Edit and Imputation : From Suspicious to Scientific Techniques

Eric Rancourt
Statistics Canada, Household Survey Methods Division
Tunney's Pasture
Ottawa, ON, Canada K1A 0T6
Eric.Rancourt@Statcan.ca

1. Introduction

In carrying out surveys the aim has always been to produce estimates based on all units of the sample. However this is rarely possible since a number of factors may lead to missing data. Since the hope is usually to work with complete files, the idea of "assigning" a value to missing fields came naturally; be it for practical reasons or because of the level of advancement of the theory. In the context of censuses, the idea of representative samples is absent. It is then even more natural to assign a value to missing data instead of adjusting (or creating) weights.

Performing these assignments is referred to as *imputation*. It has evolved from an entirely manual ad hoc procedure aimed at trying to correct what was thought to be wrong to an ensemble of almost completely automated scientific techniques.

Because of the need for large amounts of tabulations and the relative lack of (or secondary) interest from statisticians for imputation practices, these practices have always been largely ahead of the theory supporting them. However, the situation has recently started to change.

In this paper, we outline the evolution of edit and imputation with an emphasis on imputation.

2. History of imputation

Imputation began to be used outside the context of surveys, before they were carried out. One early example is Allan and Wishart (1930), where they discussed a method to estimate missing plot values in experimental field work. A similar approach can be found in Yates (1933). During that period, references on imputation methods were scarce and estimates were usually produced using complete-case approaches or by adjusting the final numbers.

In surveys, imputation methods and practices were mainly developed in the context of censuses. For instance, until the 1950's the Canadian Census used a method referred to as Deming's method to impute for missing values. This meant using the frequency distribution of the previous census and manually recording the values on cards. Then imputation would be obtained by shuffling the deck of cards and assigning the value of the top card to the missing value (see Bankier, 1996).

The year 1953 was the one when imputation was first used in a publication in its context. As mentioned in Hansen, Hurwitz and Madow (1953), a number of corrections were required in the treatment of nonresponse for the 1948 Survey of Retail Shares. Later in the book (p. 546), it is said:
"As described in Sec. A-7, "imputation" based on knowledge of previous month's sales, payroll, etc., were made for nonrespondent cases." To the best of the author's knowledge, this is the first use of the word imputation in the context of statistics. Before then, the terminology used consisted of words such as assignments, corrections or estimations. The word imputation had existed long before, but in a different context. The Oxford English dictionary gives: "The action of attributing something, usually a fault or a crime to someone." But an important use of the word imputation is in the economic theory of value when an amount is attributed to a category. This is actually the root meaning of the Latin verb *imputare* that can be found in texts of authors such as Tacitus or Pliny.

The main development that contributed to the advancement of imputation methods is the computer. Though there are earlier examples (Healy and Westmacott, 1956), the 1960 US Census

(Census of Population, 1960; Daly and Eckler, 1961) is the first survey where most of the editing was done by an electronic computer. Similarly, the 1961 Canadian Census was the first to fully use computers in the modern sense. Nordbotten (1963) presents an important discussion of the topic.

Computers then allowed for the use of enhanced methods of edit and imputation. It was now possible to use the auxiliary information "around" the missing values. Hot-deck techniques could now be used instead of cold-deck approaches.

Once the use of computers became well established, edit and imputation practices became very popular and more sophisticated (see Chapman, 1976 for an early account and Kovar and Whitridge, 1995 for a more recent one). Indeed, a number of imputation systems came to light. Examples are CANEDIT, GEIS and CANCEIS/NIM in Canada and SPEER in the US. Although these systems (and others) have been available, the theoretical background behind imputation did not follow suit. This has totally changed, since in recent years (last three decades) we have witnessed the emergence of principles that set the foundation for imputation theory.

