CHAPTER 9

GWSM AND RECORD LINKAGE

Data from different sources are increasingly being combined to augment the amount of information that we have. Often, the databases are combined using record linkage. 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 from the second file with a certain probability. Then, a decision is made on whether this link is a true link or not. Note that this process usually requires a certain amount of manual resolution.

We again consider the production of an estimate of a total of one target clustered population U^B when using a sample s^A selected from another population U^A that is linked to the first population. However, we assume that the two populations have been linked together using probabilistic record linkage. Note that this type of linkage often leads to a complex linkage between the two populations.

In this chapter, we will try to answer the following questions:

a) Can we use the GWSM to handle the estimation problems related to populations linked together through record linkage?
b) Can we adapt the GWSM to take into account the linkage weights issued from record linkage?
c) Can the GWSM help in reducing the manual resolution required by record linkage?
d) If there is more than one approach to use the GWSM, is there a "better" approach?

It will be seen that the answer is clearly yes to (a) and (b). However, for question (c), it will be shown that there is unfortunately a

price to pay in terms of an increase to the sample size, and therefore to the collection costs. For question (d), although there is no definite answer, some approaches seem to generally be more appropriate.

9.1 RECORD LINKAGE

The concepts of *record linkage* were introduced by Newcome *et al.*. (1959), and formalised in the mathematical model of Fellegi and Sunter (1969). As described by Bartlett *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 called *exact matching*, in contrast to *statistical matching*. This last process attempts to link files that have few units in common. In this case, linkages are based on similar characteristics rather than unique identifying information. To learn more about statistical matching, see Budd and Radner (1969), Budd (1971), Okner (1972) and Singh *et al.* (1993). In this chapter, we will restrict ourselves to the context of record linkage. However, the theory presented can also be used for statistical matching.

Suppose that we have two files A and B containing characteristics respectively relating to two populations U^A and U^B. The two populations are related in a way. They can represent, for example, exactly the same population, where each of the files contains a different set of characteristics of the units of that population. They can also represent different populations, but naturally linked to one another. For example, one population can be one of parents, and the other population one of children belonging to the parents, as illustrated in Figure 1.2. Note that the children usually live in households that can be viewed as clusters.

Another example is one of the creation of Statistics Canada's Whole Farm Database. This example was presented before in section 7.4.4. The first population is a list of farms from the Canadian Census of Agriculture, and the second population is a list of taxation records (or income tax reports) from the Canada Revenue Agency (CRA). In the first population, each farm is identified by a unique identifier called the FarmID and some additional variables such as the name and address of the farm operators that are obtained from the Census questionnaire. The second population consists of tax reports of individuals having declared some form of agricultural income. These individuals live in households (or clusters). The unique identifier on those records is a corporation number or a social insurance number, depending on whether or not the

business is incorporated. Note that each income tax report submitted to CRA contains similar variables (name and address of respondent, etc.) as those obtained by the Census of Agriculture.

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, coming respectively from each file, are linked together or not. With this linkage process, the probability of having a real match between two records is calculated. Based on the magnitude of this probability, it is then decided whether they can be considered as really being linked together or not.

Formally, we consider the product space $A \times B$ from the two files *A* and *B*. Let *j* indicate 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 $A \times B$, we compute a *linkage weight* reflecting the degree to which the pair *(j,k)* has a true link. The higher the linkage weight is, the more likely the pair *(j,k)* has a true link. The linkage weight is commonly based on the ratio of the conditional probabilities of having a match υ and an unmatch $\overline{\upsilon}$, given the result of the outcome of the comparison $\Delta_{\varsigma jk}$ of the characteristic ς of record *j* from *A* and *k* from *B*, $\varsigma = 1, ..., p$. Thus, the linkage weight can be defined as follows:

$$\begin{aligned} \dot{\theta}_{jk} &= \log_2 \left\{ \frac{P(\upsilon_{jk} \mid \Delta_{1jk}\Delta_{2jk}...\Delta_{pjk})}{P(\overline{\upsilon}_{jk} \mid \Delta_{1jk}\Delta_{2jk}...\Delta_{pjk})} \right\} \\ &= \dot{\theta}_{1jk} + \dot{\theta}_{2jk} + ... + \dot{\theta}_{pjk} + \dot{\theta}_{\bullet jk} \end{aligned} \tag{9.1}$$

where $\dot{\theta}_{\varsigma jk} = \log_2 \left\{ \frac{P(\Delta_{\varsigma jk} \mid \upsilon_{jk})}{P(\Delta_{\varsigma jk} \mid \overline{\upsilon}_{jk})} \right\}$ for $\varsigma = 1, ..., p$, and $\dot{\theta}_{\bullet jk} = \log_2 \left\{ \frac{P(\upsilon_{jk})}{P(\overline{\upsilon}_{jk})} \right\}$.

The mathematical model proposed by Fellegi and Sunter (1969) considers the probabilities of an error in the linkage of units *j* from *A* and *k* from *B*. The linkage weight is then defined as

$$\theta_{jk}^{FS} = \sum_{\varsigma=1}^{p} \theta_{\varsigma jk}^{FS},$$

where $\theta_{\varsigma jk}^{FS} = \log_2(\varphi_{\varsigma jk} / \overline{\varphi}_{\varsigma jk})$ if characteristic ς of pair *(j,k)* is linked, and $\theta_{\varsigma jk}^{FS} = \log_2((1-\varphi_{\varsigma jk})/(1-\overline{\varphi}_{\varsigma jk}))$ otherwise. The expressions used here are $\varphi_{\varsigma jk} = P(\Delta_{\varsigma jk} | \upsilon_{jk})$ and $\overline{\varphi}_{\varsigma jk} = P(\Delta_{\varsigma jk} | \overline{\upsilon}_{jk})$. Moreover, it is assumed that the p comparisons are independent.

The linkage weights given by (9.1) are defined on the set $\Re$ of real numbers, i.e., $\dot{\theta}_{jk} \in]-\infty,+\infty[$. When the ratio of the conditional probabilities of having a match υ_{jk} and an unmatch $\overline{\upsilon}_{jk}$ is equal to 1, we get $\dot{\theta}_{jk} = 0$. When this ratio is close to 0, $\dot{\theta}_{jk}$ approaches $-\infty$. It can however be practical to define the linkage weights on $[0,+\infty[$. This can be achieved by taking the antilogarithm of $\dot{\theta}_{jk}$. We then obtain the following linkage weight θ_{jk}:

$$\theta_{jk} = \frac{P(\upsilon_{jk} \mid \Delta_{1jk}\Delta_{2jk}...\Delta_{pjk})}{P(\overline{\upsilon}_{jk} \mid \Delta_{1jk}\Delta_{2jk}...\Delta_{pjk})}. \tag{9.2}$$

Note that the linkage weight θ_{jk} is equal to 0 when the conditional probabilities of having a match υ_{jk} are equal to 0. In other words, we have $\theta_{jk} = 0$ when the probability of having a true link for *(j,ik)* is zero.

Once a linkage weight θ_{jk} has been computed for each pair *(j,k)* of $A \times B$, we need to decide whether the linkage weight is sufficiently large to consider the pair *(j,k)* as being linked. For this, a *decision rule* is generally used. With the approach of Fellegi and Sunter (1969), we choose an upper threshold θ_{High} and a lower threshold θ_{Low} to which each linkage weight θ_{jk} is compared. The decision is made as follows:

$$D(j,k) = \begin{cases} \text{link} & \text{if } \theta_{jk} \geq \theta_{High} \\ \text{possible link} & \text{if } \theta_{Low} < \theta_{jk} < \theta_{High} \\ \text{non-link} & \text{if } \theta_{jk} \leq \theta_{Low}. \end{cases} \tag{9.3}$$

The lower and upper thresholds θ_{Low} and θ_{High} are determined by error bounds that are determined prior to the record linkage process, based on false links and false non-links. When applying the decision rule (9.3), a manual resolution is necessary to make a decision concerning the pairs whose linkage weights are between the lower and upper thresholds. This is generally done by looking at the data, and also by using auxiliary

information. In the agriculture example, variables such as date of birth, address and postal code, which are available on both files, are used for this purpose. The application of decision rule (9.3) leads to the definition of an indicator variable l_{jk} such that $l_{jk} = 1$ if the pair *(j,k)* is considered to be a link, and 0 otherwise. Note that the decision rule (9.3) does not prevent the existence of complex links such as those illustrated in Figure 2.1.

By using an automated system and by applying a probabilistic method, the record linkage process can contain some errors. This problem has been discussed in several papers, namely Bartlett *et al.* (1993), Belin (1993) and Winkler (1995). Linkage errors are out of the scope of this book, and thus will only be briefly covered in certain occasions in this chapter.

