# ESTIMATION USING THE GENERALIZED WEIGHT SHARE METHOD: THE CASE OF RECORD LINKAGE

**Pierre Lavallée and Pierre Caron**
**Pierre Lavallée, Statistics Canada, Ottawa, Ontario, K1A 0T6 CANADA, plavall@statcan.ca**

**Key words: Generalized Weight Share Method, Record Linkage, Estimation, Clusters.**

## 1. INTRODUCTION

To augment the amount of available information, data from different sources are increasingly being combined. These databases are often combined using record linkage methods. When the files involved have a unique identifier that can be used, the linkage is done directly using the identifier as a matching key. When there is no unique identifier, a probabilistic linkage is used. In that case, a record on the first file is linked to a record on the second file with a certain probability, and then a decision is taken on whether this link is a true link or not. Note that this process usually requires a certain amount of manual resolution.

Manual resolution usually requires a large of amount of resources with respect to time and employees. It might then be legitimate to see if it would be possible to evaluate if manual resolution can be reduced or even eliminated. This issue will be addressed in this paper, especially when one tries to produce an estimate of a total (or a mean) of one population when using a sample selected from another population linked somewhat to the first population. In other words, having two populations linked through record linkage, we will try to avoid any decision concerning the validity of links, but still be able to produce an unbiased estimate for a total of one of the two populations.

The problem that is considered here is to estimate the total of a characteristic of a population that is naturally divided into clusters. Assuming that the sample is obtained by the selection of units within clusters, if at least one unit of a cluster is selected, then the whole cluster will be interviewed. This usually leads to cost reductions as well as the possibility of producing estimates on the characteristics of both the clusters and the units.

The present paper will show that avoiding deciding on the validity of the links can be achieved using the Generalized Weight Share Method (GWSM) that has been described by Lavallée (1995). This method is an extension of the Weight Share Method presented by Ernst (1989). Although this last method has been developed in the context of longitudinal household surveys, it was shown that the Weight Share Method can be generalized to situations where a population of interest is sampled through the use of a frame which refers to a different population, but linked somehow to the first one.

## 2. RECORD LINKAGE

The concepts of record linkage were introduced by Newcome *et al.* (1959) and formalized in the mathematical model of Fellegi and Sunter (1969). As described by Barlett *et al.* (1993), *record linkage* is the process of bringing together two or more separately recorded pieces of information pertaining to the same unit (individual or business). Record linkage is sometimes also called *exact matching*, in contrast to *statistical matching*. This last process attempts to link files that have few units in common. Linkages are then based on similar characteristics rather than unique identifying information. In the present paper, we will restrict ourselves to the context of record linkage. However, the developed theory could also be used for statistical matching.

Suppose that we have two files A and B containing characteristics related to two populations $U^A$ and $U^B$, respectively. The two populations are somehow related to each other. The purpose of record linkage is to link the records of the two files A and B. If the records contain unique identifiers, then the matching process is trivial. Unfortunately, often a unique identifier is not available and then the linkage process needs to use some probabilistic approach to decide whether two records of the two files are linked together or not. With this linkage process, the likelihood of a correct match is computed and, based on the magnitude of this likelihood, it is decided whether we have a link or not.

Formally, we consider the product space $\boldsymbol{AXB}$ from the two files A and B. Let $j$ indicates a record (or unit) from file A (or population $U^A$) and $k$ a record (or unit) from file B (or population $U^B$). For each pair $(j,k)$ of $\boldsymbol{AXB}$, we compute a linkage weight $\theta_{jk}$ reflecting the degree to which the pair $(j,k)$ is likely to be a true link. The higher the linkage weight $\theta_{jk}$ is, the more likely the pair $(j,k)$ is a true link. The linkage weight $\theta_{jk}$ is commonly based on the ratios of the conditional probabilities of having a match $\mu$ and an unmatch $\overline{\mu}$, given the result of the outcome of comparison $C_q$ of the characteristic $q$ of the records $j$ from A and $k$ from B, $q=1,...,Q$.

