

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

Evaluating the Accuracy when Linking Records in Waves

by Abel Dasyilva and Arthur Goussanou

Release date: September 8, 2025



Statistics
Canada

Statistique
Canada

Canada

Evaluating the Accuracy when Linking Records in Waves

Abel Dasylyva and Arthur Goussanou¹

Abstract

At Statistics Canada, many data sets are linked with quasi-identifiers such as the first name, last name, or address. In such cases, linkage errors are a potential concern and must be measured. In that regard, previous studies have shown that the evaluation may be based on modeling the number of links from a given record while accounting for all the interactions among the linkage variables and dispensing with clerical reviews, so long as the decision to link two records does not involve other records. In this communication, the methodology is adapted for a class of practical strategies, which violate this constraint by linking the records in consecutive waves, where a given wave links a subset of the records that are not linked in previous waves. In particular, the linkage may be based on a deterministic wave followed by a probabilistic one.

Key Words: Entity resolution; Linkage accuracy; Data integration; Non-sampling error.

Disclaimer: The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada.

1. Introduction

At Statistics Canada, many data sets are linked with quasi-identifiers such as the last name, address or birth date, which are not unique and may be recorded with typos. Therefore, linkage errors may arise, which must be evaluated. These errors include false negatives and false positives, where a false negative is the absence of a link between records that are from the same unit, and a false positive is the presence of a link between records that are from different units. Currently, these errors are measured with clerical reviews. However, due to cost pressures, there is a growing interest in replacing these reviews by a model-based estimation procedure, using one of many available models, which are reviewed by Dasylyva and Goussanou (2024). A promising approach is based on modeling the number of links of a record, where the decision to link two records does not involve other records (Blakely and Salmond, 2002; Dasylyva and Goussanou, 2020). Unfortunately, this constraint is violated by a popular linking strategy that links the records in waves, where each wave links the records that have no link in previous waves. For example, the linkage may comprise a deterministic wave followed by a probabilistic one. While Chipperfield et al. (2018) propose an estimator for such linkages, they crucially rely on the restrictive conditional independence assumption, which may be violated. This work aims to avoid this assumption by extending the model described by Dasylyva and Goussanou (2020). The remaining sections cover the methodology, simulations and conclusion in this order.

2. Methodology

A popular strategy is to link the records in two waves, where the first wave is deterministic (e.g., linking where there is exact agreement on each variable), and the second wave is a probabilistic linkage of the records, which have no links in the first wave. Of course, the linkage may also have many more waves, where each wave links the records that are not linked by the previous waves, with a distinct linking criterion. To evaluate the linkage accuracy, a naïve solution is based on the univariate neighbor model (Dasylyva and Goussanou, 2020), which may yield biased estimates because the decision to link two records is a function of other records. A better solution is based on the multivariate neighbor model (Dasylyva, Goussanou and Nambeu, 2024).

¹Abel Dasylyva, Statistics Canada, 100 Tunney's pasture driveway, Ottawa ON, K1A0T6, abel.dasylyva@statcan.gc.ca; Arthur Goussanou, Statistics Canada, 100 Tunney's pasture driveway, Ottawa ON, K1A0T6;

The following paragraphs introduce the notations and assumptions before describing these two approaches.

Notations and assumptions: Here, we consider the imperfect linkage of two duplicate-free files, which are independent Bernoulli samples S_A and S_B from a finite population with N units, where N is known. The linkage is based on quasi-identifiers that are recorded with typos in each file. In such a situation, there is no certainty regarding the pairs of records that are from the same unit. This means that records from the same unit correspond across the two samples through a random permutation $\Pi(\cdot)$ of $\{1, \dots, N\}$, which is uniformly distributed and unknown. Without losing any generality, unit i may be associated with record V_i in S_B , and record $V_{\Pi(i)}$ in S_A ; the record values living in a record space \mathcal{V}_N . The inclusion of the units in the samples and their record values (i.e., $[(I(i \in S_A), I(i \in S_B), V_i, V_{\Pi(i)})]_{1 \leq i \leq N}$) are assumed to be independent of $\Pi(\cdot)$. Furthermore, the sample inclusion indicators and the record values are assumed to be mutually independent and identically distributed across the units, i.e., the tuples $(I(1 \in S_A), I(1 \in S_B), V_1, V_{\Pi(1)}), \dots, (I(N \in S_A), I(N \in S_B), V_N, V_{\Pi(N)})$ are assumed to be iid. Thanks to the assumptions about $\Pi(\cdot)$, no generality is lost by conditioning on the fact that it is the identity permutation (i.e., $\Pi(i) = i$ for each i) in what follows. This allows us to simplify the notation by keeping this conditioning implicit, i.e., the conditional expectation $E[\mathcal{E}|\Pi]$ is simply denoted by $E[\mathcal{E}]$ for any event \mathcal{E} of interest. The conditioning also implies that unit i is associated with records V_i and V_i' .