9.2 GWSM ASSOCIATED WITH RECORD LINKAGE

Let U^A be the population containing M^A units and U^B be the population consisting of N clusters where each cluster i contains M_i^B units. With record linkage, linked are established between the populations U^A and U^B using a probabilistic process. As mentioned previously, record linkage uses a decision rule D such as the one given by (9.3) to decide whether or not there is a link between unit j from U^A and unit ik from U^B. Once the links are established, we then have two populations U^A and U^B linked together and where the links are identified by the indicator variable $l_{j,ik}$. Recall that the decision rule (9.3) does not prevent complex links from being obtained.

Although the links can be complex, the GWSM can be used to estimate the total Y^B from population U^B using a sample s^A obtained from population U^A. Therefore, the answer is yes to question (a) expressed at the beginning of this chapter. The GWSM used with populations U^A and U^B linked together by record linkage with decision rule (9.3) will be called, in the rest of the chapter, the *classical approach*.

It should be noted that these estimates obtained by the application of the GWSM can be biased if Constraint 2.1 presented in chapter 2 is not satisfied. In this case, estimator (2.1) underestimates the total Y^B. To resolve this problem, a practical solution is to group two clusters so that at least one non-zero link $l_{j,ik}$ is obtained for each cluster i. This solution generally requires manual resolution. Another solution is to create, or

impute, a link by randomly choosing a link within the cluster. The link with the largest linkage weight $\theta_{j,ik}$ can also be chosen. Note that for a unit j from U^A, there may only be links $l_{j,ik} = 0$ with all units ik from U^B. However, this is not a problem since we are only interested in the coverage of the target population U^B, and not the one of U^A.

Now, with the classical approach, the use of the GWSM is based on links identified by the indicator variable $l_{j,ik}$. Is it necessary to establish whether or not there is positively a link for each pair *(j,ik)*? Would it be easier to use the linkage weights $\theta_{j,ik}$ (without decision rules) to estimate the total Y^B? These questions lead to question (b), that is, if it is possible to adapt the GWSM to take into account the linkage weights issued from record linkage. The answer to this question is yes, as it was shown in section 4.5 that it was possible to extend the use of the GWSM to weighted links.

Recall that by presenting the WSM in the context of longitudinal surveys, Ernst (1989) proposed the use of constants α in the definition of estimation weights. Setting $\tilde{\theta}_{j,ik} = \theta_{j,ik} / \theta_i^B$, where $\theta_i^B = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \theta_{,ik}$, a version of the GWSM was obtained in section 4.5, constructed from these constants. Coming back to the context of longitudinal surveys, we saw in section 3.3 that Kalton and Brick (1995) looked at the determination of optimal values for the constants α of Ernst (1989) by looking to minimise the variance. They concluded that: "in the two-household case, the equal household weighting scheme minimises the variance of the household weights around the inverse selection probability weight when the initial sample is an epsem[1] one". They also added that: "in the case of an approximately epsem sample, the equal household weighting scheme should be close to the optimal, at least for the case where the members of the household at time t come from one or two households at the initial wave". Recall that if s^A is a sample of persons, considering the fact that the persons represent households of size 1, the equal weighting of households and the equal weighting of persons are equivalent, which corresponds to the fair share method describe in section 3.2. This suggests that, for the version of the GWSM described in section 4.5, we should be close to the optimal values by setting the values of the constants α to zero for some units and to an

[1] "epsem" stands for equal probability selection method.

equal positive value for all other units of the cluster. As with $\alpha_{j,ik} = l_{j,ik} / L_i^B$, the desired types of values are directly obtained, and the classical approach should then produce variances close to the minimum for the estimate of the total Y^B. This result was proved in a formal way in section 4.6.3 for the case of simple random sampling.

In the present section, three different approaches are given where the GWSM uses the linkage weights $\theta_{j,ik}$. The first approach is to use all the non-zero links (i.e., with $\theta_{j,ik} > 0$) identified through the record linkage process with their respective linkage weights. The second approach is the one where we use all the non-zero links with linkage weights above a given threshold θ_{High}. The third approach consists of randomly choosing the links proportionally to $\theta_{j,ik}$.

9.2.1 Approach 1: use all non-zero links with their respective linkage weights

With the use of all non-zero links with the GWSM, it can be justified to give more importance to the links that have a larger linkage weight $\theta_{j,ik}$, compared to those that have a small linkage weight. By definition, for each pair *(j,ik)* obtained from crossing populations U^A and U^B, the linkage weight $\theta_{j,ik}$ reflects the tendency of the pair *(j,ik)* to have a true link. In this case, instead of using the indicator variable $l_{j,ik}$ identifying whether or not there is a link between unit *j* from U^A and unit *k* of cluster *i* from U^B, we can use the linkage weight $\theta_{j,ik}$ obtained in the first steps of the record linkage process. Note that this implies the elimination of the manual resolution since no decision rule is used.

The application of this approach assumes, of course, that the file with the linkage weights is available. In practice, the only file available is often the final file, once the linkage process ends, after manual resolution. In this case, the linkage weights are not generally available (only the indicator variables $l_{j,ik}$ remain) and the three proposed approaches are then no longer pertinent.

For each unit *j* selected in s^A, we identify the units *ik* of U^B that have a non-zero linkage weight with unit *j*, i.e., $\theta_{j,ik} > 0$. Let $\Omega^{RL,B} = \{i \in U^B \mid \exists j \in s^A \text{ and } \theta_{j,i} > 0\}$ with $\theta_{j,i} = \sum_{k=1}^{M_i^B} \theta_{j,ik}$ be the set of n^{RL} clusters identified by the units $j \in s^A$, where "RL" stands for record

linkage. Note that because we use all linkage weights greater than zero, we have $n^{RL} \geq n$, where n is the number of clusters identified by the classical approach.

To estimate the total Y^B of the population U^B, one can use the estimator

$$\hat{Y}^{RL,B} = \sum_{i=1}^{n^{RL}} \sum_{k=1}^{M_i^B} w_{ik}^{RL} y_{ik} \tag{9.4}$$

where w_{ik}^{RL} is the estimation weight obtained from the GWSM. This weight is obtained by directly replacing the indicator variable $I_{j,ik}$ with the linkage weight $\theta_{j,ik}$ in the steps of the GWSM described in chapter 2. The following steps are then obtained.

Steps of the GWSM for approach 1

Step 1: For each unit k of the clusters i from $\Omega^{RL,B}$, the initial weight $w_{ik}^{\prime RL}$ is calculated, that is:

$$w_{ik}^{\prime RL} = \sum_{j=1}^{M^A} \theta_{j,ik} \frac{t_j}{\pi_j^A} \tag{9.5}$$

where $t_j = 1$ if $j \in s^A$, and 0 otherwise. Note that a unit ik having no link with any unit j from U^A automatically has an initial weight of zero.

Step 2: For each unit k of the clusters i from $\Omega^{RL,B}$, we calculate

$$\theta_{ik}^B = \sum_{j=1}^{M^A} \theta_{j,ik} .$$

Step 3: The final weight w_i^{RL} is calculated:

$$w_i^{RL} = \frac{\sum_{k=1}^{M_i^B} w_{ik}^{\prime RL}}{\sum_{k=1}^{M_i^B} \theta_{ik}^B} . \tag{9.6}$$

Step 4: Finally, we set $w_{ik}^{RL} = w_i^{RL}$ for all $k \in U_i^B$.

It should be noted that because they are present in the numerator and the denominator, the linkage weights do not need to be between 0 and 1. They just need to represent the likelihood of having a link between two units from populations U^A and U^B. It is also interesting to see that

with $\tilde{\theta}_{j,ik} = \theta_{j,ik} / \theta_i^B$, where $\theta_i^B = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \theta_{j,ik}$, we obtain an equivalent formulation to the one coming from the generalisation of the estimation weight described in section 4.5.

With the classical approach, each cluster *i* of U^B is assumed to have at least one non-zero link with a unit *j* of U^A. This constraint is translated here into the need of having, for each cluster *i* of U^B, at least one linkage weight $\theta_{j,ik}$ greater than zero with a unit *j* of U^A. In theory, it is not guaranteed that this constraint will be satisfied following the record linkage process. For example, it is possible that for a cluster *i* of U^B, there is no linkage weight $\theta_{j,ik}$ greater than zero. In that case, the estimation weight (9.6) underestimates the total Y^B. To solve this problem, the same solutions proposed in the context of the indicator variables $l_{j,ik}$ can be used. That is, two clusters can be collapsed, for example, in order to get at least one linkage weight $\theta_{j,ik}$ greater than zero for the new cluster. Unfortunately, this solution may require manual intervention, which has been avoided up to now by not using a decision rule. A better solution is to impute a link by choosing one link at random within the cluster. Then, a small value $\theta_{j,ik} > 0$ can be assigned arbitrarily for the chosen link.

