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

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, the census questions were asked on paper forms, but in 2021 a combination of online forms and paper forms will be used.

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.

Census Record Matching

The 2011 UK Census estimated that in the UK there are about 65 million people (63.2m) and 25 million households (26.4m), with the CCS sampling 1% of postcodes, counting about 600,000 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 55,000 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.

Current Challenges

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 *Methods for Record Linkage* 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; recall of at least 99.75 % and precision of at least 99.9 %. In 2011, the number of matches was 649,944. This allows no more than 1,629 true matches to be missed by the matching procedure.

In 2011, 70% of the people matches were made automatically using a mixture of deterministic matching and *Fellegi-Sunter* probabilistic matching. This left 30% for manual clerical matching, which involved two processes to find the remaining matches that couldn't be made automatically. Firstly, *clerical resolution*:

deciding whether record pairs that automated methods designated as possible matches are matches, and secondly, *clerical searching*: searching for a match manually when no possible matches are initially presented. For households, a deterministic method was used that was able to match 60% automatically.

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 and the number of staff may also be lower.

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, or to speed up this process. The slowness of this procedure 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 their automated matching methods (detailed later in this document), they are still left with 9% of people records to match manually (5% for household records) when testing these methods on 2011 data. ONS predict that of these people matches, a further 8% will be found by clerical resolution, leaving 1% of matches still to make. This 1% (~5,300 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.

ONS are developing methods to generate a list of possible matches for CCS/census records that could not be automatically labeled as a definite match (or definite non-match), in order to speed up clerical resolution decisions and reduce the number of unmatched records assigned to clerical searching. One challenge here is that if any method requires training data, there will not be any available in advance of it being deployed on the matching days, due to appropriate 2021 census/CCS example records not yet existing. To complicate things further, there is no guarantee that the kinds of difficult-to-match record pairs that a learning algorithm might find useful are likely to arise early in the matching procedure.

The next section of this document will explain the improvements to the census-CCS matching methodology already made by ONS since 2011, after summarising the relevant background literature on record linkage algorithms.

Current Work

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 categorising 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 *Felligi-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 review. Each field used in the calculation is assigned a weight, computed either by an "Expectation Maximisation" algorithm or from the probabilities in training data (Murray 2018).

One key problem with probabilistic record linkage is that it assumes independence of the fields, which is typically not the case. For example, in record linkage between the census and CCS, fields such as first name and date of birth are unlikely to be conditionally independent.

Improvements in Census to CCS Record Linkage

ONS have begun to improve upon the methods used for record linkage between the 2011 census and CCS. Based on their improvements so far, they predict that in 2021, 91% of people records, and 95% of household records can be matched automatically (compared with 70% and 60% respectively in 2011). In this section of the document, the key improvements to the methodology that resulted in this performance increase will be detailed.

In order to improve upon determinstic 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 manual clerical review 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 (tenure, type of property, number of usual residents etc) together with the sets of people records that make up a household occupancy. This method has enabled ONS to make 95% of the matches on the 2011 households Gold Standard.

ONS have also looked into making improvements to the match rate for *Fellegi-Sunter* probabilistic matching. Rather than use the Expectation Maximisation algorithm, they plan to calculate the values for the weights of record fields manually (initially using the 2011 data) and then iteratively improving this using the matching (both automatic and clerical) carried out in 2021. In addition, 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. Testing this approach with 2011 data gave a pairs completeness of 97.8%. An alternative blocking pass on date of birth has also been attempted in order to capture the remaining 2.2%, but no extra matches were made using this.

Some steps have already been taken to speed up the clerical review 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 to clerical review, 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 matching procedure, ONS have developed a *Pre-search* algorithm, which is applied to the CCS and census records assigned for clerical review by the prior automated matching methods before the more laborious clerical searching is attempted. This algorithm finds potential candidates for pairing using very loose blocking, ranks them using *Fellegi-Sunter* scoring, and then presents them to the clerical reviewer who gets to make the final decision as to whether there is a match. The ultimate goal would be to be able to say with confidence that if the matching record is not amongst the top candidates presented to the reviewer, then there is no match for that record.

This method is already working well; when there is a match (as evaluated by the 2011 Gold Standard), it appears as the first record on the list 98% of the time. There is however no way to know for sure that methods that work well for 2011 data will work as well on 2021 data and ONS are keen to consider alternative ML methods for improving the *Pre-search* algorithm.

Potential Uses of Machine Learning

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-labeled 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.

A common example of machine learning in record linkage has already been discussed in this report; the use of the Expectation Maximisation 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. This method does not require training data and 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).

Another example that doesn't require training data involves the use clustering algorithms to group together similar comparison vectors (which contain information about the differences between fields in a pair of records), with the idea being that similar comparison vectors correspond to the same class (i.e. match, non-match or possible match) (Elmagarmid, Ipeirotis, and Verykios 2007).

There are a variety of classification algorithms that have been applied to record linkage that require labeled training data, including support vector machine classification and decision trees, but Christen (2012) notes that none of 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 ambiguous; the kinds that would be classed as indeterminate by a probabilistic method and sent for clerical review. In response to this problem, active learning methods have been developed that require far less training data, initially only using labeled record pairs from ambiguous cases (where the uncertainty of match/non-match classification was high). The classifier will initially work for only some un-labeled instances, but can be used to find record pairs in the un-labeled data pool which, when labeled, will improve the accuracy of the classifier at the fastest possible rate (Elmagarmid, Ipeirotis, and Verykios 2007). Those pairs can then be manually labeled, adding to the training data and progressively improving the classifier.

The next section of this report will outline some of the proposed methods not already being explored by ONS for improving both the *Pre-search* algorithm and the overall record linkage methodology.

Beyond Current Methods

Next Steps

- Needs discussion with ONS

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.

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.

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.