SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE 1979, VOL. 5 NO. 2

A PERSONAL VIEW OF HOT DECK IMPUTATION PROCEDURES[1]

Innis G. Sande[2]

A Hot Deck imputation procedure is defined to be one where an incomplete response is completed by using values from one or more other records on the same file and the choice of these records varies with the record requiring imputation.

General approaches to Hot Deck imputation are outlined, with emphasis on the interaction between the edit constraints and the imputation procedures. Distance functions can be constructed on a mixture of categorical and numeric fields, can be modified to take account of the relative importance of fields and can discriminate against less desirable donors. Matching fields may be correlated with missing fields, may be linked with missing fields by edits or may be natural stratification variables; but increasing the number of matching fields does not necessarily result in a better match. It is important to audit the imputation process and to summarize its performance.

Hot Deck procedures should be evaluated to study the bias and reliability of the estimates, donor usage and frequency of imputation failure in terms of a variety of conditions of the data and variations of the imputation procedure. It appears that the only generally available approach to evaluation is by simulation.

1. INTRODUCTION

There is a growing awareness of the use of imputation in the preparation of data files. As the files get larger, the need for automatic imputation becomes essential.

---

[1] Adapted from a paper presented to the Symposium on Incomplete Data, Washington, D.C., August 1979.

[2] Innis G. Sande, Business Survey Methods Division, Statistics Canada.

Methods of imputation vary considerably, ranging from the use of default values to the development of complex models. One class of imputation procedures is the so-called Hot Deck type, in which an incomplete response is completed by using values from one or more other records on the same file and the choice of these records varies with the record requiring imputation.

This paper describes the author's perception of Hot Deck procedures as a solution to the imputation problem. This necessitates first discussing her perception of the imputation problem, since a different viewpoint could very well result in a different assessment.

2. GENERAL OBSERVATIONS ABOUT IMPUTATION

Imputation is the process of estimating individual values in a data set. It is a direct generalization of the "missing observation" problem in Analysis of Variance and the "Incomplete Data" problem in Multivariate Analysis. Solutions of these two problems typically make use of very specific model assumptions about the data.

The need for imputation arises in two ways:

(i) a record (multivariate observation for a single case) contains one or more missing values because the data is unavailable; or

(ii) a record is inconsistent, i.e. its values do not satisfy natural or reasonable constraints (edits) and one or more values are designated for change (and are, therefore, artificially "missing").

One may reasonably ask: why impute at all? Would it not be preferable to leave the data incomplete and analyze what remains, tabulating the missing values as "unknown"? Surely, imputation is a process of delusion, giving the impression that the data are in better shape than they actually are.

There is much to be said for this argument. Imputation, by whatever method, can add no new information to the data (except, possibly, when auxiliary data are available). If badly done, it may result in serious misrepresentation of facts. However, there seem to be at least two cases where imputation is useful:

(i) imputation of a very small proportion of values, so small and by such a method that no discernible distortion of the data could result, may make the data set much easier to handle, e.g. the imputation of a few points in a time series of equally spaced observations; or

(ii) imputation where the end products are tabulations at arbitrary levels of aggregation.

Case (ii) is the one familiar to survey takers in particular. Including "unknowns" in all tables is usually thought to be untidy and deleting these cases produces inconsistent tables (the totals or marginal distributions vary). In addition, where a record consists of a fairly large number of fields, one has the feeling that some information about the missing or questionable values is contained in the portion that is good.

It is not the function of this paper to discuss all methods of imputation; but we note some of the common elements of imputation procedures.

(a) There is a close relationship between editing and imputation. If a record fails an edit, it is not always obvious which fields are faulty; yet some basis for deciding which fields to change must be established. (We are assuming that all possible cleaning up, by means of reference to original records or respondents, has been done). Complex edits make life exceedingly hard, in deciding both which fields to impute and how to impute missing or questionable fields (see [5]). This problem is frequently ignored in theoretical work on missing data.

(b) The marginal and joint distributions of responses are almost certainly different from those of the underlying population. In the case of numeric data, such distributions are unlikely to be normal. Moreover, transformations to normality (or just to less pronounced skewness) usually result in transformations of the edits which make them more difficult to deal with.

(c) The pattern of missing fields varies from record to record. In an n field record (excluding the identifiers) there are 2^n possible patterns of fields to impute. The edit procedures may reduce this number of possibilities in practice, if only to simplify imputation.

