Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Decision Rules and Error-Rate Estimation for Record Linkage Using a Probability Model

Clayton Block[1]

Abstract

Since 1997, Elections Canada has maintained the National Register of Electors, a database of Canadians aged 18 and over, used to administer federal elections. This database is updated from several federal and provincial administrative sources, linked to electors in the database using personal information such as names, date of birth, gender, and address. Initially, commercial linkage software based on Fellegi-Sunter theory was used for these linkage activities. Gradually, the methodology and software used have shifted towards custom-built solutions, providing more flexibility over how potential pairs get processed, and reducing the classification error rates associated with the linkage process. One key improvement to the methodology is a reformulation of the familiar Fellegi-Sunter decision rule, now put in terms of a probability of interest and compared to an error tolerance. For matching on personal information, the required probabilities are calculated from the observed pairs with the aid of a simple probability model for chance agreement on date of birth. The model assumptions should be quite realistic. The probabilities calculated for each pair can also be simply added up to produce estimates of the two types of matching error, requiring no specialized software and no complex mathematical procedures. The methods described will be used for various linkage processes at Elections Canada, each with different expected match rates. The believability of the resulting error rates will then be assessed. In the future, these results could be compared and contrasted with those obtained from competing, more complicated error-rate estimation methods.

Key Words: Probabilistic record linkage; Fellegi-Sunter theory; Classification error rates.

1. Introduction

1.1 Background

Since 1997, Elections Canada has maintained the National Register of Electors, a database of Canadians aged 18 and over, used to administer federal elections and referenda. In addition to information supplied directly from electors themselves, the database is maintained using updates from several federal and provincial administrative sources. In order to be used for updates, records from update sources first need to be matched to records in the Register database, a process known as record linkage. The information maintained in the Register includes names, addresses, dates of birth and gender.

1.2 Description of the Record Linkage Problem

Given a pair of records, we want to decide if the two records refer to the same entity or to different entities. Record pairs may be from the same source, as in duplicate detection, different sources, as in file matching, or both.

Once decisions about the record pairs in question have been made, it would also be useful to determine how many classification errors were made. There are two types of errors possible: accepting a pair that actually refers to different entities, and rejecting a pair that actually refers to the same entity. These two types of errors will be referred to as false + and false –, respectively.

In order to make sensible and informed decisions, it is crucial that the consequences of making each type of error be fully understood, and if possible, factored into the decisions made. In the context of Register maintenance at

[1]Clayton Block, Elections Canada, 30 Victoria St, Gatineau QC, Canada, K1A0T6 (clayton.block@elections.ca)

Elections Canada, false + pairs can cause Register records to get corrupted with false information that becomes difficult to correct, and legitimate electors that do not appear on an electoral list, or appear at an incorrect address. This results in electors being inconvenienced, which could lead to negative public perception of Elections Canada, and could even result in negative press coverage of the organization. On the other hand, false – pairs can cause legitimate updates to be missed, or cause ineligible electors or duplicate records to remain on an electoral list. This could also lead to negative public perception of Elections Canada, and perhaps negative press coverage.

Note that both types of errors can lead to negative consequences. Taking these into consideration, Elections Canada considers a false + to be more serious than a false –, not only because the implications are more serious, but also because it is more difficult to correct after the fact. In fact, many false – cases are easily corrected later, when more information becomes available about the pairs, such as address updates.

In addition to simply accepting or rejecting record pairs, a third option could always be considered. In cases where it is unclear whether to accept or reject a pair given the information available, the decision itself could be deferred until enough information can be obtained to resolve the case, for example by contacting the entities involved. Unfortunately, given the limited time, information and resources available, and the extra burden on electors this would require, this is not seen as a realistic option for Elections Canada.

2. Record Linkage Decision Rules

2.1 Record Linkage is a Probability Problem

Consider the following fictional but realistic example of a record pair:

Name	Date of Birth	Address
Robert J Smith	July 9, 1963	123 Main St, K1L5T4
Bob Smith	July 9, 1963	246 Elm Dr, R1M4T9

In the absence of any other information, there is no way to know for sure if this is one person who has moved to a new address, or two different people who just happen to have similar names and the same date of birth. When faced with this uncertainty, the appropriate tool to use is probability.

