Department of Finance and Personnel
www.dfpni.gov.uk

Data Matching Using Northern Ireland Administrative Data: A Worked Example

October 2014

Northern Ireland Statistics & Research Agency

The Northern Ireland Statistics and Research Agency

The Northern Ireland Statistics and Research Agency (NISRA) is an Executive Agency within the Department of Finance and Personnel (DFP) and has been in existence since April 1996. The Agency also incorporates the General Register Office (GRO) for Northern Ireland. NISRA's core purpose is to provide a high quality, cost effective, statistics, research and registration service that informs policy making and the democratic process and the wider public.

The overall corporate aims of NISRA are to:

• provide a statistical and research service to support decision making by Northern Ireland Ministers and Departments and to inform elected representatives and the wider community through the dissemination of reliable official statistics; and

• administer the marriage laws and to provide a system for the civil registration of births, marriages, civil partnerships, adoptions and deaths in Northern Ireland.

NISRA can be found on the internet at www.nisra.gov.uk

Northern Ireland Statistics and Research Agency
McAuley House
2-14 Castle Street
Belfast
BT1 1SA

Contents

1.0 Introduction

The paper on counting the population; "The Use of Administrative Data in Population Estimates" available from the Northern Ireland Statistics and Research Agency (NISRA) website (www.nisra.gov.uk) shows the value of using administrative data in the creation of an Administrative Data Population Estimate (ADPE). That research was based on aggregated data but concluded that there is a need to process the administrative data at individual record level to increase the accuracy. This paper reports on the methodology that NISRA will use to match administrative data sources to provide accurate linkages that will help develop the Administrative Data Population Estimates work further.

2.0 Record Matching

The aim of any matching exercise is to match records from 2 different sources as accurately as possible. The absence of a unique identifier for each record across all systems is a key issue – it means that records have to be matched based on the available information (usually name, address and date of birth).

An exact match can be created when the details on an individual are identical across two systems. However there are many reasons why the records for an individual on different sources may not be exactly the same. Information such as names and dates of birth may be incorrect or held inaccurately in one or both sources, and alternatives such as middle names in place of forenames or aliases in place of forenames. The method of collection, purpose of the data and the amount of verification applied before the data is entered onto the systems invariably mean that the records will not be identical on all systems. The differences may be minor, such as a slightly different spelling of the surname or they could be significant such as an incorrect address, date of birth or a middle name used in place of a forename. It is therefore important to ensure that any matching algorithms take account of non-exact matched records.

It is also very likely that matching two sources will not achieve a full match because of different coverage issues, capture methods and potential time lags between the disparate data sets.

For the 2011 Census, census records were matched to the 2011 Census Coverage Survey (CCS) using a combination of exact matching, score based matching, clerical matching and clerical searching. A summary of these methods is provided below:

- Exact matching – automatically linking pairs of records that are identical on all matching fields (for example name, sex, date of birth, gender and postcode);
- Score based matching – scoring pairs of records for their overall level of agreement and automatically linking those that score above a specified threshold;
- Clerical matching – the manual review of pairs of records that are classified as potential matches based on their overall agreement scores. These records have scored lower than the auto-match threshold and a trained matcher is required to make a decision based on the evidence available as to whether or not the two records should be linked; and

- Clerical searching – individually taking the 'residuals' (unmatched records) on one of the datasets and querying the database of the second dataset for a matching record. Where potential match pairs are identified, a clerical decision is made by the matcher as to whether or not to link the two records.

In the 2011 Census to CCS matching, the role of clerical matching and clerical searching was very important for the optimisation of matching accuracy. By using these methods, it ensured that the number of false positives and false negatives were kept to a minimum, thereby enabling accurate coverage adjustments to be made in the estimation process.

The focus of this research to date has concentrated on the development of new methods that perform well in a fully automated matching process and the ability to measure the quality of these matches compared to methods that would include clerical matching.

This paper outlines a matching process that gives high match rates with low levels of matching error

The matching methodology proposed is deterministic, or 'rule based', referred to as match-keys. A series of match-keys have been developed, each of which is designed to resolve a particular type of inconsistency that often occurs between records belonging to the same individual in differing administrative data sources.

Match-keys are created by putting together pieces of information to create unique keys that can be used for automated matching.

Inconsistency between matching variables can occur in a number of different forms. A single match-key alone cannot resolve all of the inconsistencies that occur between data sources, hence the need for multiple match-keys. A series of match-keys have been developed, each of which is designed to resolve particular inconsistencies between match pairs.

The highest level of matching is exact matching which links pairs of records that are identical on all matching fields (match-key 1). An example of a non-exact match-key is one constructed from the first two characters of an individual's forename and surname (Bi-grams), combined with their date of birth and postcode district (match-key 3).

The match-keys are processed in a stepwise manner starting with match-key 1 and working down to the last match-key. Records are only linked on a match-key if it is unique on both datasets (i.e. one-to-one match). If multiple records match on a particular match-key then the link is not made and candidates are passed on as a residual to the next match-key.

3.0 Testing the Matching Methodology

The match-key process was tested on several small subsets of data currently available within NISRA and the results showed that the process was both viable and accurate. To test this methodology, 2 datasets with demographic details and full coverage of the Northern Ireland population have been identified.

Data Set A has a very high coverage of a unique identifier (97.4 per cent) for the Northern Ireland population and has been collected for operational purposes.

