A Split Questionnaire Survey Design applied to Brazilian Census

Nádia Maria Coelho Rodrigues
Pontific Catholic University of Rio de Janeiro - PUC-Rio
Rua Marquês de São Vicente, 225 - Gávea
22453-900 - Rio de Janeiro - RJ - Brasil
Electrical Engineering Department
Phones: (55 21) 3114-1202 or (55 21) 3114-1203

Kaizô Iwakami Beltrão
National School of Statistics - ENCE
Rua André Cavalcanti, 106
20231-050 - Rio de Janeiro - RJ - Brasil
Phones: (55 21) 2142-4680 or (55 21) 2142-8796

E-mail: nadiacoelho@yahoo.com
kaizo@ibge.gov.br

Abstract

Brazilian Census is conducted with two questionnaires: a short and a long one. The long questionnaire contains 93 items per person and the short one 23. Respondent burden is one of the main causes for missing information. Using real data we simulate a *matrix sampling* approach: respondents are allocated only items part of one of three components, thus yielding three data sets and lessening respondent burden. Individual data from these three files are linked to create a complete new data set, called *synthetic* data file, using *statistical matching*. Every single record is concatenated to similar records from the other two files, using an imputation procedure based on *hot deck*. We find that all the selected empirical distributions of the complete data are well reproduced in the *synthetic* data sets, as well as bivariate and conditional distributions.
Nearly the same inferences can be achieved using *matrix sampling* design with a reduced cost and less respondent burden.

1 - Introduction

The design of surveys must balance many competing goals. Due to the financial burden of selecting individuals for studies, many survey questionnaires, the problem of excessively long questionnaires has been arising with increasing frequency, along with the related problems of declining response rates and increasing time investment for the respondent and the interviewers (for an previous discussions see Adams and Darwin, 1982, and Dillman, Sinclair, Clark, 1993). Since questionnaire based surveys are widespread means to gather all this information, it is hardly surprising that a variety of methods like split questionnaires with rotational elements have been developed to solve the above-mentioned problem. However, these approaches usually lead to reduced sub samples of the originally desired data sets such that in some cases the sample size for multi-dimensional analyses gets to be very small.

Raghunathan and Grizzle (1995) introduced a split questionnaire survey design where the original questionnaire is divided into several components with each component containing a roughly equal number of questions. The split approach is based on the multiple matrix sampling design which has long been used in US educational testing, achievement testing, and program evaluation, see Shoemaker (1973), Holland and Rubin (1982), or Munger and Lloyd (1988). With multiple matrix sampling, basically there are subgroups of variables created randomly and these subgroups are randomly assign to subgroups of units; these random assignments can lead to estimation problems due to non-identification and highly reduced data sets for multivariate analysis. In the split questionnaire survey design, apart from a core component with questions that are considered to be vitally important (e.g., socio-demographic questions), also only a selection of the other components is administered to every interviewee. This clearly reduces interview time, yielding lower survey costs as well as reducing the respondent burden. But unlike the matrix sampling approaches, the missing data now are imputed to finally end up again with a complete(d) data set. This approach merely requires that combinations of variables, which are to be evaluated, must be jointly observed in a small sub sample (to avoid estimation problems due to non-identification). Thus, depending on the split design any desired analysis can be carried out while retaining the original sample size. A statistical matching approach is used (Draper, D. ... [et al.], (1992)).
To illustrate this, Figure 1 shows a split questionnaire survey design where interactions of third order among the split variables are assumed to be analyzed. In our exemplary design the questionnaire is divided into three components (plus the core component with questions administered to all sample individuals).

