Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Measuring the Quality of a Probabilistic Linkage through Clerical-Reviews

Abel Dasylva[1], Melanie Abeysundera, Blache Akpoué, Mohammed Haddou and Abdelnasser Saïdi

Abstract

Probabilistic linkage is susceptible to linkage errors such as false positives and false negatives. In many cases, these errors may be reliably measured through clerical-reviews, i.e. the visual inspection of a sample of record pairs to determine if they are matched. A framework is described to effectively carry-out such clerical-reviews based on a probabilistic sample of pairs, repeated independent reviews of the same pairs and latent class analysis to account for clerical errors.

Key Words: record linkage, probabilistic linkage, linkage error, clerical-review.

1. Introduction

In probabilistic linkage, the decision to link two records, i.e. to classify them as matched or equivalently as relating to the same person, is based on the odds-ratio of observed disagreements between the records (Fellegi and Sunter, 1969). Probabilistic linkage is susceptible to errors that include false positives and false negatives. These errors occur because there is no unique key and for other reasons including typos and differences in formats. Clerical-reviews offer a viable solution for measuring these errors when linking social data including names, addresses and birthdates, e.g. in census coverage studies (Byrne et al., 2002, Dasylva et al., 2014).

A clerical-review is a visual inspection of a record-pair by a person who decides whether the records are matched. It is also called a *manual resolution*, if the goal is to manually link some record pairs. Clerical-reviews perform many valuable functions in the lifecycle of a linkage project, including measuring linkage errors; the focus of this paper. However, clerical-reviews can be subjective, error-prone and costly. They raise three important questions including how to best select a clerical-sample, how to review a selected pair and how to account for clerical errors, which the existing literature has not fully addressed.

The following sections are organized as follows. Section 2 describes the problem of linkage errors. Section 3 describes how to sample pairs for clerical-review. Section 4 describes how to review sampled pairs. Section 5 describes how to estimate the error rates. Section 6 describes how to account for clerical errors. Section 7 describes the estimation of linkage errors for a linkage between the Canadian Mortality Database (CMDB) and the Canadian Community Health Survey (CCHS). Section 8 gives the conclusion.

2. The problem of linkage errors

Probabilistic linkage is susceptible to linkage errors such as false positives and false negatives, which have various sources. However accurately measuring these errors is a challenge.

[1] Corresponding author, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A0T6, Canada (abel.dasylva@canada.ca); Melanie Abeysundera, Statistics Canada; Blache Akpoué, Statistics Canada; Mohammed Haddou, Statistics Canada; Abdelnasser Saidi, Statistics Canada.

2.1 Sources, types and impact of linkage errors

Linkage errors include false positives and false negatives. A false positive or a bad link occurs when a record is linked to an unrelated record. A false positive is further classified as an impossible or incorrect link. It is an *impossible* link when there is no matched record in the other dataset. Otherwise it is an *incorrect* link.. A false negative occurs when a record is not linked to a related record. It is also called a *missing* link. Two measures of linkage error are the False Positive rate (FPR) and the False Negative Rate (FNR). Let FP, TP, FN and TN denote the number of false positives, true positives, false negatives and true negatives, respectively. Then the FPR and FNR are computed respectively as $FP/(FP+TN)$ and $FN/(FN+TP)$. Additional error measures include the sensitivity, specificity and precision that are defined respectively as $1-FNR$, $1-FPR$ and $TP/(TP+FP)$.

2.2 Existing solutions for measuring linkage errors

Users of linked data require accurate measures of linkage errors to adjust for them. So far, proposed solutions have been based on statistical models or clerical-reviews.

In theory, model-based solutions do not require clerical-reviews. Such a solution wasproposed by Fellegi and Sunter (1969) under assumptions of conditional independence. Although the solution is fully automated and cost effective, Belin and Rubin (1995) have noted the inaccuracy of the resulting estimates. Other solutions have incorporated *training* or *truth data*, possibly via clerical-reviews, as in Armstrong and Mayda (1993), Thibaudeau (1993) and Belin and Rubin (1995). Larsen and Rubin (2001) have also described an iterative procedure with multiple rounds of clerical-reviews, which are used as training data, with an Expectation-Maximization (E-M) procedure.

