Data Linkage and Matching

6.1 Introduction

There are two main approaches to the use of administrative data in the statistical production process:

- As a direct source for statistics
- As an indirect source, in combination with other sources

If data from several administrative sources are used to supplement survey data or populate a statistical register, the national statistical organisation will need to find some way of linking those data. This will typically take the form of matching, which can be defined as the linkage of data from different sources based on common features present in those sources.

6.2 Common Identifiers?

If these common features include some sort of common reference or identification number (referred to as a common identifier from this point onwards), the process can be referred to as exact matching, and is relatively easy. In exact matching there are two possible outcomes, either two records from different sources match exactly on the basis of the common identifiers used, or they don't. In other words, a record with the identifier 123456 will match to a record in a different source with the same identifier (assuming the sources are covering the same units!), whereas it will not match to a unit with the identifier 123457.

Exact matching depends heavily on the quality of the matching variables used in each source. If there are errors in the common identifiers in at least one source, there is a high risk of either matching the wrong units, or failing to match units that should be matched. For this reason, even when common identifiers exist in all the files to be matched, it may not be sufficient to rely on exact matching alone.

Sometimes identifiers can include check digits, i.e. one or more characters that are generated according to a standard algorithm based on the other digits in the identifier. If check digits are present, they should help to guarantee a certain level of quality by eliminating most typing or reading errors.

6.3 Matching Keys and the Concept of Distinguishing Power

Where common identifiers are not present, or are not of sufficient quality to give the required level of accuracy in matching, it is necessary to consider using other variables common to the sources involved. The variables chosen are often referred to as "matching keys". Note: it is not always necessary for these variables to be present in both sources, as in some cases they can be derived (for example, see the discussion on turnover per head ratios in Box 4.4). When variables other than common identifiers are used, the matching routines tend to rely on probabilities to determine which records match.

The variables most commonly used for this sort of probabilistic matching are name, address, date of birth, occupation or economic activity code. The choice of variables to be used for matching should take into account both the "distinguishing power" of each variable. Distinguishing power relates to the uniqueness of the values of the matching key. Some variables have higher distinguishing powers than others:

- High distinguishing power: reference number, full name, full address
- Low distinguishing power: sex, age, city, nationality

Within a variable such as "full name", it is also possible for some values to have a higher distinguishing power than others. Names that are unique will have the highest distinguishing power, whereas those that are more common (e.g. John Smith in many English-speaking countries) will have a much lower distinguishing power.

Distinguishing power can also depend on the level of detail, e.g.

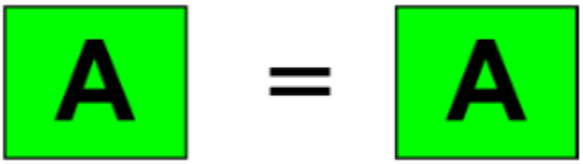
- "Born 1960, Paris" = low distinguishing power
- "Born 23 June 1960, rue de l'Eglise, Montmartre, Paris" = high distinguishing power

Therefore careful choice of matching keys, taking account of the concept of distinguishing power, can have a significant impact on the success of a matching exercise.

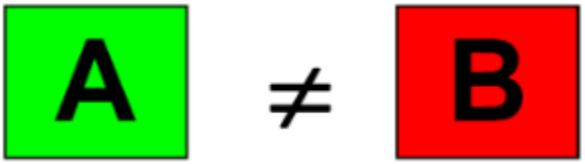
6.4 Some Basic Matching Terminology

When two records are compared, they can be referred to as a "pair". The following scenarios illustrate the main potential outcomes from applying matching techniques to that pair of records:

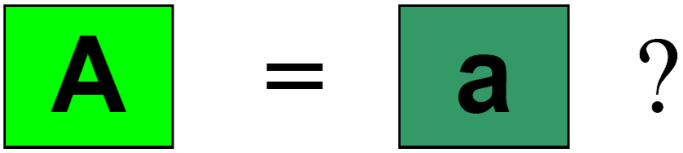
1) Match - A pair that represents the same entity in reality



