National Records of Scotland

Beyond 2011

The Ord Wood Project: A method of calculating match probabilities from record linkage output

Published on 9 January 2014

Contents

1. Background

This paper presents a method for calculating match probabilities from record linkage output. The methods described in this paper were used to link the Census 2011 to the NHS Central Register and are presented in the paper 'Matching the Census 2011 to the NHS Central Register using the Ord Wood method'.

A.1: The method of calculating match probabilities

Let M and $\overline{M}$ respectively denote that a given pair is a match and is a non-match. Since these events are exclusive and exhaustive, $P(M)=1-P(\overline{M})$ and the prior probability of the linkage data *D* can be expressed as a weighted average of its conditional probabilities given these two events. Then, dividing through by $P(M)P(D\mid\overline{M})$, Bayes' theorem yields

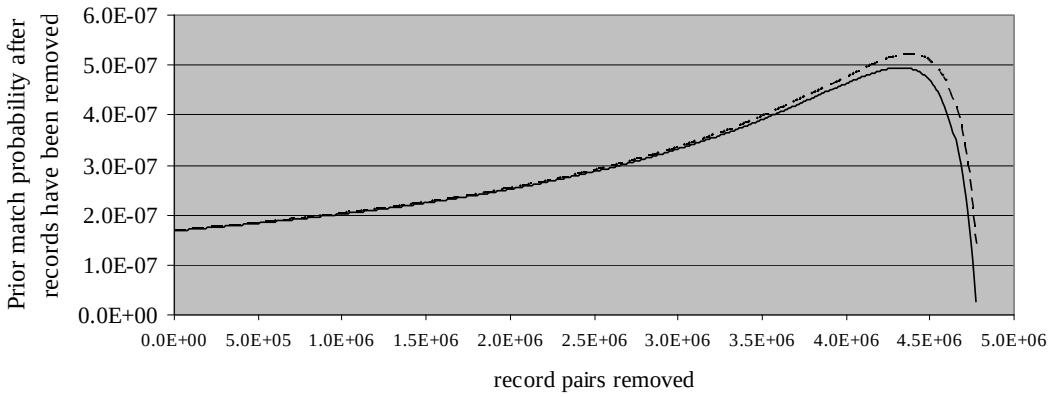
$$P(M\mid D)=\frac{P(D\mid M)P(M)}{P(D)}=\frac{P(D\mid M)P(M)}{P(D\mid M)P(M)+P(D\mid\overline{M})\{1-P(M)\}}=\frac{\frac{P(D\mid M)}{P(D\mid\overline{M})}}{\frac{P(D\mid M)}{P(D\mid\overline{M})}+\frac{1-P(M)}{P(M)}}.$$

To use this equation we need values for $P(M)$ (henceforth denoted p_0) and for the likelihood ratio (usually denoted by the symbol λ with a subscript *i* to denote that it is calculated separately for each record pair). $P(M)$ is the prior probability that a record pair will be a match and is given by N_M, the number of record pairs which are matches, divided by the number of ways in which a record pair can be chosen. This denominator is the product of the two file sizes $N_A N_B$, both of which are known. However, while the range of values within which N_M lies will probably be known, its exact value will not and an estimate must be used in its place. We assume one-to-one linkage and so N_M should not be greater than the number of records in the smaller file (which we take to be file B). The linkage procedure described below is multistage in which pairs identified as links at one stage are 'banked' and removed from the files to go to the next stage. This is done in the interests of computational efficiency but it means that the equation for p_0 must be amended. If *k* record pairs have been removed (and assuming that all pairs were true links and that no false links have been removed), the equation becomes:

$$p_0=\frac{N_M-k}{(N_A-k)(N_B-k)}. \qquad \text{[A1]}$$

Regarded as a function of *k*, and using values for the constants taken from the paper 'Matching the Census 2011 to the NHS Central Register using the Ord Wood method', this is plotted in Figure 1. It can be seen that the prior match probability actually increases as true links are removed until almost all have been removed. If the assumption that all pairs removed were true links is false, and a small proportion of them were false links, then the probability increases even faster as shown by the dotted line in Figure 1, though the effect is small unless significant numbers of false links are removed.