Clerical-reviews measure linkage errors by selecting a probability sample of record-pairs and computing design-based estimates of error measures. Such a solution has been described by Heasman (2014) and also applies to deterministic linkages. Guiver (2011) has described a quality control framework for manual resolution in the grey zone, i.e. bewteen the two thresholds (Fellegi and Sunter, 1969). Yet it may be easily adapted for measuring linkage errors. However, the existing literature is largely silent regarding the actual process for reviewing a sampled pair. Indeed none of Fellegi and Sunter (1969), Newcombe (1988), Newcombe et al. (1983), Guiver (2011) or Heasman (2014) have provided specific details. Another outstanding issue is that of potential clerical-errors that have been largely ignored starting with Fellegi and Sunter (1969). However it is a potentially serious problem that is currently undermining the credibility of clerical-reviews.

3. How to select pairs for review?

In a probabilistic linkage, a pair is assigned a linkage weight and classified as rejected, possible or definite, according to two weight thresholds. A rejected pair has a linkage weight below a first lower threshold, while a definite pair has a weight above a second upper threshold. A possible pair has a weight between the two thresholds and must be resolved manually according to Fellegi and Sunter (1969). In a fully automated solution, the two thresholds coincide. Ideally, the sampling design for the clerical sample should have a positive inclusion probability for each possible, rejected or definite pair. In practice it is necessary to exclude rejected pairs below a cut-off weight, because of the very large number of rejected pairs. The sampling frame is then comprised of the pairs above this cut-off weight.

In what follows, consider a sampling frame including N record-pairs that are labeled $i=1,\dots,N$. For pair i, let γ_i denote the observed vector of disagreements (also called vector of comparison outcomes or simply *outcomes vector*), $m(\gamma_i)$ the conditional probability of the outcomes vector given that the pair is matched, $u(\gamma_i)$ the conditional probability of the same vector given that the pair is unmatched, and $w_i=\log(m(\gamma_i)/u(\gamma_i))$ the linkage weight. Note that with G-LINK, (Chevrette, 2010), the linkage weight is instead computed as $10\log_2(m(\gamma_i)/u(\gamma_i))$. Also define M_i the *latent* variable indicating the match status of pair i, where $M_i=1$ if the pair is matched and $M_i=0$ otherwise. Of course, this match status is unknown for an arbitrary pair, except when it is selected in the clerical sample and the clerical-review process is infallible. Finally let L_i denote the linkage decision, where $L_i=1$ means that the pair is linked.

For a given sample size n and a stratification by linkage weight, a Neyman allocation (Lohr, 1999, pp. 108) may minimize the variance of some total that is related to an error measure. The simplest and most natural such total is the total number of matched pairs in the sampling frame, i.e. $\sum_{i=1}^{N} M_i$. Alternatively, the sample size may be minimized for a target variance of the same total. For a fixed sample size and H given strata denoted by $U_1, \dots, U_H$, Neyman allocation requires an estimate of the variance of M_i within each stratum. For stratum U_h with size N_h denote this variance by S_h^2. For the estimation, assume independent pairs within each stratum. Using the conditional variance formula, the stratum variance is the sum of two terms. The first term is the stratum mean of the conditional variance $var(M_i|\gamma_i)$. The second term is the stratum variance of the conditional mean $E[M_i|\gamma_i]$. Also note that conditional on γ_i, M_i has a Bernoulli distribution with probability $p_i = E[M_i|\gamma_i] = P(M_i = 1|\gamma_i)$ and variance $var(M_i|\gamma_i) = p_i(1 - p_i)$. Thus we have the following expression.