Once a linkage weight $\theta_{jk}$ has been computed for each pair *(j,k)* of ***AXB***, we need to decide whether the linkage weight is sufficiently large to consider the pair *(j,k)* a link. This is typically done using a decision rule. With the approach of Fellegi and Sunter (1969), we use an upper threshold $\theta_{High}$ and a lower threshold $\theta_{Low}$ to which each linkage weight $\theta_{jk}$ is compared. The decision is made as follows:

$$D(j,k) = \begin{cases} \text{link} & \text{if } \theta_{jk} \geq \theta_{High} \\ \text{can be a link} & \text{if } \theta_{Low} < \theta_{jk} < \theta_{High} \\ \text{nonlink} & \text{if } \theta_{jk} \leq \theta_{Low} \end{cases} \quad (2.1)$$

The lower and upper thresholds $\theta_{Low}$ and $\theta_{High}$ are determined by *a priori* error bounds based on false links and false nonlinks. When applying decision rule (2.1), some clerical decisions will be needed for those linkage weights falling between the lower and upper thresholds. This is generally done by looking at the data, and also by using auxiliary information. By being automated and also by working on a probabilistic basis, some errors could be introduced in the record linkage process. This has been discussed in several papers, namely Barlett *et al.* (1993), Belin (1993) and Winkler (1995).

The application of decision rule (2.1) leads to the definition of an indicator variable $l_{jk} = 1$ if the pair *(j,k)* is considered to be a link, 0 otherwise. As for the decisions that need to be taken for those linkage weights falling between the lower and upper thresholds, some manual intervention may be needed to decide on the validity of the links. Note that decision rule (2.1) does not prevent the existence of many-to-one or one-to-many links.

## 3. THE GENERALIZED WEIGHT SHARE METHOD

The GWSM is described in Lavallée (1995). It is an extension of the Weight Share Method described by Ernst (1989) but in the context of longitudinal household surveys. The GWSM can be viewed as a generalization of *Network Sampling* and also of *Adaptive Cluster Sampling*. These two sampling methods are described in Thompson (1992), and Thompson and Seber (1996).

Suppose that a sample $s^A$ of $m^A$ units is selected from the population $U^A$ of $M^A$ units using some sampling design. Let $\pi_j^A$ be the selection probability of unit *j*. We assume $\pi_j^A > 0$ for all $j \in U^A$.

Let the population $U^B$ contain $M^B$ units. This population is divided into *N* clusters where cluster *i* contains $M_i^B$ units. From population $U^B$, we are interested in estimating the total $Y^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$ for some characteristic *y*.

An important constraint that is imposed in the measurement (or interviewing) process is to consider all units within the same cluster. That is, if a unit is selected in the sample, then every unit of the cluster containing the selected unit will be interviewed. This constraint is one that often arises in surveys for two reasons: cost reductions and the need for producing estimates on clusters. As an example, for social surveys, there is normally a small marginal cost for interviewing all persons within the household. On the other hand, household estimates are often of interest with respect to poverty measures, for example.

With the GWSM, we make the following assumptions:

1) There exists a link between each unit *j* of population $U^A$ and at least one unit *k* of cluster *i* of population $U^B$, i.e. $L_j^A = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} \geq 1$ for all $j \in U^A$.
2) Each cluster *i* of $U^B$ has at least one link with a unit *j* of $U^A$.
3) There can be zero, one or more links for a unit *k* of cluster *i* population $U^B$, i.e. it is possible to have $L_{ik} = \sum_{j \in U^A} l_{j,ik} = 0$ or $L_{ik} = \sum_{j \in U^A} l_{j,ik} > 1$ for some $k \in U^B$.

We will see in Section 4 that in the context of record linkage, some of these assumptions might not be satisfied.

By using the GWSM, we want to assign an estimation weight $w_{ik}$ to each unit *k* of an interviewed cluster *i*. To estimate the total $Y^B$ belonging to population $U^B$, one can then use the estimator

$$\hat{Y} = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \quad (3.1)$$

where *n* is the number of interviewed clusters and $w_{ik}$ is the weight attached to unit *k* of cluster *i*. With the GWSM, the estimation process uses the sample $s^A$ together with the links existing between $U^A$ and $U^B$ to estimate the total $Y^B$. The links are in fact used as a bridge

to go from population $U^A$ to population $U^B$, and vice versa.

