Estimating Linkage Errors under Regularity Conditions

Abel Dasylva* Arthur Goussanou†

Abstract

The accurate and cost effective estimation of linkage errors remains a major challenge for the automated production and use of linked data. However this exercise is worthwhile only if the linked data are fit for use. A new model is proposed to estimate the errors without clerical reviews, training data or conditional independence assumptions, under regularity conditions that guarantee the fitness for use of the linked data. It is based on the number of records adjacent to a given record, when linking files that have few duplicate records and a nearly complete coverage of the target population. Additional benefits include the estimation of false negatives due to blocking criteria, as well as record level measures of errors; two challenges for previous models.

Key Words: big data, data integration, record linkage, entity resolution, data matching, false negative rate

Disclaimer: The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada. It describes theoretical methods that might not reflect those currently implemented by the Agency.

1. Introduction

Record linkage aims at identifying records from the same entity that may be a person, household or business. However linkage errors occur when the linkage decisions are based on non-unique quasi-identifiers that are recorded with errors or variations, e.g. names. These linkage errors include false negatives and false positives, where a false negative is not linking records from the same entity and a false positive is linking records from different entities. These errors may generate some bias in the analysis of linked data.

The accurate estimation of linkage errors is required when using linked data in the production of official statistics. However this is a challenge mainly because there is often no certainty about which records come from the same entity. Previous solutions have relied on training data, expert input or the assumption of conditional independence in the record pairs. Under this assumption, different variables have comparison outcomes, which are conditionally independent given that the records come from the same entity or not. It facilitates the estimation of linkage errors but rarely applies to actual data. As for training data and expert input, they are costly and not free from errors.

This paper describes a new error model without these limitations, which extends previous work by Blakely and Salmond (2002). The remaining sections are organized as follows. Section 2 describes the notation. Section 3 proposes regularity conditions that relate to the fitness for use of the linked data. Section 4 introduces the concept of neighbour that is strongly connected to the linkage errors. Sections 5 and 6 describe the proposed model and the related error estimators respectively. Section 7 describes the simulation study. Section 8 concludes with the future work.

*Statistics Canada, 100 Tunney's Pasture driveway, Ottawa ON, K1A0T6, abel.dasylva@canada.ca

†Statistics Canada, 100 Tunney's Pasture driveway, Ottawa ON, K1A0T6

2. Notations

It is convenient to first address the error estimation problem in the simplest setting, when linking two perfect registers of the same finite population, where perfect means without undercoverage or duplicate records. To this end, consider a finite population with N entities and two perfect registers of this population. In the first register, entity i is associated with record $\boldsymbol{v}_i$ that takes its values from the set $\mathcal{V}$. For simplicity the set $\mathcal{V}$ is assumed to be finite even if it is possibly large. Let $\boldsymbol{v}'_j$ (also in $\mathcal{V}$) denote the j-th record in the second register. The records from the two registers correspond through a uniform random permutation $\pi(.)$, such that the record $\boldsymbol{v}'_{\pi(i)}$ is from individual i. The permutation $\pi(.)$ is unknown and assumed independent of the record values. The record values from the different individuals are assumed to be independent and identically distributed. This is also saying that the sample $\left(\boldsymbol{v}_1, \boldsymbol{v}'_{\pi(1)}\right)$,..., $\left(\boldsymbol{v}_N, \boldsymbol{v}'_{\pi(N)}\right)$ is independent and identically distributed.
The two registers are linked without resolving the conflicts. This means that the decision to link two records only depends on these records. The linkage decisions are characterized by the collection $[\mathcal{B}(\boldsymbol{v})]_{\boldsymbol{v}\in\mathcal{V}}$ of subsets of $\mathcal{V}$, such that $\boldsymbol{v}_i$ and $\boldsymbol{v}'_j$ are linked if and only if $\boldsymbol{v}'_j \in \mathcal{B}(\boldsymbol{v}_i)$. The subset $\mathcal{B}(\boldsymbol{v}_i)$ is called *neighbourhood* of $\boldsymbol{v}_i$. The record $\boldsymbol{v}'_j$ is a *neighbour* of $\boldsymbol{v}_i$ if it belongs to the neighbourhood, i.e. if the two records are linked. Then a false negative occurs if $\boldsymbol{v}'_{\pi(i)} \notin \mathcal{B}(\boldsymbol{v}_i)$ for some i. A false positive occurs if $\boldsymbol{v}'_{\pi(i')} \in \mathcal{B}(\boldsymbol{v}_i)$ for some distinct i and i'. These errors are evaluated by different measures, including the False Negative Rate (FNR), the False Positive Rate (FPR) and the Positive Predicted Value (PPV). The false negative rate is the proportion of record pairs that are not linked, among the pairs where the records come from the same entity. The false positive rate is the proportion of record pairs that are linked, among the pairs where the records come from different entities. As for the positive predicted value, it is the proportion of record pairs, where the records come from same entity, among the linked pairs. In the current setting, we have