Figure 1: ***Split questionnaire design with three components***

Questionnaire Number	Core Component	Split Variables		
		Component 1	Component 2	Component 3
1	■	■		
2	■		■	
3	■			■
4	■		■	
5	■	■		
6	■	■		

■ asked

□ not asked

In the case of Figure 1 a split design generates three different components.
An example that motivated the work presented in this article is the Brazilian Census of Population and Housing conducted decennially by the Brazilian Statistical Office (IBGE). This survey has been applied two questionnaires: one short and other long. The former is a subset of the latter. The short one is applied to every household and it will be considered 'the core'. Besides the enumeration, the objective of the survey is to estimate general characteristics of the Population: Migration and displacement, Education, Labor and Income, Fertility and Nuptiality and Familial structure and ties and characteristics of Housing Units, selecting information on the structure by age, sex, situation of the housing unit, color or race, religion and levels of disability or physical or mental deficiency of the population resident in Brazil.
The goal of this article is to show that the split questionnaire survey design can offer solutions to a wide range of questionnaire based surveys that suffer either or both from high costs or a high respondent burden. This proposal is based on the work of Raghunathan and Grizzle (with some slight modifications). In the next section we will cover the component structure of the questionnaire and also discuss the rules for the assignment of questions to components. Raghunathan and Grizzle (1995) state that variables with high partial correlation coefficients should go into different components. This sounds reasonable, because following this condition you avoid that variables which explain each other very well are always jointly missing for any observation. However, in some cases the purely data generated solution of the questionnaire design must be modified. The third section describes the statistical matching algorithm to develop proper multiple imputations method for analyzing data from the split questionnaire design.
The missing data are imputed using Hot-Deck procedure; see Rubin (1986) and Little (1988). The imputation procedure itself is based on the so-called *hot deck*. In this contribution the validity of split and imputation is discussed based on the preservation of empirical distributions, bivariate associations, conditional associations and on the useful criterion *mean square error* (MSE). We find that many empirical distributions of the complete data are well reproduced in the *synthetic* data sets. Section four is divided into two parts: After a description of the application and implementation of matrix sampling design, some characteristics of

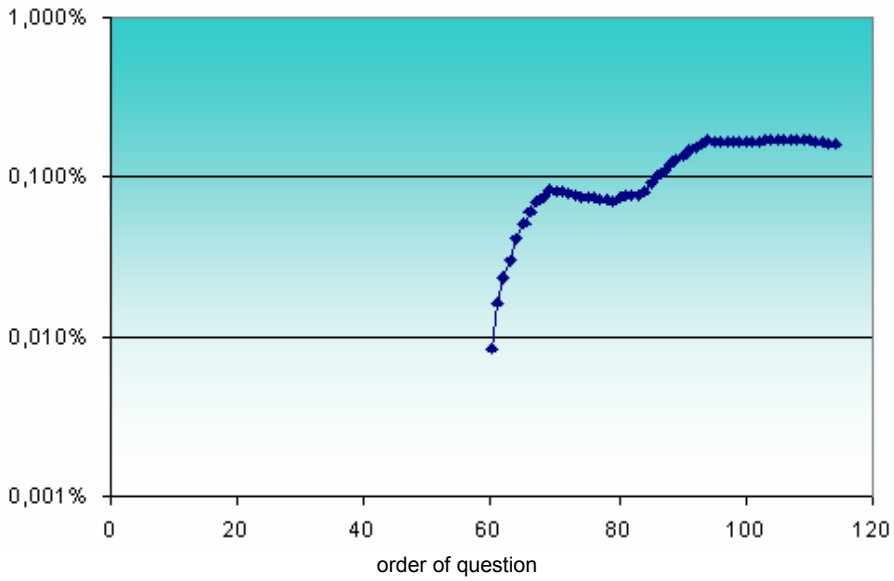
the original and the synthetic data set are compared, as well as inference procedures obtained from the split questionnaire design and the complete questionnaire design. Finally, the last section will resume some of the problems and pitfalls encountered and discuss possible solutions.