Fig 1: Prior match probability as a function of the number of record pairs removed



The treatment of the likelihood ratio λ_i depends on how it is related to w_i, the weight value calculated by the software and used to sort the pairs into a 'league table'. The user manual should specify how w_i is related to λ_i, though not all user manuals do so exactly. For example, for both the Rec Link and Link Plus packages, w_i is the logarithm of the likelihood ratio. The base used for the logarithms is not given, though it is needed for present purposes. The base used by a package can be calculated from the relationship which exists between the *m*- and *u*-parameters (respectively the probabilities of observing agreement given that the pair is a match and that it is a non-match) and the agreement and disagreement weights[1] for each field. All these are either calculated by the software, in which case they should feature somewhere in the output, or determined by the user. As an example of how the log base *b* can be calculated we note that the agreement weight is given by $A = \log_b(m/u)$ from which $b = (m/u)^{1/A}$. Applying this calculation to Rec Link always yields the exponential constant *e* for the log base so the logarithms are natural and λ_i is given by $\exp(w_i)$. Link Plus in contrast uses logarithms to the base 3 so λ_i is given by three raised to the power w_i. The Relais package does not use logarithms at all so λ_i is simply equal to w_i. In any event, the weight output by the software must be converted into a likelihood ratio and the probability p_i that the *i*th record pair is a match is given by

$$p_i = \frac{\lambda_i}{\lambda_i - 1 + 1/p_0}. \qquad [A2]$$

This equation assumes that the model is exactly specified and that the parameter estimates are exact, neither of which is likely to be the case in practice. If the weights are used only to sort the record pairs in descending value of p_i then this equation will suffice provided that the output values of w_i are monotonically related to the true values (i.e. those which would be output if the model were exactly specified). In practice, this is usually a reasonable assumption. For present purposes however it is no longer sufficient to have a value which is monotonic (or nearly so) in p_i. Now it is necessary that the w_i values are accurate and so the consequences of imperfect specification must be considered before the method can be used. This is done in sections A.2 to A.4 below. In the remainder of this paper, it will be assumed that the

Footnote

1) As defined in equation [A3] below.

w_i values are the natural logarithms of the likelihood ratios and that the linkage fields are assumed to be stochastically independent of each other conditional on match status. In this case the w_i value for record pair *i* is the sum of the field level agreement weights w_{if} for that pair over the fields *f* being used for linkage or, formally, $w_i = \sum_f w_{if}$. Section A.4 considers the effect on the method of violations of the conditional independence assumption.

A.2: Errors in parameter estimates

Inaccurate parameter estimates are one obvious source of error in the method. In the simple Fellegi-Sunter model, field level weights are calculated as

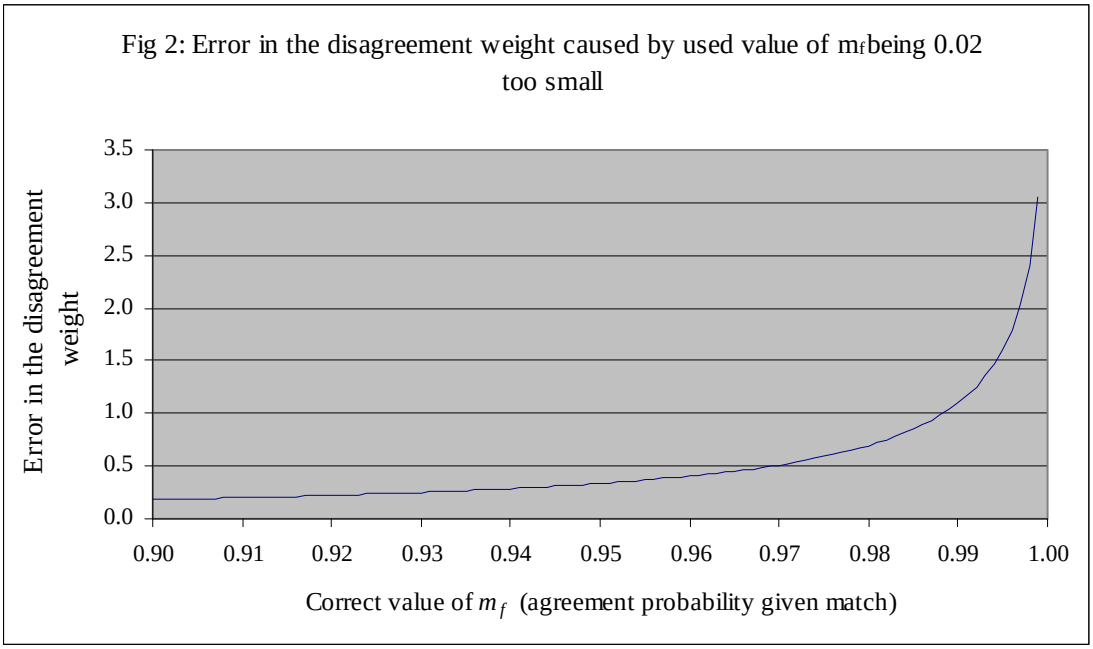
$$w_{if} = \ln\left(m_f / u_f\right) \quad \text{in the case of agreement on field } f; \quad [A3a]$$

