AN IMPUTATION ALGORITHM APPLIED THE NONRESPONSE PROBLEM

Jose Brito [*] Nelson Maculan [†] Luiz Ochi [‡] Flavio Montenegro [§] Luciana Brito [¶]

[*] ENCE, Escola Nacional de Ciências Estatísticas
Rua André Cavalcanti,106, sl 403, CEP:20231-050, Rio de Janeiro, Brazil
jose.m.brito@ibge.gov.br

[†] COPPE, Universidade Federal do Rio de Janeiro
P.O. Box 68511, 21941-972 Rio de Janeiro, Brazil
maculan@cos.ufrj.br

[‡] UFF, Universidade Federal Fluminense, Instituto de Computação
Rua Passo da Pátria 156, Bloco E, 3 andar, São Domingos, Niterói, RJ, Brazil
satoru@dcc.ic.uff.br

[§] IBGE, Instituto Brasileiro de Geografia e Estatística, DPE/COMEQ
Av.Chile, 500, 10 Andar, Centro, Rio de Janeiro, RJ, Brazil
flavio.montenegro@ibge.gov.br

[¶] UNIPLI, Centro Universitário Plínio Leite
Av. Visconde do Rio Branco, 123, Centro, Niterói, RJ, Brazil
luroquebrito@hotmail.com

ABSTRACT

This work describes an imputation algorithm to solve the nonresponse problem in surveys. The nonresponse is associated the occurrence of a missing values in at least one variable of at least registry or unit of the survey. In order to prevent the negative effects of nonresponse, an intense research has been produced in this area and many procedures have been implemented. Among these, we detach the imputation methods, that consist basically of substituting a missing value by some suitable one, according some criterion or rule. In this work we propose a new imputation algorithm that combines the clustering method and GRASP metaheuristic.To evaluete its performance we present a set of computational results considering data from Brazilian Demographic Census 2000.

1. INTRODUCTION

Nonresponse is a normal but undesirable feature of a survey [1]. It is characterized by incomplete records of a survey database, which may occur in the phase of data collection or data estimation. Nonresponse occurs when, at least for one sampling unit (household, person, etc) of the population or sample [2] of the survey, there is nonresponse to one question of a questionnaire (record) or the information given is not usable. Or else, when at least one item of a questionnaire was not completed (survey variable). Incomplete questionnaires due to nonresponse are common in surveys, but deserve attention. Therefore, a considerable amount of money has been spent in the development and improvement of procedures associated to data assessment, in order to prevent the occurrence of nonresponse or to minimize its negative effects. There has been extensive research in this field, which is reported in many studies, such as [1, 3, 4, 5]. Among the procedures being developed are those classified as imputation methods, which basically consist in replacing a missing data with an estimated value, according to a criterion or rule [1]. With the purpose of treating the nonresponse issue, the present study introduces a method that combines an imputation rule, a technique of cluster analysis [6, 7] and GRASP metaheuristics [8, 9] (Greedy Randomized Adaptive Search).

2. NONRESPONSE AND IMPUTATION

There are two types of nonresponse: (1) total nonresponse, which corresponds to the units from which no usable information was collected, and partial nonresponse, corresponding to the units from which there is at least one variable with a missing value and which are not part of the total nonresponse set. The present study has focused on the treatment of partial nonresponse. Then, the concept of nonresponse is described in greater detail, with emphasis on some procedures for the treatment of nonresponse through imputation methods. At first we may consider a set of p variables associated e.g. to the sociodemographic characteristics of a survey defined by $X_1, X_2, ..., X_p$. Such characteristics are obtained for n persons (records), which determines a matrix X_{np} that has for each input X_{ij} the value of the jth variable (characteristic) observed in

the ith $i = 1,...,n$ record. If a M_{ij} indicating variable of the observation of the corresponding data is associated to each X_{ij}, we'll have $M_{ij} = 1$, If there is a value for X_{ij} and $M_{ij} = 0$, If it is otherwise. And based on this, a M matrix that defines the pattern of the missing data is defined. In the present article, we shall treat the missing data associated to one single variable X_j (Univariate Missing Data), known as the study variable. That is, the matrix M shall have zero elements in only one of its columns. The remaining variables $(p-1)$ shall be treated as explicative variables, that is, variables correlated with the variable of interest and that can be used to predict the values of this variable.