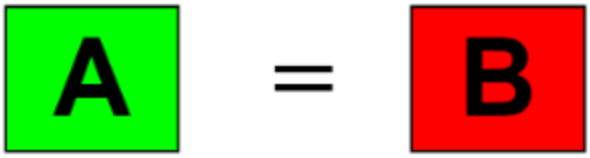
2) Non-match - A pair that represents two different entities in reality



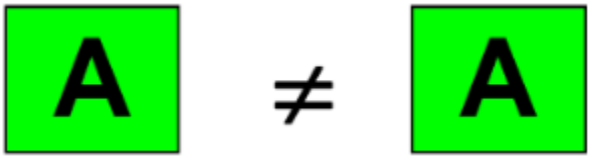
3) Possible Match - A pair for which there is not enough information to determine whether it is a match or a non-match



4) False Match - A pair wrongly designated as a match in the matching process (false positive)



5) False Non-match - A pair which is a match in reality, but is designated as a non-match in the matching process (false negative)



To get a better appreciation of matching concepts and issues in practice, please see the matching exercise at the end of this chapter. It uses made-up, but realistic data to illustrate how matching without common identifiers requires a certain amount of judgement, and how matching can often be more of an art than an exact science. Any form of probabilistic matching is likely to result in a certain proportion of false matches and false non-matches, as well as the need for further investigation of possible matches.

6.5 Matching Techniques

Matching techniques can be split into two basic categories:

1) Clerical matching – by definition this requires significant human input, so is therefore likely to be:

- **Expensive**
- **Inconsistent**
- **Slow**
- **But; intelligent**

2) Automatic matching – once operational (i.e. ignoring one-off set-up costs), this approach minimises human intervention, so is likely to be:

- **Cheap**
- **Consistent**
- **Quick**

- **But; of limited intelligence**

The best solution is therefore to use an automatic matching tool to find the obvious matches and no-matches, and to refer possible matches to specialist clerical staff. To be cost-efficient, the aim must be to maximise automatic matching rates whilst minimising clerical intervention. The remainder of this chapter considers the main features of automatic matching, and how it can be used and improved in practice.

6.6 How Automatic Matching Works

Automatic matching tools usually follow a similar sequence of steps, though depending on the particular application, some steps may be omitted or others may be added. The most common steps are:

1) Standardisation

This step is mainly used for text variables, or variables that should conform to a specific format. Examples of standardisation processes are:

- Abbreviations and common terms are replaced with standard text, for example the text string “ltd” could be converted to “limited”, and “mfg” to “manufacturing”.
- Common variations of names are standardised, for example there may be different versions of the name of a city (“Brussel” / “Bruxelles” in Belgium, “Derry” / “Londonderry” in Northern Ireland). A similar process is needed for person names where there are different spellings of the same name (“Jane” / “Jayne”) or common short versions of a name (“Bill” / “William”). This is a similar process to, and can possibly be combined with the standardisation of abbreviations.
- “Noise” words are removed – typically these are words or phrases with very low distinguishing power, examples could include “road” or “street” in addresses.
- Postal codes, dates of birth etc. are given a common format, for example “3 January 1985” could be converted to “030185”.

The process of standardisation is heavily language dependent, and may also vary according to the type of records being matched, thus the above examples only illustrate the process. Each instance of matching will require prior work, usually based on an investigation of the data, to determine which standardisation rules should be applied.

Standardisation can also be seen as a form of data cleaning, and as such, carries a risk that it could distort or reduce the quality of the data, and even in extreme cases reduce the likelihood of finding a correct match. Such risks are usually very small, and are usually due to ambiguity in the string being standardised. Examples in the English language include the abbreviation “St.” which could refer to either “street” or “saint”, and the name “Chris”, which could be a short form of “Christopher” (male) or “Christine” (female).

Another type of standardisation sometimes used as an initial step in a matching process is to check addresses against a definitive list, usually from the national postal authority. This can range from a check that the combination of postal code and town / city / region is valid, to a full check of the entire address. The success of such a check will obviously be heavily dependent on the quality of the reference file of addresses used.