3. Edit and Imputation principles

Minimum change principle. Traditionally, the edit and imputation efforts have been targeted at correcting the data until a satisfactory level of consistency was achieved. Following the 1971 Canadian Census, Fellegi and Holt (1976) laid out the foundations of a new principle guiding edit and imputation. The idea was to seek a solution that makes records pass all edits by changing the smallest possible number of fields, thereby preserving as much as possible the integrity of the information provided by the respondents. This was also an important change in that it allowed to mathematically formalise the creation and application of a set of edits.

Formal precision measure. Statisticians have always been uncomfortable with inference from samples containing missing data, whether the weights are adjusted or the data imputed. To "formalise the subjective notion" that traditional inference is not appropriate when there is imputation, Rubin (1977) developed multiple imputation. Equipped with this tool, producers of micro data files could now provide external users with properly imputed files that enable them to make correct inferences. Rubin (1996) provides a detailed account of the uses of multiple imputation.

Process improvement. In the past, edit and imputation was viewed as a correction process (Szameitat and Zindler, 1965) to improve the data. With the popularity of total quality management during the 1980's, this view has changed to become a more general one of learning and improving the process and the survey. While Granquist (1984) initially discussed the role of editing, Granquist and Kovar (1997) present new ideas about editing, where it is clear that edit and imputation are not tools to correct the data, but rather a means of improving the survey processes and the overall quality.

Imputation as a modelling activity. All the multiple imputation literature is filled with model-based imputation, but only since Kalton and Kasprzyk (1986) has there been a documented model framework to represent single imputation. In this framework, the imputation methods are represented by

$$\hat{y}_k = b_{0r} + \sum b_{jr} z_{jr} + \hat{e}_k ,$$

where y is the variable of interest, z_j the auxiliary variables, b_j the estimated parameters, $\hat{e}_k$ selected or estimated errors and r the response set. Imputation consists of constructing classes, selecting appropriate variables, identifying relationships between variables and assessing the quality of the model on respondents. Therefore, it is a modelling exercise. For instance, Särndal (1992) builds on the above model framework by making use of the idea of a superpopulation to produce a method to correctly estimate the variance in presence of imputation.

Conditionality of the response mechanism. There are two ways of representing a response mechanism in the survey context. One is to assume that it is conditional on the sample. In this case, the probability of responding is not viewed as an intrinsic characteristic of the respondents, but

rather as the result of a number of factors and stimuli directly related to the survey. Most of the theoretical work on nonresponse and imputation has been influenced by this approach (Dalenius, 1983). Another approach is to assume that the response probability is independent of the sample. Fay (1991) has discussed the idea and used it to simplify calculations of expected values. There is currently no consensus on which approach should be used, as each appears to have its own merits.

4. Imputation frameworks

Over the years, imputation has developed under a number of frameworks. It is important to realise which framework is being used when choosing an imputation strategy in order to be able to clearly spell out the underlying assumptions and understand the properties of the imputation methods. One possible classification of these frameworks is:

1) Experience-based: Experts are performing the imputation based on their knowledge;
2) Distribution-based: Distributions are estimated and imputation is obtained from them;
3) Model-based: A model is constructed, validated and used to produce values to impute;
4) Frequency-based: Under a response mechanism, values are imputed without using a model;
5) Empirical-based (donor-based): A donor is found and its values are used for imputation.

These frameworks have helped in the development of theoretical properties of imputation methods. For instance, framework 3) enabled Rancourt (1999) to provide the properties of nearest neighbour imputation.

5. Impact of imputation

If the estimation activity stopped after point estimation, then the quality of estimates would remain unknown. That is why we should estimate the variance, which also allows for producing inference about parameters. For a long time, survey statisticians have used imputation without (formally) knowing its impact. However, there now exist a number of methods to correctly estimate the variance in presence of imputation. Lee, Rancourt and Särndal (2000) present a detailed account of such methods along with an extensive comparison of the methods, both qualitative and quantitative.

