

## **Modelling Approach to Census to Census Matching**

At a previous Census External Panel Meeting, Sir Bernard Silverman suggested that we should investigate using a modelling approach for census to census matching rather than clerically reviewing all the candidate pairs that are currently sent for clerical review.

We estimate that we will have to clerically review around 35,000 candidate pairs. A reduction of around 70% of that required in 2021. This reduction of clerical work has been achieved by development of the automated checking algorithm which splits possible census duplicates into three categories: Automatically accept, automatically reject and clerical review.

Sir Bernard's idea was that we could model the proportion of candidate pairs in the clerical review category that are genuine duplicates. This may require less clerical work than reviewing all 35,000 pairs, although a certain amount of clerical work would still be required in order to train and quality assure the model.

Initial work on this led to the development of a model that could accurately classify the pairs that are automatically accepted but did not classify the pairs that are sent for clerical review so is not of benefit. Unfortunately, we have not been able to develop a model that accurately classifies the pairs in the clerical zone. We only need to get the proportions right, so we don't need a model that accurately classifies each individual pair. However, as we don't have a model that classifies the pairs in the clerical zone at all, we can't estimate the proportions either.

Unfortunately, we have not been able to progress with this work due to lack of resource (covid, hashed matching, census to CCS have taken priority). Census to census modelling has been low priority because we already have a viable method that is significantly more efficient than that used in 2011.

In addition, the census to census methodology fits in well with our overall matching strategy, because in periods when there is no census to CCS matching to be completed (due to automatic processes needing to be run) the clerical matchers will be able to get on with census to census matching rather than being idle.

We have therefore taken the decision to proceed with our original methodology and use inverse sampling coupled with the automatic checking algorithm and clerical review where necessary.