If the result is that a “cleaned” address is used, it is good practice to also keep a copy of the raw data. In several cases (including matching business data in the UK), it has been found that using cleaned addresses increases the likelihood of matching some records, but decreases it for others. Combining the results of two parallel matching exercises, one using cleaned addresses, and the other using the raw versions can often give the best results.

Two other potential consequences of using cleaned addresses should also be noted, even though they are not strictly related to matching. The first is that in some countries, postal authorities may give a discount for bulk mailing where the addresses used conform to certain standards, so this may help to offset the costs of the cleaning and matching process. On the other hand, substituting a cleaned address for that supplied by a respondent may in some cases cause annoyance to the respondent. If cleaned addresses are used for mailing statistical questionnaires, this may affect response rates. These are further arguments for storing both cleaned and raw data whenever possible.

2) Parsing

Parsing can, to some extent, be seen as an extension to standardisation. In this step, text is converted from a form that is readily recognisable by humans, to a form that is more logical for computer processing, and therefore more likely to correctly match. The resulting text strings are often referred to as matching keys. Early approaches to parsing in the English language often used the “Soundex algorithm”, first patented in 1918. This algorithm, or derivations from it, form the basis of many matching applications. However, parsing rules vary considerably between languages, and should be tuned to give the best results for the data concerned.

Examples of parsing rules could include the following:

- Converting letters or groups of letters with similar sounds to a common string, e.g. “f”, “v” and “ph” to “f”.
- Removing silent letters, e.g. the “h” in the name “Thomas”.
- Converting all characters to either upper or lower case.
- Converting vowels to a single character.
- Removing vowels at the end of a name or word.

- Replacing double letters with single letters, e.g. “Ann” becomes “An”

For example, during parsing using all of the above rules, the string “Steven Thomas Vale” could be converted to “stafan tamas fal”. The string “Stephen Tomos Vael” would also give the same result, showing how parsing can help improve matching rates by reducing the impacts of different ways of spelling a name, and of spelling errors. It should also be noted that varying the order in which the parsing rules are applied could affect the outcome.

As with standardization, however, if parsing is not sufficiently well adapted to the data to be matched, there is a risk that it could do more harm than good. At least in the initial stages, the impact of the parsing routines should be carefully analysed, and in all cases, a copy of the raw data should be retained for comparison purposes.

3) Blocking

If the file to be matched against is very large, it may be necessary to break it down into smaller “blocks” to save processing time. There are several ways to do this, for example, if the record to be matched has an address in a certain town, it may only be necessary to match it against the block containing other records from that town, rather than all records for the whole country.

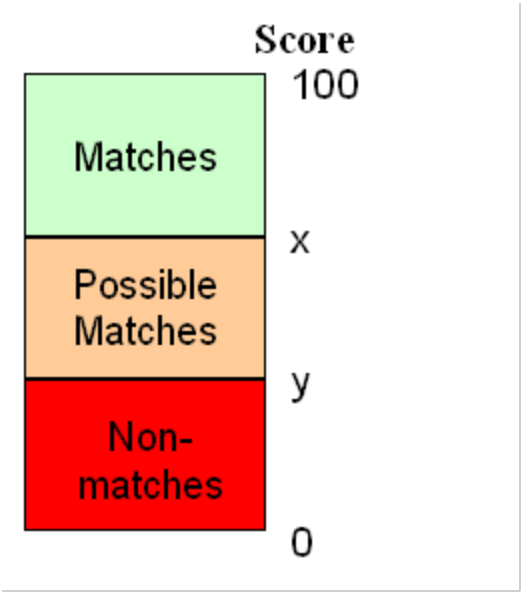
Blocking must be used with care, and is often likely to result in reductions in the overall matching rate. However, if these reductions are minimal, and the gain from faster processing times is substantial, blocking can improve the cost-efficiency of the matching process. In some cases it may even be appropriate to have two or more attempts at matching, using different blocking criteria. For example, after applying relatively restrictive matching criteria to the full dataset, it might prove advantageous to re-process the sub-set of records that do not initially match, using successively less restrictive blocking criteria.

