Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 68(1) 2014 101-108

Web Based Sample Selection for Survey Data

S.B. Lal, Anu Sharma, Hukum Chandra and Anil Rai
Indian Agricultural Statistics Research Institute, New Delhi

Received 09 September 2013; Revised 04 January 2014; Accepted 13 January 2014

SUMMARY

The basic goal of survey sampling is to make inferences about some population parameter based upon a random sample drawn from the population. Determining sample size and then drawing a sample are important components of the whole process. With the advancements in networking technologies and availability of sufficient internet bandwidth, it is possible to implement computational procedure for sample selection and make it available to interested statisticians for estimating population parameters. In this article we describe a web based software for survey sample selection (S4). This software is an attempt to make available a free online software for survey sample selection which will not only help in selection of samples but also help in management of sampling frame. Indeed, S4 allows quick and convenient platform, with inbuilt selection procedures for statisticians and surveyors. Sample selection procedure includes both equal and unequal probability based selection of sampling units. Equal probability based selection method includes simple random sampling with replacement, simple random sampling without replacement and systematic sampling whereas unequal probability based selection includes probability proportional to size with replacement. The selection methods can vary at every stage of selection up to three stages. This software is available at http://nabg.iasri.res.in/s4/.

Keywords: Sampling, Sample selection, Imputation, Survey data analysis, ASP.NET, C#, Object oriented programming, .NET technology.

1. INTRODUCTION

Sampling is faster, cheaper and reliable mean for drawing statistical inference about the population parameters of interest. Selection of representative sample from the target population is an important part of inference from survey data. However, the selection of samples in surveys rarely involves just simple random sampling. Instead, more complex sampling schemes are usually employed, involving, for example, stratification and multistage sampling. These complex sampling schemes are usually considered to reflect complex underlying population structures; for example multistage sampling scheme is generally used to reflect the geographical hierarchy of the population. Therefore, complex survey designs are used to collect survey data to obtain a representative sample from a population (see *e.g*., Chromy *et al.* 2005 and CEMCA 2008). A complex sample design often includes stratification, clustering and multiple stages of sample with equal or unequal weighting (SAS 1999). Mahajan *et al.* (2008) developed a window based software for survey data analysis on .NET framework with an object-oriented concept based programming architecture. It also consists of a set of reusable library methods for estimation of population parameters. In large scale surveys, selecting the representative samples from the population is a cumbersome task because of the fairly involved and nested computations. Many software packages for survey data analysis (for example, SUDAAN, STATA, WesVarPC, PC-CARP, CENVAR and CLUSTERS etc.) provide methods for sample selection (see *e.g.*, Lepkowski *et al.* 1996 and references therein). But, these software are costly and sometimes require

Corresponding author: Hukum Chandra
E-mail address : hchandra@iasri.res.in

extensive coding as well as domain expertise. At the same time it is also found that these software are either stand-alone or client server based. Further there is no software available exclusively for sample selection. User friendly online sample selection software can be helpful for survey samplers, researchers and various agencies in selection of sample for carrying out surveys. We developed web based software for survey sample selection (S4) from the target population which provides a free online solution for sample selection. Note that it provides both equal and unequal probability based random sample selection methods. The developed software implements sample selection methodology at every stage for stratified multi-stage random sampling which is common sample selection procedure across surveys, in particular, large scale surveys. That is, this software is capable of selecting sample up to stratified three stage sampling designs. In addition, the sampling units are selected on the basis of equal or unequal probabilities at each stages of selection. Equal probability based selection methods include simple random sampling with (or without) replacement (srswr or srswor), systematic sampling, cluster sampling etc., whereas, unequal probability based selection method includes probability proportional to size with replacement (ppswr). In particular, the developed software implements standard procedures for selecting sampling units from fixed population. It is noteworthy that with the advancement in web technology and availability of enhanced bandwidth, implementing this web application for sample selection, accessible to all user is a reality.

This software (S4) has been developed using ASP.NET (C#) programming language under .NET framework 3.0 (Haertle 2002). This software is installable on Internet Information Server using a setup program under windows environment. The Graphical User Interface (GUI) of the software is fully browser based with links to various other sections on the left pane of the browser window which is required during the process. In order to make its use convenient, most of the functionalities of the software can be used by clicking mouse button and minimum inputs are required from keyboard. In the following section we describe software structure. In section 3 we then discuss the design of the software to implement the sample selection procedures. A brief methodology used for sample selection following standard text books is given in section 4. In section 5 we provide detailed features of the software. The results are given in section 6. Finally, in section 7 concluding remarks are made.