The linkage is based on K waves, which are based on a combination of diverse linkage criteria and methods. For example with three waves, the first may be deterministic, while the second is probabilistic and the last is based on machine learning, e.g., k -means clustering. In wave $k = 1, \dots, K$, denote the indicator of a link between the records i and j by $L_{ij}^{(1,k)}$ and suppose that this decision does not involve other records beyond V_i and V_j' . The records are specifically linked as follows. In the first wave, V_i and V_j' are linked if $L_{ij}^{(1,1)} = 1$. In the second wave, the two records are linked if V_i has no link in the first wave (implying $L_{ij}^{(1,1)} = 0$) and $L_{ij}^{(1,2)} = 1$. In general, in wave $k \geq 2$, V_i and V_j' are linked if V_i has no link in all the previous waves and $L_{ij}^{(1,k)} = 1$. For what is to follow, it is convenient to introduce the decisions $L_{ij}^{(2,1)} = L_{ij}^{(1,1)}$, and for $k > 2$, let $L_{ij}^{(2,k)} = L_{ij}^{(1,k)} \prod_{t=1}^{k-1} (1 - L_{ij}^{(1,t)})$, where $L_{ij}^{(2,k)}$ only involves V_i and V_j' for each k , and $L_{ij}^{(2,1)}, \dots, L_{ij}^{(2,K)}$ are mutually exclusive, i.e., $L_{ij}^{(2,k)} L_{ij}^{(2,k')} = 0$ if $k \neq k'$. Let $L_{ij}^{(2)} = L_{ij}^{(2,1)} \vee \dots \vee L_{ij}^{(2,K)}$ (\vee denoting the binary OR operator) denote the indicator that V_i and V_j' are linked by some rule in the first set of rules, which is also the indicator that they are linked by some rule in the second set of rules, since $L_{ij}^{(1,1)} \vee \dots \vee L_{ij}^{(1,K)} = L_{ij}^{(2,1)} \vee \dots \vee L_{ij}^{(2,K)}$. Finally, let L_{ij} denote the ultimate linkage decision for V_i and V_j' . The goal is to assess the errors for this latter decision, when the linkage is based on quasi-identifiers such as names or dates.

Following Fellegi and Sunter (1969), a record pair is called *matched* if its records are from the same unit. Then a record pair is cross-classified in a confusion matrix, according to whether it is matched, and whether it is linked as shown in Table 2-1. A true positive (TP) is a matched pair that is linked, a true negative (TN) is an unmatched pair that is not linked, a false negative (FN) is a matched pair that is not linked, and a false positive (FP) is an unmatched pair that is linked, where the errors comprise the latter two kinds of pairs.

Table 2-1
Confusion matrix

	Linked	Not linked
Matched	TP	FN
Unmatched	FP	TN

For convenience, let TP , TN , FN and FP also denote the numbers of pairs of the different types, which serve to compute the recall and precision, where the recall is the proportion of matched pairs that are linked (i.e., $TP/(TP + FN)$), and the precision is the proportion of linked pairs that are matched (i.e., $TP/(TP + FP)$).

