IMPUTATION OF ESTABLISHMENT SURVEY DATA

John G. Kovar and Patricia J. Whitridge
Statistics Canada

Nonresponse to establishment surveys is an ongoing problem for agencies responsible for the collection and dissemination of economic data. This is not to say that nonresponse is of no concern to social survey statisticians. For economic surveys, however, the data are often highly skewed and quantitative in nature, which poses unique challenges in dealing with the problem of missing data. As nonrespondents may not behave in the same manner as the respondents in terms of the characteristics of interest, special care must be taken to avoid potential biases that can result from inappropriate treatment of missing or inconsistent data. While imputation is but one method of dealing with incomplete data, its use is widespread. In this chapter we discuss the reasons for and against using imputation techniques.

The sources of nonresponse are numerous and varied. Complete questionnaires may be missing due to the interviewer's inability to contact the establishment, or the establishment's unwillingness to participate in the survey. We refer to such cases of nonresponse as *total* or *unit nonresponse*. Unfortunately, many filled-out and returned questionnaires are often incomplete. This may be because the respondent (i.e., the person or persons assigned by the establishment as responsible for completing the survey) is unable, not sufficiently knowledgeable, or unwilling to respond to some questions. Missing, invalid, or inconsistent responses also may be obtained because of misinterpretation of concepts, inaccuracy or inadvertent omissions. This type of nonresponse is referred to as *item nonresponse*.

Both total and item nonresponse may occur at random or not, but this assessment is difficult to make. In most cases however, nonresponse is not random: large establishments often have better reporting arrangements, small farms tend not to appreciate fully the utility of surveys, and nonprofit institutions may feel they do not have the resources or time to comply. We distinguish here between two types of nonrandom nonresponse: ignorable and nonignorable (Rubin, 1987). Nonrandom nonresponse is deemed *ignorable* if it depends only on the observed variables on the data file. Nonresponse that depends systematically on the actual items subject to nonresponse is deemed *nonignorable*. (See Rubin (1987) for a more rigorous definition.) The choice of imputation method will depend on the practitioner's suspicions about the ignorability of the nonresponse mechanism, despite the fact that ignorability is undetectable given the data set in question.

In most instances, ignoring the missing data and basing analyses on responding units only will lead to serious biases, since the respondents rarely form a random subsample of the entire sample. In general, two broad options exist in the quest of minimizing the bias caused by nonresponse. In the case of item nonresponse, the missing values are often replaced by feasible data values to create a completed data file for analyses. This technique is referred to as *imputation*. For total nonresponse, survey weight adjustment is likely the best redress for nonresponse. Weight adjustment could be considered for item nonresponse on a variable by variable basis. This is often cumbersome as different weights would be required for each variable and imputation is then preferred. In fact, some establishment surveys even use imputation in cases of total nonresponse. This would generally be done only if there existed sufficient auxiliary information from other sources or past occasions.

Imputation techniques for establishment surveys have largely evolved from methods used in social and demographic settings. Of interest in this chapter is the adaptation of such methods to the skewed, quantitative data of business surveys. Other methods, of course, have been developed predominantly within the economic survey framework. These too are discussed in the following sections. We do not elaborate on methods that find their domain of application almost exclusively within the social survey arena. As such, without explicitly stating so, this chapter addresses imputation issues from the economic survey perspective, realizing, however, that many of the issues are common to both sides.

Historically, imputation has been a manual process based on subject matter experience and knowledge. Due to advances in technology, imputation processes have become increasingly automated. This has resulted in more objective and reproducible results. Automated imputation methods are also easier to monitor and evaluate, since useful data concerning the process can be maintained by the computer.

Imputation methods and associated software are becoming more prevalent and their use more widespread and accepted. However, the availability of automatic imputation systems can tempt the practitioner to impute in nontraditional situations such as in the case of total nonresponse, nonresponse "by design" (e.g., two-phase sampling), and in cases where the respondents are not required to report certain characteristics. The latter scenario can lead to very high rates of incomplete data, as is sometimes observed with administrative data bases. The use of weighting in these situations as an alternative to imputation must be considered very carefully and not dismissed too quickly. Its statistical justifiability is usually preferable to the operational convenience of a completed data set. On the other hand, where the propensity for nonresponse can cause serious biases and nonresponse is not random but is ignorable, imputation can be useful in correcting the bias (Colledge et al, 1978). The use of imputation on a large scale basis encountered in the above examples is known as *mass imputation*.

Once a data set has been completed by imputation, care must be taken that it is not treated as "clean". Imputed values must be properly flagged and the methods and sources clearly identified. Some analysts may choose to perform their own adjustments for nonresponse, which can only be done if the imputed data are properly identified. While imputed data sets serve well to produce ad hoc estimates of means and totals at various levels of aggregation, their use for data analysis is not without difficulty, in particular when the survey design is complex. Analyses can be misleading if the imputed values are treated as observed data, as variances and covariances may be seriously underestimated. Some recent techniques put forth to remedy the problem of variance underestimation include multiple imputation, model-assisted methods, and jackknife procedures.

Finally, the resulting imputation-revised data sets are inherently unverifiable, since the "true" data values are unknown for survey nonrespondents. Nonetheless, it is important that measures of quality and indications of success or failure be provided. The reason behind the imputation, be it nonresponse, invalid data, or values that are unusually large should be taken into account, in addition to considering which specific fields are being revised. Measures of quality should help evaluate both the impact of the imputation on the data set as well as the imputation process itself.

While it has been said that the best treatment for nonresponse is not to have any, we take the more pragmatic approach and attempt to show that imputation, despite all of its shortcomings, can be a useful and practical tool. For further insight, readers should refer to the papers by Kalton and Kasprzyk (1986) and Sande (1979; 1982). Extensive bibliographies on the topic are contained in Kalton (1983), Madow, et al. (1983), Bogestrom, et al. (1983), and more recently in Little (1988), Pierzchala (1990), and Kovar, et al. (1992).