(d) The imputer does not have a great deal of time to fiddle with the data after they come in. In fact, he often has tight deadlines. He may have little, if any, test data to work on before the data collection begins.

(e) Imputation, by any method, does not solve any specific estimation problem more satisfactorily than the usual analytical estimation techniques. That is, in order to estimate a particular quantity, θ, one can only use such

data as there are and a relevant model. However, imputation does "solve" the problem of being able to produce, very easily and in a consistent way, estimates of any population parameters (arbitrary totals, means, proportions, etc.), even those the survey was not designed to estimate, although possibly with no guaranteed precision.

What the imputer wants, therefore, is a procedure which

(i) will impute plausibly and consistently provided only that the non-missing data satisfied the edits;

(ii) will preserve the underlying distributions in the data, or at least reduce the bias in the responses, and preserve the relationships between fields as far as possible;

(iii) will work for (almost) any pattern of missing fields; and

(iv) can be set up ahead of time.

Particular techniques of imputation may vary in their dependence on particular models and their ability to stabilize estimates, reduce bias (relative to standard estimation techniques) or preserve the relationships between the variables.

3. HOT DECK PROCEDURES

We will define a Hot Deck imputation procedure to be one where an incomplete response is completed by using values from one or more other records on the same file (i.e. from the same survey) and the choice of these records varies with the record requiring imputation.

Thus, simply inserting the stratum mean of the good data in the missing field is not a Hot Deck procedure because the choice of the records used in the mean is independent of the record requiring imputation. All imputations of the same field in that stratum would be the same. Choosing a "good" record, (the donor) which resembles the "bad" record (the recipient) and using the donor to supply the values of the fields missing in the recipient, is a Hot Deck procedure.

As a simple example of a Hot Deck procedure, consider the case of a record consisting entirely of categorical data. An incomplete record requiring imputation in one or more fields is matched to a collection of complete records in the file (the Hot Deck) which have identical values in the remaining fields. One of these complete records is chosen at random and is used to donate the values of the missing fields to the incomplete record.

To formalize the above description somewhat, suppose the n fields of the record are $X_1,\ldots,X_n$. The recipient record lacks $X_1,..,X_l$, but has values for $X_{l+1},\ldots,X_n$. The recipient record before imputation will then be $\underset{\sim}{x}_R = (,,, x_{l+1}, \ldots, x_n)$, where the blanks stand for unknown values. Now a collection, $C(\underset{\sim}{x}_R)$, of complete records of the form

$$(X_1, \ldots, X_l, x_{l+1}, \ldots, x_n)$$

is identified and one is chosen at random, say

$$\underset{\sim}{x}_D = (x_1', \ldots, x_l', x_{l+1}, \ldots, x_n).$$

This is the donor record. The completed recipient is then $\hat{\underset{\sim}{x}}_R = \underset{\sim}{x}_D$.

It is easy to see that such a procedure would produce consistent imputations, would tend to preserve underlying distributions and to reduce response bias in fields where response is relatively poor. It would work in any situation and could be set up in advance.

In reality, the operation does not work quite as smoothly because (i) there are computational problems, and (ii) there are no exact matches in some cases.

The computational problems are mainly ones of sheer size: for a record with a moderate number of fields, the number of possible Hot Decks which would have to be identified is very large. In practice, compromises have to be made in order to reduce the potential number of the decks. This is usually done by matching on fewer fields and/or imputing for one or a group of fields at a time. This means that a particularly scanty record may receive data from several donors, and also that successive imputations may result in a record which fails one or more edits. To avoid the latter situation, various ad hoc procedures may be employed along the way.

There seem to be two main types of Hot Deck for categorical data which we will call sequential and random choice.

In the sequential procedure (used by the U.S. Census of Population and Current Population Survey, [9], [10], [12], [14], the data are processed one record at a time. A field A (or group of fields) is imputed by defining a cross-classification of several other related fields (B, C, D,..) on which a match is to be made. For each cell in this classification, that value of A is retained which occurred in the last record processed with the corresponding values of B, C, D,... Thus, as the file is processed, the values in the individual cells of the (B, C, D,..) matrix change. When a record which lacks a value of A occurs, it receives the value currently in the cell of the matrix

which matches its own values of B, C, D,... If two such records (lacking a value of A, but with the same values of B, C, D,...) occur consecutively, the same value of A will be imputed in each case, since no records will have been processed which could cause the value to change. The matching fields (and therefore the imputation matrix) vary with the fields to be imputed. In those cases where imputation of a single field might result in an edit failure after imputation, a set of related fields is deleted and imputed together. One other obvious problem with the sequential procedure is that each imputation matrix must be initialized.