Following the same steps as those from the proof of Theorem 4.1, the estimator $\hat{Y}^{RL,B}$ given by (9.4) can be rewritten in the following way:

$$\hat{Y}^{RL,B} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} \theta_{j,ik} z_{ik}^{RL} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j^{RL} \qquad (9.7)$$

where $z_{ik}^{RL} = Y_i / \theta_i^B$ for all $k \in U_i^B$, and $\theta_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} \theta_{j,ik}$.

With this last expression, it can be shown that the estimator $\hat{Y}^{RL,B}$ is unbiased using the same development as Corollary 4.1. Finally, by following Corollary 4.2, the variance of $\hat{Y}^{RL,B}$ is given by

$$Var(\hat{Y}^{RL,B}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j^{RL} Z_{j'}^{RL} . \qquad (9.8)$$

To estimate the variance (9.8), one of the two estimators (4.12) or (4.13) can be used by replacing the variable Z_j with Z_j^{RL}.

9.2.2 Approach 2: use all non-zero links above a given threshold

The use of the GWSM for all non-zero links might require the manipulation of very large files of size $M^A \times M^B$. This can occur if almost all pairs (j,ik) between populations U^A and U^B have non-zero linkage weights $\theta_{j,ik}$. In practice, even if this happens, it is strongly possible that most of these linkage weights will be very small or negligible. Even if the linkage weights are not non-zero, the links coming from these small linkage weights are probably not true. Indeed, looking at equation (9.2), we note that if $\theta_{j,ik}$ is very small, the conditional probability that there is a link between j and ik is then much smaller than the conditional probability that there is no link. In that case, it might be useful to only consider the links with linkage weights above a given threshold θ_{High}.

As with approach 1, we no longer use the indicator variables $l_{j,ik}$ identifying the links, but instead, we use the linkage weights $\theta_{j,ik}$ obtained in the first steps of the record linkage process. However, with approach 2, we restrict ourselves to the linkage weights greater than or equal to a threshold θ_{High}. The linkage weights below the threshold θ_{High} are considered as zeros. We therefore define the following linkage weight:

$$\theta_{j,ik}^{RLT} = \begin{cases} \theta_{j,ik} & \text{if } \theta_{j,ik} \geq \theta_{High} \\ 0 & \text{otherwise.} \end{cases} \tag{9.9}$$

For each unit j selected in s^A, we identify the units ik of U^B that have $\theta_{j,ik}^{RLT} > 0$. Let $\Omega^{RLT,B} = \{i \in U^B \mid \exists j \in s^A \text{ and } \theta_{j,i}^{RLT} > 0\}$ with $\theta_{j,i}^{RLT} = \sum_{k=1}^{M_i^B} \theta_{j,ik}^{RLT}$ be the set of the n^{RLT} clusters identified by the units $j \in s^A$, where "RLT" stands for record linkage with threshold. Note that $n^{RLT} \leq n^{RL}$. On the other hand, we have $n^{RLT} = n$ if the record linkage between U^A and U^B is done using the decision rule (9.3) with $\theta_{High} = \theta_{Low}$.

To estimate the total Y^B of population U^B, we can use the estimator

$$\hat{Y}^{RLT,B} = \sum_{i=1}^{n^{RLT}} \sum_{k=1}^{M_i^B} w_{ik}^{RLT} y_{ik}, \tag{9.10}$$

where w_{ik}^{RLT} is the estimation weight obtained from the GWSM. This weight is obtained by directly replacing the linkage weight $\theta_{j,ik}$ with the linkage weight $\theta_{j,ik}^{RLT}$ given by (9.9) in the steps of the GWSM described in section 9.2.1.

It is again interesting to see that with $\tilde{\theta}_{j,ik}^{RLT} = \theta_{j,ik}^{RLT} / \theta_i^{RLT,B}$, where $\theta_i^{RLT,B} = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \theta_{j,ik}^{RLT}$, a formulation is obtained that is equivalent to the one coming from the generalisation of the estimation weight described in section 4.5.

By definition, the number of zero linkage weights $\theta_{j,ik}^{RLT}$ will be greater than or equal to the number of zero linkage weights $\theta_{j,ik}$. The constraint that each cluster *i* of U^B must have at least one linkage weight $\theta_{j,ik}^{RLT}$ greater than zero with a unit *j* of U^A will thus be more difficult to satisfy. To solve this problem, the same solutions proposed in section 9.2.1 can be used. For example, two clusters can be collapsed in the same way to get at least one linkage weight $\theta_{j,ik}^{RLT}$ greater than zero for each cluster *i* of $\Omega^{RLT,B}$. Unfortunately, this solution can require manual intervention, which has been avoided up to now by not using any decision rule. A better solution is to impute a link by randomly choosing a link within the cluster. A value of $\theta_{j,ik}^{RLT}$ equal to the threshold θ_{High} can then be assigned to this link.

As in section 9.2.1, the estimator $\hat{Y}^{RLT,B}$ given by (9.10) can be rewritten in the following way:

$$\hat{Y}^{RLT,B} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} \theta_{j,ik}^{RLT} z_{ik}^{RLT} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j^{RLT}, \tag{9.11}$$

where $z_{ik}^{RLT} = Y_i / \theta_i^{RLT,B}$ for all $k \in U_i^B$, and $\theta_i^{RLT,B} = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} \theta_{j,ik}^{RLT}$.

With (9.11), the estimator $\hat{Y}^{RLT,B}$ can be proven to be unbiased using the same development as Corollary 4.1. Finally, by following Corollary 4.2, the variance of $\hat{Y}^{RLT,B}$ is given by

$$Var(\hat{Y}^{RLT,B}) = \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \frac{(\pi^A_{jj'} - \pi^A_j \pi^A_{j'})}{\pi^A_j \pi^A_{j'}} Z_j^{RLT} Z_{j'}^{RLT} . \qquad (9.12)$$

To estimate the variance (9.12), one of the two estimators (4.12) or (4.13) can be used by replacing the variable Z_j with Z_j^{RLT}.

9.2.3 Approach 3: choose the links randomly

In order to avoid making a decision on the links between units j from U^A and units k of clusters i from U^B, one can decide to simply choose the links at random from the set of all links with linkage weights $\theta_{j,ik}$ greater than zero. For this, it is reasonable to choose the links with probabilities proportional to the linkage weights. This can be done using *Bernoulli trials* where, for each pair(j,ik), we can decide to accept a link or not by generating a random number $v_{j,ik} \sim \mathrm{U}(0,1)$ that is then compared to a quantity proportional to the linkage weight $\theta_{j,ik}$.

In the point of view of record linkage, this approach cannot be considered as optimal. Indeed, when using the decision rule (9.3) of Fellegi and Sunter (1969), the idea is to minimise the number of false links and false non-links. The link $l_{j,ik}$ is accepted only if the linkage weight $\theta_{j,ik}$ is large (i.e., $\theta_{j,ik} \geq \theta_{High}$), or if it is moderately large (i.e., $\theta_{Low} < \theta_{j,ik} < \theta_{High}$) and has been accepted after manual resolution. The random selection of links using Bernoulli trials can lead to the selection of links that would have not been accepted through the decision rule (9.3), even though the selection probabilities are proportional to the linkage weights. Following the Bernoulli trials, some of the links accepted between the two populations U^A and U^B can be false, and some other links may have been falsely rejected. The linkage errors therefore tend to be higher if the Bernoulli trials are used. However, in the present context, the quality of the links can be considered as a secondary interest. The problem here is to estimate the total Y^B using the sample s^A selected from U^A, and not to evaluate the quality of the links. In section 9.3, the precision of the estimates of Y^B will be measured with respect to the sampling variability of the estimators, by conditioning on the linkage weights $\theta_{j,ik}$. Note that this sampling variability will take into account the random selection of the links, but not the linkage errors.

To reduce the number of non-zero links, the present approach is therefore considered as being of potential interest, even if the quality of the resulting links can be questionable.