$$\begin{aligned} S_h^2 &= \frac{1}{N_h} \sum_{i \in U_h} p_i(1 - p_i) + \frac{1}{N_h - 1} \sum_{i \in U_h} (p_i - \bar{p}_h)^2 \\ \bar{p}_h &= \frac{1}{N_h} \sum_{i \in U_h} p_i \end{aligned}$$

Neyman allocation leads to the sample size $n_h = (N_h S_h / \sum_{t=1}^{H} N_t S_t) n$ in stratum U_h. The probability p_i is related to the linkage weight when the pairs are assumed independent and identically distributed (IID) according to the mixture $\lambda m(\gamma_i) + (1 - \lambda) u(\gamma_i)$, where λ is a positive mixing proportion, which is possibly unknown. This relationship takes the logistic form $\log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{\lambda}{1-\lambda}\right) + w_i$. An estimate of the mixing proportion is a by-product of any Expectation-Maximization (E-M) procedure that is used to estimate the linkage weights. However, it is typically unknown when the linkage weights are set through a manual iterative procedure (Howe and Lindsay, 1981). In this latter case, the mixing-proportion may be estimated from the pairs and their linkage weights w_i, by the maximization of a partial log-likelihood, assuming that the specified weight is correct for each pair in the blocks. This partial log-likelihood is simply derived as follows. Suppose a *known linkage weight function* $w(\gamma)$ for each possible outcomes vector γ, and an unmatched distribution $u(;\boldsymbol{\theta})$ (also called u-distribution) parametrized by the vector $\boldsymbol{\theta}$. For example, $u(;\boldsymbol{\theta})$ may be based on a log-linear model with $\boldsymbol{\theta}$ comprising of selected interaction terms.The pairs have the mixture distribution $p(\gamma;\boldsymbol{\psi}) = u(\gamma;\boldsymbol{\theta})\left[\lambda e^{w(\gamma)} + 1 - \lambda\right]$, where $\boldsymbol{\psi} = (\lambda, \boldsymbol{\theta})$ and λ is independent of $\boldsymbol{\theta}$. Thus the log-likelihood has the following expression.

$$\log L = \underbrace{\sum_i \log\big((\lambda e^{w_i} + 1 - \lambda)\big)}_{I} + \underbrace{\sum_i \log\big(u(\gamma_i;\boldsymbol{\theta})\big)}_{II}$$

The log-likelihood separates into two parts that may be maximized independently. The maximization of the first part (part I) yields the maximum likelihood estimator of the mixing-proportion λ. This estimation procedure is nonparametric because the u-distribution parameters play no role. However it relies on two important assumptions. The first assumption is that the potential pairs, i.e. the pairs satisfying the blocking criteria, behave as IID pairs with a constant mixing-proportion. In practice this assumption may not hold because the mixing-proportion may vary across blocks. The second assumption is that the specified weights do not deviate from the true linkage weights. In practice, significant deviations are expected when the weights are modified manually. The consequence of such deviations may be an estimated mixing-proportion that falls outside the $[0,1]$ interval. Such a result also provides a diagnostic on the specified linkage weights. In such cases, a simple solution is manually setting the mixing-proportion to a “reasonable” value between 0 and 1.

To further optimize the sample allocation, strata boundaries may be optimized using different schemes, including Dalenius’ cum-$\sqrt{f}$ rule (Dalenius and Hodges, 1959; Dalenius, 1957) based on the distinct p_i values.

4. How to review selected pairs?

Codifying clerical-reviews is a difficult exercise because they are inherently subjective. However, they should be constrained by the following simple guidelines to live up to their potential. First, a reviewer should have the least amount and even no information about the linkage, such as pair linkage weights, outcomes vectors, weight thresholds, or aggregate statistics about different pair types (definite, rejected, or possible). Second, some or all pairs must be subjected to repeated reviews by at least three independent reviewers, with each reviewer required to make a yes/no decision regarding the match status, and conflicts possibly resolved by the majority decision. Although this setup requires additional clerical-resources, it does provide many important benefits. The first advantage are the improved quality of clerical decisions for the pairs that undergo repeated reviews. The second advantage is the ability to evaluate clerical errors based on observed conflicts. The third benefit is the built-in mechanism for detecting pairs that do not provide enough evidence for reliable decisions, without any need for a DON'T KNOW category. In fact, the use of such a category is strongly discouraged because individual reviewers may use it hastily. Third, each reviewer must document any reference to an external source when processing a pair. Such an external source may be a public online resource, e.g., an online obituary, phone directory or map. When cost is an issue, repeated reviews may be applied to a carefully selected subsample of the original sample.