In the random choice procedure, a single current donor matrix is not maintained, but a record is chosen at random from a deck with suitable characteristics. The choice of matching fields in both sequential and random choice procedures must be made considering likely major sources of variation and the number of eligible records available in each cell for donation. We return to the problem of matching in the next section.

In the random choice procedure used by the Canadian Census ([7]), matching is done on those fields linked to the missing fields by edit constraints, as well as fields correlated with the missing fields. This could result in a very large number of matching fields; but those fields are eliminated which do not restrict the value of the field to be imputed given the values of the data present. The procedure first attempts to impute all missing fields using a single donor. If this fails, a field-by-field Hot Deck imputation is tried.

These Hot Deck procedures involve mainly categorical data. It is easy to see that if a single numeric (quantitative) field or a whole group of such fields has to be imputed by matching on categorical fields, these systems will still work. Trouble starts when there are several quantitative fields linked by edit constraints, or matching has to be done on quantitative fields, or both. The matching problem

can occasionally be dealt with by splitting the range of the variable (e.g. age) into intervals and coding the intervals; but if several variables are involved, one may find that the data are unevenly distributed through the grid.

It appears that numeric Hot Decks are not as common as categorical. In [9], a sequential procedure is described for imputing income data; but note that all income fields are imputed even if only some of them are missing. The reluctance o use numeric Hot Decks seems to be due in large part to the difficulties of coping with the edit structure. Furthermore, for certain single-field Hot Decks, it is known that estimates based on imputed data are more variable than those based on weight-adjusted data ([1]). This is because the Hot Deck contains extreme-value as well as central-value records. Funny records are bad enough when they are real - they are no joke when they are imputed.

Conceptually, a numeric Hot Deck requires a distance function to be defined between records on the matching fields since an exact match in numeric fields is unlikely. This function need not be a metric - it need not even be symmetric in the recipient and donor records.

The Hot Deck consists of all "good" (i.e. complete in all the relevant fields) records. For a particular recipient, a donor (or "good") record in its neighbourhood is identified and the missing fields of the recipient record are supplied by transformation of the corresponding fields of the donor. In one implementation at Statistics Canada [11], the nearest m complete records to a particular recipient are identified. This requires an efficient search algorithm. An attempt to complete the deficient record using fields from one of its m neighbours is made, taking the complete records in order of nearness. The donation is successful when the completed record passes the edits. If none of the m neighbours will do the job, the imputation fails and further processing is required.

When implementing this type of system, it is advisable to consider judicious transformations of the data both for matching and for imputation (the transformations appropriate to each function need not be the same). The distributions of some numeric data become very attenuated in the tails, so that "nearness" in the untransformed data changes in different regions. It is also sometimes possible to transform the data in such a way which both conforms with the edits and facilitates a correct imputation. For example, if an edit is $A + B + C \leq E$, then division by E transforms the edit to $P_A + P_B + P_C \leq 1$ and instead of A, B, C, E as data, we have P_A, P_B, P_C, E as data. The distance function may now be defined on some transformation of P_A, P_B, P_C, E.

In one implementation of a mixed numeric and categorical Hot Deck, one nearest neighbour was identified and a set of estimates of the missing fields was specified depending on the values of fields in both recipient and donor records so as to force a consistent imputation. Thus, for example, if the field A was missing in the recipient, the imputation might be

$$\hat{A}_R = f(A_D, B_R, B_D)$$

where B is a field or set of fields present in both recipient and donor and the subscripts D and R signify donor and recipient fields respectively ([2]).

In order to reduce the variability of the numeric Hot Deck imputation, the device of averaging over neighbours (or successive records in a sequential system) has been suggested and used. This will work for single numeric field imputation and will stabilize the final estimates. However, where several numeric fields are being imputed, an averaged record will not necessarily satisfy the edits. In general, mixed numeric and categorical procedures, there is no way to average categorical data.