Applying the univariate neighbor model: The accuracy of the decision L_{ij} may be evaluated with the univariate neighbor model (Dasylyva and Goussanou, 2020) based on the number of links $n_i = \sum_{j=1}^N I(j \in S_A)L_{ij}$ of record i from S_B . According to this model

$$n_i \sim \sum_{g=1}^G \alpha_g \text{Bernoulli}(p_g) * \text{Poisson}(\lambda_g), \quad (2.1)$$

where $*$ is the convolution operation, G is the number record classes, and p_g and λ_g are respectively the expected numbers of true positives and false positives per record in class g , with $\bar{p} = \alpha_1 p_1 + \dots + \alpha_G p_G$ and $\bar{\lambda} = \alpha_1 \lambda_1 + \dots + \alpha_G \lambda_G$ denoting the expected numbers of true positives and false positives of any record (i.e. from any class) in S_B . Then, the recall and precision correspond respectively to $P(i \in S_A)^{-1} \bar{p}$ and $\bar{p}/(\bar{p} + \bar{\lambda})$ (Dasylyva, Goussanou and Nambu, 2024), which may be estimated by maximizing the likelihood of the n_i 's. However, this approach is naïve and yields biased estimators because the decision to link two records is not independent of other records here. Indeed, this is easily seen in the following simple counter example, where $K = 2$, $L_{ij}^{(1,1)}$ is based on the exact agreement on all the variables, while $L_{ij}^{(1,2)}$ is based on a looser criterion. Then, V_i is linked to $V_{j'}$ if the two records are identical, or if $V_i \neq V_{j'}$ for each $j' \in S_A$ and $L_{ij'}^{(2,2)} = 1$. Clearly, L_{ij} is a function of each record in S_A . A better solution is to apply the multivariate version of the neighbor model (Dasylyva et al., 2024).

Applying the multivariate neighbor model: The multivariate neighbor model (Dasylyva, Goussanou and Nambu, 2024) was developed to enable the joint estimation of the linkage accuracy of many mutually exclusive linking criteria, where each criterion links two records independently of other records. It is a finite multivariate mixture, where each component is based on the convolution of a multinomial distribution with a multivariate compound Poisson distribution (Wang, 1996). In general, a compound Poisson distribution is parametrized by a positive Poisson parameter λ and a severity distribution $h(\cdot)$. When $h(\cdot)$ is discrete, the probability mass function has the form

$$P(X = x) = \sum_{t=0}^{\infty} \frac{\lambda^t}{t!} e^{-\lambda} h^{*t}(x),$$

where $h^{*t}(\cdot)$ is the t -fold convolution of $h(\cdot)$ with itself. For convenience, the multivariate compound distribution with Poisson parameter λ and severity distribution $h(\cdot)$ is denoted by $MCP(\lambda, h(\cdot))$. The multivariate finite mixture arises from a convergence in distribution, when N becomes arbitrarily large under regularity conditions.

To describe how the multivariate neighbor model applies to the problem in hand, denote by $n_i^{(k)}$ the number of links of V_i based on $L_{ij}^{(2,k)}$, and let $\mathbf{n}_i = [n_i^1 \dots n_i^K]^\top$ (also denoted by $[n_i^{(k)}]_{1 \leq k \leq K}$ for convenience). For $v \in \mathcal{V}_N$, define the neighborhood of v as a set $\mathcal{B}_N(v)$ containing all the record values from \mathcal{V}_N , which are linked to v in some wave, with a positive probability, i.e.,

$$\mathcal{B}_N(v) \supset \left\{ v' \in \mathcal{V}_N \text{ s. t. } E \left[\sum_{k=1}^K L_{ii}^{(2,k)} \mid (i, j) \in S_B \times S_A, (V_i, V_i') = (v, v') \right] > 0 \right\}.$$

Finally, let \mathcal{V}_N^* denote the subset of record values that may be observed in S_B with a positive probability, and for distinct i and j , define

$$\begin{aligned} p_N^{(k)}(v) &= E[I(i \in S_A)L_{ii}^{(2,k)} \mid i \in S_B, V_i = v], \\ \lambda_N^{(k)}(v) &= E[I(j \in S_A)L_{ij}^{(2,k)} \mid i \in S_B, V_i = v], \\ \lambda_N^{(0)}(v) &= P(j \in S_A \text{ and } V_{j'} \in \mathcal{B}_N(V_i) \mid i \in S_B, V_i = v), \end{aligned}$$