6. Conclusion

In the past, edit and imputation was viewed as a second-class statistical processes. With the advent of computers and the development of the theory, they have now come to be accepted as important parts of the statistical processes involved in a survey. In fact, imputation is now one of the most active and fruitful areas of research in survey methodology as there is almost always a paper on or related to imputation in each issue of many statistical journals.

REFERENCES

Allan, F.E. and Wishart, J. (1930). A Method of Estimating the Yield of a Missing Plot in Field Experimental Work, *Journal of Agricultural Sciences*, 20, 399-406.

Bankier, M. (1996). Adjusting for Non-Response in the Canadian Census. Correspondence with Nathan Keyfitz, Statistics Canada, Ottawa.

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

Dalenius, T. (1983). Some Reflections on the Problem of Missing Data. In: *Incomplete Data in Sample Surveys*, W.G. Madow and I. Olkin eds., Vol. 3, 411-413, New York: Academic Press.

Daly, J.F. and Eckler, A.R. (1961). Applications of Electronic Equipment to Statistical Data-Processing in the U.S. Bureau of the Census. *Bulletin de l'Institut International de Statistique*, 319-327.

Fay, R.E. (1991). Design-Based Perspective on Missing Data Variance. *Proceedings of the 1991 Annual Research Conference*. US Bureau of the Census, 429-440.

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

Granquist, L. (1984). On the Role of Editing. *Statistical Review*, 2, 105-118.

Granquist, L. and Kovar, J.G. (1997). Editing of Survey Data: How much is Enough. In: *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin eds., 415-435, New York: John Wiley and Sons.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory. Vol. I and II. NewYork: John Wiley and sons.*

Healy, M. and Westmacott, M. (1956). Missing Values in Experiments Analysed on Automatic Computers. *Applied Statistics*, 203-206.

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

Kovar, G. and Whitridge, P. J. (1995). Imputation of Business Survey Data. In: *Business Survey Methods*, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott eds., 403-423, New York: John Wiley and Sons.

Lee, H., Rancourt, E., and Särndal, C.-E. (2000). Variance Estimation from Survey Data under Single Value Imputation. *Working paper*. Methodology Branch, Statistics Canada, HSMD-2000-006E.

Nordbotten, S. (1963). Automatic Editing of Individual Statistical Observations. *Conference of European Statisticians Statistical Standards and Studies*, 1-55.

Rancourt, E. (1999). Estimation with Nearest Neighbour Imputation at Statistics Canada. *Proceedings of the Survey Research Methods Section*, American Statistical Association. 131-138.

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

Rubin, D.B. (1996). Multiple Imputation after 18+ years. *Journal of the American Statistical Association*. 91, 473-489.

Särndal, C.-E. (1992). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. *Survey Methodology*, 18, 241-252.

Szameitat, K. and Zindler, H.-J. (1965). The Reduction of Errors in Statistics by Automatic Corrections. *Automatic Detection and Correction of Errors in Data Processing on Electronic Computers*, 395-416.

U.S. Census of Population 1960 : Characteristics of the Population.

Yates, F. (1933). The Analysis of Replicated Experiments When the Field Results are Incomplete. *Empire Journal of Experimental Agriculture*, 1, 129-142.

RESUMÉ

Les méthodologistes d'enquête ont eu à traiter le problème de la non-réponse et de données inacceptables depuis les premières enquêtes. Afin d'améliorer les résultats, ils ont toujours effectué certaines corrections. Avec l'avancement de la technologie, ces corrections (maintenant appelées vérification et imputation) sont devenues automatisées. Étant donné la disponibilité des techniques et outils, la pratique a presque tout le temps été en avance sur le développement des fondements théoriques. C'est seulement lors de ces dernières décennies que nous avons été témoins de l'émergence de résultats fournissant les bases théoriques des méthodes. Pendant ce temps, notre compréhension du rôle de la vérification et de l'imputation a évolué de la perspective de corriger à celle de contribuer à améliorer les processus et la qualité globale de l'enquête.