Data Set B is collected by a different data custodian for operational purposes and has full coverage of the same unique identifier. Data Set A and Data Set B have been collected by different data custodians independently, for different purposes, both include demographic details and both have the same unique identifier.

Data Set A was matched to Data Set B using the demographic details in both datasets using the match-keys methodology. The pairs of unique identifiers were then compared to check if the match made was correct. If the unique identifiers are the same the match was considered to be correct. If the unique identifiers were different the match was considered to incorrect (or a mismatch).

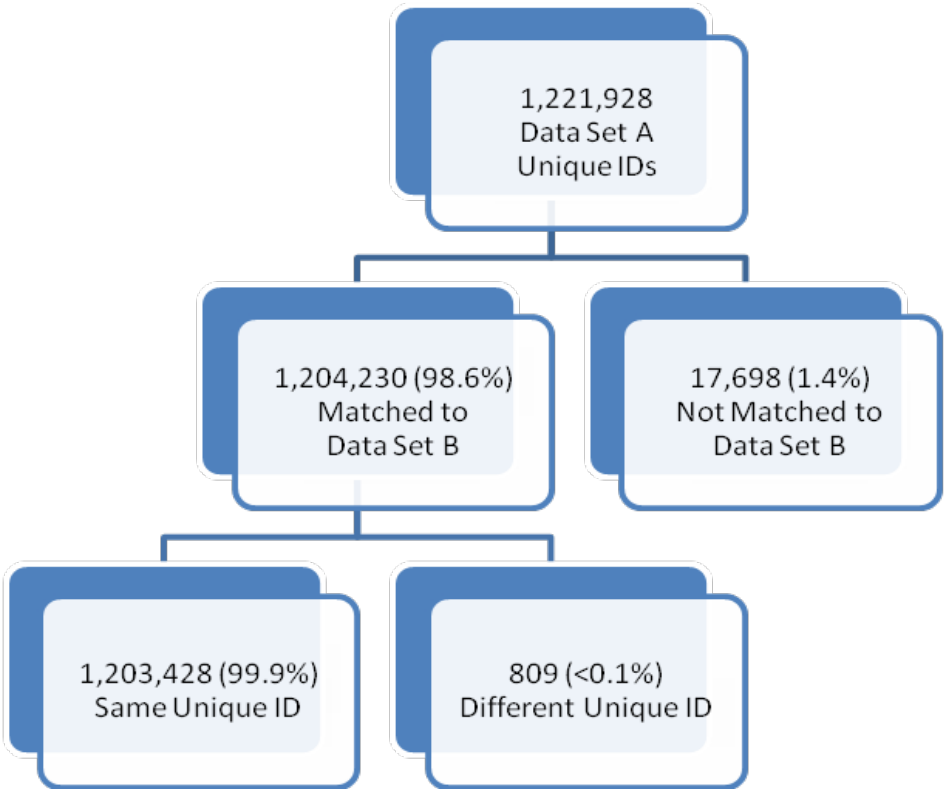
The match-keys developed for these data sets are given below:

Match-key	Description	Inconsistencies resolved by match-key
1	Forename, Surname, date of birth, Postcode	None - exact agreement
2	Forename Initial, Surname Initial, date of birth, Postcode Sector (BT30 6 or BT1 1)	Name / postcode discrepancies
3	Forename Bi Gram, Surname Bi Gram, date of birth, Postcode District (BT30 or BT1)	Name discrepancies / movers in area
4	Forename Initial, date of birth, Postcode	Name discrepancies
5	Surname Initial, date of birth, Postcode	Name discrepancies
6	Forename, Surname, Age, Postcode	Date of birth discrepancy
7	Forename, Surname, Postcode and age within 5 years of each other	Date of birth discrepancy
8	Forename, Surname, date of birth	Movers out of area
9	Surname, Forename, date of birth, Postcode (matched to Match-key 1)	Forename / surname transpositions

NOTE: Bi Gram is the first two characters of a string

The number of people on Data Set A with the unique identifier was 1,221,928. Using the match-keys methodology described above a total of 1,204,230 records were matched to a record in Data Set B. Comparing the unique identifiers of those that were matched, 1,203,428 had the same unique identifier and 809 records had a different unique identifier.

The overall match rate between Data Set A and Data Set B was 98.6 per cent with 99.9 per cent of these matches to the same unique identifier - this is a very high level of precision.



This matching methodology is similar to that used by the Office for National Statistics as part of the Beyond 2011 programme of research - see 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

This methodology can be adapted to meet the needs of differing datasets as they may not all have the same demographic data – for example the Data Set A did not contain gender. It may be possible to develop the matching methodology further to increase the coverage of matched records whilst still maintaining a high degree of confidence by limiting the number of false positives and false negatives – however the additional return will have negligible effect on the overall match rate (see above).

It is acknowledged that the data matching methodology outlined in this paper requires the provision of individual personal identifiable data from the data owner. It is beyond the scope of this paper, but NISRA is taking forward work to ensure that the provision of data is covered by appropriate legislation and protocols, and that the confidentiality of personal identifiable data is fully respected.

4.0 Conclusion

This matching methodology can achieve a high level of matching with a high degree of precision as demonstrated by the example contained within this paper. This level of matching and accuracy will be suitable for use in the development of the Administrative Data Population Estimates based on individual records.

The example stated has provided a robust method of quality-assuring the data matching methodology. The methodology has been used to match other subsets of the population from different administrative data sets and this has achieved similar levels of matching and precision. Further work is required to ascertain whether the match rate can be increased, without introducing false positives and false negatives.