The first step before performing the Bernoulli trials is to transform the linkage weights in a way such that they are contained in the [0,1] interval. By looking at the definition (9.1), it can be seen that the linkage weights $\dot{\theta}_{j,ik}$ correspond in fact to a "logit" transformation (in base 2) of the probability $P(\upsilon_{j,ik} \mid \Delta_{1j,ik}\Delta_{2j,ik}...\Delta_{pj,ik})$. In the same way, the linkage weights $\theta_{j,ik}$ given by (9.2) depend only on this same probability. Hence, one way to transform the linkage weights is simply to use the probability $P(\upsilon_{j,ik} \mid \Delta_{1j,ik}\Delta_{2j,ik}...\Delta_{pj,ik})$. From (9.1), we obtain this result by using the function $\tilde{\tilde{\theta}} = 2^{\dot{\theta}}/(1+2^{\dot{\theta}})$ and, from (9.2), by using $\tilde{\tilde{\theta}} = \theta/(1+\theta)$. If the linkage weight are not obtained through a definition similar to (9.1) or (9.2), another possible transformation is to simply divide each weight by the maximum value $\theta_{Max} = \max(\theta_{j,ik} \mid j = 1,...M^A, i = 1,...,N, k = 1,...,M_i^B)$. Note that we assume here that the linkage weights are all greater than or equal to zero, which is the case from definition (9.2), but not necessarily in general.

Once the adjusted linkage weights $\tilde{\tilde{\theta}}_{j,ik}$ have been obtained, we generate for each pair (j,ik) a random number $\nu_{j,ik} \sim \mathrm{U}(0,1)$. Then, we assign the value 1 to the indicator variable $\tilde{\tilde{l}}_{j,ik}$ if $\nu_{j,ik} \leq \tilde{\tilde{\theta}}_{j,ik}$, and the value 0 otherwise. This process provides a set of links similar to the ones used in the classical approach, with the exception that now the links are determined randomly and not through a decision process like (9.3). Note that since $E(\tilde{\tilde{l}}_{j,ik}) = \tilde{\tilde{\theta}}_{j,ik}$, the sum of the adjusted linkage weights $\tilde{\tilde{\theta}}_{j,ik}$ corresponds to the expected total number of links L from the Bernoulli trials, i.e.,

$$\sum_{j=1}^{M^A}\sum_{i=1}^{N}\sum_{k=1}^{M_i^B}\tilde{\tilde{\theta}}_{j,ik} = L. \qquad (9.13)$$

With the present approach, by randomly selecting links, it is strongly possible that Constraint 2.1 related to the GWSM will not be satisfied. To correct this problem, a link can be imputed by choosing the

link with the largest linkage weight $\tilde{\tilde{\theta}}_{j,ik}$ within the cluster. The link can also be selected randomly with a probability proportional to $\tilde{\tilde{\theta}}_{j,ik}$.

For each unit j selected in s^A, we identify here the units ik of U^B that have $\tilde{\tilde{l}}_{j,ik} = 1$. Let $\tilde{\tilde{\Omega}}^B = \{i \in U^B \mid \exists j \in s^A \text{ and } \tilde{\tilde{L}}_{j,i} > 0\}$ where $\tilde{\tilde{L}}_{j,i} = \sum_{k=1}^{M_i^B} \tilde{\tilde{l}}_{j,ik}$ be the set of the $\tilde{\tilde{n}}$ clusters identified by the units $j \in s^A$. Note that $\tilde{\tilde{n}} \leq n^{RL}$. Unfortunately, in contrast to n^{RL} and n^{RLT}, the number of clusters $\tilde{\tilde{n}}$ is hardly comparable to n, the number of clusters obtained using the classical approach.

To estimate the total Y^B of the population U^B, we can use

$$\tilde{\tilde{Y}}^B = \sum_{i=1}^{\tilde{\tilde{n}}} \sum_{k=1}^{M_i^B} \tilde{\tilde{w}}_{ik} y_{ik} \tag{9.14}$$

where $\tilde{\tilde{w}}_{ik}$ is the estimation weight obtained from the GWSM. This weight is obtained by directly replacing the indicator variables $l_{j,ik}$ with $\tilde{\tilde{l}}_{j,ik}$ in the steps of the GWSM described in section 2.1.

Steps of the GWSM for approach 3

Step 1: For each unit k of the clusters i from $\tilde{\tilde{\Omega}}^B$, the initial weight $\tilde{\tilde{w}}'_{ik}$ is calculated, that is:

$$\tilde{\tilde{w}}'_{ik} = \sum_{j=1}^{M^A} \tilde{\tilde{l}}_{j,ik} \frac{t_j}{\pi_j^A}, \tag{9.15}$$

where $t_j = 1$ if $j \in s^A$, and 0 otherwise.

Step 2: For each unit k of the clusters i from $\tilde{\tilde{\Omega}}^B$, $\tilde{\tilde{L}}_{ik}^B = \sum_{j=1}^{M^A} \tilde{\tilde{l}}_{j,ik}$ is calculated. The quantity $\tilde{\tilde{L}}_{ik}^B$ represents the realised number of links between the units of U^A and unit k of cluster i from U^B.

Step 3: The final weight $\tilde{\tilde{w}}_i$ is calculated:

$$\tilde{\tilde{w}}_i = \frac{\sum_{k=1}^{M_i^B} \tilde{\tilde{w}}'_{ik}}{\sum_{k=1}^{M_i^B} \tilde{\tilde{L}}_{ik}^B}. \tag{9.16}$$

Step 4: Finally, we set $\tilde{\tilde{w}}_{ik} = \tilde{\tilde{w}}_i$ for all $k \in U_i^B$.

By conditioning on the accepted links $\tilde{\tilde{l}}_{j,ik}$, it can be shown that the estimator $\tilde{\tilde{Y}}^B$ given by (9.14) is unbiased, assuming of course that Constraint 2.1 is satisfied. Let $E_l(.)$ be the expected value carried out in relation to all the possible realisations of links. Let $\tilde{\tilde{\mathbf{L}}}$ be the set of realised links, i.e.,

$$\tilde{\tilde{\mathbf{L}}} = \left\{ \tilde{\tilde{l}}_{j,ik} \right\}_{j=1,i=1,k=1}^{M^A,N,M_i^B}.$$

We then have

$$E(\tilde{\tilde{Y}}^B) = E_l[E(\tilde{\tilde{Y}}^B \mid \tilde{\tilde{\mathbf{L}}})]. \tag{9.17}$$

By conditioning on the set $\tilde{\tilde{\mathbf{L}}}$, the estimator (9.14) is then equivalent to the estimator (2.1) (or the estimator (4.1)). From Corollary 4.1, $E(\tilde{\tilde{Y}}^B \mid \tilde{\tilde{\mathbf{L}}}) = Y^B$ is directly obtained and therefore, the estimator (9.14) is conditionally unbiased. Consequently, this estimator is unbiased in an unconditional way. To obtain the variance of $\tilde{\tilde{Y}}^B$, we again proceed in a conditional way from

$$Var(\tilde{\tilde{Y}}^B) = E_l[Var(\tilde{\tilde{Y}}^B \mid \tilde{\tilde{\mathbf{L}}})] + Var_l[E(\tilde{\tilde{Y}}^B \mid \tilde{\tilde{\mathbf{L}}})].$$

First of all, since $E(\tilde{\tilde{Y}}^B \mid \tilde{\tilde{\mathbf{L}}}) = Y^B$, we have

$$Var_l[E(\tilde{\tilde{Y}}^B \mid \tilde{\tilde{\mathbf{L}}})] = Var_l[Y^B] = 0. \tag{9.18}$$

By conditioning on the set $\tilde{\tilde{\mathbf{L}}}$, it was already mentioned that the estimator (9.14) is equivalent to the estimator (2.1). By Corollary 4.2, the following result is thus obtained:

$$Var(\tilde{\tilde{Y}}^B \mid \tilde{\tilde{\mathbf{L}}}) = \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} \tilde{\tilde{Z}}_j \tilde{\tilde{Z}}_{j'} \tag{9.19}$$

where $\tilde{\tilde{Z}}_j = \sum_{i=1}^{N}\sum_{k=1}^{M_i^B} \tilde{\tilde{l}}_{j,ik}\tilde{\tilde{z}}_{ik}$ with $\tilde{\tilde{z}}_{ik} = Y_i/\tilde{\tilde{L}}_i^B$. The variance of $\tilde{\tilde{Y}}^B$ can therefore be written

$$Var(\tilde{\tilde{Y}}^B) = E_l\left[\sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} \tilde{\tilde{Z}}_j \tilde{\tilde{Z}}_{j'}\right]. \tag{9.20}$$

To estimate the variance (9.20), one of the two estimators (4.12) or (4.13) can be used by replacing the variable Z_j with $\tilde{\tilde{Z}}_j$.