1. INCOMPLETE DATA

Economic survey data are rarely complete. Since economic data are often more skewed than social data, significant levels of nonresponse are unacceptable. In some economic surveys, the skewness combined with large weights leads to situations where even a few nonrespondents can have a large impact on the estimates. Whether the nonresponse is due to refusals, noncontacts, or just simple inability to answer, responses may be missing for whole questionnaires, parts of questionnaires, or individual questions. Furthermore, the responses for some data items may be deleted if they are deemed incorrect as the result of an edit failure. From this point on, we will refer to all such cases as missing data.

Clearly, missing data items cannot be assumed to be zero, though some survey processing systems fill the missing fields with zeros. This practice complicates the process, in that true zeros must be distinguished from missing values later on. We recommend that the processing system provide a data file where missing values are clearly identified.

One technique for dealing with missing data is the do-nothing option that would leave missing data as such. That is, the values are identified as missing on the data file, allowing the analyst to deal with the problem in a "locally optimal" fashion. In this way, the analyst could make a number of assumptions about the missing fields, and build the uncertainty into the econometric model. However, for simple tabulations and analyses, this approach has two flaws. First, different analysts, making different assumptions, would generate different tabulations, yielding inconsistent if not contradictory results. This includes both those analysts that adjust for nonresponse, as well as those who (perhaps dangerously) base their analyses on the responding units only. For example, one could estimate the total by multiplying the mean of the responding unit values by the population size (equivalent to mean imputation), or by simply adding up the responding unit values (equivalent to imputation of zeros). Besides problems in interpretation, such inconsistencies undermine the credibility of the data, and by extension, the data collecting agency. Secondly, the treatment of missing data could often be performed based on incomplete knowledge. This problem becomes more serious the further the data are removed from the collecting agency and the more aggregated they become. Confidentiality concerns often dictate that crucial information be suppressed. In the extreme, the data are tabulated and disseminated with an "unknown" category. In this instance, users are left to their own devices in interpreting this category. Leaving missing data as missing is more appropriate when the data are being disseminated in their entirety, as a micro data file. Indications of how to deal with the missing data must be provided by the data producers. It is the data collecting agency that best knows the data and all of their limitations, in particular since it often has access to related data from other sources.

Another option is to deal with missing data at the estimation stage, through the use of survey weights. Usually this method is reserved for the case of total nonresponse, though, in theory, it can be used to address item nonresponse as well. It is generally agreed that weighting is preferable to imputation in cases of total nonresponse. Some establishment surveys stand as exceptions to this rule, since, if available, quantitative frame data can be used better through imputation to adjust for ignorable nonresponse. Weight adjustment can be done at various levels, be it stratum, domain, or unit level (Little and Rubin, 1987), but in all cases the idea is to increase the weights of the respondents to account for the nonrespondents. Raking ratio estimation and poststratification also can be used to advantage when external information is available. The literature on weight adjustment for nonresponse is plentiful; we refer the reader to Chapman (1976), Little (1988), the Hidiroglou, Särndal and Binder chapter, and for more recent discussion aimed in particular at nonresponse adjustments at the variable level, to Yansaneh and Eltinge (1993). Weight adjustment for the purposes of dealing with nonresponse has the desirable property that it is usually theoretically tractable, and thus can be evaluated rigorously. Furthermore, if carried out properly, it can go far in eliminating biases due to uneven distribution of nonrespondents in the population. By contrast, ad hoc tabulation requests may be more difficult to supply. If the weights have to be modified, the resulting tabulations may be inconsistent, leading to credibility problems. A further complication arises when weighting is used to adjust for total nonresponse in longitudinal surveys, in that the same unit may have different weights on different occasions, making longitudinal analysis difficult. In fact, the very concept of total versus item nonresponse comes to question as respondents reply to some, but not all survey occasions.

The third option is for the data collecting agency to impute for the missing items. That is, to insert plausible values in the place of the missing values, so that internally consistent records are created, and a complete "rectangular" data file is obtained. Imputation is most feasible for item nonresponse, though for establishment surveys, it is often used for total nonresponse. For example, as mentioned above, imputing large units may be preferable to weight adjustment, since the data collection agency very often has good auxiliary information about large establishments from other sources. In the case of item nonresponse, imputation can make use of additional information available about the unit through the other data items. In either case, we assume that auxiliary data are available and can thus restrict the remaining discussion to methods of imputation for item nonresponse only. With imputed data files, estimation is greatly simplified and ad hoc tabulations can be produced quickly and consistently. Good imputation techniques can be used to preserve known relationships between variables, address systematic biases and reduce nonresponse bias. In most cases, imputation can be objective and reproducible. On the other hand, imputed data files can give the users a false impression of quality (Sande, 1979; Granquist, 1984; 1990). Furthermore, there is a danger that imputation will destroy reported data to create consistent records that fit preconceived models (Kovar, 1991), models that the analyst will rediscover. At the very least, imputed data must be flagged properly.

2. IMPUTATION METHODS

Imputation methods can be classified into two broad categories, deterministic and stochastic, more or less along the lines of Kalton and Kasprzyk (1986). Some methods of imputation can be labelled deterministic since, given the sample of respondents, the imputed values are determined uniquely. Other methods are non-deterministic, or stochastic, since the imputed values are subject to some degree of randomness. Methods belonging to the deterministic set include logical imputation, historical imputation, mean imputation, sequential (ordered) hot deck methods, ratio and regression imputation, and the nearest neighbour imputation. These can be further divided into methods that rely solely on deducing the imputed value from the data available for the nonrespondent and other auxiliary data (logical and historical) and those methods that make use of the observed data of other responding units for the given survey. Use of current observed data can be made directly by transferring data from a chosen *donor* record (hot deck and nearest neighbour) or by means of models (ratio and regression). Examples of stochastic methods are the random hot deck, regression with random residuals and in fact any deterministic method with random residuals added. All of these methods are discussed in more detail below.