2 - Component Structure

In the first section we have learned that all split variables should be administered to components such that you get high partial correlations for variables in different components. To fulfill this requirement we need a complete data set to generate the component structure. An example that motivated the work presented in this article is the Demographic Brazilian Census 1991, that comprises general characteristics of population, Migration and displacement, Education, Labor and income, Fertility and Nuptiality and Families and housing units, conducted periodically by the Brazilian Statistical Office (IBGE). This decennial census was planned and conducted with two questionnaires, a short and long one. Since 1960, the survey is a probability sample of households to make the long form. The short form is applied on every household. The calculation of a suitable measurement of association with the long complete questionnaire answered, in advance in order to calculate the associations/correlations using the ordinal survey variables is done.
After obtaining the correlation matrix a cluster analysis can be used to generate the component structure. Instead of cases we cluster variables, and by using the correlation matrix rather than a distance matrix, those variables with low correlations will be put into the same cluster. This means that variables with high correlations end up in different clusters. These clusters represent the questionnaire components and statistical standard software can be used to calculate the required number of clusters/components. “Average linkage within groups” minimizes the average Euclidean distances for variables within each cluster while maximizing the distances for variables in different clusters. If the cluster sizes tend to vary heavily (and hence the length of the different questionnaires), “Ward” can be used as an alternative distance measure. Thus, a “reversed cluster analysis” is an easy way to generate the required component structure. However, some questions are required to be in the same block or even in a specific order as the following example might demonstrate: Suppose you we are interesting to tabulate the variable age, for example, then it must be in all generated components. The long questionnaire was split into three components. This paper decides to have three different components and the rules for the assignment of questions to components is based on aspects of the application of the generalized data editing and imputation software named DIA to the 1991 Population Census [of Brazil] Basic Questionnaire, (Oliveira, L. C... [et al.], (1996)). This software, developed by the Spanish National Statistical Institute, based on Fellegi & Holt methodology, handles editing and imputation of categorical data in one processing cycle and provides comprehensive information to control and assess the automatic correction process. The analysis reveals the

data quality and efficiency of the software adopted, which ensures data consistency while preserving basic distribution properties.
The long questionnaire contains 93 items per person and the short one 23. The Figure 2, shows empirical evidence that the long questionnaire have higher non-response rate than the shorter. The short had non-response rate equal zero percent.

Figure 2: ***" Medium Percentile" - Cumulative Non-response Rate to the Long Questionnaire***



3 – Statistical Matching Algorithm

Under the multiple imputation approach, we replace each set of missing components by more than one plausible set of values. Each completed data set formed by combining an imputed set of values with the observed set of components is analyzed, resulting in the estimates and the covariance matrix of the target quantities of interest. Under the assumption that the missing data are missing at random (MAR) or even missing completely at random (MCAR), as defined by Rubin (1976, 1987) and Little and Rubin (1987), and that the parameters are distinct (see Schafer, 1997) the missing data mechanism is said to be ignorable. If these assumptions hold, it is possible to suitably impute the missing data by a standard multiple imputation technique. As the assignment of components to individuals is randomized the missing data mechanism can be treated to be MCAR, or, at least, MAR. The production of one carefully imputed data set is the actual task. The imputation of missing values is carried out by predictive mean matching which is basically both regression and nearest neighbor

approach. Regression imputation (with rounding to the nearest observed value) tends to overestimate the explained sum of squares yielding a higher R2 than the original data would do. In order to avoid this, predictive mean matching (PMM) combines regression imputation with nearest neighbor approaches. Instead of rounding the regression estimate to the nearest observed value, each case with an initially missing value scans the observed values to "find" the case with the closest regression estimate to its own regression estimate and adopts the corresponding observed value (see Figure 3). Predictive mean matching is more likely to preserve original sample distributions than rounding to observed values, because outliers like the high-lighted values in Fig. 3 do not necessarily change the structure of the sample. One minor drawback of PPM in this situation is that only "observed" rather than "possible" values can be imputed. However, in our surveys items usually have a very limited number of possible categories and therefore this problem can be safely ignored. For a first run (starting solution) the computer algorithm includes all variables of the core component to generate initial estimates for all partly missing variables. Then, all variables (with exception of variables within the same component) are included in the regression, thus transporting any combination of variables implemented in the split.
After split the long questionnaire in three components, the application and implementation of matrix sampling design is in Fig. 3.