In some cases auxiliary variables are present for all data. If they are categorical variables only, they may simplify the Hot Deck procedure by guaranteeing minimal matching. If they include numerical variables which are correlated with the survey fields subject to imputation, they can be effectively used as the total basis for matching, making the search procedure much simpler. In such cases, one may argue that it would be better to use a ratio estimate rather than impute missing data; but ratio estimates (like weight-adjusted estimates) are not additive and it appears ([2]) that a suitable imputation procedure could be less biased than the ratio estimate while (more or less) preserving the variance.

For large scale imputation (imputation of large numbers of entire survey records) good auxiliary variables, possibly from administrative sources, are essential and the process can be thought of as transforming auxiliary data into survey data (e.g. [2]).

4. SPECIAL PROBLEMS

4.1 Distance Functions

One of the myths about numeric Hot Decks seems to be that choice of the distance function is critical. In fact, judging from the experience with experimental systems at Statistics Canada, the performance of the Hot Deck is not particularly sensitive to the form of the distance function, once the variables have been transformed and rescaled. However, some distance functions are easier to deal with than others, a particularly attractive one being, after transforming to uniform marginals:

$$d^N(i,j) = \sup_k |x_{ik} - x_{jk}|$$

where i and j index the records. If one of the variables is more important than another, one can incorporate this by weighting them

$$d^N(i,j) = \text{Sup } w_k \ |x_{ik} - x_{jk}| \ .$$

Categorical data can be incorporated by defining suitable resemblance functions between the classes of a categorical variable. For example, if variable A takes values $A_1, \ldots, A_K$, then

$$R(A_k, A_k) = 0$$

and $$R(A_k, A_\ell) = 1 \quad \text{if } A_k \text{ and } A_\ell \text{ are compatible,}$$

$$= 10^5 \text{ if they are not.}$$

One can now define, where A_i is the value of A taken by the ith record,

$$d^C(i,j) = \underset{A}{\text{Sup}} \ (A_i, A_j).$$

The numeric and categorical distance functions can then be combined, e.g.

$$D(i,j) = d^N(i,j) \ .(1 + d^C(i,j)).$$

Obviously, there are many ways to play this game.

If one has reservations about the jth observation, one can inflate any distance which incorporates it, and so render it less preferable than other nearby observations:

$$d'(i,j) = d(i,j)(1 + h_j) \ , \quad \text{where}$$

d (i,j) is any measure of a distance and h_k is presumably zero for most observations k. In particular, in a matching or random choice situation, when i is the recipient and j the donor record, h_j may be a function of the number of times j has already been used as a donor. This has the effect of spreading the donor usage around and avoiding the over-use of a particular donor. Whether this is an advantageous procedure is open to question. If response is poor in some region (so that donors are rare), does one necessarily want to impute using donors in a nearby region where the response is good, but the characteristics of the response may be different? Repeated use of a particular donor will inflate the variance; but equalizing donor usage may result in bias. The main reason for limiting donor usage may be the pacification of nervous clients.

4.2 Choosing the Matching Fields

When a record fails an edit which involves several fields, it is not always obvious which fields are in error. If several edits involving common fields are failed, there are some intuitive grounds for casting suspicion on one or more of the common fields. Depending on the circumstances, one may believe that certain fields are more prone to error than others. The decision about which fields to impute is an editing decision which has little to do with the method of imputation, except insofar as it facilitates the imputation, and we will not deal with it here.

The question we do address is: given that the decision has already been made as to which fields are missing (to be imputed), which of the remaining fields are used for matching? The natural candidates seem to be (i) fields correlated with missing fields, to ensure a good imputation, (ii) fields linked by edits to the missing fields, to avoid edit failure after imputation, and (iii) natural stratification variables employed in the survey design, which may influence the missing data as in (i).

In the case of a record with many variables (such as the Census of Population), the collection of all reasonable fields may be so large that implementation is difficult and there is no guarantee of a match. In the case of a mixed categorical and numeric match, an exact match on the categorical variables may force a poor match on the numeric variables.

Increasing the number of matching variables may not result in a better match. One should give some hard thought to what compromises are acceptable in terms of grouping classes (so that, for example, a recipient in industry I may be imputed from a donor in a compatible industry J) and eliminating variables so that a donor pool of suitable size is available.

