

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

Model-based Threshold Selection for Agricultural Linkages

by Christian Arsenault

Release date: September 8, 2025



Statistics
Canada

Statistique
Canada

Canada

Model-based Threshold Selection for Agricultural Linkages

Christian Arsenault¹

Abstract

Optimal threshold selection is a critical challenge in probabilistic linkage, with significant implications for the accuracy and reliability of linked datasets. This paper analyzes the performance of the neighbour model, a recently proposed error model which models linkage errors by the number of links from each record. Three threshold selection algorithms utilizing the neighbour model were assessed, highlighting the strengths and limitations of each. Their performance was assessed through simulation studies, which demonstrated that methods using the neighbour model achieved lower relative bias compared to two established methods for threshold selection. Additionally, the practical utility was validated through goodness-of-fit tests conducted on four agricultural datasets, showing the potential of the model for use in real-world applications.

Key Words: Statistical matching; Probabilistic modeling; Linkage errors.

1. Introduction

Record linkage is a crucial component to the infrastructure of the Agriculture program at Statistics Canada. For the Census of Agriculture and surveys, the use of administrative data has become increasingly important. Many linkage tasks are completed using G-Link, Statistics Canada's generalized system for probabilistic linkage, which is implemented based on the methodology by Fellegi and Sunter (1969). One of the main challenges related to linkage projects is assessing the linkage quality and setting the similarity threshold to accept links. Without ground truth information to identify true matches, thresholds are often set manually, and clerical reviews are used to evaluate linkage accuracy. With clerical review being a costly and inefficient endeavor, there is a clear need for research in this area. This paper analyzes the approach outlined by Dasylyva and Goussanou (2022) for setting linkage thresholds without the need for training data or manual review, from both a theoretical and practical perspective. This method will be tested in a simulation study where the true match status is known for each pair, and it will also be tested on four agricultural administrative linkages to assess its viability in practice.

2. Theory

2.1 Primer on Record Linkage

The goal of record linkage is to find which records refer to the same entity in two distinct files. This task is trivial when we have a primary key for linkage such as a Business Number or Social Insurance Number. However, difficulties arise when we only have variables which may not be unique to a given entity or may be recorded with errors, such as business names or dates of birth. Suppose we have two complete registers A and B which refer to the same population with K linkage variables. If we consider all the possible pairs in the cartesian product of $A \times B$, then each pair is referred to as matched if its records refer to the same entity, and unmatched otherwise. When sampling a pair, we denote the event that a pair is matched by M , and the event that a pair is unmatched by U . Fellegi and Sunter (1969) described the probabilistic method of record linkage, where two records are linked if their similarity exceeds a

¹ Christian Arsenault, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A0T6, christian.arsenault@statcan.gc.ca

threshold. This similarity is based on the vector of agreements of a pair, $(\gamma^{(1)}, \dots, \gamma^{(K)})$, which is modeled with a log-linear mixture,

$$P(\gamma) = P(M)m(\gamma) + P(U)u(\gamma) \quad (2.1)$$

where $m(\gamma) = P(\gamma|M)$ and $u(\gamma) = P(\gamma|U)$ are called m - and u - conditional probabilities respectively. Under the conditional independence assumption, each of these conditional probabilities is the product of the corresponding marginal probabilities of $\gamma^{(1)}, \dots, \gamma^{(K)}$, and the linkage weight is the sum of weights associated with the different variables. The linkage weight, $w(\gamma) = \log(m(\gamma)/u(\gamma))$ is then used to determine the optimal linkage decision. Fellegi and Sunter proposed that the linkage rule be determined by an upper threshold θ and lower threshold θ' with respective probabilities of linking τ and τ' , the calculation of which has been a subject of ongoing research, and the main topic of this report. However, we will focus on the problem of assigning a single threshold θ , above which pairs are considered to be matched, as this is what is typically done in practice.

2.2 Previous work

Threshold calculations and error estimation in probabilistic record linkage have been discussed for some time. There have been many contributions in this space, each with their limitations for practical use. Two of these methods will be discussed in this report, the method by Belin and Rubin (1995), and that by Sariyar, Borg, and Pommerening (2011), which is based on extreme value theory. The method by Belin and Rubin sets the threshold by modelling the linkage weights as a mixture of two distributions, where the first distribution represents the false positives, and the second distribution represents the true positives. Although this method avoids making the conditional independence assumption, training data is required for a calibration process. In practice, this training data may be based on clerical reviews that are costly. The method by Sariyar et al. (2011) is designed to address some of the limitations of the above method, including the need for training data. This method requires expert input to determine the interval of the mean excess plot where it is linear, which can be a difficult task depending on the weight distribution.