Blocking is clearly most appropriate for very large datasets, such as individual records from a population census, but as a matching technique, it is likely to decline in value as computer power and processing speed increases.

4) Scoring

Most automatic matching routines use some form of scoring to assess the likelihood of a match between two records. Scores are allocated based on how closely the matching variables agree. These scores can be used to determine whether a pair of records is considered to be a definite match, a possible match or a non-match. Figure 5.1 shows how the different categories can be assigned based on threshold scores x and y, in this case expressed as scores out of 100.

Figure 6.1 The Use of Threshold Scores to Determine Category of Match



The next logical question is to how to determine the values of x and y. One option is to use a model based approach, as proposed by Fellegi and Sunter[1], though in practice, a trial and error method is equally likely to be used.

For repeated matching exercises, data quality will vary over time, so a periodic re-assessment of the values of x and y is needed. Similarly, changes in the requirements for the matched data, or in the resources available for matching can lead to revised thresholds. Thresholds may also vary significantly between different datasets.

In setting the values of x and y, it is also necessary to consider the impact of different types of matching error. If a false match is likely to lead to disclosure of statistical information about one unit to another, then the value for x should be set sufficiently high to make the risk of a false match

acceptably low. However, if there is no risk of disclosure, and the results of the matching will be used in a study where a certain proportion of false matches is unlikely to have a significant impact on the results, the value of x can be lower.

The availability of clerical resources to investigate possible matches will often in practice place a constraint on the distance between x and y. In all such cases, clerical intervention should be prioritised. This may be by score, so that those possible matches with the highest score are checked first, as this could be assumed to give the most benefit. Alternatively, some other characteristic of the units involved (e.g. number of employees for businesses) can prioritise the clerical work so that it minimises the impact of potential duplication.

6.7 Matching Applications in Practice

Although they often work in rather different ways to that described above, the matching applications most familiar to many people are Internet search engines. They take a text string (typed in by the user) and then search for web pages related to that string, often scoring the results and returning them in order of perceived relevance. Some form of parsing may also be apparent in the results, or through the suggestion of alternative spellings.

Internet search engines also provide a good demonstration of the concept of distinguishing power. For example, at the time of writing, a search on www.google.com for the text string “matching” returned around 700 million results, whereas “statistical matching” returned about 30 million results, and “parsing techniques in statistical matching” returned around 1.6 million results. More detail clearly helps to focus the search.

In the world of official statistics, there have been two main approaches to developing data matching applications:

- Using “off the shelf” commercial software, e.g. Informatica Identity Resolution (incorporating SSAName3)[2]. It should be noted however, that some form of customization is likely to be needed before any commercial package can be used to its full potential.
- Developing matching routines in house, e.g. software developed by the US Census Bureau[3], Statistics Canada[4] and ISTAT, the Italian statistical office[5].

An alternative approach to matching is the “trigram” method, which works by splitting text strings into groups of three characters, and then calculating the proportion of identical groups between two strings.

For example, matching the string “Steven Vale”:

Ste/tev/eve/ven/en /n V/ Va/Val/ale

To the string “Stephen Vale”:

Ste/tep/eph/phe/hen/**en /n V/ Va/Val/ale**

Results in six matching trigrams (shown in bold), out of a total of thirteen unique trigrams from both strings, thus giving a score of 6/13 or 0.46. Parsing of the strings may help to improve the score, but as discussed above, may also introduce errors[6].

Box 6.1 – Case Study – Extracts from “Matching Records Without a Common Identifier - The UK Experience” by Steven Vale and Mike Villars

This text is derived from the full paper, which can be found at: http://www1.unece.org/stat/platform/download/attachments/56230020/matching+paper.pdf?version=1

The UK statistical business register uses data from several administrative and statistical sources, the most important of which are Value Added Tax records and Pay As You Earn income tax records. There is a considerable overlap in the coverage of these two sources, so to minimise duplication it is essential to check that new units from each source are genuine, and haven't already been added from the other source. Each source has its own system of unit identifiers, which means that matching based on names and addresses is the best solution.