If *M* denotes a true match, *U* denotes a true non-match, and *Outcome* summarizes everything pertinent that we can observe about a given pair, an appropriate decision rule would be

Reject pair if	$P_M < tolerance\ for\ false+$
Accept pair if	$P_U < tolerance\ for\ false-$
Defer decision	Otherwise

where
$$P_M = \Pr(M|Outcome) \quad (1)$$
$$P_U = \Pr(U|Outcome) = 1 - P_M \quad (2)$$

If deferral of the decision is not an option, one of the tolerances is simply dropped, the decision becomes accept or reject based on the tolerance retained, and one type of error must be left uncontrolled.

The sample space used to define the probabilities of interest is simply the set of all possible record pairs that can be formed from the records available, and all the information that is available about these records and record pairs.

2.2 General Considerations for Decision Rules

The probabilities required for making sound record linkage decisions are not known, so have to be somehow estimated. The resulting estimates will not be perfect, and will depend heavily on what information is used. Like

any decisions, subjectivity should play a part if desired. For logical consistency, in addition to calculated probabilities, the following should be considered general guiding principles for record linkage decisions:

1. All available information about the records in the pair that is pertinent to the decision should be utilized.
2. Subjective aspects of the process should be set in advance if possible, and should be logically integrated with more objective considerations.
3. Any pairs with common *Outcome* should receive the same decision.
4. Any pair with *Outcome* 'better' than another pair that was accepted should also be accepted.
5. Any pair with *Outcome* 'worse' than another pair that was rejected should also be rejected.

Determining which pairs are 'better' or 'worse' than others is where the objective and subjective aspects of the problem can come into conflict, perhaps never getting resolved perfectly.

2.3 Probabilistic Record Linkage

One commonly used approach to this problem is known as probabilistic record linkage (Fellegi and Sunter, 1969). Instead of using the probability of interest P_M directly, the suggested decision rule is

Reject pair if	$R < lower\ threshold$
Accept pair if	$R > upper\ threshold$
Defer decision	Otherwise

where $$R = \Pr(Outcome|M)/\Pr(Outcome|U) \quad (3)$$

If all the probabilities are correctly specified, using Bayes' Rule it can be easily shown that this is mathematically equivalent to the rule specified in section 2.1 above.

To calculate *R*, the suggested approach is to limit *Outcome* to the results of individual comparisons of the relevant fields on the records themselves. A key simplifying assumption is that the comparison outcomes for each field included are all independent of each other.

The probabilities are estimated iteratively from a subset of all possible pairs, obtained by requiring strict agreement on several different combinations of fields or field components. Instead of using the estimate of *R* directly, it is transformed into a weight for each pair, which is compared to weight thresholds. The weights and thresholds are typically adjusted after examining results based on samples of records.

2.4 Drawbacks of Probabilistic Record Linkage

When established in 1997, Elections Canada's Register maintenance program used commercial software to carry out probabilistic record linkage using the approach described above. Over time, several drawbacks to the method were observed, the most important of these violating some of our desired guiding principles:

1. Tolerance thresholds are not specified directly, but rather loosely controlled by adjusting threshold values with no meaning outside the linkage process. This makes them very subjective, where they should ideally be entirely objective.
2. Decisions based on calculated weight alone invariably lead to many pairs getting accepted while being demonstrably worse than some rejected pairs, and other pairs getting rejected while being demonstrably better than some accepted pairs. Much intervention is required to restore logical consistency.
3. Because of the complexity of name variation in use, it is useful to clerically review cases with some levels of partial agreement. Calculated weights alone are of little use in determining which pairs require such review. Much intervention is required to avoid adding unnecessary subjectivity to the process.
4. The most important drawback is the very serious under-use of available information pertinent for making linkage decisions, elaborated upon in section 3.

3. An Alternative to Probabilistic Record Linkage

3.1 Giving Up Complete Generality

Note that the probabilistic record linkage approach described above is completely general, in that it does not require knowledge of the types of data fields being used. This generality comes at a very high cost, as there is much value in knowing something about the fields being used for linkage.