9.2.4 Some remarks

The three proposed approaches do not use the decision rule (9.3). They also do not require any manual resolution. Consequently, the answer to question (c) is yes. That is, the GWSM can help in reducing the manual resolution required by record linkage. Note that there is however a price to pay for avoiding manual resolution.

First, with approach 1, the number n^{RL} of clusters identified by the units $j \in s^A$ is greater than or equal to the number n of clusters identified by the classical approach, i.e., when the decision rule (9.3) is used to accept the links or not. This happens because we use all non-zero links, and not just the ones satisfying the decision rule (9.3). As a consequence, the collection costs with approach 1 are greater than or equal to the ones related to the classical approach. It needs then to be checked which ones are the most important: the collection costs or the costs of manual resolution. Note that if the precision resulting from the use of approach 1 is much higher than the one from the classical approach, it can be more advantageous to choose approach 1 than the classical approach.

With approach 2, we have $n^{RLT} \leq n^{RL}$ and therefore the collection costs of this approach are less than or equal to the ones of approach 1. If the precision of approach 2 is comparable to the one of approach 1, then approach 2 will certainly be more advantageous than approach 1. By comparing approach 2 with the classical approach, it can be seen that the collection costs can be almost equivalent if the value of the threshold θ_{High} is chosen to be relatively close to the lower and upper thresholds of the decision rule (9.3).

Note that approach 2 does not use any manual resolution. If the precision of approach 2 is at least comparable to the one of the classical approach, then approach 2 is more advantageous. Note that if $\theta_{High} = \theta_{Low}$, the two approaches differ only in the definition of the estimation weights ensuing from the GWSM. Approach 2 uses the linkage weights $\theta_{j,ik}^{RLT}$, while the classical approach uses the indicator variables $l_{j,ik}$. Setting $\theta_{High} = \theta_{Low}$, it is certainly of interest to know which approach has the highest precision.

With approach 3, the number of selected links is less than or equal to the number of non-zero links used by approach 1, i.e., $\tilde{\tilde{n}} \leq n^{RL}$. Hence, the collection costs of approach 3 are less than or equal to the ones of approach 1. As mentioned before, unlike n^{RL} and n^{RLT}, the number of clusters $\tilde{\tilde{n}}$ is hardly comparable to n. The two depend on different parameters: the classical approach depends on the thresholds θ_{Low} and θ_{High}, while approach 3 depends on the adjusted linkage weights $\tilde{\tilde{\theta}}_{j,ik}$.

9.3 SIMULATION STUDY

We performed a simulation study to evaluate the approaches presented in this chapter, including the classical approach. For this study, we compared the precision obtained for the estimation of the total Y^B for five different approaches:

Approach 1: use all non-zero links with the linkage weights $\theta_{j,ik}$

Approach 2: use all non-zero links above a threshold

Approach 3: choose the links randomly

Approach 4: classical approach

Approach 5: use all non-zero links with the indicator variables $l_{j,ik}$.

Approach 5 is a mixture of approach 1 and the classical approach. It consists of first accepting all links of the pairs *(j,ik)* that have a linkage weight greater than zero, i.e., assign $l_{j,ik} = 1$ for all pairs *(j,ik)* where $\theta_{j,ik} > 0$, and $l_{j,ik} = 0$ otherwise. The GWSM described in chapter 2 is then used to estimate the total Y^B from the estimator (2.1). Approach 5 was added to the simulations to verify the effect of using the indicator variables $l_{j,ik}$ instead of the linkage weights $\theta_{j,ik}$ when using all non-zero links. As with all the other approaches, according to Corollary 4.1, approach 5 is unbiased. Since the five approaches yield unbiased estimates of the total Y^B, we compared them with the *standard error* (the square root of the variance), and more specifically with the *coefficient of variation* (the standard error divided by the expected value of the estimator).

9.3.1 Data used

The simulation study was performed using the agricultural data presented in section 7.4.4. Thus, the study again is inspired by Statistics Canada's Whole Farm Data Base. Recall that this database has information on livestock, crops and the income and expenditures (tax data) of Canadian farms (Statistics Canada, 2000a). The data used for the simulations come from Québec and New Brunswick. Although the simulations were inspired by the Whole Farm Data Base, some processes and data were changed for reasons of confidentiality, and also to not needlessly complicate the discussion. However, we believe that these changes do not affect the conclusions drawn from the simulations.

The population U^A is a list of M^A farms coming from the 1996 Farm Register. This list essentially comes from the 1991 Canadian Census of Agriculture, with different updates that have been made since 1991. The units j from U^A thus represent farms, but note that each farm j can have many farm operators. In addition to the FarmID, the Farm Register contains a farm operator number together with some demographic variables related to the farm operators.

The target population U^B is a list of M^B tax records (or income tax reports) from the Canadian Revenue Agency (CRA). This second list is the 1996 CRA Unincorporated Business File that contains tax data for the persons declaring at least one farming income. This file contains a household number (only for a sample), a tax filer number, and also demographic variables related to the tax filers. The units k are thus the tax reports that are completed by the different members of households i (or clusters). The target population U^B has N households. The respective sizes of the populations U^A and U^B are given in Table 7.1.

For the simulations, linkage has been performed for the two populations U^A and U^B (in fact, linkage of the files A and B related to these populations). To do this, a linkage process was used based on the matching of five variables. It was performed using the MERGE statement in SAS®. The records on both files were compared to one another in order to determine whether or not there is a match. The record linkage was performed using the following five key variables common to both files:

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 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 typing error.

Records that matched on all five variables were given the highest linkage weight ($\theta = 60$). Records that matched on only a subset of at least two of the five variables received a lower non-zero linkage weight ($\theta = 2$). Pairs of records that did not match on any combination of key variables were considered as pairs having no possible links, which is equivalent to having a linkage weight of zero.

Two different threshold were chosen for the simulations: $\theta_{High} = \theta_{Low} = 15$ and $\theta_{High} = \theta_{Low} = 30$. The upper and lower thresholds, θ_{High} and θ_{Low}, were set to be the same to avoid the grey area where some manual intervention is needed when applying the decision rule (9.3).

Following the linkage process, the constraint requiring that each cluster i of the target population U^B have at least one non-zero link was not satisfied for all clusters. To correct the situation, we imputed a link by choosing the link with the largest linkage weight $\theta_{j,ik}$ within the cluster. In the case where all linkage weights are zero, we chose a link at random.

The record linkage process used here does not exactly correspond to the one used to construct the Whole Farm Data Base. For more information on the exact process, refer to Lim (2000). We believe that the changes, however, do not affect the conclusions drawn from the simulations. Recall that the main goal of the simulations is to evaluate the different approaches for the estimation of Y^B, and not to solve the problems related to the construction of the Whole Farm Data Base.

Following record linkage, it turns out that the populations U^A and U^B are linked by complex links. Indeed, a farm j sometimes has many operators and each operator returns one tax report k to the CRA. There is then a "one-to-many" link since we have one farm j linked to many tax reports k. On the other hand, an operator who deals with more than one farm j can return a single tax report k for the set of farms that he operates. Therefore, this type of link is "many-to-one" since there are many farms j

linked to a single tax report k. Finally, there are also situations of complex links where the operators deal with more than one farm and where each farm has a number of different operators. The populations U^A and U^B as well as their links can be represented by Figure 2.1.

9.3.2 Sampling plan

For the simulations, the sample s^A was selected from U^A (Farm Register) using simple random sampling without replacement, without any stratification. We also considered two sampling fractions: 30% and 70%. The variable of interest y for which we want to estimate the total Y^B is the total farming income. Since we have the entire populations of farms and tax records, it was possible to calculate the value of Y^B and the variances from the theoretical formulas developed for this approach. Furthermore, because a simple random sampling without replacement was performed, these theoretical formulas can be simplified. For example, in the case of approach 1, the variance of $\hat{Y}^{RL,B}$ given by (9.8) can then be written in the following form:

$$Var(\hat{Y}^{RL,B}) = M^A \frac{(1-f^A)}{f^A} \frac{1}{M^A - 1} \sum_{j=1}^{M^A} (Z_j^{RL} - \overline{Z}^{RL})^2 , \qquad (9.21)$$

where $f^A = m^A / M^A$ is the *sampling fraction* and $\overline{Z}^{RL} = \frac{1}{M^A} \sum_{j=1}^{M^A} Z_j^{RL}$.