The GWSM allocates to each sampled unit a final weight established from an average of weights calculated within each cluster $i$ entering into $\hat{Y}$. An *initial weight* that corresponds to the inverse of the selection probability is first obtained for unit $k$ of cluster $i$ of $\hat{Y}$ having a non-zero link with a unit $j \in s^A$. An initial weight of zero is assigned to units not having a link. The *final weight* is obtained by calculating the ratio of the sum of the initial weights for the cluster over the total number of links for that cluster. This final weight is finally assigned to all units within the cluster. Note that the fact of allocating the same estimation weight to all units has the considerable advantage of ensuring consistency of estimates for units and clusters.

Formally, each unit $k$ of cluster $i$ entering into $\hat{Y}$ is assigned an initial weight $w'_{ik}$ as :

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A} \qquad (3.2)$$

where $t_j = 1$ if $j \in s^A$ and 0 otherwise. Note that a unit $k$ having no link with any unit $j$ of $U^A$ has automatically an initial weight of zero. The final weight $w_i$ is given by

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}} \qquad (3.3)$$

where $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik}$. The quantity $L_{ik}$ represents the number of links between the units of $U^A$ and the unit $k$ of cluster $i$ of population $U^B$. The quantity $L_i = \sum_{k=1}^{M_i^B} L_{ik}$ then corresponds to the total number of links present in cluster $i$. Finally, we assign $w_{ik} = w_i$ for all $k \in i$ and uses equation (3.1) to estimate the total $Y^B$.

Now, let $z_{ik} = Y_i / L_i$ for all $k \in i$. As shown in Lavallée (1995), $\hat{Y}$ can also be written as

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} z_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \qquad (3.4)$$

Using this last expression, it can easily be shown that the GWSM is design unbiased. The variance of $\hat{Y}$ is directly given by $Var(\hat{Y}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'}$ where $\pi_{jj'}^A$ is the joint probability of selecting units $j$ and $j'$ (See Särndal, Swensson and Wretman (1992) for the calculation of $\pi_{jj'}^A$ under various sampling designs).

## 4. THE GWSM AND RECORD LINKAGE

With record linkage, the links $l_{j,ik}$ have been established between files A and B, or population $U^A$ and population $U^B$, using a probabilistic process. As mentioned before, record linkage uses a decision rule $D$ such as (2.1) to decide whether there is a link or not between unit $j$ from file A and unit $k$ from file B. Once the links are established, we have seen that it is then possible to estimate the total $Y^B$ from population $U^B$ using a sample obtained from population $U^A$. One could then ask if it is necessary to make such a decision. That is, is it necessary to establish whether there is positively a link for the given pair $(j,k)$, or not? Would it be easier to simply use the linkage weights $\theta_{jk}$ (without using any decision rule) to estimate the total $Y$ from $U^B$ using a sample from $U^A$? If this were the case, it is easy to see that reducing the amount of clerical intervention required in the record linkage process could save time and resources. In the present section, we will see that the GWSM can be used to answer the previous question. Three methods will be considered.

Method 1: Use all possible links with their respective linkage weights

When using all possible links with the GWSM, one wants to give more importance to links that have large linkage weights $\theta$ than those that have small linkage weights. We no longer use the indicator variable $l_{j,ik}$ identifying whether there is a link or not between unit $j$ from $U^A$ and unit $k$ of cluster $i$ from $U^B$. Instead, we use the linkage weight $\theta_{j,ik}$ obtained in the first steps of the record linkage process. By doing so, we do not need any decision to be taken to establish whether there is a link or not between two units.

By definition, for each pair $(j,ik)$ of ***AXB***, the linkage weight $\theta_{j,ik}$ reflects the degree to which the pair $(j,ik)$ is likely to be a true link. This linkage weight can then directly replace the indicator variable $l$ in equations (3.2) and (3.3) that define the estimation weight obtained through the GWSM. We then get the estimation weight $w_{ik}^{RL}$.