When incomplete records are found in a given database, that is, when there is missing information on one of the variables of the database, data can be imputed. Imputation is a procedure through which the missing values for one or more study variables "are filled" with estimated values [1]. These "replacements" must be performed according to a rule. The imputed values can be classified into three main categories: (i) values constructed using a device for automatic imputation of missing values, considering an imputation statistical rule (ii) values observed for elements with similar response; (iii) values constructed by expert opinion or "by the best possible estimate" [1]. The categories (i) and (ii) can be called statistical rules because they use a statistical method aimed to produce a replacement value reasonably close to the original value. The (i) category is frequently based on regression prediction [1]. Imputation is especially used in the treatment of partial nonresponse, which concerns the simulations presented in this article, although it can also be used in the treatment of total nonresponse.

There are several methods of imputation [1, 5], such as: (1) Nearest Neighbor Imputation: a function of the distance between the complete and incomplete records is calculated considering the explicative variables $(p-1)$. The value of the observed unit with the smallest distance to the non-respondent unit will be substituted for the missing item. (2) Hot Deck Imputation: the variable X_j associated to an incomplete record is substituted for a value obtained from a distribution estimated from the available data (complete records). A complete record (donor) is selected in order to provide values for the missing information in the incomplete record (recipient). This method is typically implemented in two stages: in the first stage, a set of data is distributed into k groups (imputation classes) considering the explicative variables $(p-1)$ associated to the study variable. Once the groups k are defined, in the second stage, the group of each incomplete record is identified. The complete records of a group are used to estimate the unknown values in the incomplete records. (3) Mean imputation: it is a simple method applicable to continuous variables. It substitutes the missing values with the general mean for the variable.

3. METHODOLOGY

The present study shall treat the problem of nonresponse with the type of imputation classes used in the Hot Deck method, expanding the use of these classes to the cases of mean imputation (which is then based on records associated to each one of these classes). Since the definition of the imputation classes has direct impact on the incomplete records, a new methodology for the definition of the classes shall be proposed in this study, with the application of the cluster analysis, a technique widely used to solve the problem of obtaining homogeneous groups (clusters) from a database with special characteristics or attributes [7]. The clusters formed are characterized as follows: the objects of one cluster are very similar and the objects or different clusters are very dissimilar, considering the objective function (that aggregates the distances) shown in the equation below.

$$f = \sum_{l=1}^{k} \sum_{\forall o_s, o_r \in C_l} d_{sr} \tag{1}$$

The function presented in the equation 1 considers for each cluster $C_l, l = 1,...,k$ the sum of all the objects that are part of the group. Therefore, minimizing f consists in allocating all the objects to the clusters in such a way that the total sum of the distances (dissimilarities) between two objects from each one of the clusters is minimum.

Regardless the objective function considered or other distance functions, this is not a simple task because of the combinatorial nature of this type of problem (see also [10, 11]). If a process of exhaustive search is used to obtain an optimal solution, all solutions shall be enumerated, that is, all the possibilities of combination of the objects n in groups k. In general, the m number of possibilities grows exponentially as a function of n [6]. Such characteristic makes it impracticable to obtain the exact resolution of average and large instances of these problems. A previous study on metaheuristics applied to cluster problems [12, 13, 14, 15] suggests that it is a good alternative for the resolution of several clustering problems. In general, with the application of metaheuristics, feasible solutions of higher quality than those from heuristics (local minimums) are obtained.

Considering the last observation, and with the purpose of constructing the classes used in the imputation of data, a cluster algorithm that uses GRASP meta-heuristics was developed [9] and whose objective function is the 1 equation. The GRASP is an iterative greedy heuristic to solve combinatorial optimization problems. Each iteration of the GRASP algorithm contains two steps: construction of a local search. In the construction, a feasible solution is built using a randomized greedy algorithm, while in the next step a local search heuristic is applied based on the constructed solution.

3.1. Grasp Algorihtm

Construction Procedure: Considering a D set formed by objects n (records of a database) and a fixed number of clusters k, k objects of D are selected, with each object allocated to a cluster $C_l, l = 1,..,k$. Then, in each construction iteration, each one of the $(n-k)$ objects is allocated considering their proximity to the objects o_j that are already part of each group C_l. That is, in each iteration, there is a list of candidates LC composed of objects o_i not yet allocated to a cluster and two vectors q and g. Each position q contains the number of the cluster where the closest object o_j is located (using the 1 equation of each object o_i). The vector g corresponds to the distance of the object o_j in the database located at the shortest distance from each object o_i. Based on the referred information, a LCR restricted candidate list is constructed, which is formed by the o_i objects, so that $g_i \leq g_{min} + \alpha(g_{max} - g_{min})$. Being g_{max} and g_{min}, respectively the maximum and minimum distances found in g. Then, an object LCR (element) is randomly selected and allocated to one of the clusters considering the information stored in q. Every time a new object is inserted in one of the clusters, the candidate list is updated. And when $LC = \emptyset$ all the objects shall be allocated to one of the clusters k.