2.3 The Neighbour model

As described in Dasylyva and Goussanou (2022), a method was proposed to calculate the linkage threshold and resulting linkage error without the need for user input or training data. It estimates the linkage accuracy by modeling the number of links of a record with a model called the neighbour model. Some notation is required to describe this model. First, we assume a population of N entities with two complete files for the population, each free of duplication. The model also applies when the files are not censuses, but it is convenient to describe the model in this simpler case. Each population then has N records, where we denote the values from the first file as v_i where i corresponds to entity i , and takes its values from the finite set V . We then consider the second file to be an unknown random permutation of the i s, denoted $\pi(\cdot)$, where the record values for the same entity in the two files are independent and identically distributed. The *neighbourhood* of record v_i , denoted $B(v_i)$, refers to the subset of V such that v_i and v'_j are linked. Following these definitions, we can define various linkage accuracy measures. A false negative occurs when $v'_{\pi(i)} \notin B(v_i)$, and a false positive occurs when $v'_{\pi(i')} \in B(v_i)$ for distinct i and i' . There are many accuracy measures that can be defined based on these, but we will focus on the precision and recall. The precision is the proportion of linked records that are true matches, and the recall is the proportion of true matches that were linked. They can be described in terms of the false negative rate (FNR) and false positive (FPR) rate of the linkage. More precisely,

$$\begin{aligned} \text{FNR} &= \frac{1}{N} \sum_{i=1}^N I(v'_{\pi(i)} \notin B(v_i)) \\ \text{FPR} &= \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{i' \neq i}^N I(v'_{\pi(i')} \in B(v_i)) \\ \text{Precision} &= \frac{1 - \text{FNR}}{1 - \text{FNR} + (N-1)\text{FPR}} \end{aligned} \quad (2.2)$$

Now that the neighbour concept is established, we can show how they can be used to estimate linkage errors and set the threshold. Having made the assumptions that we have two complete files that are duplicate free, denote the number

of neighbours from a record v_i by n_i , which is used to model the linkage errors. The following two equations show the relationship between n_i and the false negatives (FN) and false positives (FP) involving v_i :

$$\text{FN} = \begin{cases} 1 & \text{if } n_i = 0, \\ 0 \text{ or } 1 & \text{if } 1 \leq n_i \leq N - 1, \\ 0 & \text{if } n_i = N \end{cases}, \quad \text{FP} = \begin{cases} 0 & \text{if } n_i = 0, \\ n_i - 1 \text{ or } n_i & \text{if } 1 \leq n_i \leq N - 1, \\ N - 1 & \text{if } n_i = N \end{cases} \quad (2.3)$$

With this information, we can model the linkage errors. We can view n_i as the sum of the true positives and false positives. The former is based on whether we link v_i to the matched record in the other file. It is simply a Bernoulli trial with probability $p_N(v_i)$ conditional on v_i . As for the false positives, they follow a Binomial distribution based on $N - 1$ trials (one for each unmatched record in the other file) and probability of success $\lambda_N(v_i)$ conditional on v_i . If N is large, the second distribution converges in distribution to a Poisson with parameter λ . Therefore, the number of neighbours of a record can be modelled by the following finite mixture with G classes:

$$n_i \sim \sum_{g=1}^G \alpha_g \left(\text{Bernoulli}(p_g) * \text{Poisson}(\lambda_g) \right) \quad (2.4)$$

The parameters of the mixture can be estimated by maximum likelihood estimation, and an estimator for the precision of the linkage follows from the law of large numbers.

$$\text{Precision} \rightarrow \left(1 + \frac{\sum_{g=1}^G \alpha_g \lambda_g}{\sum_{g=1}^G \alpha_g p_g} \right)^{-1} \text{ as } N \rightarrow \infty \quad (2.5)$$

Having an estimator for the precision of the linkage not only gives us an indication of the quality, but also provides the inputs to allow us to set the threshold.

2.4 Algorithms to Select the Threshold Using the Neighbour Model

Now that we have an error model, we can make use of it to set the linkage threshold. The threshold is set by providing a target precision if the goal is to maximize the recall, or by providing a target recall if the goal is to maximize the precision. For simplicity, we will assume that a target precision has been set. Three algorithms were considered for setting the threshold which are outlined in the following three sections.