2. SOFTWARE ARCHITECTURE

Three-tier client-server architecture has been implemented on *Web Based Sample Selection Software*. Fig. 1 shows the architecture of this Software. There are three layers of this software namely, User Interface or Client Side Interface Layer (CSIL), Business Logic Layer or Server Side Application Layer (SSAL) and Data Access Layer or Database Layer (DBL).

1. Client Side Interface Layer (CSIL): CSIL has been implemented using CSS (Cascading Style Sheets), HTML and Javascript. Web pages have been developed using ASP.NET web forms with CSS for styling and Javascript for validations on client side input controls.

2. Server Side Application Layer (SSAL): SSAL implementation has been done using (ASP.NET) which provides the web developers a framework to create dynamic content on the server that is secure and fast. ASP.NET technology works under the .NET framework, a software framework from Microsoft. It supports object oriented programming language C# which has been used for developing code behind pages for various web forms.

3. Database Layer (DBL): Database layer is implemented using MS-Access database for storing user's profile information (*i.e.* login name, password, designation, institute name and address etc.).

This software has been developed using Visual Studio 2008, an Integrated Development Environment for developing ASP.NET based web applications. The database connectivity has been done using ADO.NET technology which has the capability to communicate across heterogeneous environments, to serve a growing number of clients without degrading system performance and to quickly develop robust data access applications using its rich and extensible component object model. The communication between server and the client computer passes through a firewall to prevent unauthorized access to the software.

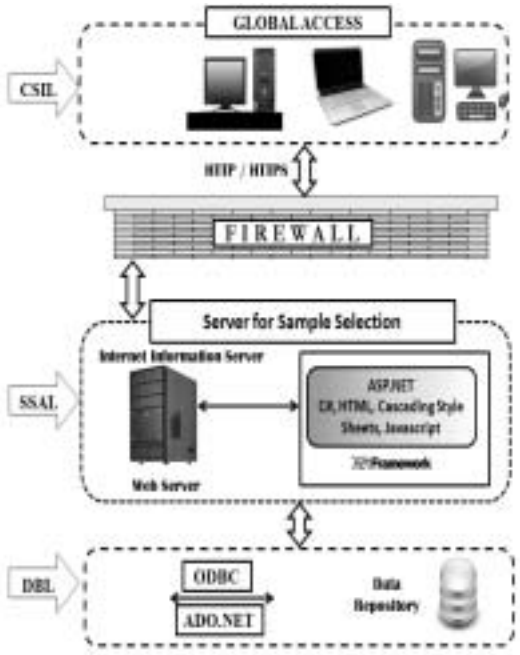

Fig. 1. Architecture of web based sample selection software.

3. DESIGN OF SOFTWARE

This Software is designed in such a way that it divides overall procedure for sample selection in two parts. The first part is data management. This module is responsible for user registration and input data file management. Logged in users can upload, download and delete input files into their specific folder specified by the login name on server. Upload option allows uploading text and excel file formats only. In the second part sample selection within each stratum as well as each stages of selection can be performed depending on sampling design of survey. The structural chart in Fig. 2 shows the data management module of the software.

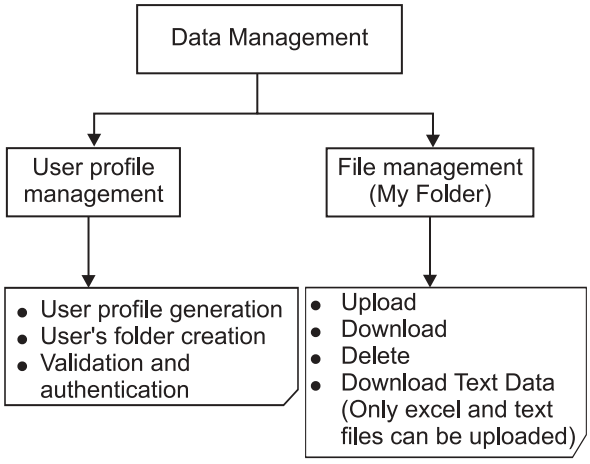

Fig. 2. Features of data management module.