Local Search Procedure: At this step, the reallocation of objects between the clusters k is sought, in order to reduce the

value of the equation (1), and consequently, produce more homogeneous clusters (classes) for performing the imputation. Considering the solution obtained in the construction step, in each iteration of this procedure, two clusters C_r and C_l are selected from the clusters k defined in the construction step. Afterwards, various (random) selections of an object $o_i \in C_r$ and an object $o_j \in C_l$ are performed, and in each selection the distances d_i, d_{il}, d_j, d_{jr} are calculated. The values for d_i and d_j correspond respectively to the sum of the distances from object o_i to the other objects C_r and the sum of the distances from object o_j to the other objects C_l. And d_{il} represents the sum of the distances from object o_i to the other objects C_l. An equal definition is applied to d_{jr}, though considering the sum of the distances between the object o_j and the objects C_r. After the calculation of the distances d_i, d_{il}, d_j, d_{jr}, three types of reallocations are assessed: (1) The object o_i is allocated to cluster C_l and the object o_j is allocated to cluster C_r and $d = -d_i + d_{il} - d_j + d_{jr}$ is calculated. (2) The object o_i is allocated to cluster C_l and $d = -d_i + d_{il}$ is calculated (3) The object o_j is allocated to cluster C_r and $d = -d_j + d_{jr}$ is calculated. The reallocation that produces the greatest reduction (lowest value of d) in the objective function given by (1) shall be applied in the current solution. Such reallocations are performed until the improvements w (reductions) in the value of the objective function are obtained, or until the number of replacement attempts is equal to a value of $n_{C_r} * n_{C_l}$. Being n_{C_r} and n_{C_l}, respectively the number of objects in clusters C_r and C_l. When at least one of the conditions is satisfied, we get back to the main loop and select two new clusters. At the end of the local search, the new candidate solution generated is checked and compared to the best results obtained so far, considering previous GRASP iterations.

3.2. Imputation Algorithm

The imputation algorithm considers, as input, a database with n records, with complete information for the $(p-1)$ explicative variables, $X_1, X_2, ...X_{p-1}$. Besides, the missing information for the study variable X_p in a given number $n* < n$ of records, or else, a percentage of nonresponse. Then, the two basic steps of the algorithm are described:

• The algorithm GRASP is applied in the determination of the imputation classes considering the number of clusters equal to k. The objective function presented in the equation 1 and used in the GRASP considers, for cluster purposes, the distances between the explicative variables $(p-1)$.

• Once the classes are constructed, the procedure of mean imputation is applied in each one of the incomplete records $n*$ in relation to X_p. This implies determining to each class C_l $(l = 1, ..., k)$ each incomplete record i is allocated and assign a value $\bar{X}_l$ that corresponds to the mean (in class l) complete records in relation to variable X_p. Thus, $\bar{X}_l = \sum_{i \in C_l} \frac{x_{ip}}{n_l *}$, being $n_l *$ the number of complete records in cluster C_l and x_{ip} the value of the variable X_p in the nth complete record that is part of the cluster C_l.

4. RESULTS

The present section contains a few computational results obtained with the application of the imputation algorithm, implemented in Delphi language (version 6.0) and run on Windows 7. All the computational experiments are performed in a 16 GB RAM I7 PC with a 2.93 GHz I7 processor. Prior to the presentation of the results, a small description of the data used in the study is made, as well as of the nonresponse mechanism [1, 5, 16] considered for the database used in the experiments.

4.1. Data

In order to perform the experiments, a real database, more specifically, a file of the Sample of the 2000 Brazilian Demographic Census (state of Rio Grande do Sul) was used. Based on this file, five weighted areas (WAs) were drawn for the simulations with the imputation algorithm. A weighted area is a small geographical area formed by a mutually exclusive enumeration areas (cluster of census segments), which comprise, each one of them, a set of records of households and people [17]. We decided to work with the file of people, where each record is related to the individual characteristics of each inhabitant. And of the variables available in these records, six variables $X_1, ..., X_6$ were selected to be considered in the imputation, as follows: sex, relationship with the responsible person, age in years, highest completed level of education, schooling years and the gross earnings from the main occupation. The five first variables (all categorical) are explicative and correlated to the earnings in reais (quantitative), which was the study variable considered.