5. How to estimate rates of linkage errors?

Let s denote the probability sample and π_i the inclusion probability for unit i. Then point estimates for the different error measures may be computed with the following simple ratio estimators:

$$\begin{aligned} \widehat{FNR} &= \frac{\sum_{i \in s} \pi_i^{-1} M_i (1 - L_i)}{\sum_{i \in s} \pi_i^{-1} M_i} \\ \widehat{FPR} &= \frac{\sum_{i \in s} \pi_i^{-1} (1 - M_i) L_i}{\sum_{i \in s} \pi_i^{-1} (1 - M_i)} \end{aligned}$$

Variance and confidence intervals may be computed through resampling, or linearization. The Rao-Wu bootstrap (Rao and Wu, 1988) is well suited to this application given the single-stage sampling design and the typically small sampling fraction in each stratum.

6. How reliable are clerical-reviews?

Clerical errors are likely to occur and cause *conflicts* when the same pair is reviewed by many independent reviewers. At the same time, conflicting decisions clearly indicate the occurrence of an error. A simple *conflict resolution* strategy is based on the *majority rule* with an odd number of reviewers. However more elaborate schemes may account for each reviewer's performance. The reliability of clerical-reviews may be measured by assuming that a pair match status is never truly known but that reviewers' errors are conditionally independent given the pair match status and other covariates. Some of these covariates may be reviewer-specific, such as education or experience. The idea of repeated reviews goes back to Newcombe et al. (1983). In traditional surveys, repeated interviews have provided the same function and have been used to evaluate measurement errors (Biemer, 2012).

Consider three independent reviewers for each pair. For convenience, define $s_h = s \cap U_h$ the subsample from stratum h. Also for $h = 1, \dots, H$, and $j = 1,2,3$, let FNR_{hj} and FPR_{hj} respectively denote the FNR and FPR for reviewer j in stratum h. The assumption is made that reviewers make conditionally independent errors given a pair match status and its stratum. In the simplest case, each stratum is a weight interval. However, it may also incorporate additional pair-specific information.

Let FNR_h and FPR_h respectively denote the FNR and FPR for the majority decision in stratum h. For a sampled pair, let C_i denote the majority decision about pair i match status, where $C_i = 1$ if the pair is declared matched and $C_i = 0$ otherwise. In a similar manner let let C_{ij} denote reviewer j's decision about pair i match status, where $C_{ij} = 1$ if the pair is declared matched and $C_{ij} = 0$ otherwise. When the FNR and the FPR are small for all the reviewers in each stratum, and it is safe to assume that the majority decision is infallible (i.e. $C_i = M_i$), reviewer j's FNR (defined as $P(C_{ij} = 0 | M_i = 1)$) and FPR (defined as $P(C_{ij} = 1 | M_i = 0)$) may be simply estimated as follows.

$$\begin{aligned} \widehat{FNR}_{hj} &= \frac{\sum_{i \in s_h} \pi_i^{-1} C_i (1 - C_{ij})}{\sum_{i \in s_h} \pi_i^{-1} C_i} \\ \widehat{FPR}_{hj} &= \frac{\sum_{i \in s_h} \pi_i^{-1} C_{ij} (1 - C_i)}{\sum_{i \in s_h} \pi_i^{-1} (1 - C_i)} \end{aligned}$$

Reviewer j's performance across all the strata is estimated as follows.