Many of these methods can be applied to the whole data set at once, or independently within imputation classes. For example, one might choose to impute independently in each province, or within different standard industrial classifications (SIC). Imputation classes are used to reduce the impact of nonresponse bias, and to improve the accuracy of the imputation. The classes are constructed using control variables available for all units, so that, ideally, the variables to be imputed are homogeneous within the groups. That is, in such a way that within the classes, respondents and nonrespondents, if they were available, would be similar. Often a proxy for the propensity to respond can be used as a control variable. For example, small and large businesses often respond at different rates. Constructing imputation classes based on a measure of business size can ensure that all sizes are properly represented. In practice, the imputation methods are rarely used in exclusion, as no method is perfect for all situations. The imputation system often comprises several methods to be used in a predefined sequence. Thus if one method fails, there is a backup strategy in place. For example, a monthly survey might use historical imputation as its preferred method, but for newly rotated units would instead use mean value imputation.

2.1 Deterministic Methods

Logical (or *deductive*) *imputation*, often performed during data collection or at the early stages of imputation, refers to any method that establishes the value of a missing item with certainty, by means of the constraints and the remaining reported values for that data record. Examples include cases when only one subcomponent of a total is missing, or when some of the reported values correspond to the extremes of the edit bounds. For example, barring negative values, if the edits specify that 'expenses' are to be less than 'income', then when an 'income' of zero is reported, zero 'expenses' can be deduced. The number of logical relationships between questionnaire items is generally high for economic surveys. However, logical imputation should be based only on exact relationships, not approximate ones. For example, while it may be acceptable to impute zero wages and salaries to records with zero employees, it would be unwise to do the opposite and impute zero employees to records with zero wages and salaries, since the employees may be taking some form of leave without pay. If used properly, logical imputation can be a useful tool in the early stages of data grooming and is often performed as part of the editing process (see Granquist's chapter).

Historical imputation is most useful in repeated economic surveys, in particular for variables that tend to be stable over time (e.g., number of employees). Essentially, it consists of using values reported on previous occasions by the same unit to impute for missing data for the current occasion. Clearly this method attenuates the size of trends and the incidence of change. Variants of the method would adjust the previous values by some measure of a trend, often based on other (reported) variables on the record or within the imputation class. The method is most effective when the relationship between occasions is stronger than the relationship between units. When this is the case, and when previous values are available, historical imputation can be very appropriate as it is relatively unaffected by the nonresponse mechanism, even in cases of nonignorable nonresponse.

Mean value imputation replaces the missing value with the mean of the reported values for that imputation class, and as such, can only be contemplated for quantitative variables. This method is often used as a last resort. It preserves the respondent means, but it destroys the distributions and multivariate relationships, by creating an artificial spike at the imputation class mean value. Mean value imputation performs poorly when nonresponse is not random, unless the imputation classes are chosen exceptionally well. It yields the same results for means and totals as weighting class adjustment would if the same classes were used. However, for statistics other than means and totals, the results can be disastrous. For this reason, mean value imputation is not an imputation method of choice, but it can be useful as a backup strategy for cases when other imputation techniques fail, provided that the imputation classes are sufficiently fine.

Sequential hot deck methods replace the missing data item by a corresponding value from the last responding unit preceding it in the data file. There exists a number of variants on the method (Kalton, 1983; Sande, 1983). In general, the data file is processed sequentially, alternately either storing values of a clean record for later use, or using previously stored values to impute a record with missing data items. The sequential hot deck methods have their origins in social and demographic survey applications. Their adaptation to economic surveys has usually taken advantage of the continuous nature of the data. Thus, while in social applications the data files are usually divided into imputation classes, economic surveys often order the files according to some measure of size (possibly within broader imputation classes), or perhaps geographically. For example, under the hypothesis that neighbouring farms tend to behave alike, ordering an agricultural data file geographically will increase the likelihood of donors and recipients being spatially close. Care must be taken when imputing from sorted files that a systematic bias is not introduced by forcing the donors to be always "smaller" or "larger" than the recipients, depending on the direction of processing. Colledge, et al., (1978) have addressed this problem by modifying the sequential nature of the procedure so that donors that lie on either side of the recipient would be considered. The choice of classing versus sorting variables needs to be made carefully. In general, continuous variables that measure the "size" of the establishment are usually better used for sorting. Discrete variables related more to the "type" of the establishment are better reserved for classing, especially whenever it is important that class boundaries not be crossed during imputation.

An often cited disadvantage of sequential hot deck imputation is the fact that the same donors can be used a number of times (Kalton and Kasprzyk, 1986) whenever a "run" of nonrespondents is encountered. This can artificially create spikes in continuous distributions. To remedy this problem, some hot deck systems store three or four good records at a time, and cycle through them whenever two or more recipients have to be imputed in a row.

An appealing property of the sequential hot deck method is that actual, observed data are used for the imputation. Thus no invalid values can be imputed, as can be the case with mean, ratio, or regression methods (e.g., imputing nonintegral values); and the distributions are reasonably well preserved, as are means and totals. However, imputed values may be inconsistent when combined with existing responses. The robustness of the sequential hot deck against nonrandom nonresponse is directly related to the choice of the imputation classes and sorting variables. The sequential hot deck has become a popular imputation technique, as it is easily incorporated into existing data validation programs, and allows for data adjustment during and after imputation. Herein, however, also lies its weakness: many imputation systems incorporating sequential hot deck methods can become very large and therefore difficult and costly to maintain.

Nearest neighbour imputation, just as the sequential hot deck, uses data from clean records to impute missing values of recipient records. The difference lies in that the donors are chosen in such a way so that some measure of distance between the donor and the recipient is minimized. We note here that the distance measure is not a spatial one, but rather some multivariate measure based on the reported data, and thus the method is more appropriate for use with continuous data. The nearest neighbour method shares some of the sequential hot deck and regression imputation properties. Like the hot deck method, it uses actual, observed data and tends to preserve distributions. (For this reason it is sometimes classified as a hot deck method.) Since all missing variables for a given record are usually imputed from the same donor record, multivariate relationships are better preserved. The nearest neighbour method can control the effect of nonresponse bias due to nonrandom, but ignorable nonresponse. However, like the hot deck, it can use donors repeatedly when the nonresponse rate within an imputation class is high. Systems that incorporate the nearest neighbour method often allow for parametric specification of the minimum number of donors within an imputation class (Kovar and Whitridge, 1990) that must be available before an imputation is performed. Alternately, the number of times a record has been used as a donor can be incorporated as a penalty measure into the distance function (Colledge, et al., 1978). In this latter case, however, unlike the pure nearest neighbour method, the implementation would be sensitive to the file ordering.