$$w_{if} = 0 \quad \text{in the case of a missing value in either record; and} \quad [A3b]$$

$$w_{if} = \ln\left\{\left(1 - m_f\right)/\left(1 - u_f\right)\right\} \quad \text{in the case of disagreement on field } f. \quad [A3c]$$

where m_f and u_f are respectively the probabilities of observing agreement on field *f* given a match and given a non-match. The use of frequency-specific weights complicates this somewhat by adding a second term to the right hand side of [A3a] but it does not change the position materially. Both *m* values and *u* values must be estimated. Accurate values for the latter can be derived as follows. The probability p_0 that a record pair selected at random from the original files will be a match will in practice be very small. Its upper bound is the reciprocal of the number of records in the larger file. It is possible therefore to sample pairs at random and assume that they are all non-matches since any matches which are included will have negligible impact on the estimates. This technique is used by Link Plus and gives *u* estimates which agree very well with the expected values derived from probability theory. We can assume then that the values u_f in [A3a] and [A3c] are not problematic.

It is more difficult to estimate m_f values accurately since this would require a set of record pairs which is representative of all matches and it is not easy to see how this could be compiled. In practice, m_f values are estimated using the fact that $1 - m_f$ is the probability that agreement will *not* be observed on field *f* even if the records refer to the same person. This in turn equals the error rate, defined in broad terms, for that field and this will often be known reasonably accurately on the basis of experience accrued with the two data sets being linked. We also note that, unless the data quality of one of the input files is unusually poor, m_f will in practice lie in the range from 0.90 to 0.99. Thus an error in its value will have little effect on the agreement weight as the logarithm of m_f will stay close to zero. Similarly, since u_f is very small for most fields (gender being the obvious exception), an error in its value will have little effect on the disagreement weight. Therefore the agreement weight is sensitive to changes in u_f and the disagreement weight to changes in m_f . Since the latter parameters are the more difficult to estimate, we can concentrate on the effect of error in m_f on the disagreement weight.



We can do this as follows. Let us assume that the true value m_f is 0.99 but the value used was 0.97 (0.02 too small). The effect on the disagreement weight of an error of 0.02 in m_f (one-fifth of its likely range from 0.90 to 0.99) is plotted in Figure 2. It can be seen that the effect on the disagreement weight of an error of 0.02 rises above 0.50 if m_f is above 0.97 and rises above 1.00 if it is above 0.99. For comparison, it is shown in section A4 below that the effect of a large violation of the conditional independence assumption (first name and gender given a non-match) is around 0.67. Thus significant effects due to error in parameter estimates will only occur if for example an error rate of one in 30 is mistaken for an error rate of one in a hundred. General experience with data sets should enable errors of this size to be avoided.

A final point on parameter error is that if m_f is incorrect, the direction of the error will depend on whether the value used is too large or too small. As the errors are additive then if they occur in both directions, they will cancel each other to some extent. The extent of this cannot however be gauged without knowing the true values, knowledge which would of course defeat the purpose of the calculation.

A.3: The record pair-level weight w_i

Consider equation [A2] above. If its value is very close to one or very close to zero then the error in p_i is likely to be small since there is less scope for variation very close to the boundary of the outcome space. This will happen if λ_i is several orders of magnitude larger or smaller than $1/p_0$, which will be the case respectively high up, and low down, in the output table. The error occurs in the more problematic centre section where neither of the two values is negligible with respect to the other. In practice, $1/p_0$ will be larger (and perhaps significantly larger) than the number of records in file A. In the application reported below, its initial value is about six million. If we are concerned about w_i values which differ from $1/p_0$ by less than a factor *f* or more than a factor $1/f$, then the range of values which contribute the error is given by:

$$\ln\left(\frac{f}{p_0}\right) < w_i < \ln\left(\frac{1}{fp_0}\right).$$

Let us assume for example that the range of w_i values which are of concern are those where w_i is less than 100 times greater than, or more than 100 times smaller than $1/p_0$. If the value of $1/p_0$ is 500,000, the range is from 8.52 to 17.73, a range which fits well with experience of conducting clerical reviews using the natural log likelihood ratios of Rec Link (but with smaller files than those used in the application reported below).