Through this software, sample can be randomly selected from the population in two ways, by equal or unequal probability based selection. In case of equal probability based selection methods, each element has an equal probability of being selected from a list of all population units. The selection can be made by simple random sampling with (or without) replacement (srswr or srswor), systematic sampling, cluster sampling etc. Additional information is to be supplied by the user in case he/she wants to make selections with unequal probability. This additional variable is either an auxiliary variable for computation of probability of selection of every unit in the population or the probability value itself. A hierarchical chart showing sample selection methods of the software is shown in Fig. 3.

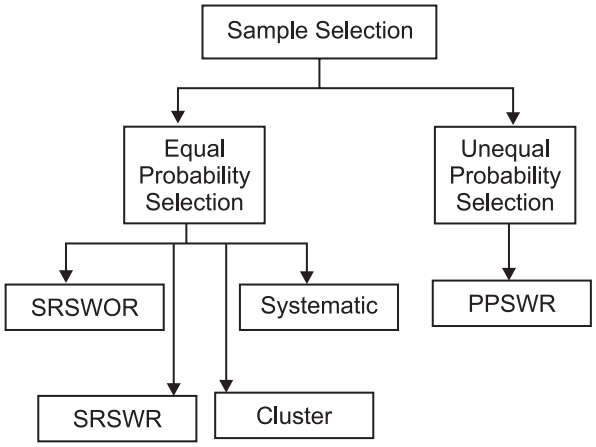

Fig. 3. Hierarchical structure chart for carrying out sample selection in the software.

4. SAMPLE SELECTION METHODOLOGY

4.1 Sample Selection Procedure

Following Cochran (2002) standard procedure for sample selection has been implemented for developing this software. In addition, we also reviewed various other literatures for developing the programming logic of the software. See Sukhatme (1984), Therese (2004) and Raghunathan *et al.* (2007). Details of implementation of different procedures of sample selection have been described below.

4.1.1 Equal probability based selection

The selection of sampling units based on equal probability is simple random sampling with or without replacement (srswr or srswor) and systematic sampling.

In simple random sampling each element has an equal probability of being selected from population sampling frame (without and with replacement). The algorithm for implementing this method of sample selection is given below:

1. Create an array containing the list of the population units from the selected excel file.
2. Get the desired sample size as input from the user.
3. Assign all individuals on the list a consecutive number from zero to the required number. Each individual must have the same number of digits as each other individual.
4. Select an arbitrary number in the table of random numbers.
5. For the selected number, look only at the number of digits assigned to each population member.
6. If the number corresponds to the number assigned to any of the individuals in the population, then that individual is included in the sample.
7. Go to the next number in the column and repeat step #7 until the desired number of individuals has been selected for the sample.

In systematic sampling an ordered sampling frame is used for selection of elements. The most common form of systematic sampling is an equal probability method, in which every *k*th element in the frame is selected, where *k* is sampling interval. Only the first unit is selected at random. The process involves the following steps:

1. Create an array containing the list of the population units from the selected excel file.
2. Determine the desired sampling fraction, say 50 out of 1000; and also the number of the *k*~unit. [$k = N/n = 1000/50 = 20$].
3. Starting with a randomly chosen number between 1 and *k*, both inclusive, select every *k*~unit from the list. If in the above example the randomly chosen number is 4, the sample shall include the 4th, 24th, 44th, 64th, 84th units in each of the series going upto the 984th unit.

4.1.2 Unequal probability selection

In probability proportional to size (pps) sampling (in particular, pps sampling with replacement), auxiliary information (*x*) present on the frame is used to calculate the probabilities of each unit in the population. Each member of the survey population has a chance of being included in the sample. In this method, unit with bigger size has higher chance of being included in the sample. The auxiliary variable available on the frame is used for selecting the units by cumulative frequency method. The algorithm for the procedure of selection a sample of size *n*, given below, has been followed in this software. Let $x_1, \ldots, x_N$ are the positive integers proportional to the probabilities assigned to the *N* units in the population, then

1. Compute cumulative totals for the sizes x_i; $i = 1, \ldots, N$, that is, $T_i = \sum_{j=1}^{i} x_j$.
2. Chose a random number *r*, such that $1 \le r \le T_N$, where $T_N = \sum_{j=1}^{N} x_j$.
3. Select the i^{th} population unit if $T_{i-1} \le r \le T_i$ with probability $\frac{x_i}{T_N}; i = 1, \ldots, N$.
4. For selecting a sample of *n* units with pps with replacement, repeat the method *n* times, where *n* is the number of units to be selected from the population

5. SOFTWARE FEATURES

5.1 Data Management

The software accepts MS Excel files for providing input values at various stages of sample selection. Depending on the survey design being implemented, the user needs to provide input information to the software at each stage of sample selection. The excel file containing the information about the input to be supplied, can be uploaded on the server using the link provided on the web form. The uploaded file by the user is always placed in user's folder and visible on "MyFolder" link. In case the user has not uploaded the file in "MyFolder" section, a facility has been provided to upload the file at each stage of sample selection.