A *Monte Carlo study* was also conducted to empirically calculate the bias and the variance under the different approaches. Note that for approach 3, only the Monte Carlo study was used. For the Monte Carlo study, 500 samples s^A from U^A were selected for each sampling fraction 30% and 70%, and for each threshold 15 and 30. The empirical bias and the empirical variance of each estimator (represented by $\hat{Y}$) were calculated using

$$\hat{Bias}(\hat{Y}) = \hat{E}(\hat{Y}) - Y^B = \frac{1}{500} \sum_{s^A=1}^{500} \hat{Y}_{s^A} - Y^B \qquad (9.22)$$

$$\hat{Var}(\hat{Y}) = \frac{1}{500} \sum_{s^A=1}^{500} (\hat{Y}_{s^A} - \hat{E}(\hat{Y}))^2 . \qquad (9.23)$$

The coefficients of variation (CV) were then calculated by using

$$\hat{CV}(\hat{Y}) = 100 \times \frac{\sqrt{\hat{Var}(\hat{Y})}}{\hat{E}(\hat{Y})}. \qquad (9.24)$$

The Monte Carlo study was, among other things, performed to verify in an empirical manner the accuracy of the theoretical formulas given in section 9.2. The results all indicated that the theoretical formulas are exact.

9.3.3 Results and discussion

The results of the simulations are given in Figures 9.1 to 9.4, Table 9.1 and Figure 9.5. Figures 9.1 to 9.4 provide histograms of the CVs obtained for each of the five approaches. Eight histograms are shown, corresponding to the eight cases obtained by crossing the two provinces Québec and New Brunswick, the two sampling fractions 30% and 70%, and the two thresholds 15 and 30.

On each bar of the histograms, one can see the number of non-zero links between U^A and U^B for each of the five approaches. For approach 3, it is in fact the expected number of non-zero links. Note that the number (expected or not) of non-zero links does not change from one sampling fraction to another.

Table 9.1 shows, for each of the eight cases, the average number of clusters surveyed for each approach. This average is calculated with respect to the 500 samples s^A used for the simulations. The numbers is parentheses represent the standard error for the number of surveyed clusters. The standard errors are relatively small compared to the averages and therefore, the number of clusters surveyed do not vary greatly from one sample to another.

Figure 9.5 gives, for each of the eight cases, a graph of the obtained CVs for the five approaches as a function of the average number of surveyed clusters.

By looking at Figures 9.1 to 9.4, it can be seen that in all cases, approaches 1 and 5 give the smallest CVs for the estimation of total farming income. Therefore, using all non-zero links produces estimates with the greatest precision. Looking at Table 9.1, we note however that these approaches are the ones for which the number of surveyed clusters is the highest. In fact, we can see that the greater the number of surveyed clusters, the greater the precision of the estimates is. This result is shown in Figure 9.5 where we can see that the CVs tend to decrease as the

average number of surveyed clusters increases. Although this observation is well known in the classical sampling theory, it is not necessarily evident in the context of indirect sampling. As we can see from equations (4.11a) and (4.11b), it is not the sample size of s^A that increases, but rather the homogeneity of the derived variables Z_j.

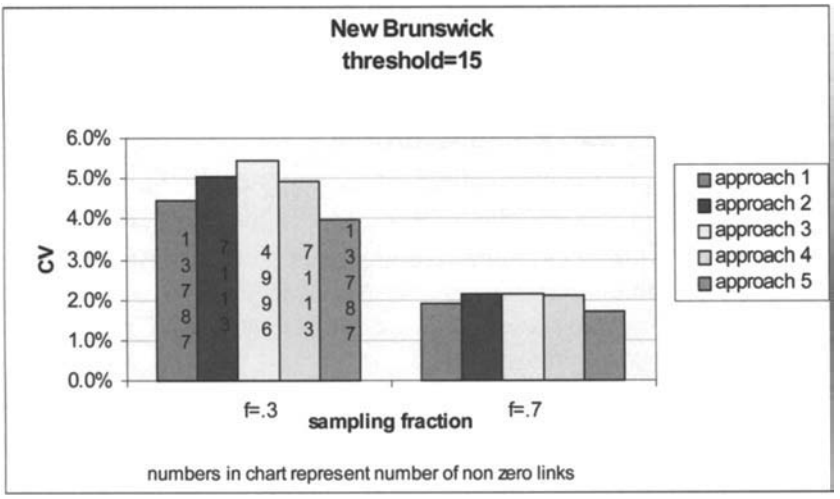


Figure 9.1: *CVs for New Brunswick* (*with* $\theta_{High} = \theta_{Low} = 15$)

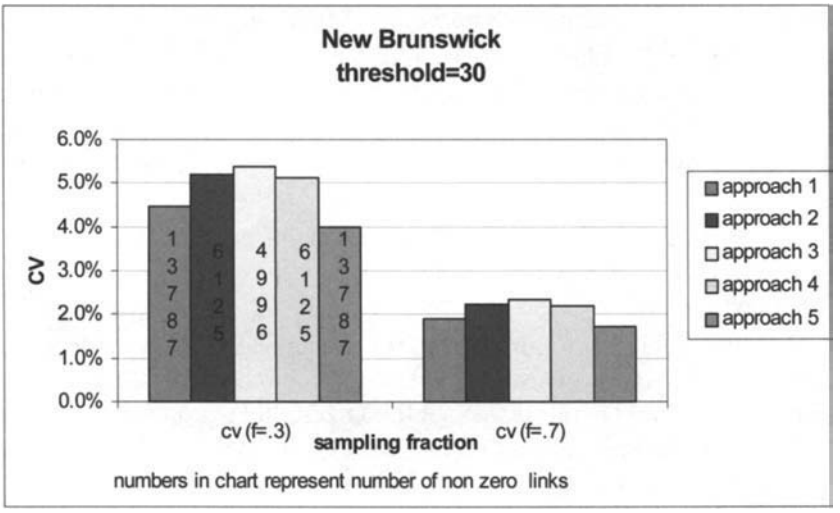


Figure 9.2: *CVs for New Brunswick* (*with* $\theta_{High} = \theta_{Low} = 30$)

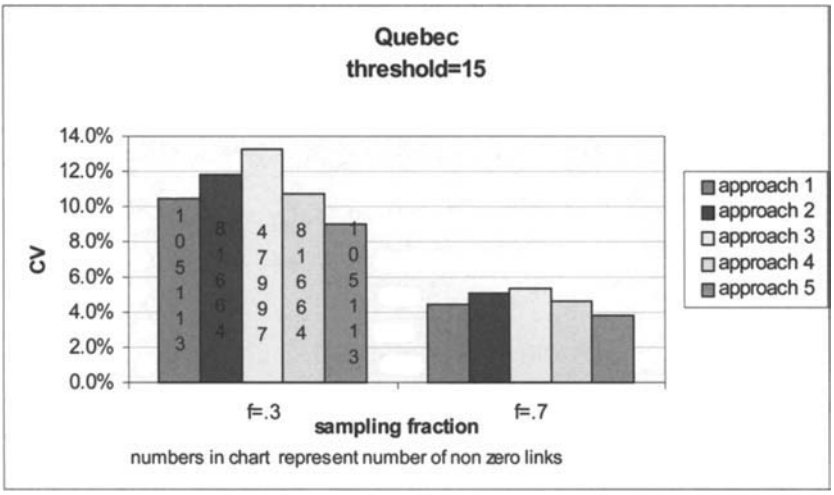


Figure 9.3: *CVs for Québec* (*with* $\theta_{High} = \theta_{Low} = 15$)

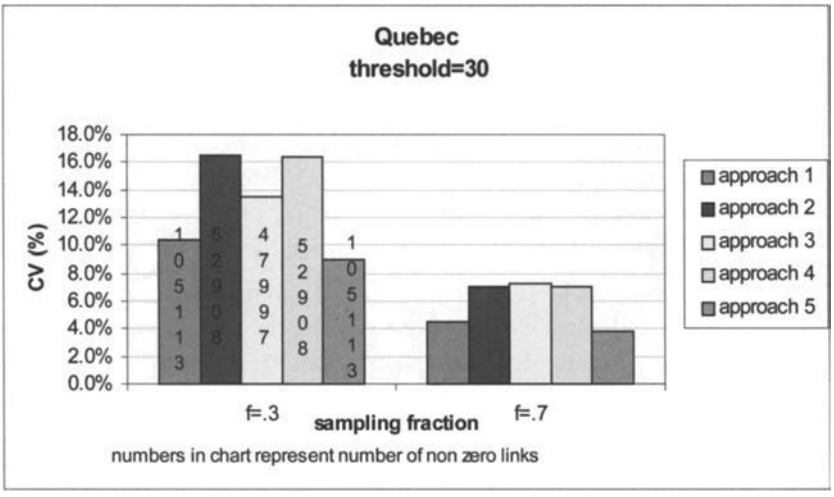


Figure 9.4: *CVs for Québec* (*with* $\theta_{High} = \theta_{Low} = 30$)