2.4.1 Exhaustive Search

This method involves evaluating each candidate pair weight among the observed pair weights in the linkage to determine if they meet the provided target precision. The algorithm is as follows.

1. Sort the pairs by descending weight.
2. For each unique pair weight, calculate the estimated precision.
 - If the estimated precision is less than or equal to the target precision, return the current pair weight.

This method is simple and easy to implement, but it can be slow depending on the number of unique pair weights.

2.4.2 Binary Search

Binary search has a logarithmic run time because the search space is halved with each iteration, but has the inconvenience of requiring a tolerance parameter. The algorithm is as follows:

1. Initialize an optimality tolerance for which the target precision can differ from the estimated precision.
2. Set the median pair weight for the current search space as the threshold and calculate the target precision.
 - If the estimated precision is within the optimality tolerance, set this as the final threshold.
 - If the estimated precision is less than the target, recurse on the upper values of pair weight.
 - Otherwise, recurse on the lower values of pair weight.

2.4.3 Recursive Partitioning

The third algorithm involves using recursive partitioning to build a decision tree using the estimated linkage accuracy from the neighbour model to estimate the heterogeneity of each node. If we consider our vector of comparison outcomes, $\gamma = (\gamma^{(1)}, \dots, \gamma^{(K)})$, then we can build a decision tree which aims to create homogeneous nodes in terms of the conditional match probability (i.e., the conditional probability that a pair is matched given the comparison outcomes) by splitting on the possible linkage variables $\gamma^{(k)}$. If splitting on the linkage variable $\gamma^{(k)}$ results in two nodes R_1, R_2 corresponding to the agreement and disagreement on the linkage variable, then the form of the cost function for split t is:

$$\hat{J} = \left(\sum_i I(X_i \in R_1(t)) \right) \text{Var}(Y|X \in R_1(t)) + \left(\sum_i I(X_i \in R_2(t)) \right) \text{Var}(Y|X \in R_2(t)) \quad (2.6)$$

The two variances in this weighted sum can be written in terms of the precision to make the connection back to the neighbour model:

$$\text{Var}(Y|X \in R(t)) = E[Y|X \in R(t)](1 - E[Y|X \in R(t)]) \quad (2.7)$$

$$E[Y|X \in R] = \text{Precision} \rightarrow \left(1 + \frac{\sum_{g=1}^G \alpha_g \lambda_g}{\sum_{g=1}^G \alpha_g p_g} \right)^{-1} \text{ as } N \rightarrow \infty \quad (2.8)$$

We can now minimize the cost function using an estimated precision of the neighbour model to determine the optimal splits. The algorithm can be summarized as follows:

1. Recursively partition the data using the estimated error from the neighbour model as the splitting criterion.
2. Sort the terminal nodes of the resulting decision tree by increasing linkage weights. For each terminal node:
 - If the current cumulative precision of the linkage is greater than the target, classify the pairs as matched and continue to the next node.
 - If the current cumulative precision of the linkage is equal to or less than the target and the precision in the previous node is greater than the target, classify records in this node as matched with a probability such that the target precision is met. Set the threshold of the linkage as the greatest pair weight in this node.
 - Otherwise, classify the pairs as unmatched in the node.

This method provides information about the pair-level quality of the linkage since each pair will be assigned to a node with an estimated precision. Also, the linkage variables used in splitting provide insight into which variables were most effective in splitting the pairs into homogeneous subgroups, essentially showing which variables were important in the linkage. However, the algorithm is slow when there are many linkage variables and many unique pair weights, making it difficult to apply in practice.

2.5 Goodness of Fit

The last piece of theory that we will use in the results is the goodness-of-fit test. This method was outlined by Dasylyva and Goussanou (2024), and it is a modified chi-squared test to test the validity of the neighbour model. Since the neighbour model indicates that the neighbours from each record should follow a mixture where each component is the convolution of a Bernoulli distribution with a Poisson distribution, this means that we have expected frequencies of neighbours that we should observe for a given linkage. In this case, it is tempting to use a chi-squared test, but such a test requires independence, which is not met due to correlated observations. A solution proposed by Dasylyva and Goussanou (2024) was to use a subset of m_N and n_i values, such that the observations are approximately independent, where the choice $m_N \propto N^{2/5}$ is optimal. Then, for a given number of classes, we can test if the model applies to a given dataset. The test is based on how many records have no link, one link, two links, ..., $d - 1$ links, and d or more links, and our resulting test statistic is approximately chi-square distributed with $d - 2G$ degrees of freedom, where G denotes the number of classes of the finite mixture. In this goodness-of-fit test, our null hypothesis is then rejected

if the test statistic is greater than the critical level $\chi_{d-2G,1-\alpha}^2$. In our tests, however, we will use a parametric bootstrap procedure to estimate the critical level, which is robust to small values of m_N .