Table 1. Format of data to be supplied by the user for sample selection at stratum level.

Str ID	Total No. of Blocks	No. of blocks to be selected
1	25	6
2	23	5

Table 2. Format of data to be supplied by the user for sample selection at stage one.

Str ID	IDs of blocks selected	Total No. of Villages	No. of Villages to be selected
1	1	12	4
1	2	13	5
1	3	14	2
1	4	22	5
1	5	21	4
1	6	21	3
2	1	19	3
2	2	18	3
2	3	13	2
2	4	19	3
2	5	20	3

Table 3. Format of data to be supplied by the user for sample selection at stage two.

Str ID	Block ID	Village ID	Total No. of HH	No. of HH to be selected	Str ID	Block No.	Village No.	Total No. of HH	No. of HH to be selected
1	1	1	10	2	2	1	1	9	1
1	1	2	7	1	2	1	2	22	3
1	1	3	9	2	2	1	3	29	4
1	1	4	4	1	2	2	1	20	3
1	2	1	5	1	2	2	2	29	4
1	2	2	10	2	2	2	3	20	3
1	2	3	12	3	2	3	1	22	4
1	2	4	10	2	2	3	2	8	1
1	2	5	9	2	2	4	1	18	2
1	3	1	17	3	2	4	2	28	4
1	3	2	10	2	2	4	3	7	1
1	4	1	5	1	2	5	1	20	3
1	4	2	11	2	2	5	2	22	4
1	4	3	19	3	2	5	3	8	1
1	4	4	24	4					
1	4	5	10	1					
1	5	1	11	2					
1	5	2	20	3					
1	5	3	10	2					
1	5	4	9	2					
1	6	1	7	1					
1	6	2	22	3					
1	6	3	18	2					

During the process of sample selection using this software, whenever input from the user is required, a dropdown list containing the names of excel files present in "MyFolder" has been provided. For carrying out sample selection at every stage, there are mainly two locations where, the input files need to be specified by the user. The user specified excel file is supposed to contain the values for total size of the strata/stages and the number of the units to be selected from those stratum/stages as shown in Table 1, Table 2 and Table 3.

In case of unequal probability selection method, an auxiliary variable or a variable specifying probability of selection for every unit is required, which can also be chosen or uploaded from/to "MyFolder". Table 4,

Table 4. Format of auxiliary variable for unequal probability selection method (Stratum Level).

Str ID	Block IDs	Auxiliary Variable
1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	46, 38, 87, 87, 37, 98, 44, 117, 58, 66, 56, 90, 45, 23, 40, 93, 67, 34, 86, 81, 79, 56, 70, 49, 18
2	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23	30, 78, 26, 59, 48, 60, 43, 31, 55, 42, 46, 33, 126, 59, 88, 63, 105, 126, 78, 77, 47, 44, 45

Table 5. Format of auxiliary variable for unequal probability selection method (Stage I).

Str ID	Block IDs	Village IDs	Auxiliary Variable
1	1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	30, 42, 13, 25, 36, 41, 22, 16, 32, 32, 22, 33
1	2	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	26, 17, 21, 19, 37, 31, 32, 49, 11, 12, 44, 54, 34
1	3	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	22, 8, 13, 21, 19, 17, 26, 31, 21, 28, 22, 26, 32, 25
	and so on		
and so on			

Table 5 and Table 6 shows the format of the file in case the auxiliary variable is selected. In this example, the stratum has been identified by "Str ID", stage I as blocks, stage II as villages and households (HH) as stage III for better understanding.

Table 6. Format of auxiliary variable for unequal probability selection method (Stage II).