*Figure 3**. Imputation algorithm***

Form number	**Core**	**Component 1**
1		
2		
3		
4		
5		
6		
7		

Form number	**Core**	**Component 2**
1		
2		
3		
4		
5		
6		
7		

Form number	**Core**	**Component 3**
1		
2		
3		
4		
5		
6		
7		

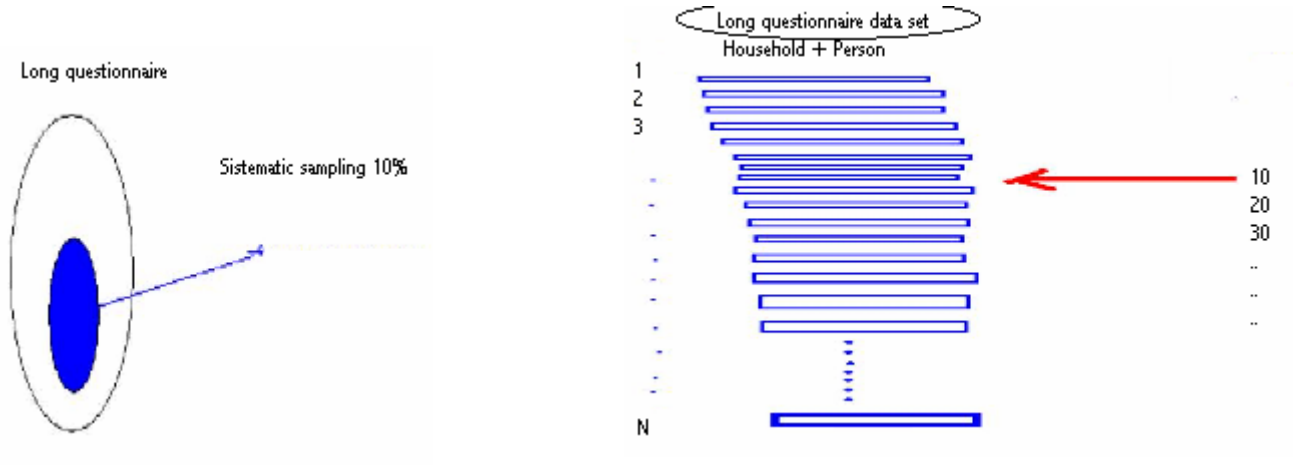
Synthetic form	4	Core	**Component 1**	6	Core	**Component 2**	1	Core	**Component 3**

Synthetic form	Core	**Component 1**	**Component 2**	**Component 3**

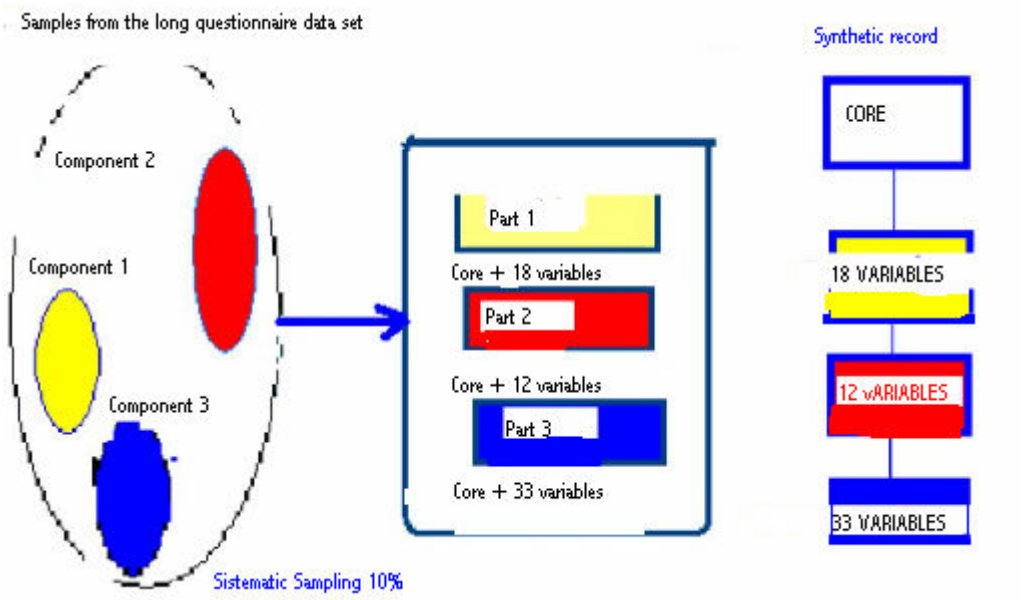
.

4 - Results

In this work, there is only one matrix sampling design. Before, we resample the original reconstructed file three times using a systematic sampling design. Then, for the simulation we split each long questionnaire into three different components files.