3. Results

This section presents the outcomes of using the neighbour model in both simulated and real-world contexts. Section 3.1 will detail a simulation study done to compare the neighbour, EVT and BR methods for threshold selection. The estimated precision of each method will be compared to the ground truth of the simulations to determine the efficacy of each method. In section 3.2, the modified chi-squared test will be applied to a variety of agricultural datasets to gauge the applicability of the neighbour model in practice.

3.1 Simulations

For our simulation study, 100 populations of $N = 10000$ units were generated, where each unit had $K = 10$ categorical features which were uniformly drawn from $\{1,2,3\}$ as well as a blocking variable $B \in \{1, \dots, 5000\}$. From each simulated population, two censuses were generated with typos recorded for each categorical linkage variable. Typos were inserted with probability proportional to k , specifically, $e_k = 0.05 + \frac{k}{10}(0.2 - 0.05)$, such that each linkage variable was recorded without error with probability $1 - e_k$. If there was a typo, a value was chosen randomly from the other two possible values for that variable. To simulate the effect of blocking, potential pairs were created by only comparing records which met the blocking criteria. For those that did meet the blocking criteria, we compared the linkage variables by creating outcome vector γ_k , and a linkage weight was assigned of the form $w = 10\log_2 B + w_1 + \dots + w_K$, where

$$w_k = \gamma_k \log_2 \left(\frac{m_k}{u_k} \right) + (1 - \gamma_k) 10 \log_2 \left(\frac{1 - m_k}{1 - u_k} \right), \quad (3.1)$$

$$m_k = (1 - \varepsilon_k)^2 + \varepsilon_k^2 / (d - 1), \quad (3.2)$$

$$u_k = 1/d \quad (3.3)$$

Note that since the results are simulated, w is the actual probabilistic linkage weight, which can be calculated directly from the m - and u - probabilities. For each of the simulations, we calculated the linkage threshold according to two of the candidate neighbour methods, as well as the BR and EVT methods. The BR method requires training data for a calibration process, and the EVT method requires selection of a critical parameter, both of which were given by randomly sampling 10% of the pairs, which is the ideal case for both methods if the clerical reviews are error free as is the case here. The other two methods are based on the neighbour model, including exhaustive search (neighbour-1 in the figures) and binary search (neighbour-2 in the figures). Recursive partitioning was omitted for the simulation study due to long run times. The following two figures show how each algorithm performed in terms of the relative difference between the achieved precision based on the selected weight threshold and target precision for a target precision of 0.9 and 0.8 respectively

Figure 3.1-1
Relative Difference Between Achieved and Target Precision - Target Precision = 0.9

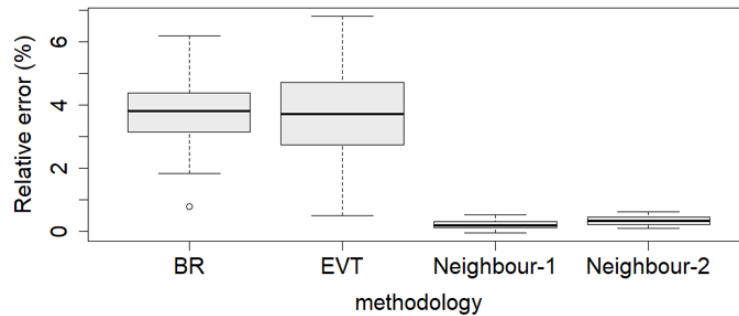
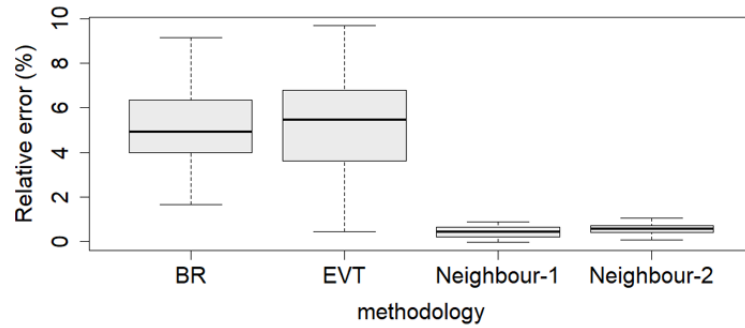