4.2. Mechanisms that Lead to Missing Data and the Generation of Incomplete Records

As in any other study aimed to assess whether the method of imputation produces good estimates for the imputed variable [2], the nonresponse mechanism must be considered. That is, since information on a given study variable is missing, these values shall be imputed on a subset of records. In particular, concerning the earnings, it is known that the loss of information is greater for classes with higher income, which characterizes a mechanism of nonresponse called Not Missing at Random (NMAR). This means that the probability of non-information of each input in the nth column of X shall depend on the values observed for the variable X_p in matrix X (see section two). Such mechanism was used to perform the simulations considering a database where all the records contain the information for the study variable (original records). With the application of the nonresponse mechanism, subsets of incomplete records in relation to the gross earnings from the set can be generated, and consequently apply imputation to these records. The number of incomplete records generated in the simulation depends on the rate of nonresponse considered.

One possible procedure for the generation of incomplete records consists in assigning a previous value pr $(0 \leq pr \leq 1)$ that corresponds to the probability of nonresponse (missing information) to the study variable in each original record. In the present study, in particular, such probability was obtained considering the variables relationship with the responsible person (11 categories), highest completed level of education, (10 categories) and schooling years (four categories). According to the category informed for each one of these variables, a probability pr of 0.1, 0.2 or 0.3 of the earning value (X_6) not being informed was attributed to each record. The more the category is related to high earnings, the greater the probability is [16]. Once this probability is defined, a value between 0 and 1 is drawn for each record, and this value is compared to the probability of nonresponse (pr) of the record. If the probability of the record is lower than the value drawn, such record shall have the gross earning value informed at the incomplete database, and, otherwise, it shall be considered a missing data on this database.

With the use of this procedure, r replicas can be generated from the complete database, which correspond to the database with different incomplete records.

4.3. Computational Experiments

Initially, for the application of the imputation algorithm to the records associated to the five files of people (WAs) (see section 4.1), a rate of nonresponse of 10% was defined and $r = 100$ replicas of the original databases were generated with different subsets of incomplete records for each r replica. Applying mean imputation to the incomplete records, we obtain for each replica the complete records and the imputed records. Considering such information, the values $\bar{X_m}^r$ e $\bar{X_c}^r$ were calculated, which correspond to the means associated to X_p considering: all the records of each replica (complete and imputed) and only the complete records. It is also said that the same classes of imputation (clusters) were used in all the replicas. In this particular experiment, the algorithm GRASP was applied considering the values k equal to 4, 6 and 8. Still concerning the GRASP, the number of iterations was fixed in 50, improvements equal to 20 and the parameter α equal to 0.5.

Table (1) shows the results obtained with the application of the imputation algorithm to the records of the five instances used in the simulations. The first column contains the number of the instance and column two contains the records of each WA. Column three contains the number of constructed clusters (classes of imputation). Columns four and five contain the value of the objective function (1) and the processing time (seconds) to construct the clusters, generate the 100 replicas and apply the imputation. Columns six, seven and eight contain the values of $\bar{X_p}$, $\bar{X_m}$ e $\bar{X_c}$ that correspond, respectively, to the mean of the incomes of all records (original database) and the mean of the means of $\bar{X_m}^r$ and $\bar{X_c}^r$ considering the 100 replicas, that is: $\bar{X_m} = \frac{\sum_{r=1}^{100} \bar{X_m}^r}{100}$ $\bar{X_c} = \frac{\sum_{r=1}^{100} \bar{X_c}^r}{100}$. Finally, column nine contains the value of ρ that corresponds to the relative mean deviation between $\bar{X_p}$ and $\bar{X_m}^r$: $\rho = \sum_{r=1}^{100} \frac{|\bar{X_p} - \bar{X_m}^r|}{\bar{X_m}^r}$.

