2. Where there must be exact agreement on a variable/set of variables for the record pair to be compared. [↑](#footnote-ref-2)