Str ID	Block IDs	Village IDs	Household IDs	Auxiliary Variable
1	1	1	1	77
1	1	1	2	73
1	1	1	3	75
1	1	1	4	74
1	1	1	5	63
1	1	1	6	58
1	1	1	7	65
1	1	1	8	92
1	1	1	9	98
1	1	1	10	89
1	1	2	1	90
				and so on…

5.2 Sample Selection using the Software

This software helps researcher to select the sampling units from the population as per envisaged sampling design. In case of stratified sampling design, the population is divided into a number of strata. Further, these strata may have a number of clusters and at each stage of selection of sampling units can be done by equal or unequal probabilities. These methods have been included in the software and these are shown in Table 7.

Table 7. Sampling designs and sample selection methods available in the software.

S. No.	Sampling designs	Probability
1.	Simple Random Sampling Without Replacement (SRSWOR)	Equal
2.	Simple Random Sampling With Replacement (SRSWR)	Equal
3.	Systematic Sampling	Equal
4.	Probability Proportional to Size With Replacement (PPSWR)	Unequal
5.	Cluster Sampling	Equal/ Unequal

Sample selection process using the software has been shown in Fig. 4. This software is capable of carrying out the sample selection for stratified sampling up to three stages. The procedure of using this software involves mainly the following activities.

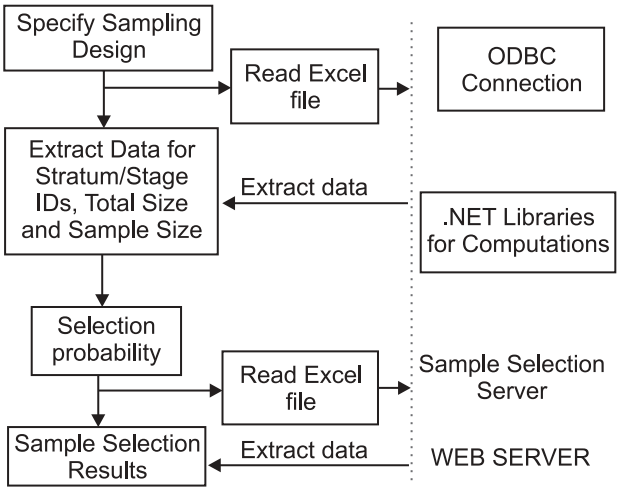

Fig. 4. Sample selection process diagram

5.2.1 Specification of sample and population sizes

In case of stratified sampling, the number of strata present in the population can be specified from a dropdown list available on the web form. Depending on the selected number of strata, it is essential to select an excel file as input file, where stratum id, stratum size and number of units to be selected from each stratum is mentioned as shown in Table 1. Sample excel file can be downloaded to the local drive to have better understanding of the format of this file, which is required to be uploaded. In the next step, selection of excel sheet name, the column names for stratum id, its total size and number of units to be selected can be specified as shown in Fig. 5 and 6. By clicking on "Read Data and Proceed" button, the control moves to get the input for the sample selection method.

Fig. 5. Sample selection using the software (Stage 1)

Fig. 6. Sample selection using the software (Stage 2)

5.2.2 Choice of sample selection methods

As the Fig. 4 shows sample selection at stage 1 where the selection for equal or unequal probability and corresponding variables can be provided. In case of equal probability there are three methods of sample selection available as shown in Table 7. For unequal probability selection method an auxiliary variable or probability values for all population units is required for computation of sample units. An excel file can be specified or uploaded for providing these values to the software. The button "Show Selection" can be clicked to carry out the computation based on the selection made and see the results of sample selected on the bottom right of the screen.

6. RESULTS

The supply of inputs for specifying stratum and stage details are similar in all three stages of sample selection, except the names of columns to be specified. An additional column name is needed to be specified for stage 1 id in case of sample selection at stage 2, and, even one more column name is to be specified for stage 2 id in case of sample selection at stage 3.

Results of the sample selection at each stage are shown on bottom right corner of the screen while user clicks on the button "Show Selection". A "Finish" button has been provided to specify that all the inputs for sample selection has been made and results are needed to be displayed. The web form shows the results of sample selection combined in a single page as shown in Fig. 7. Option has also been provided for saving the results to an excel file to the client computer.

