ONS Census Coverage Scoping Project

November 2019

Contents

This document presents the results of a short, joint, scoping exercise carried out in October 2019 by a small team from the Alan Turing Institute and the Office for National Statistics (ONS).

It was originally intended that this document would provide the foundation for a longer project to explore the application of machine learning to the matching of records between the 2021 UK census and the Census Coverage Survey (CCS).

Our main aim below is to specify the challenges faced with regards to this task and provide sufficient background information. The next section contains a brief summary of the context in which the research is being carried out and the background information on the UK census and CCS.

Background

The UK Census and Census Coverage Survey

In the UK, the national census is carried out every 10 years, in order to measure the population size and demographics. In some countries, the census count itself is published; the UK aims to provide a census estimate, adjusted for the "undercount" and "overcount" occurring when people are missed or counted multiple times. In 2011, most people filled out the census on provided paper forms, with 16 % of respondents choosing to do so online. A key difference in 2021 will be that for most people, online forms will be the default, with paper forms only being provided upon request.

To calculate the census estimates, an independent enumeration of a sample of 1 % of postcodes known as the Census Coverage Survey (CCS) is carried out. This takes place after the census and involves in-person interviews carried out at the selected addresses; data from the occupants is obtained for a small selection of the core census fields: first name, surname, date of birth, sex, marital status, address and occupation.

The 2011 UK Census estimated that in the UK there are about 65 million people (63.2 m) and 25 million households (26.4 m), with the CCS sampling 1 % of postcodes, counting about 650 000 people and 340 000 households.

In the postcodes sampled by the CCS, about 95 000 individuals counted by the Census were not matched in the CCS; likewise, there were about 60 000 individuals counted by the CCS who were not matched in the Census. These figures are higher than the final estimates of under-enumeration because the sample postcodes were weighted towards areas where high under-enumeration was expected.

Calculating the census estimate relies on records from the CCS being correctly paired with census records that correspond to the same person (the same goes for matching households). This is the central challenge the methods discussed in this document seek to address.

The challenge of record matching

Difficulties in matching the CCS records with census records from the same person or household occur when there is missing/incomplete information in one of the records, or differences due to spelling mistakes, scanning errors and other mistakes. As such, this problem can be considered a "record linkage" problem. See the *Review of the "Record Linkage" Problem* section of this report for a longer summary of the record linkage problem and the algorithms used to tackle it.

Record matching between the CCS and census is subject to strict precision and recall criteria; precision of at least 99.9 % and recall of at least 99.75 %.

In 2011, 70% of the people matches were made automatically using a mixture of deterministic matching and *Fellegi-Sunter* probabilistic matching (for households, a deterministic method was used which matched 60 % automatically). This left 30 % of people matches to be made manually via a clerical matching procedure involving two distinct manual processes. Firstly, *clerical resolution* involved people deciding whether record pairs that automated methods designated as possible matches were indeed matches. Secondly, *clerical searching* involved searching for a match manually amongst all the unmatched records. These two processes are referenced throughout this document; *clerical matching* refers to both of these.

The clerical matching procedure took the equivalent of 30 full-time staff all working for 30 weeks in 2011. In 2021, the deadline for completing the census matching will be only 8 weeks from when all the census and CCS returns are in.

Thus, ongoing work at ONS aims to minimise (to the greatest degree possible) the need for clerical searching as part of the 2021 matching methodology and to speed up all clerical matching. The slowness of clerical searching in 2011 owes much to the fact that in order for a CCS record for which it is suspected there will be a match to be ruled out and considered a non-match, it must first be checked against every single census record for which there is currently no match.

Even after improvements ONS have already made to the automated matching methods (detailed later in the *Progress since the 2011 Census* section), ~10 % of people records were still left to match manually (~5 % for household records) when testing these methods on 2011 data. ONS predict that of these people matches, a further ~7.5 % will be found by clerical resolution, leaving ~2.5 % of matches still to make. The remaining matches could be included anywhere amongst the unmatched CCS records (55 000 in 2011) and unmatched census records from CCS areas (95 000 in 2011). Any method used to replace clerical searching for these very difficult matches should ideally also declare when records do not have a match, in order to avoid clerical searching still being required.

The next section of this document summarises some of the relevant literature on record linkage.