Ratio and regression imputation methods make use of auxiliary variable(s), and replace the missing values with the corresponding ratio or regression predicted value. These are excellent imputation methods for establishment surveys, especially in cases where the auxiliary variable is both highly correlated with the (continuous) variable to be imputed, and available for all (or at least a high proportion) of the sampled units. For these methods to be effective, the response variable needs to be continuous, as does the auxiliary variable used with the ratio imputation method. The independent regression variables can be both continuous or discrete, making use of dummy variables in the discrete case. In most establishment surveys, ratio type relationships between questionnaire items are abundant, and should be put to good use. This is particularly true if it is suspected that the zero-one response indicators are correlated with the independent variable, that is in cases of nonrandom but ignorable nonresponse. Estimates of means and totals based on the ratio or regression imputed data sets are the same as those that would be obtained using weight adjusted ratio or regression estimators, provided that the same classes were used. The methods perform well in cases of random nonresponse, as well as nonrandom, but ignorable nonresponse situations, that is, in cases when the nonresponse propensity is related to the auxiliary variable(s) used by the ratio or regression. Despite the advantages, some distributional problems persist in cases where the independent (auxiliary) variables can be identical for several units, since in those cases the imputed values will be the same. As well, time and effort are required to develop the model, which then must be verified or adjusted regularly.

2.2 Stochastic Methods

Deterministic methods often reduce the variation of the variable of interest and sometimes also distort the distributions. Most stochastic imputation techniques have been introduced in an attempt to preserve the distribution and the variability of the data set (Little, 1988). As such, many of these methods are variations of those described above. As will be seen in Section 4 below, all imputation procedures introduce, to various degrees, an extra component of variation that must be accounted for. The use of stochastic methods in itself is not sufficient to establish correct variance estimates (Särndal, 1990).

Following the framework introduced by Kalton and Kasprzyk (1986), we let $\hat{y}_{mi}$ be the imputed value for the ith missing observation. A large number of imputation methods can be described, at least approximately, by the general model

$$\hat{y}_{mi} = b_{r0} + \sum_j b_{rj} x_{mij} + \hat{e}_{mi} ,$$

where x_{mij} are the values of the auxiliary variables (indexed by j) for the ith observation, b_{r0} and b_{rj} are the coefficients of a regression between y and x based on the responding units, and the $\hat{e}_{mi}$ are the residuals chosen in a prespecified manner.

The ratio and regression methods introduced above fit into this framework in an obvious way, by setting the residuals to zero. By letting the x variables be dummy variables representing the imputation classes, class mean imputation is described. The *random hot deck method*, in which donors are chosen completely at random with replacement from the entire sample of respondents within imputation classes, is represented in this framework by adding to the class mean a residual equal to the difference of the donor value and the class mean. Similarly, many other imputation methods can be explained within this framework.

Kalton and Kasprzyk classify imputation methods as deterministic or stochastic according to whether the residuals are set to zero or not. As such, for example, the above described mean imputation, ratio imputation, and regression imputation are considered deterministic. The random hot deck, on the other hand, is classified as stochastic.

Clearly, any deterministic method can be made stochastic by adding judiciously chosen residuals that satisfy the usual condition of zero expectation. In particular, regression imputation with the addition of a randomly chosen residual from the set of observed residuals has been studied recently by Lee, et al. (1991). To the best of our knowledge, no such method is used widely, perhaps because it would decrease the precision in the estimates of means, increase the complexity of the software, and would not by itself account for the variance due to imputation.

The random hot deck method, even with the use of imputation classes, is used less frequently in economic surveys than in social surveys. This is likely because some information is deemed lost by not exploiting the continuous nature of the data as is the case with ordered, sequential hot deck. Unlike the sequential hot deck, however, the random hot deck makes all clean records eligible for selection as donors at any one time. Because of this, the distributions are better preserved, and multiple use of donors is limited. An alternative known as the *weighted random hot deck method* that controls the number of times a donor is used, and ensures that all donors have a nonzero probability of selection, was proposed by Cox (1980). In fact, Cox proposed a way of including survey weights in the hot deck method, so that for each imputation class weighted estimates of means using imputed data are the same in expectation as weighted estimates of means using respondents only. The weighted hot deck procedure is the imputation analogue to weighting class adjustment (Cox and Cohen, 1985, Chapter 8).

2.3 Implementation

The last three decades have seen a large number of imputation systems developed. These range from highly customized and survey specific systems, to fairly generalized and reusable software. Initially, the imputation systems were just automating the manual, sequential, "detect and correct" rules, often taking advantage of the computer only in so far as adding more rules. This often resulted in systems so large and complex that no record could pass all the edits. In the mid 70's, Fellegi and Holt (1976) proposed that edit and imputation systems should adhere to four principles (see Pierzchala's chapter). A number of automatic imputation systems have since been developed adhering very closely to the Fellegi-Holt principles. The early efforts concentrated on applications with qualitative data, and gave rise to systems such as the Canadian CANEDIT, the Hungarian AERO and the Spanish DIA. Refer to Economic Commission for Europe (1992) for more information on these systems and further bibliography. In the mid to late eighties, two systems that deal with quantitative, economic data emerged. These are the U.S. Census Bureau's SPEER (Greenberg, 1987; Greenberg and Petkunas, 1990; Draper, et al., 1990), and the Statistics Canada's GEIS (Kovar, et al., 1988a; 1988b). Pierzchala describes the latter two automatic systems in his chapter. The availability of such generalized software has broadened the potential scope for imputation, and may tempt practitioners to use imputation for less conventional applications, such as mass imputation.

3. MASS IMPUTATION

To save resources, many statistical agencies are resorting to two-phase sampling of administrative records. Administrative record data for the first-phase sample are then used to select an efficient subsample for which additional information is collected by a sample survey. Classical estimates based upon the subsample require the derivation of secondary weights. An alternative is to impute the missing parts of the nonsampled primary units to create a complete rectangular file (Whitridge, et al., 1990b); this technique is known at Statistics Canada as mass imputation. The sampling rate for the second phase is often between 10 and 30%, although it can vary. As such the amount of missing data is quite high: 70-90%. Classical imputation methods, designed for low nonresponse rates, must be used with caution.