The assumptions of Section 3 concerning the GWSM still apply. For instance, the existence of a link between each unit $j$ of population $U^A$ and at least one unit $k$ of population $U^B$ is translated into the need of having a non-zero linkage weight $\theta_{j,ik}$ between each unit $j$ of $U^A$ and at least one unit $k$ of cluster $i$ of $U^B$. The assumption that each cluster $i$ of $U^B$ must have at least one link with a unit $j$ of $U^A$ translates into the need of having for each cluster $i$ of $U^B$ at least one non-zero linkage weight $\theta_{j.ik}$ with a unit $j$ of $U^A$. Finally, there can be a zero linkage weight $\theta_{j,ik}$ for a unit $k$ of cluster $i$ of population $U^B$. In theory, the record linkage process does not insure that these constraints are satisfied. This is because the decision rule (2.1) does not prevent to have many-to-one or one-to-many links, or no link at all. For example, it might turn out that for a cluster $i$ of $U^B$, there is no non-zero linkage weight $\theta_{j,ik}$ with any unit $j$ of $U^A$. In that case, the estimation weight $w_{ik}^{RL}$ underestimates the total $Y^B$. To solve this problem, one practical solution is to collapse two clusters in order to get at least one non-zero linkage weight $\theta_{j,ik}$ for cluster $i$. Unfortunately, this solution might require some manual intervention, which we try to avoid. A better solution is to force to have a link by choosing one link at random within the cluster.

Method 2: Use all possible links above a given threshold

Using all possible links with the GWSM as in Method 1 might require the manipulation of large files of size $M^A \times M^B$. This is because it might turn out that most of the records between files A and B have non-zero linkage weights $\theta$. In practice, even if this happens, we can expect that most of these linkage weights will be relatively small or negligible to the extent that, although non-zero, the links are very unlikely to be true links. In that case, it might be useful to only consider the links with a linkage weight $\theta$ above a given threshold $\theta_{High}$.

For this method, we again no longer use the indicator variable $l_{j,ik}$ identifying whether there is a link or not, but instead, we use the linkage weight $\theta_{j,ik}$ obtained in the first steps of the record linkage process and above the threshold $\theta_{High}$. The linkage weights below the threshold are considered as zeros. We therefore define the linkage weight: $\theta_{j,ik}^T = \theta_{j,ik}$ if $\theta_{j,ik} \geq \theta_{High}$, 0 otherwise. The estimation weight $w_{ik}^{RLT}$ is then directly obtained by replacing the indicator variable $l$ in equations (3.2) and (3.3) by $\theta_{j,ik}^T$.

The number of zero linkage weights $\theta^T$ will be greater than or equal to the number of zero linkage weights $\theta$ used for Method 1. Therefore, the assumption of having a non-zero linkage weight $\theta_{j,ik}^T$ between each unit $j$ of $U^A$ and at least one unit $k$ of $U^B$ might be more difficult to satisfy. The assumption that each cluster $i$ of $U^B$ must have at least one non-zero linkage weight $\theta_{j,ik}^T$ with a unit $j$ of $U^A$ can also possibly not be satisfied. In that case, the estimation weight $w_{ik}^{RLT}$ underestimates the total $Y^B$. To solve this problem, one solution is to force the selection of the link with the largest linkage weights $\theta$ within the cluster. This will lead to accepting links with weights $\theta^T$ below the threshold. If there is still no link, then choose one link at random within the cluster is a possible solution.

Method 3: Choose the links by random selection

In order to avoid taking a decision on whether there is a link or not between unit $j$ from $U^A$ and unit $k$ of cluster $i$ from $U^B$, one can decide to simply choose the links at random from the set of possible links. For this, it is reasonable to choose the links with probabilities proportional to the linkage weights $\theta$. This can be achieved by Bernoulli trials where, for each pair *(j,ik)*, we decide on accepting a link of not by generating a random number $u_{j,ik}$ that is compared to the linkage weight $\theta_{j,ik}$.

The first step before performing the Bernoulli trials is to rescale the linkage weights in order to restrict them to the [0,1] interval. This can be done by dividing each linkage weight $\theta_{j,ik}$ by the maximum possible value $\theta_{Max}$. Although in most practical situations, the value $\theta_{Max}$ exists, it is not the case in general. When this is not possible, one can then use a transformation such as the inverse logit function $f(x) = e^x/(1+e^x)$ to force the adjusted linkage weights $\tilde{\theta}$ to be in the [0,1] interval. The chosen function should have the desirable property that the adjusted linkage weights $\tilde{\theta}$ sum to the expected total number of links $L$ in ***AXB***, i.e.

$$\sum_{j=1}^{M^A}\sum_{i=1}^{N}\sum_{k=1}^{M_i^B}\tilde{\theta}_{j,ik} = L .$$

Once the adjusted linkage weights $\tilde{\theta}_{j,ik}$ have been obtained, for each pair *(j,ik)*, we generate a random number $u_{j,ik} \sim U(0,1)$. Then, we set the indicator variable $\tilde{l}_{j,ik}$ to 1 if $u_{j,ik} \leq \tilde{\theta}_{j,ik}$, and 0 otherwise. This process provides a set of links similar to the ones used in the original version of the GWSM, with the exception that now the links have been determined randomly instead of through a decision process comparable to (2.1). The estimation weight $\tilde{w}_{ik}$ is then directly obtained by replacing the indicator variable $l$ in equations (3.2) and (3.3) by $\tilde{l}_{j,ik}$.