Review of the "Record Linkage" problem

There are many databases containing records that refer to real-world entities, such as people. There are also a variety of problems for which information on the same entity must be gathered from multiple databases. In order to combine or compare information on these entities from different databases, there must be a robust method for determining which records refer to the same entity. In cases like that of the census and CCS, the challenge is complicated by the reality

of missing or inaccurate data in records that should be matched i.e. those that refer to the same person.

The task of matching non-identical records from different databases that refer to the same entity is known as *record linkage*. In scientific literature it is also described by a variety of alternative names depending on the research community, including *instance identification*, *name matching*, *database hardening*, *merge-purge* and (when applied to a single database) *duplicate detection* (Elmagarmid, Ipeirotis, and Verykios 2007).

Record linkage problems deal with records that reference complex real world entities like people, with multiple data fields. The challenge is therefore greater than simply matching a single field, where commonly used string distance metrics such as the Levenshtein edit distance or Jaro-Winkler are suitable. Such metrics can however be used to compute a distance metric for the equivalent fields of two records, which has shown to be useful in matching census names with typographical errors (William E. Yancey 2005).

To avoid comparing every record in one database with every one in the other, there are a variety of different methods used to filter out extremely unlikely matches that vary in their performance and scalability. A common example is *blocking*, where all record pairs that disagree on a blocking key are initially discarded. This key can be a particular field or combination of multiple fields (Christen 2012).

The methods used for the problem of record linkage fall into the three general categories; deterministic, probabilistic and learning based methods. All of these methods work on the general premise of classifying record pairs as matches, as non-matches and in some cases as indeterminate.

Deterministic methods use a set of rules based on the constituent fields of each record pair called a "Matchkey" to classify matches. Pairs that don't match according to those rules are classified as non-matches. For example, a Matchkey for a pairing of records that have two equivalent fields could be: Field1 must be an exact match and Field2 must have an edit distance < 3.

Probabilistic methods (most commonly the *Fellegi-Sunter algorithm*) use a Bayesian approach to calculate the probability of each record pair being a match or non-match, based on the product of the set of probabilities of corresponding fields being matches or non-matches between the two records. Pairs falling below a match threshold and above a lower non-match threshold are classified as indeterminate and sent out for clerical matching. Each field used in the calculation is assigned a weight, computed either by an "Expectation Maximisation" (EM) algorithm or from the probabilities in training data (Murray 2018).

Alternative machine learning methods for record linkage are discussed later in this document; the next section explains the improvements to the deterministic and probabilistic census-CCS record linkage methods already made by ONS since 2011.

Progress since the 2011 Census

ONS have begun to improve upon the methods used for record linkage in 2011. As of November 2019, ONS show that ~90 % of people records (and ~95 % of household records) from the 2011 census and CCS record data can be matched automatically (compared with 70 % and 60 % respectively with 2011 methods). In this section of the document, the key improvements to the methodology that resulted in this performance increase are detailed and evaluated. Additionally, improvements intended to speed up clerical matching are discussed.

Improvements in Census to CCS Record Linkage

In order to improve upon deterministic matching of people, a set of matchkeys have been developed using the 2011 Census data as test data. These include derived field variables that account for common errors in name fields such as those caused by scanning (of paper forms), spelling errors or transposition errors. For example, rearranging the letters of names into alphabetical order can deal with transposition errors (Alphaname method) and use of the Jaro-Winkler edit distance or a phonetic algorithm based on English pronunciation similarity (Soundex) can deal with phonetic and spelling errors. Comparison with the 2011 Gold Standard (record pairing decisions made by all methods including clerical matching in 2011) shows that the matchkeys find ~85 % of the matches made in 2011. It should however be noted that this Gold Standard is not perfect, with duplicates being a recurring issue with using it to verify new methods.

A new set of matchkeys have also been developed for household record pairing, using household information such as tenure, type of property, number of usual residents and of particular importance, the derived variable UPRN (Unique Property Reference Number). Together with the sets of people records that make up a household occupancy, these matchkeys have enabled ONS to make ~95 % of the matches on the 2011 households Gold Standard.

ONS have also made improvements to the match rate for *Fellegi-Sunter* probabilistic matching. Changes have been made to the blocking carried out before matching; a single blocking pass is used, bringing together record pairs that match on the postcode field. All other CCS fields are therefore available for use in the actual matching. Of the matched record pairs from 2011 data, 97.8 % agreed on the postcode and were scored with *Fellegi-Sunter*. The remaining 2.2 % is expected to be captured by other methods (e.g. deterministic and associative).