WA	n	k	Time	F_{OBJ}	$\bar{X_p}$	$\bar{X_c}$	$\bar{X_m}$	ρ
1	178	4	18	2369.3	561.5	559.1	561.5	3.5
		6	6	1262.9		556.1	561.3	3.0
		8	3	783.5		555.2	559.5	3.6
2	289	4	77	7260.7	373.6	367.6	372.5	2.7
		6	24	4012.4		366.7	371.9	3.1
		8	11	2695.6		367.0	372.0	2.8
3	334	4	113	9268.9	355.3	349.5	354.1	1.7
		6	36	4932.8		350.2	354.2	1.4
		8	17	3349.6		350.2	354.8	1.3
4	410	4	215	12248.0	1174.6	1162.9	1171.1	1.5
		6	64	6808.8		1161.5	1172.9	1.7
		8	30	4359.1		1165.2	1176.1	1.6
5	539	4	485	21402.2	440.2	438.3	439.2	1.1
		6	153	11655.5		435.3	438.2	1.4
		8	71	7591.6		437.4	440.5	1.3

Table 1: Results for the Imputation Algorithm

The analysis of the results of columns 6, 7 and 8 of table (1) shows that the application of the imputation algorithm has made it possible to obtain good estimates for the mean, considering the 100 replicas. In particular, the values between 1% and 4% in column nine indicate that the means in relation to the imputed records were reasonably close to the real mean value $\bar{X_p}$.

Based on the results obtained, and despite the need for a greater number of experiments, the combination of GRASP and cluster analysis with an imputation method can be a good alternative to the treatment of the problem of nonresponse and produce good quality estimates for databases with incomplete records. In order to improve this procedure in the future, we intend to adapt it to the treatment of categorical variables. Also, we intend to use other objective functions for the construction of the clusters, as well as other metaheuristics such as ILS or Genetic Algorithms [9].

5. ACKNOWLEDGEMENTS

The FAPERJ (project APQ1) (www.faperj.br/) and CNPQ (project 474051/2010-2) (www.cnpq.br) for the financial suport.

6. REFERENCES

[1] C. E. Sarndal and S. Lundstrom, *Estimation in Surveys with Nonresponse*. John Wiley and Sons Ltd, 2005.

[2] S. L. Lohr, *Sampling: Design Analysis*. Brooks/Cole, Cengage Learning, 2010.

[3] J. G. Bethlehem and H. M. P. Kersten, "On the treatment of nonresponse in sample surveys," *Journal of Official Statistics*, vol. 1, no. 3, pp. 287–300, 1985.

[4] J. G. Bethlehem, "Reduction of nonresponse bias through regression estimation," *Journal of Official Statistics*, vol. 4, pp. 251–260, 1988.

[5] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. John Wiley and Sons Ltd, 2002.

[6] A. R. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Prentice Hall. Fifth Edition, 2002.

[7] H. C. Romesburg, *Cluster Analysis for Researchers*. Lulu Press, 2004.

[8] T. A. Feo and M. G. C. Resende, "Greedy randomized adaptive search procedures," *Journal of Global Optimization*, vol. 6, pp. 109–133, 1995.

[9] F. Glover and G. Kochenberger, *Handbook of Metaheuristics*. Kluwer Academic Publishers, 2003, pp. 219–249.

[10] P. Hansen and B. Jaumard, "Cluster analysis and mathematical programming," *Mathematical Programming*, vol. 79, pp. 191–215, 1997.

[11] P. A. L. J. Hubert and J. J. Meulman, *Combinatorial Data Analysis: Optimization by Dynamic Programming*. Philadelphia: Society for Industrial and Applied Mathematics, 2001.

[12] M. C. G. Guojun and W. Jianhong, *Data Clustering: Theory, Algorithms and Applications*. ASA-SIAM Series on Statistics and Applied Probability, 2007.

[13] M. J. Brusco and D. Steinley, "A comparison of heuristics procedures for minimum within-cluster sums of squares partitioning," *Psychometrika*, vol. 72, pp. 583–600, 2007.

[14] W. Sheng and X. Liu, "A genetic k-medoids clustering algorithm," *Journal of Heuristics*, vol. 12, pp. 447–446, 2006.

[15] M. C. V. Nascimento, F. M. B. Toledo, and A. C. P. L. F. Carvalho, "Investigation of a new grasp-based clustering algorithma applied to biological data," *Computers and Operations Research*, vol. 37, pp. 1381–1388, 2010.

[16] S. Albieri, "A ausência de respostas em pesquisas: Uma aplicação de métodos de imputação. dissertação impa."

[17] Www.censo2010.ibge.gov.br/altera_idioma.php?idioma=_EN.