$\mathbf{p}_N(v) = [p_N^{(k)}(v)]_{1 \leq k \leq K}$ and $\boldsymbol{\lambda}_N(v) = [\lambda_N^{(k)}(v)]_{1 \leq k \leq K}$. The extended model is based on the convergence in distribution of \mathbf{n}_i under the following conditions (Dasylyva, Goussanou and Nambu, 2024, Lemma 2).

$$\sup_{v \in \mathcal{V}_N^*} (N-1)\lambda_N^{(0)}(v) \leq \Lambda, \quad (2.2)$$

$$(\mathbf{p}_N(V_i), (N-1)\lambda_N(V_i)) \xrightarrow{d} F, \quad (2.3)$$

where Λ and F are independent of N , and F is the atomic distribution, which puts the probability mass α_g on $(\mathbf{p}_g, \lambda_g)$, with $\mathbf{p}_g(v) = [p_g^{(k)}(v)]_{1 \leq k \leq K}$, $\lambda_g(v) = [\lambda_g^{(k)}(v)]_{1 \leq k \leq K}$ and $g = 1, \dots, G$. To detail this limiting distribution, let $\mathbf{0}_K$ denote the K -dimensional column vector of zeros, let \mathbf{e}_k denote the K -dimensional column vector with a 1 in position k and zeros elsewhere, and for $\mathbf{z} \in \mathbb{N}^K$ let $\delta_{\mathbf{z}}(\cdot)$ denote the distribution that puts all the mass at \mathbf{z} . Then

$$\mathbf{n}_i \xrightarrow{d} \sum_{g=1}^G \alpha_g \mathbf{p}_g(\cdot) * MCP(\lambda_g, h_g(\cdot)), \quad (2.4)$$

$$p_g(\cdot) = \left(1 - \sum_{k=1}^K p_g^{(k)}\right) \delta_{\mathbf{0}_K}(\cdot) + \sum_{k=1}^K p_g^{(k)} \delta_{\mathbf{e}_k}(\cdot), \quad (2.5)$$

$$h_g(\cdot) = \sum_{k=1}^K \left(\frac{\lambda_g^{(k)}}{\lambda_g}\right) \delta_{\mathbf{e}_k}(\cdot), \quad (2.6)$$

$$\lambda_g = \lambda_g^{(1)} + \dots + \lambda_g^{(K)}. \quad (2.7)$$

Note that the compound distribution $MCP(\lambda_g, h_g(\cdot))$ corresponds to the product of K independent Poisson distributions, while the distribution $p_g(\cdot)$ is also called *incomplete multinomial distribution* in previous work (Dasylyva, Goussanou, and Nambu, 2024). The model parameters comprise α_g , $\mathbf{p}_g = [p_g^{(k)}]_{1 \leq k \leq K}$ and $\lambda_g = [\lambda_g^{(k)}]_{1 \leq k \leq K}$ for each g , and they are estimated consistently by maximizing likelihood of the \mathbf{n}_i 's (Dasylyva, Goussanou, and Nambu, 2024, Theorem 2), including the selection of G selected according to the minimum Akaike information criterion.

With these parameters, the recall and precision of $L_{ij}^{(2,k)}$ are respectively derived as $P(i \in S_A)^{-1} \sum_{g=1}^G \alpha_g p_g^{(k)}$ and $\sum_{g=1}^G \alpha_g \lambda_g^{(k)}$, for each one of the K mutually exclusive linkage criteria (i.e., $L_{ij}^{(2,1)}, \dots, L_{ij}^{(2,K)}$) while accounting for the heterogeneity of V_i in a manner that is consistent across these criteria. This is the main advantage of the multivariate model, when comparing to the separate application of the univariate model (Dasylyva and Goussanou, 2020) to $L_{ij}^{(2,k)}$ for each k . The next step is deriving the recall and precision when the records are linked in K successive waves as described above.