It is immediately clear that for w_i values in this range, it is likely (though, as shown below, not quite certain) that the terms in the series of which it is the sum have a higher number of field level agreements than disagreements. When using the Rec Link package, a value of 17.85 typically results from agreement on first name, postcode, day of birth, month of birth, year of birth and gender but disagreement on middle initial and last name. Hence there are agreements on six fields and disagreements on two. A value of 9.68 typically results from agreement on first name, last name, year of birth and postcode but disagreement on middle initial, day of birth, month of birth and gender. Hence there are agreements on four fields and disagreements on four but the sum is still positive since the fields on which there is disagreement have lower power[2] than those on which there is agreement.

Agreements are therefore likely to predominate in the record pairs but the effect of parameter error on disagreement weights cannot be ignored. As an illustration let us assume that a weight w_i of 17.77 is 0.67 greater than it should be due to violation of the conditional independence assumption (as illustrated in section 2.4 below) and 1.23 greater than it should be due to incorrect m_f estimates, a total of 1.90. On the basis of the evidence presented below, this is likely to be one of the larger errors which can occur in w_i. The correct value of the log likelihood ratio is therefore 15.87 and, if p_0 is one in 500,000, the calculated match probability will be 0.991 as opposed to a correct value of 0.940. This is an error of 0.051. If the reported weight is 7.98 (and hence the true value is 6.08) then the calculated match probability is 0.006 as opposed to a true value of 0.001, an error of 0.005. These errors are not negligible but they are small.

A.4: Violation of the conditional independence assumption

An important way in which misspecification can occur is violation of the conditional independence assumption. This assumption is computationally convenient in that it allows the weight w_i for a record pair to be expressed as the sum of the w_{if} for that pair. This greatly reduces the number of parameters to be estimated and improves the stability and robustness of the method. However in practice there are violations, the most extreme example of which is first name and gender.

Assuming that a record pair is not a match, the probability of agreement on gender is close to 0.50 since there are four equally likely combinations of which two are agreements and two disagreements. However now add the information that there has been an agreement on first name (both are David or both are Margaret). The probability of a match on gender now increases to a value close to unity. The

Footnote

2) The power of a linkage field can be measured in different ways. A common one is the difference between the agreement and disagreement weights i.e. the difference between [A3a] and [A3c].

(incorrect) agreement weight which will be used is $\ln(m_f / 0.5) = \ln(2m_f)$ whereas the correct weight is closer to $\ln(m_f)$, the difference being just less than $\ln(2)$ which is 0.693. To be correct, the weight should be based only on information which is new and has not already been introduced by agreement or disagreement on previous fields. If the fields were independent, all the information would be new and the weight would be correct. If they are not, the weight used will be too large because information is being double counted. To allow for this, the value of w_i should be reduced by say 0.67 when agreement has been observed on both first name and gender. The reduction in other cases of dependence is not so easy to quantify but should not be so great. The first name / gender combination is the most obvious departure from independence, though there are others.

Assuming that a record pair is a match, the probability of disagreement on last name is greater if the gender is female than if it is male since marriage and divorce are more likely to change female last names than male last names. Also, disagreement on either first name or last name increases the probability of disagreement on the other since both can be caused by a complete misallocation of name parts to name fields (e.g. the entry 'Mrs' for first name increases the probability that the first name will be entered in the second name field and so lead to disagreement on both fields). The effects of violations of this type are of course more difficult to quantify.

The conclusion drawn from this is that few of the errors would be large enough to jeopardise the procedure as a whole; that errors resulting from incorrect parameters values are as likely to be in one direction as in the other; that correcting errors resulting from conditional dependence requires more detail about the way in which *m*-values are calculated than is usually supplied by user manuals; and that hence the interests of minimising the expected error in the w_i values would be served by making no changes to the output values of w_i. Accordingly, in the application reported above, the likelihood ratios calculated by the Link Plus package were not changed in the light of the points raised in this section and the last two.

A.5: Using the link probabilities

Having derived the value for p_i, we require a method for using it to estimate how many record pairs would be accepted as matches by clerical reviewers. If p_i is the probability of a match then clearly $1 - p_i$ is the probability of a false link or false positive. Then a set of records pairs can be considered to be a sequence of heterogeneous Bernoulli trials. These are independent as they feature different record pairs (and hence different records as we assume that no record can feature in more than one pair). Then the number of accepted links follows a heterogeneous binomial distribution of which the mean and variance are $\sum_i p_i$ and $\sum_i p_i(1 - p_i)$ respectively.