Table 9.1: *Surveyed clusters for Québec and New Brunswick*

Threshold θ_{High}	Approach	Average number of surveyed clusters (s.e.)			
		Québec		New Brunswick	
		f^A=0.3	f^A=0.7	f^A=0.3	f^A=0.7
15	1	15752 (58)	21106 (30)	1709 (18)	2100 (7)
	2	14281 (49)	20593 (34)	1310 (17)	1966 (13)
	3	10930 (50)	18881 (47)	1123 (14)	1869 (14)
	4	14281 (49)	20593 (34)	1310 (17)	1966 (13)
	5	15752 (58)	21106 (30)	1709 (18)	2100 (7)
30	1	15752 (58)	21106 (30)	1709 (18)	2100 (7)
	2	11310 (45)	19139 (37)	1215 (17)	1924 (15)
	3	10930 (50)	18881 (47)	1123 (14)	1869 (14)
	4	11310 (45)	19139 (37)	1215 (17)	1924 (15)
	5	15752 (58)	21106 (30)	1709 (18)	2100 (7)

Now, by comparing approaches 1 and 5, it can be seen that approach 5 always provided smaller CVs than approach 1. This suggests using the indicator variable $I_{j,ik}$ instead of the linkage weight $\theta_{j,ik}$ when all the links are considered to be non-zero. Note that it seems this result can be generalised when we note that the same phenomenon is produced for approaches 2 and 4 (classical approach). Recall that because $\theta_{High} = \theta_{Low}$, the two approaches differ only in the definition of the estimation weights obtained by the GWSM; approach 4 uses the indicator variable $I_{j,ik}$ and approach 2, the linkage weight $\theta_{j,ik}$. This result is particularly important because it corresponds to the conclusions of Kalton and Brick (1995) and the ones in section 4.6.3, namely that by using $\tilde{\theta}_{j,ik} = I_{j,ik} / L_i^B$ in the version of the GWSM described in section 4.5, we should then approach minimal variances for the estimation of the total Y^B.

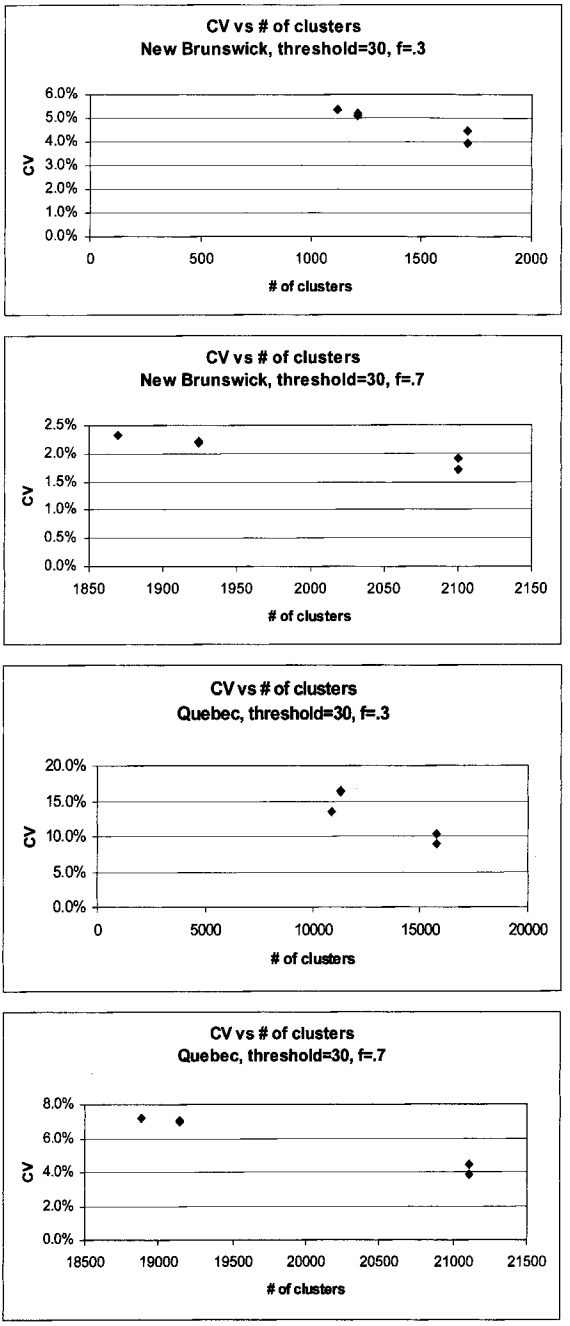
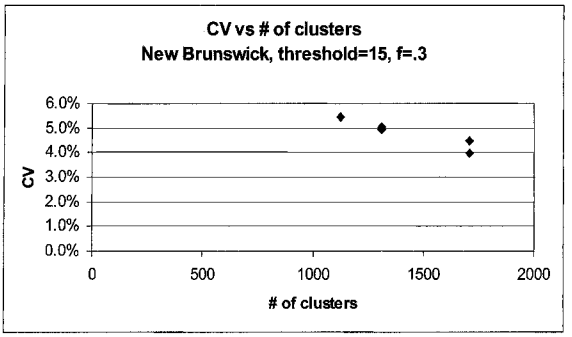
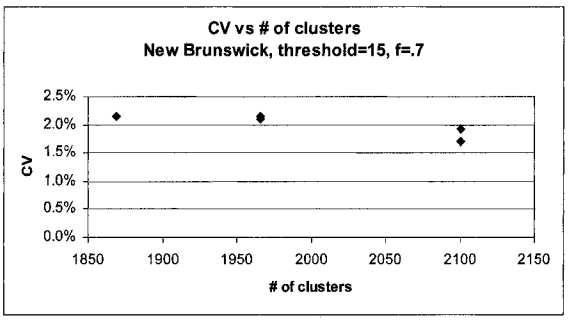
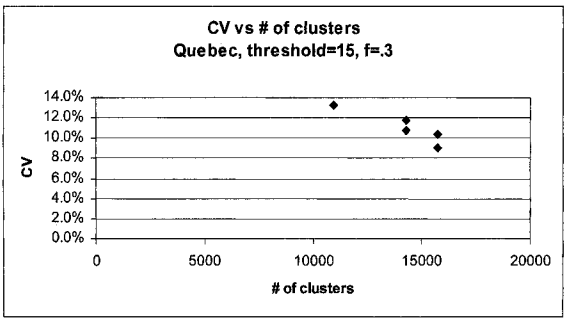


Figure 9.5: *Graphs of CVs versus average number of surveyed clusters*







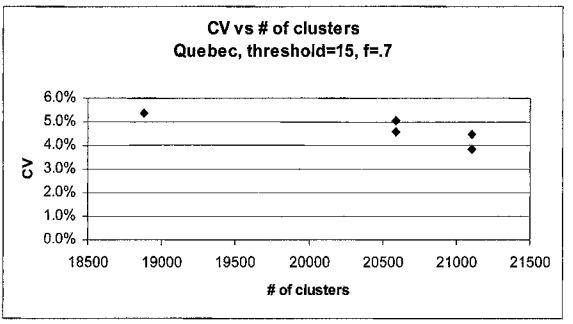


Figure 9.5 (continued): *Graphs of CVs versus average number of surveyed clusters*

Now consider approach 3. For seven out of the eight histograms from Figures 9.1 to 9.4, approach 3 produced the highest CVs. It should be noted however that this approach is the one that is based on the lowest number of non-zero links, and also the lowest number of surveyed clusters. Therefore, the poor performance of approach 3 is not surprising.

Recall that the number of non-zero links used by approach 3 does not depend on the threshold θ_{High}, and thus the CVs obtained for thresholds 15 and 30 are the same. For θ_{High}=15, the CV obtained for Québec for approach 3 proves to be higher than the ones obtained for approaches 2 and 4, and these two approaches use more non-zero links and more surveyed clusters. For θ_{High}=30, the CV obtained for approach 3 proves to be lower than the ones obtained for approaches 2 and 4, but these two approaches still used more non-zero links and more surveyed clusters. Therefore, there are intermediate situations where, with $15<\theta_{High}<30$, we get equal CVs for approaches 3 and 2, and equal CVs for approaches 3 and 4. As a result, to get equal CVs for approach 3 and each of approaches 2 and 4, more links (and more surveyed clusters) must be used by approaches 2 and 4. This suggests that approach 3 can, in some cases, be more worthwhile than approaches 2 and 4 because it produces estimates with the same precision but with lower collection costs.