To do so, let $TP_i^{(k)}$ and $FP_i^{(k)}$ denote the number of true positives and false positives for V_i in wave k , and let $TP_i = TP_i^{(1)} + \dots + TP_i^{(K)}$ and $FP_i = FP_i^{(1)} + \dots + FP_i^{(K)}$ denote the total number of true positives and false positives for the same record across all the waves. Note that $TP_i^{(k)}$ and $FP_i^{(k)}$ are not the numbers of true positives and false positives under $L_{ij}^{(2,k)}$. For convenience, redefine \bar{p} and $\bar{\lambda}$ as $\bar{p} = \lim_{N \rightarrow \infty} E[TP_i^{(k)} | i \in S_B]$ and $\bar{\lambda} = \lim_{N \rightarrow \infty} E[FP_i^{(k)} | i \in S_B]$, i.e., the limiting values of the expected numbers of true positives and false positives per record in S_B across all the waves. The following lemma gives the expressions for these parameters. The proof is found in the appendix.

Lemma 1:

$$\bar{p} = \sum_{k=1}^K \sum_{g=1}^G \alpha_g \left(I(k=1) p_g^{(1)} + I(k>1) p_g^{(k)} \exp\left(-\sum_{t=1}^{k-1} \lambda_g^{(t)}\right) \right), \quad (2.8)$$

$$\bar{\lambda} = \sum_{k=1}^K \sum_{g=1}^G \alpha_g \left(I(k=1) \lambda_g^{(1)} + I(k>1) \lambda_g^{(k)} \left(1 - \sum_{t=1}^{k-1} p_g^{(t)}\right) \exp\left(-\sum_{t=1}^{k-1} \lambda_g^{(t)}\right) \right). \quad (2.9)$$

The recall and precision of L_{ij} converge in probability to $P(i \in S_A)^{-1}\bar{p}$ and $\bar{p}/(\bar{p} + \bar{\lambda})$, respectively, under the following additional regularity conditions (Dasylyva, Goussanou and Nambeu, 2024, Theorem 1 and Corollary 1) for distinct i and i' from $\{1, \dots, N\}$.

$$NP(\mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \emptyset | \{i, i'\} \subset S_B) \leq c, \quad (2.10)$$

$$NP(V_{i'} \in \mathcal{B}_N(V_i) | \{i, i'\} \subset S_B, \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \emptyset) \leq c, \quad (2.11)$$

where c is a positive constant independent of N . Eq. (2.10) means that unmatched records in S_B have disjoint neighborhoods, with a high probability. This is also an indication that the two records are very dissimilar. Eq. (2.11) means that matched records are sufficiently similar. Thus, the recall and precision may be estimated consistently by $P(i \in S_A)^{-1}\hat{p}$ and $\hat{p}/(\hat{p} + \hat{\lambda})$, where \hat{p} and $\hat{\lambda}$ are based on the maximum likelihood (Dasylyva, Goussanou and Nambeu, 2024, Theorem 2).

So far, it has been assumed that $L_{ij}^{(1,k)}$ only depends on V_i and V_j and not on other records. However, this condition is violated if the linkage decision is a function of decision parameters that depend on all the records from the two sources. For example, this occurs when $L_{ij}^{(1,k)}$ is based on the probabilistic method; the decision parameters comprising the linkage weights and a weight threshold, where there are no clerical resources. This is also the case if the decision is based on two-means clustering, where the decision parameters comprise the centers of the two clusters. Yet, the linkage accuracy can be estimated consistently if the decision parameters are estimated consistently.

3. Simulations

The methodology is evaluated with simulations that are based on 100 repetitions. In each repetition, a fictitious population is generated with a last name and birth date for each individual, where the distributions of these variables are based on the counts from the 2010 US census (US Census Bureau, 2016, 2020). From this finite population, two files are created by drawing independent Bernoulli samples with the same inclusion probability of 0.9, where the last name and birth date are perturbed as follows. The last name string is modified by many typos, where the number of typos follows the Poisson distribution with mean 0.1. With many typos, the different typos are applied in sequence, i.e., a typo is applied to the result of modifying the last name by the preceding typos. Each typo is stochastic and equally likely to be a deletion or an insertion, at a uniformly random position in the string. For an insertion, the inserted letter is drawn uniformly from the alphabet. For the birthdate, independent perturbations are applied for year on one hand and for the day and month on the other hand. The year is recorded without typos with probability 0.9, else it is increased or decreased by 1 with equal probability for each option. Similarly, the day and month are recorded without typos with probability 0.9. Otherwise, the two date components are transposed.