Input files are processed in four phases:

- Cleaning - This routine edits the name string, removing special characters and replacing lower-case with upper-case.
- Formatting - This routine edits the name string into separate words, removing “stop words”; replacing selected words and concatenating prefix words.
- Standardisation - This routine “standardises” the name, for example removing double characters.
- Key generation - This generates codes based on the input text, e.g. if the input is “Steven Vale” the keys produced are:

STEVEN ® STAFAN ® XJXM$$$; and VALE ® VAL ® YLVO$$$$

YLVO$$$$ is the key for the last part of the name, and is used as the major key. It is checked against a table of namekeys generated from the names of each record held on the register, to find potential matches. The input name, address and postcode are compared with the name, address and postcode of each of these potential matches and given a score out of 100. If the score is >79 then the pair is considered a definite match. If the score is between 60 and 79 then it is a possible match. Any lower score is regarded as a non-match.

Duplicates on the definite match list are removed, as well as records on the possible match list that also appear on the definite match list. The records on the definite match list are then linked automatically to their corresponding units on the register. The records on the possible match list, and larger non-match records are reported for clerical checking. For a typical update, around 37% of records are definite matches and 35% are possible matches (of which approx. 80% can be matched clerically).

One problem encountered was the use of "Trading as" or "T/A" in names e.g. “Mike Villars T/A Mike’s Coffee Bar”. In this case, “Bar” would be used as the major key, but has a low distinguishing power as there are many bars in the UK. The solution was to split the name so that the last word prior to "T/A" i.e. Villars is the major key.

Annex to Chapter 6 - Matching Exercise

This exercise contains five examples where a new record has been automatically matched against an existing set of records. No definite matches have been found, but the five highest scoring possible matches are presented for clerical checking. These data are realistic, but are not actually real. Please choose the best match for the new record. Alternatively, if none of the possible matches seem close enough, you can decide that there is no match. Answers are given after Example 5.

Example 1

New record		Possible matches	
Name:	Bob the Butcher	1	Bob Daley Butchers
Address:	16 Lawrence Street		17 Barwick Green
	Southfleet		Sidcup
	Gravesend		Kent
Postcode:	DA11 7ZP		DA15 8HP

		2	Brian Dunn Brians Family Butchers 16 Pembroke Close Pembroke Street Dover Kent DA6 1FB
		3	Mr B Dunn and Mrs V Dunn Brian's Family Butcher Pembroke Street Gravesend Kent DA6 1AA
		4	B & B Butchers Mr B Jones 3 Clive Road Dartford Kent DA1 5RH
		5	B Washbrook Bob the Butcher 16 Lawrence Drive Castle Lane Southfleet Gravesend Kent DA11 7ZF

Example 2

New record		**Possible matches**	
Name: Address: Postcode:	Cars of Southfleet 3-5 Old Hill Southfleet Dartford DA1 9KT	1	Fleet Motors 31-35 Old Dover Road Dartford Kent DA15 7JF

		2	Southwold Cars 1A Southwold Close Greenhithe Kent DA23 9BC
		3	Mr D Crane T/A Southeast Cars 12A Old South Road Greenhithe Gravesend Kent DA2 9BN
		4	Mr C James & Mr G Smith Fleet Motors 29-35 Old Dover Road Fleet Kent DA15 9XX
		5	Southfleet Cars 33 Old Hill Southfleet Dartford Kent DA1 9XT

Example 3

New record		**Possible matches**	
Name:	Retail Co-operative Limited	1	Mr A Cooper Paintcraft
Address:	35, Station Parade Station Road Dartford		Unit 132 Greenway Estate Lower Station Lane Welling Kent
Postcode:	DA1 7ED		DA18 6GT