So as to better compare approach 3 and approaches 2 and 4, we made the expected number of non-zero links to be the same as the number of non-zero links used by approaches 2 and 4. To do this, we have transformed the linkage weights $\theta_{j,ik}$ into new weights $\tilde{\tilde{\theta}}_{j,ik}$ such that

$$\sum_{j=1}^{M^A}\sum_{i=1}^{N}\sum_{k=1}^{M_i^B}\tilde{\tilde{\theta}}_{j,ik} = L_0\,, \tag{9.25}$$

where L_0 is the desired number of non-zero links. The transformation used was the following:

$$\tilde{\tilde{\theta}}_{j,ik} = \begin{cases} \theta_{j,ik}/\theta_\bullet & \text{if } \theta_{j,ik}/\theta_\bullet \le 1 \\ 1 & \text{otherwise} \end{cases} \tag{9.26}$$

where $\theta_\bullet$ was determined iteratively so that constraint (9.25) is satisfied. The use of approach 3 with the transformed linkage weights by (9.26) was called approach 6. The results of the simulations are presented in Figures 9.6 to 9.9.

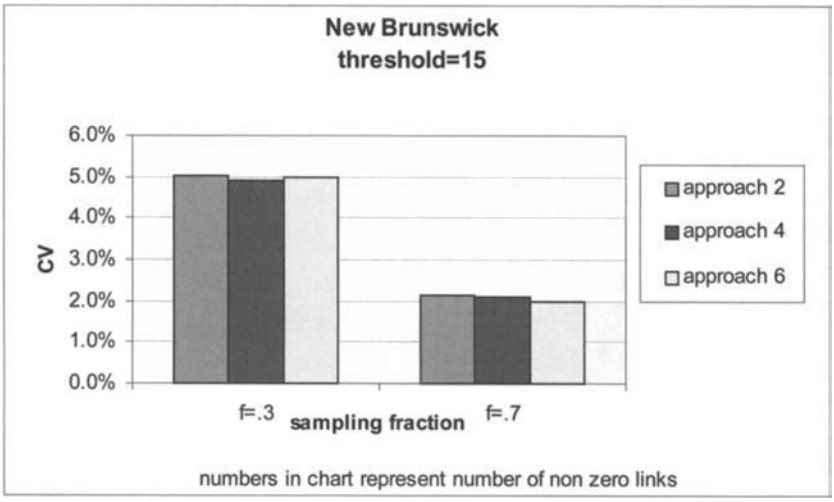


Figure 9.6: *CVs for New Brunswick* (*with* $\theta_{High} = \theta_{Low}$ =15)

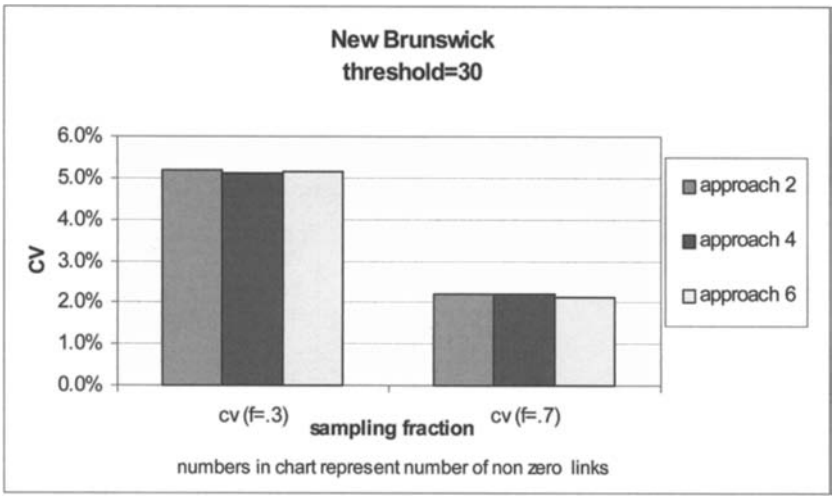


Figure 9.7: *CVs for New Brunswick* (*with* $\theta_{High} = \theta_{Low}$ =30)

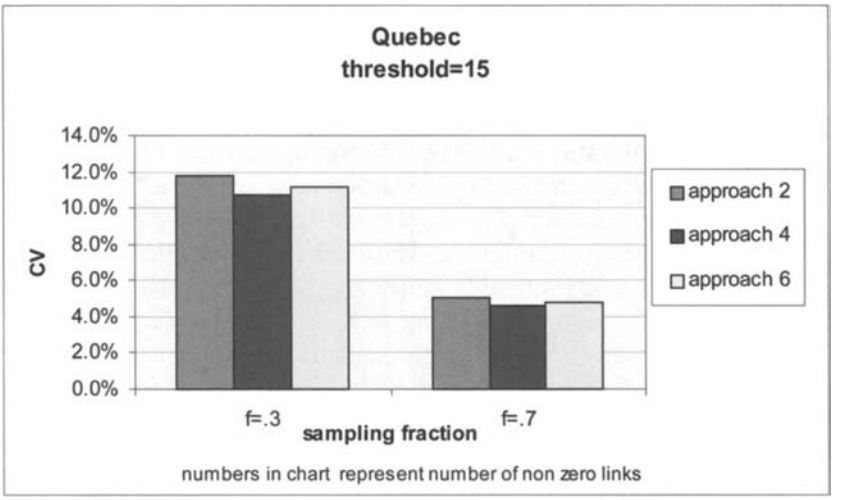


Figure 9.8: *CVs for Québec* (*with* $\theta_{High} = \theta_{Low}$ =15)

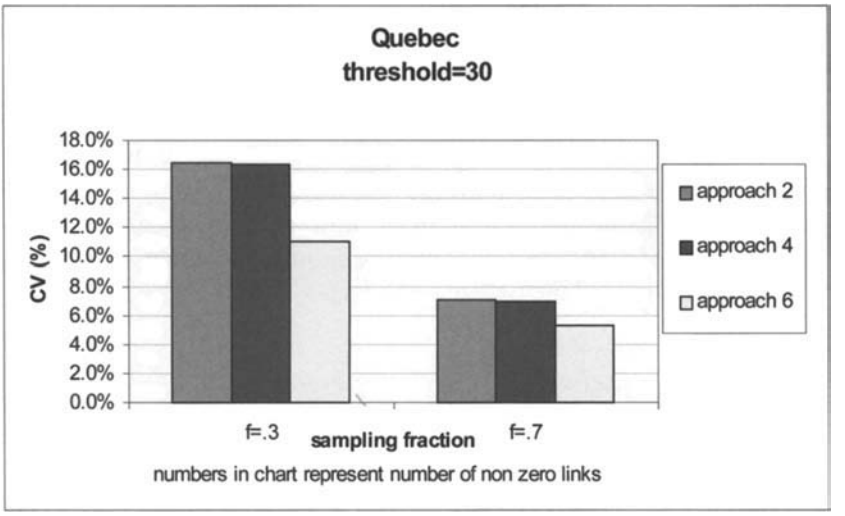


Figure 9.9: *CVs for Québec* (*with* $\theta_{High} = \theta_{Low}$ =30)

As we can see, approach 6 produced the smallest CVs for half of the cases. For the other half, approach 4 yielded the best precision. Note that this result was not obtained for a specific province, or a specific sampling fraction, or a specific threshold. It would therefore be difficult in practice to determine in advance which of approaches 4 or 6 would be likely to produce the smallest CVs. Furthermore, using the decision rule (9.3) to determine the links, it was shown that the number of false links and false non-links are minimised. Thus, if the quality of the links proves to be a concern, it is preferable to use approach 4 because the random selection of links suggested by approach 3 can lead to the selection of links that would not be acceptable through the decision rule (9.3), even if the selection probabilities of the links are proportional to the linkage weights. For these reasons, it seems preferable to choose approach 4 instead of approaches 2 and 6 (or approach 3).

In conclusion, if the number of links and the number of surveyed clusters do not pose a problem, it is suggested to use approach 5, i.e., to consider all the links of pairs *(j,ik)* that have a linkage weight $\theta_{j,ik}$ greater than zero, and to use the GWSM described in chapter 2 to estimate the total Y^B from the estimator (2.1). If the number of surveyed clusters proves to be too large because, for example, it leads to collection costs that are too high, approach 4 can be seen as a reasonable choice. Recall that the use of the threshold θ_{High} (and also the threshold θ_{Low}) is useful to reduce the number of non-zero links to manipulate. By reducing the number of non-zero links, we reduce at the same time the number of clusters identified through the sample s^A and therefore also the collection costs associated to the measurement of the variable of interest y. By reducing the number of links, however, the precision of the estimates is reduced. Thus, a compromise must be made between the desired precision and the collection costs.