An alternative use for the p_i values is to accept all pairs with a value in excess of some criterion. A criterion of 0.5 would be equivalent to accepting all pairs which were more likely to be accepted than rejected at clerical review (the 'balance of probabilities' criterion in civil law). A criterion of 0.99 or thereabouts would be equivalent to accepting all pairs which would confidently be accepted at clerical review (the 'beyond reasonable doubt' criterion in criminal law). The criterion would be driven by the relative costs attached to accepting a false link and rejecting a match.

At this point it is worth considering the distinction between $\sum_i p_i$ and N_M. The latter is the number of matches assumed to exist for the purposes of calculating $P(M)$. The former is the estimate of the number of these matches which have been found by a clerical review covering the records pairs in all stages of the linkage and is therefore a subset of N_M. It would therefore be inconsistent if $\sum_i p_i$ were to be greater than N_M. It ought to be less than N_M (how much less depending on data quality) and this should be checked when applying the method.

A.6: Implementation details

This section gives some technical details to link the general description of the linkage exercise given in the paper 'Matching the Census 2011 to the NHS Central Register using the Ord Wood method' with the technical specification of the method in sections A1 to A5.

Table 1: Parameter values for the calculation of λ_i

Field	m_f	u_f	Agree	Disagree
first name	0.96	0.00410	5.46	-3.21
middle initial	0.95	0.07887	2.49	-2.91
last name	0.97	0.00099	6.89	-3.51
birth year	0.99	0.01226	4.39	-4.59
birth month	0.98	0.08345	2.46	-3.82
birth day	0.98	0.03249	3.41	-3.88
gender	0.99	0.49922	0.68	-3.91
postcode	0.95	0.00003	10.36	-3.00

In step 1a, the value of λ_i was determined by using the agreement weights in table 1. The *m*-values were those put into the Link Plus run in step 2 and were based on general experience of error rates for these data sets. The *u*-values were those calculated by Link Plus in stage 2 and applied retrospectively.

The value of $\log \lambda_i$ for perfect agreement was 35.47. The value for p_0 was calculated from equation [A1] using the original values for N_A and N_B, putting N_M equal to $0.96N_B$ and setting *k* equal to zero.

For steps 1b and 1c, again the values in table 1 were used giving $\log \lambda_i$ values of 25.11 and 22.11 respectively, corresponding to match probabilities of 100.00% (to two decimal places of the percentages) and 99.94% respectively. The numbers of record pairs identified and removed at step 1a for step 1b, and at steps 1a and 1b for step 1c, were used as the value of *k* in [A1] to calculate p_0 for each step. Again, $1 - p_i$ was multiplied by the number of pairs in each step to give an estimate of the number of false positives accepted.

In step 2, the blocking criterion used for the Link Plus software was first name initial OR last name initial. The linkage fields and associated m_f and u_f values were as reported in table 1 except that middle name initial was not used as a linkage field since there are many missing values. The m_f values were inserted manually while the u_f values were calculated by Link Plus using the random pair selection technique described in section A2. After manual examination of the output, a $\log \lambda_i$ value of

18.0 was selected as the cut-off (corresponding to a match probability of 99.5%) and pairs were accepted if they had a value greater than this. The procedure for calculating p_0, λ_i and hence p_i as described above was applied again in order to estimate the number of false links at this step.

In step 3c, the details of the checks made were as follows. For gender, day of birth, month of birth and year of birth, the same method was used as in stage 2.

For names, the middle initial was used as per Table 1. Also, the first name was checked against the middle and last names, as well as first name, in the other file. For first names, the agreement weight was given if (i) the two first names were within an edit distance of one or (ii) the longer first name started or ended with the shorter one or (iii) either the middle or the last name in the other file started with the same characters as the first name. Last names were checked against the first and middle names, as well as last name, in the other file. The agreement weight was given if (i) the two last names were within an edit distance of one or (ii) the longer last name started or ended with the shorter one or (iii) either the first or the middle name in the other file ended with the same characters as the last name. An example of this would be if 'JohnSmith' were being compared with 'John' or 'Smith' in which case it would be assumed that the names did in fact agree but had been incorrectly parsed into the first and last name fields.

The postcode distance measure given below was calculated and a weight was used equal to the log likelihood ratio corresponding to that distance which was taken from the record pairs in stages 1a, 1c and 2.

A final point is that the total number of links either found in stages 1, 2 or 3 or estimated in stage 4 is 4,766,501 which is 95.31% of the number of records in the census file. As noted in section A5, we should check that this is less than the figure for N_M used to calculate p_0 in equation [A1]. As this was 96%, we have verified that N_M is greater than $\sum_i p_i$, as required.