A closely related observation is that it is often not possible or even desirable to do all imputation in a single pass (so that each recipient requires only one donor). The number of complete records (potential donors) may be relatively few so that matches would be poor, no use would be made of information in partially complete records and the same donors could be used repeatedly. The matching variables and distance functions appropriate for imputing some variables may not be suitable for imputing others. The imputation is therefore broken up into several stages, with certain sets of fields being imputed at each stage. Different records would be available as potential donors at each stage since they would only be required to be complete in the current matching and imputation fields. A result of this approach is that several donors may be involved in completing a deficient record. On the other hand, imputed fields can be used in matching and donation in succeeding stages.

4.3 Auditing

Some effort should be made to keep track of what the imputation process is doing. At the end of the process, one would like to know:

a) How many times a particular record has been used as a donor in a particular stage.

b) How many attempts had been made to achieve a successful imputation for a particular deficient record (this would not apply to some procedures).

c) Which donors contributed what fields to which recipients. This is important in tracing the sources of peculiar imputations. By analyzing the transfer of information from particular donors to specific recipients, one may trace and remedy problems in the imputation procedures. Remedies may consist of changing the matching variables, the method of estimating missing fields or the definition of a possible donor by excluding those which appear to be outliers although they might be acceptable records.

d) If the imputation of a field is conditional on the values of other fields (in either the recipient or the donor) which condition prevailed at the time of imputation.

e) The value of the distance function at each donation. A relatively large value could signal a problem.

Useful summaries of the run are:

i) the number of records eligible as donors,

ii) the number of records requiring imputation,

iii) the number of records eligible neither as donors nor as recipients,

iv) a frequency distribution of the number of times each donor was used over all donors (see (a) above),

v) a frequency distribution of the number of attempts to achieve a successful donation over all recipients (see (b) above),

vi) frequencies of the condition flags (see (d) above),

vii) a listing of all records for which imputation failed, and

viii) a distribution of the value of the distance function (see (e) above).

Distributions should be for records in fairly homogeneous strata.

5. EVALUATION OF HOT DECKS

An imputer with a new and shiny Hot Deck system naturally wants to know how good it is, and so do the users of the data which the Hot Deck produces. Some of the questions which arise are: how are

i) the bias and reliability of the principal estimates,

ii) donor usage, (the distribution of the frequency with which records are used as donors), and

iii) the frequency of imputation failure,

affected by

i) the size of the data set,

ii) the frequency of missing data,

iii) "non-response" bias (where the non-response may be caused by deletion of fields due to edit failure),

iv) the underlying distributions of the data,

v) the choice of matching fields,

vi) the distance function, and

vii) the particular parameters of the imputation procedure?

A little theoretical work in very restricted situations has been done on reliability and bias ([1], [12]). Part of the difficulty in extending theoretical work lies in the edit structures and part in the sources of variation. Given the sample, numerical matching procedures are generally deterministic. Sequential procedures depend on the ordering of the file which is seldom completely random.

It appears then that the only generally available approach to evaluation is by simulation, using either real or artificial data. Real data, presumably culled from the good records of previous surveys, have the advantage of being realistic. On the other hand, fake data, produced by some modelling process, are subject to more manipulation so that one can vary distributions of and relationships between variables. In either case, fields are designated as missing by some random process which can be replicated and the variation over these replications is observed and analyzed ([2], [6]).

In addition, several empirical studies have been carried out comparing Hot Deck and other procedures with respect to estimation and costs ([1], [2], [3], [4], [8]).

6. THE LAST WORD

In this paper we have attempted to outline what we believe to be the general approaches to Hot Deck imputation, with emphasis on the interaction between the edit constraints and the imputation procedure.

As a method of imputation, Hot Deck has some attractive features in comparison with its competitors, not the least of which is that no strong model assumptions need be made in order to estimate the individual values. The Hot Deck procedure can be viewed as a sort of non parametric regression. Although there may be an increase in the variability of some estimates (depending on the Hot Deck methodology), it does appear that there is a reduction in non-response bias due to partial responses or where auxiliary information is available, at least under normal survey conditions.

There are also many problems associated with Hot Deck procedures, mainly involving accommodation of the edit structure or constraints on the data, and we have tried to discuss these (or rather, those we are aware of) in a general way.

We have not attempted any discussion of the implementation of these procedures, because as far as we know, the implementation tends to be tailored to the application and, in any case, we would be well out of our depth in pretending any knowledge.