Some steps have already been taken to speed up the clerical matching process via a proposed associative people matching method, which also increases the number of automatic matches. Unmatched people in households where the household record has already been matched are given a score using *Fellegi-Sunter*. Any candidate people record pairs who score above a threshold are accepted automatically (note that this threshold can be lower than that set for

the initial people matching algorithm). Matched households that still contain unmatched people are then sent for clerical resolution, giving the reviewer a household view that clearly shows those people matches already made within the household.

In starting to address the key objective of speeding up the clerical searching procedure, ONS have developed a *Pre-search* algorithm, which is applied to the CCS records for which automated methods could not find a matching census record, before the laborious clerical searching procedure is attempted. This algorithm finds potential candidates for pairing using very loose blocking, ranks them using *Fellegi-Sunter* scoring, and then sends them for clerical resolution, with the human matcher making the final decision as to whether there is a match and which of the top 20 ranked candidates it is. The ultimate goal would be to be able to say with confidence that if the matching record is not amongst the top 20 candidates presented to the human matcher, then there is no match for that record.

The *Pre-search* algorithm is already working well; when there is a match predicted for a given 2011 CCS record assigned to *Pre-search* by prior methods, it appears as the first record on the list 89 % of the time and in the top 20 98 % of the time (to the nearest percentage points, as evaluated by the 2011 Gold Standard).

How close are we to full automation?

In order to determine when the improved record linkage methods being researched are good enough to be considered ready for the 2021 census, we evaluate them here to assess how close they are to meeting the strict precision and recall requirements of 99.9 % and 99.75 % respectively. They are tested on 2011 census/CCS data and the performance evaluated against the 2011 Gold Standard. The caveat here; there is no *guarantee* that methods meeting the precision/recall requirements on 2011 data will do so on 2021 data. It is therefore important that ONS are confident these methods are not overfitted to 2011 data when their performance is evaluated.

This evaluation can be used to judge to what extent clerical searching and resolution will be required in 2021, given the constraint of the shorter time period meaning fewer clerical matching man-hours than 2011. It's important to note that some of the methods used for clerical searching in 2011 can't possibly be performed by an algorithm. For example, some searchers made use of Google, including searching for name relations that were not obvious (e.g. Pepik as a nickname for Josef in Czech) and searching Google Maps to see if they could spot an additional property at an address. Other searchers took a closer look at the paper census forms, finding some relevant information had been written outside of the response boxes and missed when the forms were first scanned. Any methods that do not utilise clerical searching could therefore be missing matches that can *only* be made this way, increasing the number of false negatives.

It's worth noting that these kinds of manual methods could also be useful for confirming (or rejecting) the candidate matches suggested by the *Pre-search* algorithm in 2021.

Since a big part of the challenge is confidently ruling out those records without a match, it's especially important to know how many false negatives can be permitted. In 2011, the number of matches on the Gold Standard was 649 944. ONS expect automated methods to give zero false positives (matches not on the Gold Standard) due to their being conservative, with ambiguous match/non-match pairs being sent to clerical matching. ONS found that that the small numbers of false positives that were found (according to the Gold Standard) looked like good matches that were missed in 2011. For example, some of these were due to the existence of duplicates, with the false positive being a match to a different copy than the one on the Gold Standard. For these reasons, we take the false positives to be zero for all methods evaluated in this document. Since the total number of unmatched CCS records in 2011 was 59 527, we can consider the number of true negatives in the confusion matrices below (Tables 1-7) to be 59 527, when considering false positives zero.

Given the assumption that all matches are true positives (TP), we can rearrange the recall equation to calculate the permitted false negatives (FN) given the 99.75 % recall threshold (R):

$FN = (TP/R)$ - TP = (649 944 / 0.9975) - 649 944 = 1 629 (to the nearest whole).

In other words, when we are evaluating the performance of improved methods on 2011 data, it's crucial that we see fewer than 1 629 Gold Standard matches being missed. Since we know that 70 % of the matches in 2011 were found using automatic methods, we can use this as a baseline from which to evaluate the performance increases in 2019 methods, given the goal of maximising the extent to which record linkage can be done automatically. Table 1 shows that the 2011 automatic methods missed 194 983 matches that were ultimately found by clerical matching.