When it comes to record linkage, complete generality is also not really very useful. The vast majority of linkage projects fall into two broad categories: those for records containing information about personal entities, such as names, addresses, and dates of birth, and those for records containing information about business entities. Linkage approaches for these broad categories, and any other smaller categories of interest, may share much in common, but certainly need not be identical.

3.2 Integration of Common Sense Knowledge

All of Elections Canada's record linkage activities involve records of personal information, and decisions about record pairs are based to a large degree on how well this information agrees. Therefore there is much to be gained by incorporating into these decisions an understanding of why fields might disagree for true matches, and why they might happen to agree for true non-matches.

The method described in Section 2 takes what is observed about the pair, the *Outcome*, and uses this to derive a summarizing value, the total weight, which is used to make a decision about the pair. Reasons leading to imperfect agreement of personal information are complex. A single summarizing measure throws away much pertinent information. Instead, it is proposed that, whenever possible, the *Outcome* itself be used to directly decide if a pair should be accepted, rejected, or kept for further investigation, as illustrated in the simplified table below.

Table 3.2-1
Decision Rule Based Directly on Observed *Outcome*

Level of Agreement			Decision
Name/Gender	Date of Birth	Address	
High	High	High	Accept
High	High	Low	Investigate further
High	Low	High	Investigate further
High	Low	Low	Reject
Low	High	High	Investigate further
Low	High	Low	Reject
Low	Low	High	Reject
Low	Low	Low	Reject
Not Seriously Considered due to Insufficient Agreement			Reject

Even with several levels of partial agreement for each field, the large number of possible combinations will include relatively few that would be 'acceptable' under at least some circumstances. The rest, which would comprise the vast majority of possible pairs, could be safely rejected. In other words, the probability that these pairs are true matches given the observed *Outcome* can be safely assumed to be zero.

Of course, in the end, the combinations of *Outcome* deemed unacceptable are subjective. However, this can still be based on objective criteria, and applied in an automated fashion to ensure consistency. These criteria may be partly based on business requirements. For example, the need for voters to provide proof of identification at the polls might limit how much name disagreement is permissible for accepted matches in some linkage applications.

3.3 An Alternative Probabilistic Approach

The biggest drawbacks of the probabilistic record linkage approach described in section 2 are that the thresholds used are subjective, and more importantly, that available information, pertinent to the decisions, is not easily taken into account. The following approach avoids both of these drawbacks.

In record linkage involving personal records, names and dates of birth are the key fields required to identify individuals. For true matches, date of birth is the only field available that cannot disagree for legitimate reasons. Compared to names, date of birth also has a relatively small number of useful levels of partial agreement.

Let k represent the various level of partial date of birth agreement. For example, we could allow three levels, agrees, partially agrees or disagrees. Of course, the different levels would need to be clearly defined. Isolating this from everything else we know about the record pair, we have

$$Outcome = Outcome\ k_{DOB} \cap Outcome_{other} \tag{4}$$

Suppose further that it is possible to observe all pairs with $Outcome_{other}$. That is, we did not throw them away even if they were already rejected. We could then simply count the number of pairs with each date of birth outcome to obtain

$$t_k = number\ of\ pairs\ with\ Outcome\ k_{DOB} \cap Outcome_{other} \tag{5}$$
$$r_k = t_k / \sum_k t_k \tag{6}$$

If we could also somehow know

$$x_k = number\ of\ true\ matches\ with\ Outcome\ k_{DOB} \cap Outcome_{other} \tag{7}$$

then the probability of interest could be calculated by definition.