$$FNR = \frac{1}{N}\sum_{i=1}^{N} I\left(\boldsymbol{v}'_{\pi(i)} \notin \mathcal{B}(\boldsymbol{v}_i)\right), \tag{1}$$

$$FPR = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{i'\neq i} I\left(\boldsymbol{v}'_{\pi(i')} \in \mathcal{B}(\boldsymbol{v}_i)\right), \tag{2}$$

$$PPV = \frac{1-FNR}{1-FNR+(N-1)FPR}. \tag{3}$$

These three measures go by many different names in the literature. For example the positive predicted value is also called precision.

3. Regularity conditions

Estimating the linkage errors is of interest mostly when the linked data are fit for use. What this precisely means differ according to the applications of the linked data. Yet all official statistics applications require a false negative that is bounded away from 1 and a positive predicted value that is bounded away from 0, overall and for each domain of interest. This latter requirement is notable because the ability to produce estimates over small domains is an important driver for the linkage of administrative data and big data. In the following paragraphs, the proposed regularity conditions ensure the fitness for use in the above general sense. In order to describe these conditions, consider $\boldsymbol{v} \in \mathcal{V}$ and let

$$p_N(\boldsymbol{v}) = P\left(\boldsymbol{v}'_{\pi(i)} \in \mathcal{B}(\boldsymbol{v}_i) \middle| \boldsymbol{v}_i = \boldsymbol{v}\right), \tag{4}$$

$$\lambda_N(\boldsymbol{v}) \quad = \quad P\left(\boldsymbol{v}'_{\pi(i')} \in \mathcal{B}(\boldsymbol{v}_i) \middle| \boldsymbol{v}_i = \boldsymbol{v}\right),\ i' \neq i. \tag{5}$$

Then the following regularity conditions are assumed to hold.

$$\inf_{\boldsymbol{v}\in\mathcal{V}} p_N(\boldsymbol{v}) \quad \geq \quad \delta, \tag{6}$$

$$\sup_{\boldsymbol{v}\in\mathcal{V}} (N-1)\lambda_N(\boldsymbol{v}) \quad \leq \quad \Lambda, \tag{7}$$

$$(p_N(\boldsymbol{v}_i), (N-1)\lambda_N(\boldsymbol{v}_i)) \quad \sim \quad F(.,.), \tag{8}$$

where δ is positive, and δ, Λ and $F(.,.)$ do not depend on N. If $FNR \xrightarrow{p} E[FNR] = E[p_N(\boldsymbol{v}_i)]$ and $(N-1)FPR \xrightarrow{p} E[(N-1)FPR] = E[(N-1)\lambda_N(\boldsymbol{v}_i)]$ (these laws of large numbers follow from the regularity conditions), the above conditions imply that

$$FNR \quad \leq \quad 1-\delta, \tag{9}$$

$$PPV \quad \geq \quad \frac{\delta}{\delta+\Lambda}, \tag{10}$$

with high probability, when N is large. A similar result applies for any reasonable domain that is defined in terms of $\boldsymbol{v}_i$. The above conditions characterize the fitness for use of the linked data. The parameters δ and Λ may be chosen to cover the range of intended applications. However, they are not required by the estimation procedure that is described subsequently.

4. Neighbours and errors

The concept of neighbour is crucial when discussing linkage errors. Let n_i denote the number of neighbours of $\boldsymbol{v}_i$. Table 1 shows the connection between n_i and the linkage errors involving the corresponding record. Indeed, without looking at the records or the linkage decisions, it is known that each record has at most one false negative and between 0 and $N-1$ false positives. However, when there are no neighbours, with certainty, it is known that there is one false negative but no false positive. When there is a single neighbour, no information is gained about the false negatives but much information is gained about the false positives because they are known to be in the range $\{0,1\}$ instead of the much wider range, which extends from 0 to $N-1$. In general, when $2 \leq n_i \leq N-1$, much information is gained about the false positives, the number of which is known to be $n_i - 1$ or n_i, even if no additional information is obtained about the occurrence of a false negative. Finally, when there are N neighbours, with certainty, it is known that there is no false negative but $N-1$ false positives. All these observations suggest that the error rates may be estimated by modeling the n_i distribution.