By conditioning on the accepted links $\tilde{l}$, it can be shown that the resulting estimator is conditionally design unbiased and hence, unconditionally design unbiased. Note that by conditioning on $\tilde{l}$, this estimator is then equivalent to (3.1). To get the variance of $\hat{\tilde{Y}}$, again conditional arguments need to be used.

With the present method, by randomly selecting the links, it is very likely that one or more of the three assumptions of Section 3 will not be satisfied. For example, for a given unit *j* of population $U^A$, there might not be a link with any unit *k* of population $U^B$. Again, in practice, we can overcome this problem by forcing a link for all unit *j* of population $U^A$ by choosing the link that has the highest linkage weight $\theta_{j,ik}$. The constraint that each cluster *i* of $U^B$ must have has at least one link with a unit *j* of $U^A$ can also be not satisfied. Again, we can force to have a link by choosing the one with the highest linkage weight $\theta_{j,ik}$ within the cluster. If there is still no link, it is possible to choose one link at random within the cluster. It should be noted that this solution preserves the design unbiasedness of the GWSM.

## 5. SIMULATION STUDY

A simulation study has been performed to evaluate the proposed methods against the classical approach (Fellegi-Sunter) where the decision rule (2.1) is used to determine the links. This study was made by comparing the design variance obtained for the estimation of a total $Y^B$ using four different methods: (1) use all links; (2) use all links above a threshold; (3) choose links randomly using Bernoulli trials; (4) Fellegi-Sunter. Given that all four methods yield design unbiased estimates of the total $Y^B$, the quantity of interest for comparing the various methods was the standard error of the estimate, or simply the coefficient of variation (i.e., the ratio of the square root of the variance to the expected value).

For the record linkage step, data from the 1996 Farm Register (population $U^A$) was linked to the 1996 Unincorporated Revenue Canada Tax File (population $U^B$). The Farm Register is essentially a list of all records collected during the 1991 Census of Agriculture with all the updates that have occurred since 1991. It contains a farm operator identifier together with some socio-demographic variables related to the farm operators. The 1996 Unincorporated Revenue Canada Tax File contains data on tax filers declaring at least one farming income. It contains a household identifier, a tax filer identifier, and also socio-demographic variables related to the tax files.. For the purpose of the simulations, the province of New Brunswick was considered. For this province, the Farm Register contains 4,930 farm operators while the Tax File contains 5,155 tax filers.

The linkage process used for the simulations was a match using five variables. It was performed using the statement MERGE in SAS®. All records on both files were compared to one another in order to see if a potential match had occurred. The record linkage was performed using the following five key variables common to both sources: (1) first name (modified using NYSIIS); (2) last name (modified using NYSIIS); (3) birth date; (4) street address; (5) postal code. The first name and last name variables were modified using the NYSIIS system. This basically changes the name in phonetic expressions, which in turn increases the chance of finding matches by reducing the probability that a good match is rejected because of a spelling mistake or a typo.

Records that matched on all 5 variables received the highest linkage weight ($\theta$ =60). Records that matched on only a subset of at least 2 of the 5 variables received a lower linkage weight (as low as $\theta$ =2). Records that did not match on any combination of key variables were not considered as possible links, which is equivalent as having a linkage weight of zero. A total number of 13,787 possible links were found.

Two different thresholds were used for the simulations: $\theta_{High} = \theta_{Low}$ =15 and $\theta_{High} = \theta_{Low}$ =30. The upper and lower thresholds, $\theta_{High}$ and $\theta_{Low}$, were set to be the same to avoid the gray area where some manual intervention is needed when applying the decision rule

(2.1) of Fellegi-Sunter.