That is,

$$P_M = \Pr(M|Outcome\ k_{DOB} \cap Outcome_{other}) \equiv {x_k}/{t_k} \tag{8}$$

$$P_U = \Pr(U|Outcome\ k_{DOB} \cap Outcome_{other}) \equiv {(t_k - x_k)}/{t_k} \tag{9}$$

Now suppose that, short of knowing the x_k, we at least know

$$p_k = \Pr(Outcome\ k_{DOB}|M \cap Outcome_{other}) \equiv {x_k}/{\sum_k x_k} \tag{10}$$

$$q_k = \Pr(Outcome\ k_{DOB}|U \cap Outcome_{other}) \equiv {(t_k - x_k)}/{\sum_k (t_k - x_k)} \tag{11}$$

Putting equations (6) and (10) into equation (11), solving for x_k, and putting this into equation (8) yields

$$P_M = \Pr(M|Outcome) = \frac{\frac{q_k - r_k}{r_k}}{\frac{q_k - p_k}{p_k}} = \frac{relative\ distance\ of\ q_k\ from\ r_k}{relative\ distance\ of\ q_k\ from\ p_k} \tag{12}$$

For valid probabilities, it is implied that r_k must always lie between p_k and q_k. Since r_k is observed from all pairs, a mixture of true matches and true non-matches, it should usually be true if p_k and q_k are known reasonably accurately.

3.4 Specifying Required Probabilities

Calculation of probabilities in equation (12) requires the values of p_k and q_k specified in equations (10) and (11), respectively.

For true matches, date of birth would only disagree due to inaccuracies in this field. If the accuracy of dates of birth in the Register were measured, this would provide estimates of the values of p_k required. In fact, such estimates were produced in 2014, based on a small sample of 49,000 records, and are shown in the table below.

For true non-matches, level of agreement on date of birth will be assumed to occur purely by chance, independently of any other considerations. For any given date of birth, the number of individuals on the Register of Electors that have dates of birth that agree fully, agree partially, or disagree can be counted, and expressed as relative frequencies. These give the probability that a new individual has a specific level of date of birth agreement with an individual selected at random from the Register. The values for a typical example are shown in the table below.

Table 3.4-1
Estimates of p_k and q_k for Typical Date of Birth (July 9, 1963)

Level of Agreement	True match (p_k)	True non-match (q_k)
Agrees	98.77%	0.01%
Partially Agrees	1.16%	0.29%
Disagrees	0.07%	99.70%

3.5 Final Decision Rules

A preliminary decision for pairs not yet rejected should be based on a combination of *Outcome* observed and the resulting calculated probability P_M. Some pairs may have values of *Outcome* that warrant clerical review before arriving at this preliminary decision. Others may need to be accepted for operational reasons, despite a probability that would suggest otherwise. Finally, the values of *Outcome* and the preliminary decisions made should be checked for logical consistency, arriving at final decisions for each pair.

3.6 Inclusion of All Pertinent Information

Note that $Outcome_{other}$ has so far been loosely described as everything known about the pair apart from the level of agreement on date of birth. For traditional probabilistic linkage, probabilities need to be estimated for every field included, with the outcomes for these fields assumed to be independent. Removing these two requirements allows other relevant information about the pairs to be incorporated into the decision rules.

For example, level of address agreement can be specified more precisely by including all relevant address fields, without worrying about violations of independence. Limiting the number of levels of agreement for individual fields is not required, and should include the relevant frequencies, to discount chance agreement when warranted. Other fields, perhaps believed to be of lesser importance for linkage, such as status (eg. active or deceased) can also be included without any extra effort.

Most importantly, information about other pairs can also now be easily incorporated. In trying to decide if a given pair should be accepted, it would certainly be relevant to know that the records involved were also involved in other 'better' pairs. All relevant facts such as these can simply be added to the definition of $Outcome_{other}$.

3.7 Error-Rate Estimation

Once a pair has been accepted or rejected, only one of the two classification errors is possible. A simple way to estimate the number of classification errors made is by simply adding up the probability of the relevant probabilities over all pairs.

That is,

$$Number\ of\ false+ \cong \sum_{accepted} P_U = \sum_{accepted} (1 - P_M) \tag{13}$$

$$Number\ of\ false- \cong \sum_{rejected} P_M \tag{14}$$

It is hoped that the relevant probabilities can be estimated well enough with this method to produce believable estimates of linkage classification error rates, for a wide variety of record linkage projects. Once this error-rate

estimation method has been made operational, it is hoped that the results can be compared and contrasted with competing error-rate estimation methods.

References

Fellegi, I. P., and A. B. Sunter (1969), "A Theory of Record Linkage", *Journal of the American Statistical Association*, 64, pp. 1183-1210.