Mass imputation techniques have been applied at Statistics Canada to data from the Census of Construction since 1978 (Colledge, et al., 1978). This survey uses a sample of income tax records for the first phase information, then selects a subsample to collect detailed second-phase information. In this example, approximately 30% of the original tax sample is subsampled, and thus 70% is imputed. Similar methods were applied by Hinkins and Scheuren (1986) for a sample of US corporate tax returns, and most recently by Clogg, et al. (1991) for industry and occupation classification. These examples had imputation rates between 75 and 90%. Another study by Cox and Folsom (1981) attempted to apply mass imputation to data from the National Medical Care Expenditure Survey. The results showed that mass imputation was successful for imputing discovered medical visits. However, further examination revealed that significant biases could be introduced for variables that were not controlled in the imputation process. It was concluded that mass imputation was not feasible for this survey.

Mass imputation can also be useful when large amounts of data are missing for operational reasons. For example, Statistics Canada uses agricultural income tax data to produce balance sheet estimates. However, farmers are not legally required to file a balance sheet. The resulting data set is thus missing large amounts of data for many farmers, but the missing farmers do not form a random subsample of all farms.

For many imputation methods, there exists a corresponding weight adjustment method (Folsom, 1981; Whitridge, et al, 1990a). For example, in the case of simple random sampling, if the sample mean is used for imputation, then mass imputation is equivalent to the direct expansion estimator. Using a ratio estimator with auxiliary data to mass impute for variables in the subsample would be equivalent to using the ratio estimator at the estimation stage. If a more complex imputation method is chosen, such as nearest neighbour imputation, then this is implicitly equivalent to an expansion estimator with variable weights, where each record is assigned a weight corresponding to the number of times it was used as a donor. Weighting has the distinct advantage that it is based on classical statistical theory and hence it is easily defensible and its properties and behaviour are well known. It is a simple and efficient way of estimating for information of a subsample that is relatively independent of the sample. Weighting is to be recommended when estimation of variances, covariances and correlations is routinely considered by users. However, mass imputation has a place in survey processing for cases such as those where quick ad hoc estimates are needed or where second-phase sample weights are difficult to calculate, as is the case when information is missing for operational reasons.

Originally, mass imputation techniques were defended using "missing at random" arguments, since it is known that most imputation methods perform well under such conditions. That is, since the subsample was chosen at random, it is argued that those units not chosen form a random sample of "missing" records. In addition, the equivalence between weighting and mass imputation is also comforting. However, further studies (Colledge, et al., 1978) suggest that mass imputation can be useful in the presence of ignorable nonresponse in that it may help attenuate the bias. These results were further born out by Michaud (1987). For nonrandom, ignorable nonresponse, mass imputation might be preferable to weighting, since it can make more extensive use of auxiliary information when imputing the missing second phase segments. The choice of imputation method is hence very important. Mass imputation might also be recommended for situations where there are high correlations between variables in the sample and those in the subsample. Mass imputation can better preserve the relationships between variables, depending upon the imputation method chosen, as it can make more extensive use of auxiliary variables. For example, nearest neighbour imputation lends itself well to mass imputation, since it can take advantage of the multivariate relationships between variables in the sample and those in the subsample. By contrast, in the presence of nonignorable nonresponse, most imputation methods perform poorly in traditional settings (Rancourt, et al., 1992), let alone in cases of mass imputation.

Typical mass imputation applications involve imputing large amounts of data from a small subsample. Since this is the case, it is especially important that all imputed data be flagged and that users of the data be aware of the imputation. In the extreme case, two separate files could be kept: the complete rectangular file for tabulation and the original file for analysis. With such a large amount of imputed data, the evaluation of the impact of imputation becomes critical. Methodology to evaluate the imputation must be developed carefully. When the "nonresponse" is incurred by design and the underlying model is known, simulation studies can be helpful in evaluating mass imputation.

4. DATA ANALYSIS

Rightly or wrongly, imputation methods are often devised with the aim of predicting the correct response. As a result, the imputed data sets provide good estimates of means and totals. With care, the distributions also can be preserved reasonably well. The situation is not as favorable when it comes to estimates of variances and correlations. A number of studies (Santos, 1981; Kalton and Kasprzyk, 1982; 1986; Little, 1986) show that imputation tends to attenuate correlations between the imputed variables to various degrees, though the situation is improved if good auxiliary variables are used at the imputation stage. Treating the imputed values as observed values leads to underestimation of variances of the estimated means and totals if standard formulas are used (Rubin, 1978). This leads to confidence intervals that are too short, and the tendency to declare significance when none exists. The problem becomes more serious as the proportion of missing items increases (Särndal, 1990). In the case of stochastic imputation methods, for which there exists an associated model for nonresponse, a model-based estimate of variance can be calculated.

It can be shown that the true variance of the estimator of the mean, V_{Tot} , can be written as $V_{Tot} = V_{Sam} + V_{Imp} + V_{Mix}$ (Särndal, 1990), where V_{Sam} is the sampling variance component, V_{Imp} is the variance introduced by the imputation method in question, and V_{Mix} is a covariance term between V_{Sam} and V_{Imp} which in most cases is negligible or zero. An estimator of V_{Sam} could be obtained by adding a term to the usual variance formula to correct for the fact that the usual formula understates the sampling variance component when there are imputed values in the data set. This adjustment would depend on the method used, being relatively high in the case of mean imputation and negligible in the case of the random hot deck method. Unfortunately, the interest rarely lies in estimating what the variance would have been, had there been no nonresponse; but rather we are interested in the variance of the estimates based on the present, imputed file, V_{Tot}. To estimate V_{Tot}, an additional component of variance due to the imputation mechanism, V_{Imp}, must be estimated. Empirical evidence suggests that, depending on the imputation method, the underestimation when using the usual variance formulas can be of the order of 2 to 10% in the case of 5% nonresponse rate, but as high as 10 to 50% in the case of 30%

nonresponse rate (Kovar and Chen, 1993). Three general methods that estimate the variance due to imputation are presented below.