The linkage is also implemented in SAS. It is based on two waves, where the first wave is based on having the same last name and birth date, while the second wave is based on the probabilistic method. For this second wave, a simple probabilistic linkage is implemented, which reflects current practices, including the working assumption of conditional independence (i.e., the assumption that agreements on different variables are independent given that a pair is matched or unmatched). For this linkage, blocking is based on having the same year and meeting at least one of the following criteria: the same last name and day, the same last name and month, the same first letter for the last name and the same day and month, or the same first letter for the last name and a cross agreement on the day and month. By cross agreement on the day and month, we mean that the day of the first record agrees with the month of the second record, and that the month of the first record agrees with the day of the second record. The probabilistic linkage is simple and based on three comparisons, including whether the Jaro-Winkler similarity of the last name is larger than 0.8, whether the day is the same, and whether the month is the same. The linkage parameters are estimated numerically with SAS OPTMODEL under the conditional independence assumption. Finally, a record pair is linked where the estimated conditional match probability (i.e., the conditional probability that a pair is matched given the observed comparison outcomes) is equal to or greater than $1/2$, to minimize a linear combination of the false positives and false negatives. Yet, the resulting decision is likely to be suboptimal because the conditional independence assumption is violated, and the conditional match probability is estimated with some bias. However, this does not interfere with our main goal, which is to evaluate our ability to accurately measure the achieved linkage accuracy, regardless of its optimality. Besides, the proposed methodology does not rest on the conditional independence assumption. The error estimation

procedure is implemented in R based on the restricted version of the multivariate model where $p_1^{(k)} = \dots = p_G^{(k)} = p^{(k)}$. For comparison, the linkage accuracy is also estimated naïvely with the restricted univariate model under the constraint $p_1 = \dots = p_G = p$. In the simulations, the population size is increased from 100,000 to 200,00. By increasing the population size, the number of false positives is increased, and the precision is decreased. The results appear in Table 3-1. While the variance is smaller with the univariate model, the relative bias and mean square error (MSE) are much smaller with the proposed estimator. The drop in MSE is quite sharp when going from the univariate to the multivariate model. Additionally, the results show that the MSE increases with N for each estimator. Overall, the multivariate model performs better than the univariate model as expected.

Table 3-1
Simulation results.

Measure	Estimator	$N = 100,000$			$N = 200,000$		
		Relative bias (%)	Variance ($\times 10^{-6}$)	MSE ($\times 10^{-3}$)	Relative bias (%)	Variance ($\times 10^{-6}$)	MSE ($\times 10^{-3}$)
Precision	Univariate	7.92	1.42	4.79	14.11	1.38	11.96
	Multivariate	-1.79	3.90	0.25	-3.63	2.55	0.79
Recall	Univariate	7.92	2.87	2.83	14.11	1.85	8.96
	Multivariate	-1.92	5.11	0.17	-3.89	3.44	0.68

4. Conclusions

For applications of linked data to official statistics, it is critical to measure the linkage accuracy. This is also true when using a popular strategy, which consists in linking records in waves, where each wave links a subset of the records that are not linked in all the previous waves. However, evaluating the resulting accuracy has been a challenge so far, because the decision to link two records depends on other records. In this paper, the problem is finally addressed by modeling the number of links from a record, without making the conditional independence assumption or requiring clerical reviews. In future work, this approach will be extended to other linkage strategies.

References

- Blakely, T., and Salmond, C. (2002), “Probabilistic record linkage and a method to calculate the positive predicted value”, *Journal of Epidemiology*, 31, pp. 1246-1252.
- Chipperfield, J., Hansen, N., and Rossiter, P. (2018), “Estimating precision and recall for deterministic and probabilistic record linkage”, *International Statistical Review*, 86, pp. 219-236.
- Dasylva, A. and Goussanou, A. (2020), “Estimating linkage errors under regularity conditions”, *Proceedings of the Survey Methods Section, American Statistical Association*, pp. 687-692. Available at <http://www.asasrms.org/Proceedings/y2020/files/1505346.pdf>.
- Dasylva, A., Goussanou, A., and Nambu, C.-O. (2024), “Models of linkage error for capture-recapture estimation without clerical reviews”, *Survey Methodology*, 50, pp. 375-408.
- Fellegi, I., and Sunter, A. (1969), “A theory of record linkage”, *Journal of the American Statistical Association*, 64, pp. 1183-1210.
- US Census Bureau. (2016)., File b: surnames occurring 100 or more times. <https://www2.census.gov/topics/genealogy/2010surnames/names.zip> , Accessed: 2020-10-17