Figure 3.1-2**Relative Difference Between Achieved and Target Precision - Target Precision = 0.8**

The results for a target precision of 0.9 and 0.8 were similar, both having a relative bias of under 10% for all methods in all simulations. However, compared to the BR and EVT methodologies, both methods using the neighbour model saw a smaller bias and variance, meaning they offered better performance in the simulations without requiring human input or clerical reviews. While exhaustive search and binary search had a similar performance, binary search is preferred in practice due to its faster run-time, making it a more practical choice when there are many unique pair weights in the linkage problem.

3.2 Real Data

In the simulations, the binary search method is promising because it is fast and sets the threshold accurately in terms of the difference between the target precision and the achieved precision. Yet, it is equally important that the neighbour model fits the data on which it is to be used in practice, because there is no clerical review. To check this requirement, four agricultural datasets were identified for the purpose of testing the neighbour model for threshold selection.

1. Crop Insurance: List of farms participating in a federal crop insurance program.
2. AgriStability: List of farms participating in federal income stabilization program for farmers.
3. Maple Producers: List of maple producers in Quebec
4. Organic Producers: List of organic farms in Quebec

Each list was linked to Statistics Canada's Business Register which generated pairs with their probabilistic weights to calculate a threshold. Additionally, we were able to run the modified chi-squared goodness-of-fit tests described in section 2.5, the results of which are shown in the following table.

Table 3.2-1**Goodness-of-fit Test Results for Agriculture Datasets**

Dataset	Test Statistic	Critical Level	Reject H_0
Maple Producers	0.558	3.194	False
Federal AgriStability	2.757	3.337	False
Maple Producers	1.659	3.558	False
Organic Producers	0.073	5.425	False

Each dataset was tested for one class and subsets of $m_N = 50$ records, which satisfied the recommended size of $N^{2/5}$ for each dataset while still providing ample data to reliably estimate the parameters of the model. In each case, the test statistic was less than the critical level estimated in the bootstrap procedure. Thus, we were unable to reject the null hypothesis that the observed frequencies followed those specified by the neighbour model.

4. Discussion and Future Work

The neighbour model proved to be an effective tool in both our simulation study and with real data. With simulated data, the neighbour model provided a linkage threshold with an estimated precision closer to the ground truth than either the Belin-Rubin or Extreme Value Theory methods when trained on a sample of simulated data. In terms of practical use, when applied with a binary search algorithm, it ran quickly in the simulation study as well as on four agricultural datasets, and it also passed the modified chi-squared goodness-of-fit test in each case. Although using recursive partitioning was challenging for large datasets with more varied pair weight distributions, this method provides utility in assessing quality of the links the pair level. Given the potential shown of the neighbour model, it will be considered for use in our process for agricultural linkages. It is recommended to validate that the model assumptions are met, and apply the modified chi-squared test to validate the model's applicability to the specific use case.

References

- Belin, T., and Rubin, D. (1995), "A Method for Calibrating False-Match Rates in Record Linkage", *Journal of the American Statistical Association*, 90, pp. 694-707.
- Dasylda, A., and Goussanou, A. (2022), "On the consistent estimation of linkage errors without training data", *Japanese Journal of Statistics and Data Science*, 5, pp. 181-216. Paper available at <https://doi.org/10.1007/s42081-022-00153-3>.
- Dasylda, A., and Goussanou, A. (2024), "Making statistical inferences about linkage errors", *Japanese Journal of Statistics and Data Science*, 7, pp. 17-56. Paper available at <https://doi.org/10.1007/s42081-023-00228-9>.
- Fellegi, I., and Sunter, A. (1969), "A theory of record linkage", *Journal of the American Statistical Association*, 64, pp. 1183-1210.
- Sariyar, M., Borg, A., and Pommerening, K. (2011), "Controlling false match rates in record linkage using extreme value theory", *Journal of Biomedical Informatics*, 44, pp. 648-654.