Table 1: Error information from the neighbours

n_i	False negatives	False positives	Error information
0	1	0	full
$1 \leq n_i \leq N-1$	0 or 1	$n_i - 1$ or n_i	partial
N	0	$N-1$	full

5. Neighbour model

The proposed model arises from a convergence in distribution when the population gets arbitrarily large, under the regularity conditions. Indeed, the number of neighbours n_i

is the sum of two contributions, including the number of neighbours from the same entity and that from different entities, where the two contributions are independent conditional on $\boldsymbol{v}_i$, with parameters that are functions of $\boldsymbol{v}_i$. The first contribution follows the $Bernoulli(p_N(\boldsymbol{v}_i))$ distribution conditional on $\boldsymbol{v}_i$. The second contribution follows the $Binomial(N-1, \lambda_N(\boldsymbol{v}_i))$ distribution, conditional on $\boldsymbol{v}_i$. When N becomes large and $\boldsymbol{v}_i$ is such that $p_N(\boldsymbol{v}_i) = p$ and $(N-1)\lambda_N(\boldsymbol{v}_i) = \lambda$, the second contribution converges in distribution to the Poisson distribution with parameter λ (see Theorem 23.2 in Billingsley (1995)). Thus

$$n_i | \{p_N(\boldsymbol{v}_i) = p, (N-1)\lambda_N(\boldsymbol{v}_i) = \lambda\} \xrightarrow{d} Bernoulli(p) * Poisson(\lambda),$$

where $*$ denotes the convolution operator. One obtains a finite mixture if the functions $p(.)$ and $\lambda(.)$ are piecewise constant with latent (i.e. unobserved) level sets, where each component is the sum of a Bernoulli variable with an independent Poisson variable, i.e.

$$n_i \sim \sum_{g=1}^{G} \alpha_g \left(Bernoulli(p_g) * Poisson(\lambda_g) \right), \tag{11}$$

where G is the number of classes (or latent level sets). This model is the limiting form of the model by Blakely and Salmond when $N \to \infty$ and the n_i distribution is heterogeneous. The underlying parameters (α_g, p_g and λ_g for $g = 1, \ldots, G$) may be estimated by maximizing the composite likelihood of the n_i's.

6. Estimators

The neighbour model provides the basis for estimating the error rates. Indeed, when $N \to \infty$ under the regularity conditions, we have the following laws of large numbers.

$$FNR \xrightarrow{p} E[FNR] = 1 - \sum_{g=1}^{G} \alpha_g p_g, \tag{12}$$

$$(N-1)FPR \xrightarrow{p} E[(N-1)FPR] = \sum_{g=1}^{G} \alpha_g \lambda_g, \tag{13}$$

$$PPV \xrightarrow{p} \left(1 + \frac{\sum_{g=1}^{G} \alpha_g \lambda_g}{\sum_{g=1}^{G} \alpha_g p_g} \right)^{-1}. \tag{14}$$

For the FNR, the above law of large numbers applies because it is an iid sum. For the FPR, the result is based on the following observation.

$$\begin{aligned}
(N-1)FPR &= \frac{1}{N} \sum_{i=1}^{N} \frac{N-1}{N-1} \sum_{i' \neq i} P\left(\boldsymbol{v}'_{\pi(i')} \in \mathcal{B}_N(\boldsymbol{v}_i) \middle| \boldsymbol{v}_i \right) + \\
&\quad \frac{1}{N} \sum_{i=1}^{N} (N-1) \underbrace{\left(\frac{\sum_{i' \neq i} \left(I\left(\boldsymbol{v}'_{\pi(i')} \in \mathcal{B}_N(\boldsymbol{v}_i) \right) - P\left(\boldsymbol{v}'_{\pi(i')} \in \mathcal{B}_N(\boldsymbol{v}_i) \middle| \boldsymbol{v}_i \right) \right)}{N-1} \right)}_{O_p(N^{-2})} \\
&= \frac{1}{N} \sum_{i=1}^{N} (N-1)\lambda_N(\boldsymbol{v}_i) + O_p\left(N^{-1} \right).
\end{aligned} \tag{15}$$

The convergence of the PPV follows by continuity from the convergence of the FNR and FPR. The above expressions also apply to the false negatives due to blocking when the

linkage decision reduces to the blocking criteria. The neighbour model also provides a basis for measures of accuracy at the record level such as the probability of a false positive given that there is a single neighbour.

$$P\left(\bigcup_{i'\neq i}\left\{\boldsymbol{v}_{\pi(i')}\in\mathcal{B}(\boldsymbol{v}_i)\right\}\middle| n_i=1\right)=1-\frac{\sum_{g=1}^{G}\alpha_g e^{-\lambda_g}p_g}{\sum_{g=1}^{G}\alpha_g e^{-\lambda_g}\left(p_g+(1-p_g)\,\lambda_g\right)} \tag{16}$$

The above expression shows that this is a positive probability.