US Census Bureau. (2020), Annual state resident population estimates for 6 race groups (5 race alone groups and two or more races) by age, sex, and Hispanic origin: April 1, 2010 to July 1, 2019. <https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/asrh/sc-est2019-alldata6.csv> . Accessed: 2020-10-17

Wang, Y.H. (1996), “On the multivariate compound distributions”, *Journal of Multivariate Analysis*, 59, pp. 13-21.

Appendix

Proof of Lemma 1: In the first wave, a true positive occurs if and V_i and V_i' are linked, and a false positive occurs if V_i and V_j' are linked for $j \neq i$. Consequently,

$$\begin{aligned} E \left[TP_i^{(1)} \mid i \in S_B, V_i = v \right] &= p_N^{(1)}(v), \\ E \left[FP_i^{(1)} \mid i \in S_B, V_i = v \right] &= (N-1)\lambda_N^{(1)}(v), \end{aligned}$$

and

$$\begin{aligned} E[TP_i^{(1)} \mid i \in S_B] &= E[p_N^{(1)}(V_i) \mid i \in S_B] \\ &= \int p^{(1)} dF(\mathbf{p}, \boldsymbol{\lambda}) + \int p^{(1)} d(F_N(\mathbf{p}, \boldsymbol{\lambda}) - F(\mathbf{p}, \boldsymbol{\lambda})) \\ &= \sum_{g=1}^G \alpha_g p_g^{(1)} + \underbrace{\int p^{(1)} d(F_N(\mathbf{p}, \boldsymbol{\lambda}) - F(\mathbf{p}, \boldsymbol{\lambda}))}_{o(1)}. \end{aligned}$$

In a similar manner

$$\begin{aligned} E[FP_i^{(1)} \mid i \in S_B] &= E[(N-1)\lambda_N^{(1)}(V_i) \mid i \in S_B] \\ &= \int \lambda^{(1)} dF(\mathbf{p}, \boldsymbol{\lambda}) + \int \lambda^{(1)} d(F_N(\mathbf{p}, \boldsymbol{\lambda}) - F(\mathbf{p}, \boldsymbol{\lambda})) \\ &= \sum_{g=1}^G \alpha_g \lambda_g^{(1)} + \underbrace{\int \lambda^{(1)} d(F_N(\mathbf{p}, \boldsymbol{\lambda}) - F(\mathbf{p}, \boldsymbol{\lambda}))}_{o(1)}. \end{aligned}$$

In wave $k \geq 2$, a true positive occurs if $I(i \in S_A)L_{ii}^{(2,k)} = 1$ and $I(j \in S_A)\sum_{t=1}^{k-1} L_{ij}^{(2,t)} = 0$ for each $j \neq i$. Besides, there is a mutual independence among the variables that comprise $I(i \in S_A)L_{ii}^{(2,k)}$ and $I(j \in S_A)\sum_{t=1}^{k-1} L_{ij}^{(2,t)}$ for each $j \neq i$, when conditioning on the event $\{i \in S_B\} \cap \{V_i = v\}$. Hence

$$E[TP_i^{(k)} \mid i \in S_B, V_i = v] = p_N^{(k)}(v) \left(1 - \sum_{t=1}^{k-1} \lambda_N^{(t)}(v) \right)^{N-1}.$$