The assumption that there are no false positives means that the precision for each of the methods in Tables 1-7 is always 100 % and the recall (TP / (TP + FN)) can be used to say what percentage of the Gold Standard matches are found after each method is tested. The recall values for automatic record linkage methods are summarised in the following list:

1. Deterministic: 84.871 % (see Table 2)
2. Probabilistic: 88.157 % (see Table 3)
3. Associative : 90.072 % (see Table 4)

To get the recall up to 99.75 % using purely automated methods, an extra 62 898 of the 2011 Gold Standard matches would therefore need to be found (64 527 - 1 629). Could we get closer to meeting the recall requirement by relaxing the match threshold for probabilistic record linkage, allowing for some loss of

precision via false positives? Figure 1. suggests that this is not possible. The recall remains below 99.75 % even when the probabilistic match threshold is lowered beyond the point that the precision requirement of 99.9 % is no longer met.

We can also evaluate the performance of the *Pre-search* algorithm, given the numbers of additional 2011 Gold Standard matches that it presents as candidates when applied to the records assigned by prior automated methods (*Fellegi-Sunter* and household-associative). To do this, we assume that all the decisions sent to clerical resolution (before *Pre-search*) are made correctly as per the Gold Standard (see Table 5). Table 6 and Table 7 show the number of further matches that can be found when clerical matchers are presented with a single highest scoring possible match to decide on and when they are able to choose from the top 20. The recall values given these considerations are summarised in the following list:

1. Clerical Resolution: 97.514 % (see Table 5)
2. *Pre-search* (top candidate only): 99.397 % (see Table 6)
3. *Pre-search* (top 20 candidates): 99.596 % (see Table 7)

These results show that whilst clerical resolution is clearly still necessary to find many matches, ONS are close to eliminating the need for clerical searching with only 994 (2 623 - 1 629, see Table 7) additional Gold Standard matches to be found in order show that current methods are good enough to meet the precision and recall requirements on 2011 data.

Future methods developed in advance of the 2021 deadline can be evaluated in a similar manner and the best performing methods should be selected for use in 2021 (see *Next Steps*). In the next section of the document, record linkage methods that are as yet untested by ONS are discussed.

Table 1: Matches and Non-Matches found by automated record linkage methods used in 2011, including deterministic and probabilistic methods (Predicted) as evaluated by the 2011 Gold Standard (GS).

	Predicted Match	**Predicted Non-Match**
GS Match	*TP*: 454 961	*FN*: 194 983
GS Non-Match	*FP*: 0	*TN*: 59 527

Table 2: Matches and Non-Matches found by improved deterministic record linkage methods in 2019 on 2011 Census/CCS records (Predicted) as evaluated by the 2011 Gold Standard (GS).

	Predicted Match	**Predicted Non-Match**
GS Match	*TP*: 551 613	*FN*: 98 331

	Predicted Match	**Predicted Non-Match**
GS Non-Match	*FP*: 0	*TN*: 59 527

Table 3: Matches and Non-Matches found by improved probabilistic and deterministic record linkage methods in 2019 on 2011 Census/CCS records (Predicted) as evaluated by the 2011 Gold Standard (GS).

	Predicted Match	**Predicted Non-Match**
GS Match	*TP*: 572 966	*FN*: 76 978
GS Non-Match	*FP*: 0	*TN*: 59 527

Table 4: Matches and Non-Matches found by improved probabilistic, deterministic and household-associative record linkage methods in 2019 on 2011 Census/CCS records (Predicted) as evaluated by the 2011 Gold Standard (GS).

	Predicted Match	**Predicted Non-Match**
GS Match	*TP*: 585 417	*FN*: 64 527
GS Non-Match	*FP*: 0	*TN*: 59 527

Table 5: Matches and Non-Matches found by improved probabilistic, deterministic and household-associative record linkage methods in 2019, plus the additional matches that would be found if all match decisions sent to clerical resolution were made correctly for 2011 Census/CCS records (Predicted), as evaluated by the 2011 Gold Standard (GS). Assume that the 2011 Gold Standard match and non-match labels are correct.

	Predicted Match	**Predicted Non-Match**
GS Match	*TP*: 633 785	*FN*: 16 159
GS Non-Match	*FP*: 0	*TN*: 59 527

Table 6: Matches and Non-Matches found by improved probabilistic, deterministic and household-associative record linkage methods, the matches that would be found through clerical resolution and the additional matches made if a clerical matcher chose all matches shown as the top scoring candidate by the *Pre-search* algorithm on 2011 Census/CCS records (Predicted), as evaluated by the 2011 Gold Standard (GS). Assume the matcher is presented with only the top candidate in each case and declares a match cannot be found if this looks wrong, but will choose *all* those that are a match on the Gold Standard.