Rubin (1977; 1978; 1986; 1987), proposed the technique of multiple imputation to estimate the variance due to imputation by replicating the process a number of times, and estimating the between-replicate variation. A number of variants of this approach have been put forth in the literature. Theoretically, the method is defensible, though care must be taken that the imputation method be "proper" in Rubin's (1987) sense. Rubin defines "improper" methods as those that do not show enough variability between replicates to provide appropriate variance estimates. (In the extreme, all deterministic methods are obviously "improper", in that they produce identical imputed files in each replicate. Note also that even the random hot deck method is "improper".) Operationally, the multiple imputation method is unattractive as it incurs high computer costs, necessitates the maintenance of multiple files, and complicates data dissemination. The relatively high cost of imputation, even for moderately large files, tempts practitioners of multiple imputation to use a low number of replicates. This results in poor precision of the estimator of the variance due to imputation. Furthermore, recent results suggest that multiple imputation can produce inconsistent variance estimates when the inference and imputation classes cross (Fay, 1992; 1993). As such, "it is questionable whether the multiple imputation approach is feasible for routine analyses. It may be best reserved for special studies" (Kalton and Kasprzyk, 1986).

More recently, Särndal (1990) outlined a number of model-assisted estimators of variance, while Rao and Shao (1992) proposed a method that corrects the usual jackknife variance estimator. These methods are appealing in that only the imputed file (with the imputed fields flagged) is required for variance estimation. Särndal's model-assisted approach requires different variance estimators for each imputation method, but yields consistent variance estimates, provided the model holds. Several estimators have been proposed and evaluated empirically with very positive results (Lee, et al., 1991). On the other hand, the Rao and Shao adjusted jackknife method requires the implementation of only one estimator, though the temporary adjustment of the imputed values depends on the method of imputation. The actual adjustments for a number of imputation methods can be found in Rao (1992) and Rao and Shao (1992). The method is design consistent (p-consistent) under uniform nonresponse irrespective of the model, as well as design-model unbiased (pm-unbiased) under the usual linear model and any ignorable nonresponse mechanism (Rao, 1992). Empirical results show that in cases of uniform nonresponse the adjusted jackknife method essentially eliminates any underestimation of the variance, for both simple as well as complex survey designs (Kovar and Chen, 1993). In cases of nonrandom, but ignorable nonresponse, the ratio imputation method performs equally well. Rancourt, et al. (1993) have recently studied the adjusted jackknife estimator when more than one imputation method is used for the same data set.

In cases where the nonresponse mechanism is not ignorable, all imputation methods tend to produce severely biased point estimates. As such, variance estimation is of minimal interest, as the real problem lies in estimating the mean squared error. That is, more attention needs to be concentrated on improving the point estimates and their bias. Some preliminary results on this front have been put forth by Rancourt, et al. (1992). We note also that the performance of any of the above three techniques is less than satisfactory in the particular case of nearest neighbour imputation, except under ideal conditions: that is, when the auxiliary variables are very highly correlated with the response variables, and the nonresponse is uniform (Kovar and Chen, 1993).

5. EVALUATION OF IMPUTATION

The success or failure of the imputation process is difficult to evaluate, but audit trails and status reports can help. Theoretical results are scarce, though Little and Rubin (1987) provide some, along with good references. Once an imputation strategy has been implemented, it is very important to evaluate the impact the imputation has on the final data. Despite the fact that naively calculated estimates of sampling variances seem to decrease after imputation, this does not indicate an improvement in the quality of the data. However, the basic objective of imputation is to reduce the bias due to nonresponse. If the bias is indeed reduced, and the increase in the variation due to imputation does not counteract this reduction, then the quality of the estimates may be improved in the mean square error sense. However, there are other aspects that should be considered when evaluating the impact of imputation (Sande, 1982). Performance measures, or evaluation statistics, provide a standard set of measures that can be used by a manager to plan, monitor and control all aspects of the survey process. Such statistics can be used not only to provide information about the quality of the data, but to provide a basis for improvements for the next survey occasion, as described in Granquist's Chapter.

Performance measures at the edit and imputation stage reflect the quality of the data being edited, the quality of the edits themselves, and the magnitude of the changes brought about by imputation. They assist in the evaluation of the difference between the data that are actually collected and valid data. Any changes made to the data, such as those due to imputation, should be flagged. Complete documentation of all survey processes is essential for later evaluation.

Sande (1981) outlines a set of descriptive statistics that can be used to monitor the edit and imputation process. These statistics include counts of the number of times certain fields were identified for imputation, the number of times they were imputed, as well as comprehensive statistics specific to the nearest neighbour imputation method (e.g., distance between donor and recipient, transformed values of matching fields etc.). Other performance measures for imputation might include counts of the number of records that required at least one imputation, the percentages with which certain methods were used on given fields, the number of records with exactly a given number of fields imputed, as well as specific counts that depend on the imputation methods chosen. For example, if donor imputation is used, then the identifier of the donor should be retained so that statistics about how many times certain donors were used can be examined. Once these performance measures have been gathered during the imputation process, it is important that they be analyzed as fully as possible. These imputation performance measures reflect not only the data, but the survey design as well. For example, very high imputation rates for a certain field may mean that the question was poorly understood by the respondents, and may suggest that improvements are necessary to the questionnaire design or to the data collection and training procedures. Quality assurance principles need to be applied at all steps when establishing an imputation strategy. Statistics such as those described here help monitor the quality of the imputation process.