7. Simulation study

The simulations are based on a population of $N=2^K+1=129$ entities with $K=7$ dichotomous linkage variables. In the first register, $\boldsymbol{v}_i=\left(v_i^{(1)},\ldots,v_i^{(K)}\right)$, where the distribution of $\boldsymbol{v}_i$ is of the form

$$P(\boldsymbol{v}_i)=2^{-(K-1)}\left(\alpha_1 I\left(v_i^{(1)}=v_i^{(2)}\right)+\alpha_2 I\left(v_i^{(1)}\neq v_i^{(2)}\right)\right),$$

for some $\alpha_1\in(0,1)$ and $\alpha_2=1-\alpha_1$. In the second register, $\boldsymbol{v}'_j=\left(v_j^{(1)\prime},\ldots,v_j^{(K)\prime}\right)$ with $\boldsymbol{v}'_{\pi(i)}$ generated by adding errors to $\boldsymbol{v}_i$ based on $\nu\leq\mu\leq 1$, $u_i\sim Bernoulli(\mu)$, $\tau_i^{(3)},\ldots,\tau_i^{(K)}$ iid according to $Bernoulli(\nu/\mu)$, $e_i^{(1)}=e_i^{(2)}=0$, $e_i^{(k)}=u_i\tau_i^{(k)}$ for $k\geq 3$ and

$$v_{\pi(i)}^{(k)\prime}=v_i^{(k)}+e_i^{(k)}\left(1-2v_i^{(k)}\right).$$

Two records are linked if they agree on all the variables. The n_i distribution follows the neighbour model with two classes, i.e. $G=2$. However when $\alpha_1=\alpha_2=1/2$, the n_i distribution is homogeneous such that the neighbour with one class applies. Conditional independence applies when $\alpha_1=\alpha_2=1/2$ and $\mu=1$.

Four simulation scenarios are considered, which are labeled from 1 to 4. In scenario t, $\alpha_1=1/8+(2t-1)(15/31-1/8)/7$, $\nu=1/50$ and $\mu=\nu+(2t-1)(1-\nu)/7$ for $t=1,\ldots,4$. The scenarios are such that the departure from conditional independence is greater in scenario t than in scenario $t+1$. This ordering of the scenarios also applies with respect to the homogeneity of the neighbour distribution, i.e. the n_i distribution is more homogeneous in scenario $t+1$ than in scenario t. For each scenario, the simulations are based on 1,000 repetitions.

Table 2 shows the mean squared error for the different estimators. It can be seen that the neighbour model tends to have a smaller mean squared error than the model by Blakely and Salmond. It also tends to have a smaller mean squared error than the model by Fellegi and Sunter, when there is more correlation.

8. Future work

For applications, extensions are required regarding undercoverage, duplicate records and conflicts. Another challenge is how to perform valid statistical inferences because the n_i's are correlated.

REFERENCES

Billingsley, P. (1995), *Probability and measure*, New York: Wiley.

Blakely, T., and Salmond, C. (2002), "Probabilistic record linkage and a method to calculate the positive predicted value", *International Journal of Epidemiology*, 31, 1246–1252.

Fellegi, I., and Sunter, A. (1969), "A theory of record linkage", *Journal of the American Statistical Association*, 64, 1183–1210.

Table 2: Mean squared error for all the estimators.

		Scenario			
		1	2	3	4
FNR	Blakely and Salmond	7.02E-04	4.56E-03	4.64E-03	4.86E-03
	Neighbour model with $G = 2$	8.34E-04	4.23E-03	3.09E-03	2.55E-03
	Fellegi and Sunter	1.06E-02	3.96E-03	3.24E-03	3.76E-03
FPR	Blakely and Salmond	1.56E-05	9.47E-06	5.45E-06	5.44E-06
	Neighbour model with $G = 2$	2.02E-06	1.89E-06	1.44E-06	1.07E-06
	Fellegi and Sunter	2.03E-05	5.98E-06	1.07E-06	8.67E-08
PPV	Blakely and Salmond	1.10E-02	9.87E-03	6.91E-03	6.47E-03
	Neighbour model with $G = 2$	8.44E-04	1.79E-03	1.78E-03	1.61E-03
	Fellegi and Sunter	8.48E-03	4.22E-03	1.14E-03	4.22E-04