Each three component files are linked into the synthetic file, using a Statistical Matching by Hot Deck procedure. We have 4 samples, the first one is kept and the 3 others are linked by similar core to obtain the synthetic file. The process is repeated 20 times with 80 samples. Here, similar means the nearest neighbor approach. I used only data from Rio de Janeiro state. The variables race, levels of disability or physical or mental deficiency and religion of the population are kept in the core. All computation discussed in this article were performed on a PC using *SAS* programming language. The procedure "Survey Select" do *SAS* made the sampling and the statistical matching.



After a description of the application and implementation of matrix sampling design, some characteristics of the original and the synthetic data set are compared. Instead of homogeneity test or association measures using χ^2 test, we are using the useful criterion *mean square error* (MSE) to calculate the inference procedures obtained from the split questionnaire design and the complete questionnaire design.

This step will be repeated until the results for the imputed values converge to a certain level. The algorithm contains a stochastic component that converges in any of the conducted test. One explanation might be the imputation of missing values using data set components with at least 120,000 registers, each one.

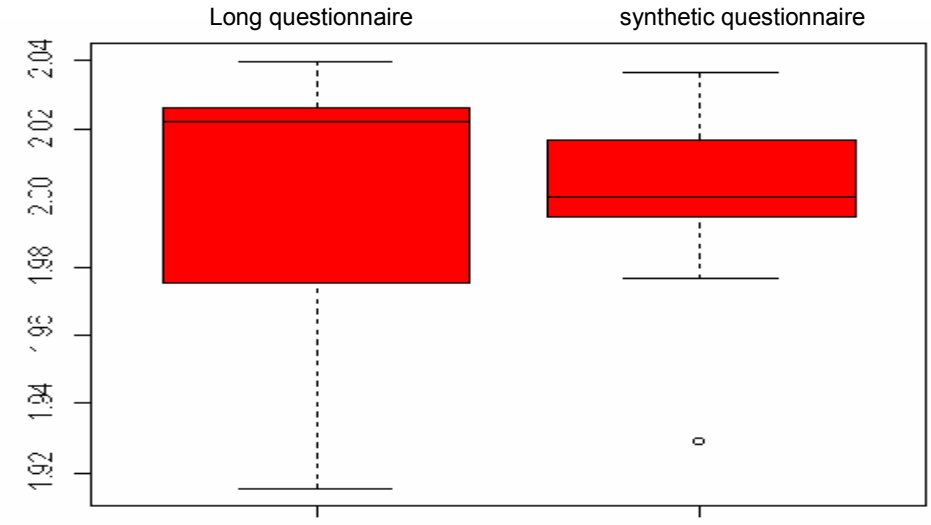
Figure 4 shows exemplary mean estimates of a split variable based on imputations, available cases and complete data set.

Using this formula to calculate MSE:

$$MSE = \sum_{i=1}^{n} \frac{(p_i - P)^2}{n}$$

where the difference is between each proportion of each cell of each matrix obtained from the split questionnaire design (p_i) and the complete questionnaire design (P) and n the size of data set components.

Figure 4 - **Proportions of a split fertility variable based on imputations, available cases and complete data set.**



Number of registers
Long questionnaire: 122,362
Synthetic questionnaire: 366,680

Most of the proportion estimates partly based on imputed values were closer to the complete data set estimator than the available case estimator. We compare the deviations between imputation and complete data estimator for one considered estimator selecting the one imputation that yields a total minimum deviation sum of squares. However, the complete data estimator is, apart from the test situation, unknown, so our idea was to derive a measure, which yields the deviation sum of squares to the proportion over all imputation proportion.
A special SAS based tool (programming by Rodrigues, N. 2003) is used to generate the imputed data sets and link them randomly. Notice that the whole imputation procedure as presented herein is random; values are sampling using a systematic design and they are ordered by core at any time.
In order to avoid that deviations of parameter estimates based on small sample sizes have a strong influence on the value of the measure, a limit of 10% is sampling from the original data set is randomly.
Thus, we obtain a MSE measure yielding weighted deviations from the Multiple Imputation(MI) estimator of the considered proportion estimates. The imputed data set so chosen was always either the “best” or, at least, among the “best” imputations.