		2	Retail Co-op Ltd 030001 35 Station Street Dartford DA1 7DH
		3	Co-operative Funeral Services 362 Longfield Street Dartford DA1 1HD
		4	Co-operative Funeral Services Ltd, CFS (No14) Ltd & CFS Pension Fund 29 Station Street Bexleyheath Kent DA32 4RH
		5	Arts Co-operative 62 Highfield Street Dartford DA21 8JD

Example 4

New record		**Possible matches**	
Name: Address: Postcode:	Dr James Johnson Griffons Penny Lane Eynsford Dartford DA46 8FF	1	Mr James John Cunningham 35 Griffin Drive Darenth Dartford Kent DA4 6FF
		2	Mr John Jameson 56 Whinfell Road Gravesend Kent DA21 8GF

		3	Mr James Johnson 123 Penny Lane Aynsford Kent DA46 3JF
		4	John James 23 Perry Lane Dartford Kent DA28 3PF
		5	Mr James John Smith 18 Cornfield Lane Eynsford Dartford Kent DA46 8FF

Example 5

New record		**Possible matches**	
Name: Address: Postcode:	Redipure Ltd 26A Queens Rd Welling DA13 8RS	1	Redipure Limited Perseverence House 36A Cross Road Howley Dartford Kent DA27 8RR
		2	Eradicure Ltd Perseverence House Cross Rd Howley Dartford Kent DA27 8RT

		3	Redpull Ltd 152 Lower Wickham Lane Wellington Kent DA13 8ED
		4	Redpull Ltd 12 Lower Wickham Welling Kent DA13 3ED
		5	Redipure Holdings Ltd Crossroads Howley Dartford DA12 3LF

Answers

This exercise shows that there is rarely 100% certainty in matching. The answers below reflect the greatest likelihood of a match according to clerical matching experts.

Example 1 – The most likely match is with existing record number 5. The trading style of this record matches the name of our new record and the addresses are fairly similar. There is one character different in the postcode, “P” instead of “F”, which could easily be a transcription error in one of the records.

Example 2 – Again, the most likely match is with existing record number 5. The names and addresses are sufficiently similar, and as in example 1, there is only one character different in the postcode. This case also highlights an interesting issue, in that the existing records 1 and 4 may also be a match. This could indicate duplication amongst existing records, and shows the value of matching a dataset against itself periodically to reduce the risk of such duplication.

Example 3 – The closest match is with existing record number 2. The main differences concern the use of abbreviations in the name of the existing record (Ltd – Limited; Co-op – Co-operative). This suggests that the automatic matching routine is not sufficiently good at relating abbreviations to their full versions. Abbreviations such as these are often specific to a language or even to a data set, and show the value of being able to customize automatic matching tools according to the types of data being matched.

Example 4 – The closest match is with existing record number 3. Record number 5 is an exact match for most of the address, as well as the postcode, and would potentially score higher in automatic matching. This illustrates the risk of setting the positive match threshold too low.

Example 5 – Based purely on the evidence here, there does not seem to be a match. However, this case illustrates the value of using additional knowledge in clerical matching. The abbreviation “Ltd” in the name of the new record indicates that it is a limited liability company. In many countries, limited liability companies must, by law, have unique names. This suggests that if the matching is intended to link units in the same enterprise, the new record should be linked to existing record number 1. The different addresses may simply refer to different sites (local units or establishments) that this company operates from. Improving the automatic matching routine to recognise corporate businesses, and put a much greater emphasis on the name in such cases, could therefore help to improve the automatic matching rate. This strategy was successfully adopted in the matching routines used for the UK statistical business register.

---

[1] See A Theory for Record Linkage, Ivan P. Fellegi and Alan B. Sunter, http://www.jstor.org/stable/view/2286061
[2] http://www.informatica.com/products_services/identity_resolution/Pages/index.aspx
[3] http://www.census.gov/srd/papers/pdf/rr2001-03.pdf
[4] http://www1.unece.org/stat/platform/display/msis/G-Link

[5] http://forge.osor.eu/projects/relais/

[6] A practical application of this method, programmed as SAS code, was demonstrated by Statistics Finland within a Eurostat project to develop statistics on business demography.