Beyond producing tables of counts of changes in data values due to imputation, it is important to evaluate or validate the final data themselves. To do this, evaluation tables consisting of the number of times a reported value was increased, or decreased, or remained unchanged, and the corresponding estimates should be considered. Typically, such tables would be produced at several different levels of aggregation, perhaps corresponding to the estimates to be tabulated. When the estimates appear suspicious, effort can then be spent ensuring that large or important units that have a significant impact on the estimates are correct, rather than manually verifying all imputed records, which can be labor intensive and expensive. Such a technique uses a "macro" approach, rather than the more commonly adopted "micro" approach, in that it looks at the impact of imputation on the estimates, rather than on specific imputed records. This method involves the same principles that form the basis for the selective editing discussed by Granquist in his chapter. It ensures that the global estimates will be reasonable, even though all underlying relationships in the data might not hold for all individual records. One way to implement this approach is to produce and examine tables of the records with the highest values for specific fields, or of the records with the largest weighted contribution to the estimates.

Another strategy to evaluate imputed data would be to examine and compare them against the expected results. This evaluation could involve external data, perhaps from administrative sources or other surveys in a comparative exercise. Probably the most effective method of evaluating the imputation is through follow-up, reinterview studies. Respondents could be contacted to resolve inconsistent responses and to complete missing responses. This would allow the analyst to estimate the bias and hence the mean square error under alternate imputation approaches. Attempts to evaluate imputation through simulation studies are usually of questionable value, since too many assumptions must be made with respect to the nonresponse mechanism. This is in direct contrast to the mass imputation situation when data are missing by design.

Evaluating the impact of imputation is often difficult, since it requires a prespecified notion of the "true" values, as well as how much imputation is acceptable. These are very subjective measures. Acceptable imputation rates depend upon the response rates, the reasons for imputation, and which specific fields are being imputed. For certain key variables that are always reported, the imputation rate should likely be low, since any imputation represents an actual change to reported data. However, for variables that tend to be missing on many questionnaires, higher imputation rates may be acceptable. The reason for imputation, be it due to nonresponse or to inconsistent data, should be considered when the impact of imputation is being evaluated, since it will have different effects upon the data. For example, imputing for nonresponse will result in a positive change to the unweighted estimates of totals, since all fields are necessarily increased (assuming positive data). Imputing for inconsistent data should not have a large effect upon the estimates themselves, since changes can either increase or decrease the values. In the end, once the imputation-revised data have been evaluated, it is important to step back and consider the costs and benefits of the imputation exercise. Was there a decrease in the nonresponse bias, and at what cost in terms of both resources and increase in total variance?

6. CONCLUDING REMARKS

All establishment surveys suffer from the effects of nonresponse. Whether the nonresponse is due to errors, misconceptions, noncontacts, or refusals, it is rarely random. Systematic nonresponse patterns can be responsible for serious biases in the survey estimates. Methods of dealing with such biases include survey weight adjustment and imputation. In this chapter we have briefly described some common imputation methods and provided suggestions on the appropriateness of their use in various situations. We note with satisfaction that most of these methods produce imputed data files from which good first order estimates can be produced. However, some methods, such as the mean value imputation, should be used with caution, because of their negative impact on distributional properties and multivariate relationships. On the other hand, methods such as the ratio, regression and nearest neighbour imputation that make use of auxiliary information, are well suited to establishment survey data and can go far in reducing nonresponse bias. The type of auxiliary information and its quality needs to be considered when choosing the appropriate method. The assumptions regarding the randomness of the underlying response mechanism, though usually unverifiable, must also be evaluated and the robustness of the imputation method against departures from these assumptions must be considered.

The availability of an imputation-revised data file facilitates the production of consistent, ad hoc tabulations. However, caution must be exercised when large scale imputation is attempted and when data analyses using imputed files are performed. We provided some guidance on the suitability of the mass imputation technique, and cautioned the would be practitioners against its use without a proper evaluation.

While mass-imputed data sets may be operationally more convenient, traditional weighting methods are likely preferable. However, when frequent, ad hoc estimates have to be produced in a consistent fashion, a mass-imputed data file may serve the purpose well. Furthermore, when there is a reason to believe that the units to be imputed are ignorably nonrandom, mass imputation may, in fact, help eliminate biases that weighting would not. Using imputed data files for data analyses, significance testing, and variance estimation in particular poses new challenges and requires the analyst's attention. We presented a brief overview of several recently developed techniques for this purpose. Multiple imputation, model-assisted methods, and adjusted jackknife techniques were reviewed.

Finally, the necessity of flagging all imputed data cannot be overstressed. All imputation methods fabricate data to some extent, and the user must be made aware of this. Only the collecting agency can annotate the data properly! The maintenance of accurate audit trails and a complete documentation of the entire imputation process is essential for its ultimate success.

ACKNOWLEDGEMENTS

The authors are grateful for the many useful comments provided by the section editor.

REFERENCES

Bogestrom, B., Larsson, M., and Lyberg L. (1983), "Bibliography on Nonresponse and Related Topics," in W.G. Madow, I. Olkin, and D.B. Rubin, (eds.) *Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies*, pp. 479-567, New York: Academic Press.

Chapman, D.W. (1976), "A Survey of Nonresponse Imputation Procedures," *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 245-251.

Clogg, C., Rubin, D., Schenker, N., Schultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, **86**, pp. 68-78.

Colledge, M.J., Johnson, J.H., Paré, R. and Sande, I.G. (1978), "Large Scale Imputation of Survey Data," *Survey Methodology*, **4**, pp. 203-224.

Cox, B.G. (1980), "The Weighted Sequential Hot Deck Imputation Procedure," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 721-726.

Cox, B.G. and Cohen, S.B. (1985), *Methodological Issues for Health Care Surveys*, New York: Marcel Dekker.

Cox, B.G. and Folsom, R.E. (1981), "An Evaluation of Weighted Hot-deck imputation for Unreported Health Care Visits," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 412-417.

Draper, L., Greenberg, B. and Petkunas, T. (1990), "On-line Capabilities in SPEER," *Proceedings of Symposium 90: Measurement and Improvement of Data Quality*, Statistics Canada, pp. 235-243.

Economic Commission for Europe (1992), *Statistical Data Editing Methods and Techniques*, Volume 1, New York: United Nations, Conference of European Statisticians.

Fay III, R.E. (1992), "When are Inferences from Multiple Imputation Valid?," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 227-232.

Fay III, R.E. (1993), "Valid Inferences From Imputed Survey Data," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. ??-??.

Fellegi, I.P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, **71**, pp. 17-35.