and

$$\begin{aligned}
E[TP_i^{(k)} | i \in S_B] &= E \left[p_N^{(k)}(V_i) \left(1 - \frac{1}{N-1} \sum_{t=1}^{k-1} (N-1) \lambda_N^{(t)}(V_i) \right)^{N-1} \middle| i \in S_B \right] \\
&= \int p^{(k)} \exp \left(- \sum_{t=1}^{k-1} \lambda^{(t)} \right) dF(\mathbf{p}, \boldsymbol{\lambda}) + \\
&\quad \int p^{(k)} \underbrace{\left(\left(1 - \frac{1}{N-1} \sum_{t=1}^{k-1} \lambda^{(t)} \right)^{N-1} - \exp \left(- \sum_{t=1}^{k-1} \lambda^{(t)} \right) \right)}_{o(1/N)} dF(\mathbf{p}, \boldsymbol{\lambda}) + \\
&\quad \underbrace{\int p^{(k)} \left(1 - \frac{1}{N-1} \sum_{t=1}^{k-1} \lambda^{(t)} \right)^{N-1} d(F_N(\mathbf{p}, \boldsymbol{\lambda}) - F(\mathbf{p}, \boldsymbol{\lambda}))}_{o(1)} \\
&= \sum_{g=1}^G \alpha_g p_g^{(k)} \exp \left(- \sum_{t=1}^{k-1} \lambda_g^{(t)} \right) + o(1).
\end{aligned}$$

In the same wave, for distinct i and j , a false positive between V_i and V_j' occurs if $I(j \in S_A) L_{ij}^{(2,k)} = 1$, $I(i \in S_A) \sum_{t=1}^{k-1} L_{ii}^{(2,t)} = 0$ and $I(j' \in S_A) \sum_{t=1}^{k-1} L_{ij'}^{(2,t)} = 0$, for each $j' \neq i, j$, with the mutual independence among the variables that comprise $I(j \in S_A) L_{ij}^{(2,k)}$, $I(i \in S_A) \sum_{t=1}^{k-1} L_{ii}^{(2,t)}$ and $I(j' \in S_A) \sum_{t=1}^{k-1} L_{ij'}^{(2,t)} = 0$, for each $j' \neq i, j$, when conditioning on the event $\{i \in S_B\} \cap \{V_i = v\}$. Therefore

$$E[FP_i^{(k)} | i \in S_B, V_i = v] = (N-1) \lambda_N^{(k)}(v) \left(1 - \sum_{t=1}^{k-1} p_N^{(t)}(v) \right) \left(1 - \sum_{t=1}^{k-1} \lambda_N^{(t)}(v) \right)^{N-2},$$

and

$$\begin{aligned}
E[FP_i^{(k)} | i \in S_B] &= E \left[(N-1) \lambda_N^{(k)}(v) \left(1 - \sum_{t=1}^{k-1} p_N^{(t)}(v) \right) \left(1 - \sum_{t=1}^{k-1} \lambda_N^{(t)}(v) \right)^{N-2} \middle| i \in S_B \right] \\
&= \int \lambda^{(k)} \left(1 - \sum_{t=1}^{k-1} p^{(t)} \right) \exp \left(- \sum_{t=1}^{k-1} \lambda^{(t)} \right) dF(\mathbf{p}, \boldsymbol{\lambda}) + \\
&\quad \int \lambda^{(k)} \left(1 - \sum_{t=1}^{k-1} p^{(t)} \right) \underbrace{\left(\left(1 - \frac{1}{N-1} \sum_{t=1}^{k-1} \lambda^{(t)} \right)^{N-2} - \exp \left(- \sum_{t=1}^{k-1} \lambda^{(t)} \right) \right)}_{o(1/N)} dF(\mathbf{p}, \boldsymbol{\lambda}) + \\
&\quad \underbrace{\int \lambda^{(k)} \left(1 - \sum_{t=1}^{k-1} p^{(t)} \right) \left(1 - \frac{1}{N-1} \sum_{t=1}^{k-1} \lambda^{(t)} \right)^{N-2} d(F_N(\mathbf{p}, \boldsymbol{\lambda}) - F(\mathbf{p}, \boldsymbol{\lambda}))}_{o(1)} \\
&= \sum_{g=1}^G \alpha_g \lambda_g^{(k)} \left(1 - \sum_{t=1}^{k-1} p_g^{(t)} \right) \exp \left(- \sum_{t=1}^{k-1} \lambda_g^{(t)} \right) + o(1).
\end{aligned}$$

QED