We know of no example of a "pure" Hot Deck being used on data of any great complexity. Hot Deck systems appear to be used in conjunction with other imputation methodologies (such as Cold Deck) in order to achieve consistency and reasonable efficiency.

No generalized Hot Deck system has been developed. The CANEDIT system ([7]) is an attempt at one for categorical data; but it has limitations. A generalized numerical Hot Deck system is being developed at Statistics Canada, which deals with linear edits only ([11]). Both these systems involve both edit and imputation phases, using the edit phase to decide which fields to impute on the basis that as few fields as possible should be changed. A generalized, integrated numerical and categorical data edit and imputation system is seen as being feasible, although there are formidable mathematical and algorithmic problems involved.

RESUME

La méthode d'imputation dite du 'hot deck' est celle où l'on complète une réponse incomplète avec des données provenant d'un ou de plusieurs autres dossiers du même fichier; le choix de ces dossiers varie selon le dossier devant faire l'objet d'une imputation.

Le document décrit la méthode générale du 'hot deck', en insistant sur l'interaction entre les contraintes de vérification et les procédures d'imputation. À partir d'une combinaison de zones catégoriques et numériques, il est possible de construire des fonctions de distance, de les modifier de manière à tenir compte de l'importance relative des zones et de défavoriser des donneurs peu désirables. Des zones correspondantes peuvent être corrélées avec des zones manquantes, raccordées à des zones manquantes par vérification ou peuvent être des variables naturelles de stratification; cependant, le fait d'augmenter le nombre de zones correspondantes ne donne pas nécessairement un meilleur appariement. Il importe de contrôler l'imputation et de résumer sa performance.

Il faut évaluer la méthode dite du 'hot deck' pour étudier le biais et la fiabilité des estimations, de l'utilisation des donneurs et de la fréquence de l'échec de l'imputation dans diverses conditions des données et la variation de la procédure d'imputation. Il semble que la simulation soit la seule approche d'évaluation qui soit généralement disponible.

REFERENCES

[1] Bailar, J.C. III and Bailar, B.A., "Comparison of Two Procedures for Imputing Missing Survey Values". Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978, pp. 462-467.

[2] Colledge, M.J., Johnson, J.H., Pare, R. and Sande, I.G., "Large Scale Imputation of Survey Data". Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978. pp. 431-436. Also, Survey Methodology, Statistics Canada, 1978, Vol. 4, No. 2, pp. 203-224.

[3] Cox, B.G. and Folsom, R.E., "An Empirical Investigation of Alternate Item Non-Response Adjustments". Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978, pp. 219-223.

[4] Ernst, L.F., "Weighting to Adjust for Partial Non-Response". Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978. pp. 468-472.

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

[6] Ford, B.L., "Missing Data Procedures: A Comparative Study". Proceedings of the Social Statistics Section, American Statistical Association, 1976, pp. 324-329.

[7] Hill, C.J., "A Report on the Application of a Systematic Method of Automatic Edit and Imputation to the 1976 Canadian Census". Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978, pp. 474-479. Also, Survey Methodology, Statistics Canada, 1978, Vol. 4 pp. 178-202.

[8] Nordbotten, S., "The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with Other Means of Improving the Quality of Statistics". Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Vol. 16, 1965, pp. 417-441.

[9] Ono, M. and Miller, H.P., "Income Non-Response in the Current Population Survey". Proceedings of the Social Statistics Section, American Statistical Association, 1969, pp. 277-288.

[10] Pritzker, L., Ogus, J. and Hansen, M.H., "Computer Editing Methods - Some Applications and Results". Bulletin of the International Statistical Institute. Proceedings of the 35th Session, Vol. 16, 1965, pp. 442-465.

[11] Sande, G., "Numerical Edit and Imputation". International Association for Statistical Computing, 42nd Session of the International Statistical Institute. December 1979.

[12] Spiers, E.F. and Knott, J.J., "Computer Method to Process Missing Income and Work Experience Information in the Current Population Survey". Proceedings of the Social Statistics Section, American Statistical Association, 1969, pp. 289-293.

[13] Szameitat, K. and Zindler, H.J., "The Reduction of Errors in Statistics by Automatic Corrections". Bulletin of the International Statistical Institute. Proceedings of the 35th Session, Vol. 16, 1965, pp. 395-417.

[14] U.S. Bureau of the Census. 1970 Census of Population and Housing, Procedural History. Chapter 15, Appendix A.