To check for effects due to the number of observations the data were combined to three data sets of different size (n = 122117, n = 122265 and n = 122298). Table 4 provides a general overview of the results for these data sets.

Table 4 **– *Deviations between complete and imputed data sets***

Descriptive measures		
	Original data set	Imputed data set
Means	1.996.105	2.000.903
Median	2.022.045	2.000.268
Variance	0,00201	0,0005099

Font: Long Questionnaire from Brazilian Census 1991

The descriptive statistics yielded satisfying results for both data sets, with smaller average deviations for the imputed data set. However, the results for the MSE tests are comparable to a limited extent, because they are based on different sample sizes.
The conducted tests are by no means an optimal choice. However, they have been widely used for many years in the field of research.

5 - Conclusions

Split questionnaire surveys have turned out to show several positive effects: They are especially useful for cost-cutting reasons if the interview time is taking a high share of the total interview costs. In complex surveys, the imputation method may require an additional trouble of difficulty implementing split designs into the survey. The method must be investigated for each application.

A second benefit of conducting studies with split questionnaire designs is the reduced respondent burden which should lead to less unit non-response and therefore to a better sample quality. The other way round, instead of reducing the respondent burden, split designs also allow to include more questions without increasing it.

The results of the previous section suggest that it is possible to reduce the respondent burden while retaining at least marginal distributions and bivariate distributions of split and core component variables. Even better results are to be expected when proper multiple imputation methods are applied. However, the reduction cannot be extended indefinitely because the remaining information is also getting less. Besides, while more components do result in further cost-cuttings and an even lower respondent burden, they also increase the complexity of the questionnaire design and reduce the sample size for every single questionnaire. Hence, an appropriate balance has to be found for the trade-off between these effects.

The survey we examined is conducted using DIA methodology, with a Brazilian Census long form, with order of questions; the answer behavior was affected by the split questionnaire design. In the scope of this project, tests were conducted and split questionnaire results were compared with the corresponding complete questionnaire results. The split questionnaire design had better results, using three components.

However, it was beyond the scope of this project to conduct additional field tests where split questionnaire results are compared with the corresponding complete questionnaire results.

References

Adams, L. M., and Darwin, G. (1982). Solving the quandary between questionnaire length and response rate in Higher Education. *Research in Higher Education,* 17, p. 231-240.

Counting People in the Information Age (1994) Duane L. Steffey and Norman M. Bradburn, Editors; *Panel to Evaluate Alternative Census Methods*, National Research Council, p. 178-202.

Dillman, D. A., Sinclair, M. D. and Clark, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed Census mail surveys. *Public Opinion Quarterly,* 57, p. 289-304.

Holland, P.W. and Rubin, D.B. (1982*) Test Equating*. Academic Press, New York.

Little, R.J.A. (1988) Missing-Data Adjustments in Large Surveys, *Journal of Business & Economic Statistics*, 6, 3, 287-297.

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

Munger, G.F. and Lloyd, B.H. (1988) The use of multiple matrix sampling for survey research, *The Journal of Experimental Education,* **56**, 187-191.

Oliveira, Luís C. de S.; Indá, Laura B.; Lima, Rita L. A.; Bianchini, Zélia M**.** *Using System DIA software for the detection and automatic correction of errors in the data compiled using the basic questionnaire from the 1991 population census.*

Raghunathan, T.E. and Grizzle, J.E. (1995) A Split Questionnaire Survey Design, *Journal of the American Statistical Association*, 90, 54-63.

Rässler, H. (2001) Split Questionnaire Survey. *Funktionale Spezifikation zur Software SQS 1.0*, Raessler automation & consulting.

Rubin, D.B. (1976) Inference and Missing Data, *Biometrika,* **63**, 581-592.

Rubin, D.B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, *Journal of Business and Economic Statistics*, **4**, 87--95.

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

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman and Hall, London.

Shoemaker, D.M. (1973) *Principle and Procedures of Multiple Matrix Sampling*. Ballinger, Cambridge, MA.

Tennstädt (1987) Are Interviews too Long? *Readership Research: Theory and Practice*, 361-369.