	Predicted Match	**Predicted Non-Match**
GS Match	*TP*: 646 025	*FN*: 3 919
GS Non-Match	*FP*: 0	*TN*: 59 527

Table 7: Matches and Non-Matches found by improved probabilistic, deterministic and household-associative record linkage methods, the matches that would be found through clerical resolution and the additional matches that would be found if a clerical matcher chose all matches shown as one of the top 20 candidates by the *Pre-search* algorithm on 2011 Census/CCS records (Predicted), as evaluated by the 2011 Gold Standard (GS). Assume the matcher is presented with the top 20 candidates in each case and declares a match cannot be found if they all look wrong, but will correctly choose *all* those that are the Gold Standard match.

	Predicted Match	**Predicted Non-Match**
GS Match	*TP*: 647 321	*FN*: 2 623
GS Non-Match	*FP*: 0	*TN*: 59 527

Can "Machine Learning" help?

As an alternative to the probabilistic and deterministic methods already discussed, a variety of ML algorithms have been applied to record linkage problems. Broadly, these methods can be grouped as follows: those that require large amounts of training data in the form of record pairs pre-labelled as matches and non-matches, those that find the record pairs for which labelling will improve match/non-match classification and those that do not require any training data.

Machine Learning approaches to Record Linkage

There are a variety of classification algorithms that have been applied to record linkage that require labelled training data, including support vector machine (SVM) classification and decision trees, but Christen (2012) notes that none of

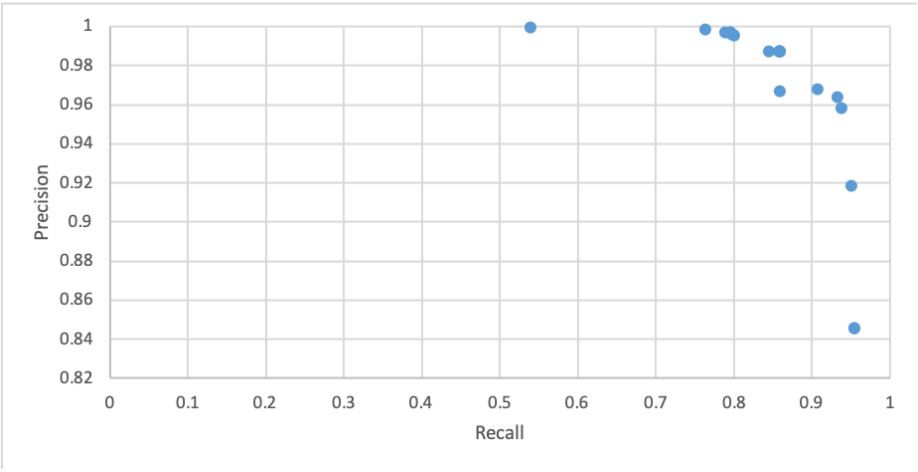

Figure 1: Precision against Recall for Census to CCS probabilistic matching. Data points show the results of different thresholds for considering record pairs a match based on *Fellegi-Sunter* scoring.

these methods have consistently outperformed probabilistic methods, especially for applications with tens of millions of records. By contrast, methods that rely on neural networks such as single layer perceptrons have been reported to outperform traditional probabilistic methods in some cases (Wilson 2011).

A key difficulty with these methods is that in order for a classifier to become highly accurate, the training data would need to include many examples of matches and non-matches and crucially, examples of both that are relatively ambiguous; the kinds that would be classed as indeterminate by a probabilistic method and sent for clerical matching. In response to this problem, active learning methods have been developed that require far less training data, initially only using labelled record pairs from ambiguous cases (where the uncertainty of match/non-match classification was high). The classifier will initially work for only some un-labelled instances, but can be used to find record pairs in the un-labelled data pool which, when labelled, will improve the accuracy of the classifier at the fastest possible rate (Elmagarmid, Ipeirotis, and Verykios 2007). Those pairs can then be manually labelled, adding to the training data and progressively improving the classifier.

The next section of this report will outline some of the methods being researched (or already implemented) by ONS that can be considered examples of ML.