$$\widehat{FNR}_{.j} \quad = \quad \frac{\sum_{h=1}^{H}\left(\sum_{i\in s_h}\pi_i^{-1}C_i\right)\widehat{FNR}_{hj}}{\sum_{h=1}^{H}\sum_{i\in s_h}\pi_i^{-1}C_i}$$

$$\widehat{FPR}_{.j} \quad = \quad \frac{\sum_{h=1}^{H}\left[\sum_{i\in s_h}\pi_i^{-1}(1-C_i)\right]\widehat{FPR}_{hj}}{\sum_{h=1}^{H}\sum_{i\in s_h}\pi_i^{-1}(1-C_i)}$$

With homogeneous reviewers, there is no *reviewer's effect* such that $FNR_{hj} = FNR_{h.}$ and $FPR_{hj} = FPR_{h.}$ for all j. Then the common reviewer error rates may be estimated by $\widehat{FNR}_{h.} = \left(\sum_{j=1}^{3}\widehat{FNR}_{hj}\right)/3$ and $\widehat{FPR}_{h.} = \left(\sum_{j=1}^{3}\widehat{FPR}_{hj}\right)/3$. The common reviewer's FNR and FPR across all the strata are estimated by $\widehat{FNR}_{..} = \left(\sum_{j=1}^{3}\widehat{FNR}_{.j}\right)/3$ and $\widehat{FPR}_{..} = \left(\sum_{j=1}^{3}\widehat{FPR}_{.j}\right)/3$.

The infallibility of the majority decision is challenged in strata where pairs provide little discriminating information, e.g., in the middle of the grey zone. Better estimates may be computed by analyzing the clerical-review results with a latent class model and a separate E-M procedure in each stratum, as suggested by Biemer (2011). In this case, a pair outcomes vector is comprised of the yes/no decisions by the individual reviewers.

7. CCHS-CMDB linkage

The above methodology has been applied to a linkage between the Canadian Community Health Survey (CCHS) and the Canadian Mortality Database (CMDB), see Sanmartin et al. (2015). The CCHS dataset is comprised of 2.3M records from collection periods beginning in 2000, 2003, 2005, and between 2007 and 2011. As for the CMDB dataset it is comprised of 3.6M records from 2000 to 2011. The two files have been linked using the probabilistic method with G-LINK, using the variables birthdate, sex, surname, given name and postal code. The application of blocking criteria has produced 418 millions pairs including 114K definite pairs and 22K possible pairs using a first set of thresholds. Some of the possible pairs were resolved manually. The final linkage decision has used a single threshold set at 92 for all pairs that were not resolved manually. For the clerical-review, the sampling frame was comprised of pairs with a linkage weight no smaller than 1. This frame was stratified into eight equally spaced weight intervals, and a sample size of 1,000 was uniformly allocated across the strata. The sample allocation is summarized in Table 7-1.

Table 7-1
Sample design for the clerical-reviews.

Stratum	Frequency	Percent	Weight interval	Sampling weight	Sample size
1	880,515	73.58	1.51 – 23.51	7,044.12	125
2	277,757	23.21	23.52 – 49.51	2,222.06	125
3	5,447	0.46	49.52 – 73.51	43.57	125
4	2,405	0.20	73.52 – 97.51	19.24	125
5	3,274	0.27	97.52 – 123.51	26.19	125
6	2,699	0.23	123.52 – 149.51	21.59	125
7	21,198	1.77	149.52 – 163.51	169.58	125
8	3,347	0.28	163.51 – 194.52	26.77	125

Each sampled pair was reviewed by three independent reviewers with the majority decision rule. The estimated error rates are given in Table 7-2, including an FNR of 2.43%, an FPR of 0.04% and a precision of 98.64%. These results demonstrate the good quality of the linkage and are in line with previous studies, see Da Silveira and Artmann (2009).

Table 7-2
Rates of linkage errors.