Folsom, R.E. (1981), "The Equivalence of Generalized Double Sampling Regression Estimators, Weight Adjustments and Randomized Hot-deck Imputations," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 400-405.

Granquist, L. (1984), "On the Role of Editing," *Statistisk Tidskrift*, **2**, pp. 105-118.

Granquist, L. (1990), "A Review of some Macro-Editing Methods for Rationalizing the Editing Process," *Proceedings of Symposium 90: Measurement and Improvement of Data Quality*, Statistics Canada, pp. 225-234.

Greenberg, B. (1987), "Edit and Imputation: A discussion," *Proceedings of the Third Annual Research Conference*, U.S. Bureau of the Census, pp. 204-210.

Greenberg, B. and Petkunas, T. (1990), "SPEER (Structured Program for Economic Editing and Referrals)," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 95-104.

Hinkins, S. and Scheuren, F. (1986), "Hot Deck Imputation Procedure Applied to a Double Sampling Design," *Survey Methodology*, **12**, pp. 181-196.

Kalton, G. (1983), *Compensating for Missing Survey Data*, Survey Research Center, University of Michigan.

Kalton, G. and Kasprzyk, D. (1982), "Imputing for Missing Survey Responses," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 22-31.

Kalton, G. and Kasprzyk, D. (1986), "The Treatment of Missing Survey Data," *Survey Methodology*, **12**, pp. 1-16.

Kovar, J.G. (1991), "The Impact of Selective Editing on Data Quality," Conference of European Statisticians, Work Session on Statistical Data Editing, Geneva, Switzerland, Working Paper No. 5.

Kovar, J.G. and Chen, E.J. (1993), "Jackknife Variance Estimation under Imputation: An Empirical Investigation," Statistics Canada, Methodology Branch Working Paper No.METH-93-007E. An invited paper presented at the Annual Meeting of the Statistical Society of Canada, Edmonton, Alberta, June 1992.

Kovar, J.G., MacMillan, J. and Whitridge, P. (1988a), "Overview and Strategy for the Generalized Edit and Imputation System," Statistics Canada, Methodology Branch Working Paper No. BSMD 88-007E/F.

Kovar, J.G., Mayda, J., Dumičić, K. and Dumičić, S. (1992), "Selected Bibliography on Data Editing and Imputation and Related Topics," Conference of European Statisticians, Work Session on Statistical Data Editing, Washington, D.C., Working Paper No. 12.

Kovar, J.G. and Whitridge, P. (1990), "Generalized Edit and Imputation System: Overview and Applications," *Revista Brasileira de Estatística*, **51**, pp. 85-100.

Kovar, J.G., Whitridge, P. and MacMillan, J. (1988b), "Generalized Edit and Imputation System for Economic Surveys at Statistics Canada," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 627-630.

Lee, H., Rancourt, E. and Särndal, C.E. (1991), "Experiments with Variance Estimation from Survey Data with Imputed Values," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 690-695.

Little, R.J.A. (1986), "Survey Nonresponse Adjustments for Estimates of Means," *International Statistical Review*, **54**, pp. 139-157.

Little, R.J.A. (1988), "Missing-Data Adjustments in Large Surveys," (With discussion) *Journal of Business and Economic Statistics*, **6**, pp. 287-301.

Little, R.J.A. and Rubin, D.B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

Madow, W.G., Olkin, I., and Rubin, D.B. (eds.) (1983), *Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies*, New York: Academic Press.

Michaud, S. (1987), "Weighting vs Imputation: a Simulation Study," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 157-161.

Pierzchala, M. (1990), "A Review of the State of the Art in Automated Data Editing and Imputation", *Journal of Official Statistics*, **6**, pp. 355-377.

Rancourt, E., Lee, H. and Särndal, C.E. (1992), "Bias Corrections for Survey Estimates from Data with Imputed Values for Nonignorable Nonresponse," *Proceedings of the 1992 Annual Research Conference*, U.S.Bureau of the Census, pp. 523-539.

Rancourt, E., Lee, H. and Särndal, C.E. (1993), "Variance Estimation Under More than One Imputation Method," *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, pp. ??-??.

Rao, J.N.K. (1992), "Jackknife Variance Estimation under Imputation for Missing Survey Data," Statistics Canada, unpublished report.

Rao, J.N.K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation," *Biometrika*, **79**, pp. 811-822.

Rubin, D.B. (1977), "Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, **72**, pp. 538-543.

Rubin, D.B. (1978), "Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 20-34.

Rubin, D.B. (1986), "Basic Ideas of Multiple Imputation for Nonresponse," *Survey Methodology*, **12**, pp. 37-47.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Sande, G. (1981), "Descriptive Statistics Used in Monitoring Edit and Imputation Process," paper presented at the 13th Symposium on the Interface, Pittsburgh, Pennsylvania.

Sande, I.G. (1979), "A Personal View of Hot-deck Imputation Procedures," *Survey Methodology*, **5**, pp. 238-258.

Sande, I.G. (1982), "Imputation in Surveys: Coping with Reality," *American Statistician*, **36**, pp. 145-152.

Sande, I.G. (1983), "Hot-deck Imputation Procedures," in W.G.Madow and I.Olkin, (eds.) *Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium*, New York: Academic Press.

Santos, R.L. (1981), "Effects of Imputation on Regression Coefficients," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 140-145.

Särndal, C.E. (1990), "Methods for Estimating the Precision of Survey Estimates when Imputation has Been Used," *Proceedings of Symposium 90: Measurement and Improvement of Data Quality*, Statistics Canada, pp. 337-350.

Whitridge, P., Bureau, M. and Kovar, J.G. (1990a), "Mass Imputation at Statistics Canada," *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, pp. 666-675.

Whitridge, P., Bureau, M. and Kovar, J.G. (1990b), "Use of Mass Imputation to Estimate for Subsample Variables," *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 132-137.

Yansaneh, I.S. and Eltinge, J.L. (1993), "Construction of Adjustment Cells Based on Surrogate Items or Estimated Response Propensities," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. ??-??.

SUGGESTED INDEX HEADINGS