Uses of Machine Learning by ONS

One example of machine learning for record linkage has already discussed in this report; the use of the EM algorithm to estimate the match and non-match class

probabilities from the set of probabilities of corresponding fields being matches or non-matches between the two records, in probabilistic record linkage. This method is considered to be of particular use in scenarios when the record fields cannot be considered conditionally independent, especially when the data contain a relatively large percentage of matches (more than 5 percent) (Elmagarmid, Ipeirotis, and Verykios 2007).

ONS also plan to further improve their *Pre-search* algorithm for 2021 usage by calculating values for the weights of record fields from 2011 record pairing decisions, then iteratively improving these weights with incoming data from matching (both automatic and clerical) carried out in 2021. Ideally, active learning will be used to pick those indeterminate record pairs assigned to clerical matching (and *Pre-search*) that improve its accuracy fastest rate. If this isn't feasible to implement, it could also be useful to utilise ONS's domain knowledge on the census to pick the most likely useful records to label on an ad hoc basis. Over time, the candidate matches for difficult to match records that the *Pre-search* algorithm recommends will improve, reducing the amount of time required for clerical resolution and the likelihood of clerical searching being required.

In the next section, suggestions for further improvement of ONS's record linkage methodology are made, including the use of active learning and other ML methods.

Potential extensions and new approaches

There are several observations that can be made about the record linkage methodology being researched by ONS that have come out of this collaboration project with The Alan Turing Institute, and recommendations on how to further improve these methods in order to meet the challenges and goals specified earlier in this document.

Part of the collaboration discussions involved thinking about the reliance on the *Naive Bayes* assumption in probabilistic record linkage; the fields considered for match scoring are conditionally independent. This is unlikely to strictly be the case for CCS records. For example, date of birth could be linked to some of the other fields like first name, with the popularity of some names being higher in particular years, or marital status, with older people more likely to be married.

Whilst some ML algorithms using training data (e.g. SVM, neural networks or gaussian processes) would not rely on the conditional independence assumption, these methods are unsuitable for the very reason that they rely on large amounts of training data, as already discussed. Also already discussed, is the suitability of setting the field weights with EM in probabilistic matching to avoid reliance on the conditional independence assumption.

The collaboration discussions also yielded several key recommendations for further improvement of the record linkage methods, which are outlined in the

following paragraphs and can be summarised as follows:

1. Utilising information from census fields not present in the CCS to aid matching
2. Using machine learning to create improved distance metrics for field matching
3. Enhancing the *Pre-search* algorithm with active learning

Making use of the full structure of the census data, including fields that are not present in the CCS, could potentially reduce the impact that missing, incomplete or corrupted field data has on record linkage. This could involve writing deterministic rules about associations between other census fields and those fields present in the CCS or using 2011 census records as training data for machine learning to uncover these associations. For example, missing data on age (date of birth) could be inferred roughly from other fields such as marital status or occupation.

Another possibility that involves ML could be to replace the distance scoring metrics like the edit distance used for field matching, with a novel algorithm that is more specific to the particular field in question. This could be a ML algorithm trained with 2011 Gold Standard census and CCS field data, which learns the common types of differences found between corresponding fields in matched record pairs. These differences would therefore be penalised (by lowering the score) less by the algorithm than more unusual mismatches when it is used to score a previously unseen field pairing.

Finally, a key recommendation is to improve the *Pre-search* algorithm using an active learning system. Doing this could offer the advantage of being able to train with 2021 data, reducing the risk of other methods being overfit to 2011 data. One example of this approach being taken comes from Sarawagi and Bhamidipaty (2002), who developed an active learning method to significantly reduce the size of the training set needed for de-duplication (record linkage within a single dataset). Their method starts with a small subset of record pairs that have been labelled as matches or non-matches for training a base classifier, with the classifier being improved by active learning. They tested this with both scientific citation and address records, finding that one to two orders of magnitude fewer pairs were required to be labelled than when random selection was used to decide which pairs should be labelled.