For the simulations, we assumed that the sample from $U^A$ (i.e. the Farm Register) would be selected using Simple Random Sampling Without Replacement (SRSWOR), without any stratification. We also considered two sampling fractions: 30% and 70%. The quantity of interest $Y$ to be estimated is the Total Farming Income. It was possible for us to calculate the theoretical variance for these estimates for various sampling fractions. We could also estimate this variance by simulations (i.e. performing a Monte-Carlo study). Both approaches were used. For the simulations, 500 simple random samples were selected for each method for two different sampling fractions (30% and 70%). The two thresholds (15 and 30) were also used to better understand the properties of the given estimators. The results of the study are presented in Figures 1 and 2.

***Figure 1. CVs with $\delta_{High} = \delta_{Low}$ =15.***

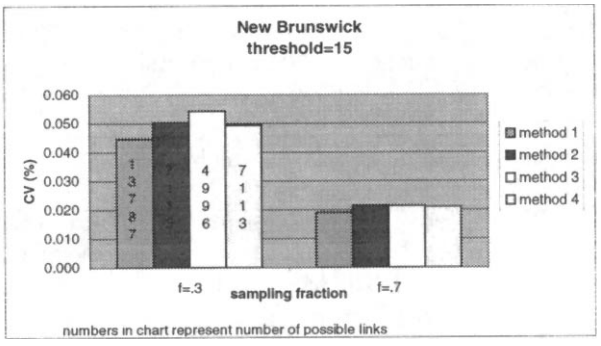
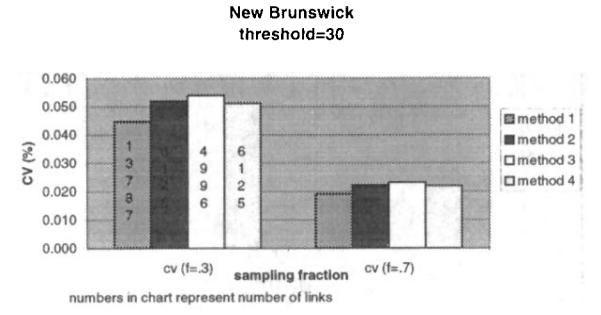



***Figure 2. CVs with $\delta_{High} = \delta_{Low}$ =30.***




By looking at the above figures, it can be seen that in all cases, Method 1 provided the smallest design variances for the estimation of the Total Farming Income. Therefore, using all possible links leads to greatest precision. This results is important since it indicates that, in addition to saving resources by eliminating manual interventions needed to decide on the links, we also gain in the precision of the estimates. This conclusion also seems to hold regardless of the sampling fraction, the threshold and the province.

As previously mentioned, the number of links to handle using Method 1 might be quite large. Therefore, a compromise method such as Method 2 (use all links above a threshold) or Method 3 (choose links randomly using Bernoulli trials) might be appealing. The precision of Method 2 seems to be comparable to Method 4 (Fellegi-Sunter). Using Method 2 can then be chosen because, unlike Method 4, it does not involve any decision rule in terms of the validity of the links.

Method 3 turned out to have the largest variance. Therefore, choosing the links randomly using Bernoulli trials does not seem to help with respect to precision. However, Method 3 is the one that used the least number of links. If this is of concern, this method can turn out to be appealing in some situations.

## BIBLIOGRAPHY


Barlett, S. , Krewski, D., Wang, Y., Zielinski, J.M. (1993). Evaluation of Error Rates in Large Scale Computerized Record Linkage Studies. *Survey Methodology*, Vol. 19, No. 1, pp. 3-12.

Belin, T.R. (1993). Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment. *Survey Methodology*, Vol. 19, No. 1, pp. 13-29.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P., Editors), John Wiley and Sons, New York, pp. 135-159.

Fellegi, I.P., Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, Vol. 64, pp. 1183-1210.

Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method, *Survey Methodology*, Vol. 21, No. 1, pp. 25-32.

Newcome, H.B., Kennedy, J.M., Axford, S.J., James, A.P. (1959), Automatic Linkage of Vital Records. *Science*, Vol. 130, pp. 954-959.

Särndal, C.-E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Thompson, S.K. (1992), *Sampling.* John Wiley and Sons, New York.

Thompson, S.K., Seber, G.A. (1996), *Adaptive Sampling.* John Wiley and Sons, New York.

Winkler, W.E. (1995). Matching and Record Linkage. In *Business Survey Methods* (Cox, Binder, Chinnappa, Colledge and Kott, Editors), John Wiley and Sons, New York, pp. 355-384.