	Linked (L)	Not Linked (NL)	
Matched	34,298	855.09	**FNR=2.43%**
Unmatched	473.57	1,161,015	**FPR=0.04%**
	Pr=98.64%		

Rates of clerical errors were also estimated under the assumption that the majority decision is infallible and that there is no reviewer's effect. Overall, the estimated reviewer's error rates are $\widehat{FNR}_{..} = 2.97\%$ and $\widehat{FPR}_{..} = 0.15\%$. These results support the thesis that clerical-reviews offer a viable option for estimating linkage errors. The sampling design for the clerical sample could have used a more optimal stratification or sample allocation. Yet, the decision was made to proceed with it because the resulting sample had already been processed by a first reviewer.

8. Conclusion

Clerical-review is a viable option for measuring linkage errors when linking social data with name, birthdate and address information. However, fully automated model-based solutions are also required, when clerical-reviews are impossible, such as when linking anonymized data.

References

Armstrong, J., and Mayda, J. (1993), "Model-based estimation of record linkage error rates", *Survey Methodology*, 19, pp. 137-147, 1993.

Belin, T.R. and Rubin, D.B. (1995). "A Method for calibrating false-match rates in record linkage", *JASA*, 90, pp. 694–707.

Byrne, R., Beaghen, M., and Mulry, M., (2002). "Clerical review of census duplicates", Accuracy and Coverage Evaluation (ACE) Revision II Report #PP-43, Washington:US Census Bureau.

Biemer, P. (2011). *Latent Class Analysis of Survey Error*. New Jersey:John Wiley.

Chevrette A. G-Link (2010). "Constructing an Avatar". *In Proceedings of the 2010 International Methodology Symposium*, Ottawa: Statistics Canada.

Dalenius, T. (1957). *Sampling in Sweden*. Stockholm: Almquist and Wiksell.

Dalenius, T., and Hodges, J.L. (1959), "Minimum variance stratification", *JASA*, 54, pp. 88-101.

Dasylva A, Titus RC, Thibault C. (2014). "Overcoverage in the 2011 Canadian Census". *In Proceedings of the 2014 International Methodology Symposium*, 29-31 October 2014, Ottawa: Statistics Canada. Available at http://www.statcan.gc.ca/sites/default/files/media/14269-eng.pdf

Da Silveira DP, Artmann E (2009). "Accuracy of probabilistic record linkage applied to health databases: systematic review", *Rev Saude Publica*, 43, pp. 875-882.

Fellegi, I.P., and A.B. Sunter (1969). "A Theory of Record Linkage", *JASA*, 64, pp. 1183-1210.

Guiver, T. (2011). "Sampling-Based Clerical Review Methods in Probabilistic Linking", unpublished report. Australia:Australian Bureau of Statistics. Available at http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.034

Heasman, D. (2014). "Sampling a matching project to establish the linking quality". *Survey Methodology Bulletin*, 72, pp. 1-16.

Howe GR, Lindsay JA (1981). "A generalized iterative record linkage computer system for use in medical follow-up studies". *Computers and Biomedical Research*, 14, pp. 327-340.

Larsen, M., and Rubin, D. (2001). "Iterated automated record linkage using mixture models", *JASA*, 96, pp. 32-41.

Lohr, S (1999). *Sampling: Design and Analysis*. New York: Duxbury.

Newcombe, H.B., Smith, M.E., and Howe, G.R. (1983). "Reliability of computerized versus manual death searches in a study of the health of eldorado uranium workers", *Computers in Biology and Medecine*, 13, pp. 157-169.

Newcombe, H.B. (1988). *Handbook of Record Linkage*. New York: Oxford University Press.

Rao, J.N.K., and Wu, C.F.J. (1988). "Resampling inference with complex survey data", *JASA*, 83, pp. 231-241.

Sanmartin C, Y Decady, R Trudeau, A Dasylva, M Tjepkema, P Fines, R Burnett, N Ross, D Manuel (2015). “Linking the Canadian Community Health Survey to the Canadian Mortality Database: A national resource to study mortality in Canada”, to appear in *Health Reports*.

Thibaudeau Y (1993). “The discrimination power of dependency structures in record linkage”, *Survey Methodology*, 19, pp. 31-38.