The methods developed by Sarawagi and Bhamidipaty (2002) work as follows: record pairs whose match status the base classifier found difficult to determine were considered to be in a "confusion region", similar to the "reject region" in the *Fellegi-Sunter* algorithm. This region is initially large, but by picking the most ambiguous record pairs for clerical resolution first, those indeterminate as clearly match or non-match which are closest to the mid-point of the confusion region, the algorithm quickly learned the peculiarities of a datasets. Sarawagi and Bhamidipaty (2002) note that the optimal base classifiers to use in tandem with active learning were Decision Trees, which out-perform SVMs and Naive

Bayes, as evaluated by precision and recall values. Similarly, Tejada, Knoblock, and Minton (2001) developed a method using multiple Decision Tree classifiers trained using varying data and parameters, to detect consistently ambiguous record pairs to be prioritised for clerical resolution. In a more recent example, Kasai et al. (2019) report that a deep learning classifier in tandem with active learning can out-perform other base classifiers for record linkage on citation datasets, with comparable performance to deep learning models that use the full training data (they use only 6 % of this training data).

To build in an active learning system into the *Pre-search* algorithm, ONS could use the existing *Fellegi-Sunter* scoring to decide which record pairs are most important to label first, which they are already planning to do as previously discussed (see *Uses of Machine Learning by ONS*). This approach involves adjusting the field weights (initially set with 2011 training data) as new record pair decisions are made in 2021. This aims to increase the likelihood of correct matches for ambiguous (difficult to match) records being presented in the top 20 match candidates list presented to clerical matchers by the probabilistic *Pre-search* algorithm. The active learning component here would be the selection of most ambiguous records for clerical matching staff to resolve first, perhaps by picking those closest to the midpoint of the *Fellegi-Sunter* match and non-match thresholds.

As an alternative to *Fellegi-Sunter* scoring, ONS could use a similar method to those developed by Sarawagi and Bhamidipaty (2002) and others discussed here, such as a using a Decision Tree base classifier in tandem with active learning. Since much of the 2021 record linkage will be done in advance of the *Pre-search*/clerical resolution stage by the determinisitic, probabilistic and household-associative methods, these labelled record pairs (match and non-match) could be used as training data for the base classifier, avoiding overfitting to the peculiarities of 2011 data.

Next Steps

Following on from the scoping exercise summarised in this document, the next steps will be for ONS to continue improving the existing record linkage methods and possibly expand upon them with the recommendations suggested here. As has been made clear in this document, there have already been substantial improvements to the record linkage methods used by ONS since 2011. It may well be that spending a lot of time and effort on the alternative methods recommended here would be less cost effective than further refinement of the improved methods currently being used.

Any new methods that are implemented can be evaluated using 2011 census/CCS data and the 2011 Gold Standard in the same way as the current methods (see *How close are we to full automation?*).

In advance of the 2021 census being carried out, a decision should be made as to which specific methods will be used. This decision should be based on three criteria; the minimisation of clerical matching requirements, the precision and recall percentages (although any that do not meet the requirements on 2011 data can be ruled out) and the perceived likelihood methods not being overfitted to 2011 data.

References

Christen, Peter. 2012. "A survey of indexing techniques for scalable record linkage and deduplication." *IEEE Transactions on Knowledge and Data Engineering* 24 (9): 1537–55. https://doi.org/10.1109/TKDE.2011.127.

Elmagarmid, Ahmed K., Panagiotis G. Ipeirotis, and Vassilios S. Verykios. 2007. "Duplicate record detection: A survey." *IEEE Transactions on Knowledge and Data Engineering* 19 (1): 1–16. https://doi.org/10.1109/TKDE.2007.250581.

Kasai, Jungo, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. "Low-resource Deep Entity Resolution with Transfer and Active Learning," 5851–61. https://doi.org/10.18653/v1/p19-1586.

Murray, Jared S. 2018. "Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering." *Journal of Privacy and Confidentiality* 7 (1): 1–22. https://doi.org/10.29012/jpc.v7i1.643.

Sarawagi, Sunita, and Anuradha Bhamidipaty. 2002. "Interactive deduplication using active learning." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 269–78. https://doi.org/10.1145/775085.775087.

Tejada, Sheila, Craig A. Knoblock, and Steven Minton. 2001. "Learning object identification rules for information integration." *Information Systems* 26 (8): 607–33. https://doi.org/10.1016/S0306-4379(01)00042-4.

William E. Yancey. 2005. "Evaluating string comparator performance for record linkage." *Statistical Research Division*, 3905–12. http://www.amstat.org/sections/srms/Proceedings/y2006/Files/JSM2006-000855.pdf.

Wilson, D. Randall. 2011. "Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage." *Proceedings of the International Joint Conference on Neural Networks*, 9–14. https://doi.org/10.1109/IJCNN.2011.6033192.