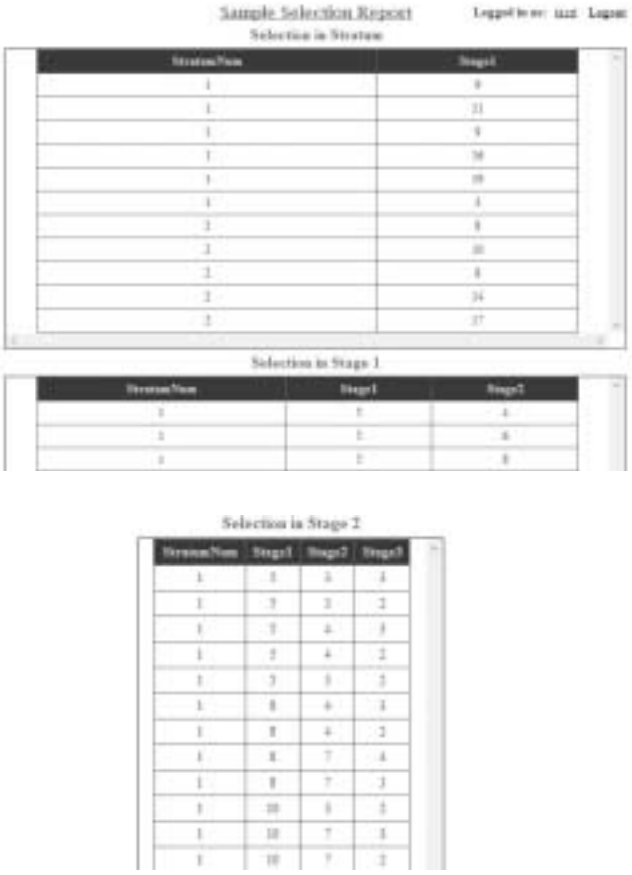

Fig. 7. Sample selection report

7. CONCLUSION

A web based software for sample selection has been developed using client and server architecture. This is accessible through World Wide Web (WWW) and requires an internet browser to run (http://nabg.iasri.res.in/S4). S4 has been developed using ASP.NET and C#. This software incorporates the standard methods of sample selection, which also includes random selection of sample units based on equal and unequal probabilities. The major features of S4 include data management, file upload, my folder management, sample selections from three stages and results section. S4 would be very useful for assist researchers in selection of random sample using complex survey design without spending time on computational aspects in the sample selection process. It is noteworthy that in case of ppswr S4 used

cumulative frequency method for sample selection. However, other method, for example Lahiri's method would be included in later version of S4.

ACKNOWLEDGEMENTS

We would like to thank the chair editor and the referee for their constructive comments. Their suggestions have led to a substantially improved final version of this paper.

REFERENCES

CEMCA (2008). Sampling (CEMCA). (Commonwealth Educational Media Centre for Asia) *Retrieved from Manual for Educational Media Researchers: Knowing your Audience*: http://www.cemca.org/books/Chapter 13.pdf

Chromy, J.R. and Abeyasekera, S. (2005). *Statistical Analysis of Survey Data*. Household Surveys in Developing and Transition Countries: Design, Implementation and Analysis. United Nations Statistics Division.

Cochran, W.G. (2002). *Sampling Techniques*. John Wiley and Sons, Inc., New York.

Haertle, R. (2002). *OOP with Microsoft Visual Basic .NET and Microsoft Visual C# Step by Step*. Microsoft Press.

Lepkowski J. and Bowles J. (1996). Sampling error software for personal computers. *The Survey Statistician*, **35**, 10-17.

Mahajan, V.K., Lal, S.B. and Sharma, A. (2008). *Software for Survey Data Analysis*. Project Report. Indian Agricultural Statistics Research Institute, New Delhi.

P.C. CARP (1986, 1989). *Users Manual*. Statistical Laboratory Iowa State University, Ames, Iowa.

SAS, I. (1999). Design Information for Survey Procedures. (SAS Institute Inc.) *Retrieved from Introduction to Survey Sampling and Analysis Procedures*: http://v8doc.sas.com/sashtml/stat/chap11/sect4.htm.

Stata: *Data Analysis and Statistical Software.* http://www.stata.com.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Application*. Iowa State University and Indian Society of Agricultural Statistics, New Delhi.

Therese, M. (2004). *Instructions for Probability Proportional to Size Sampling Technique.* RHRC Consortium Monitoring and Evaluation ToolKit. Columbia University.

Raghunathan, T.E., Solenberger, P.W. and Hoewyk, J.V. (2007). *IVEware: Imputation and Variance Estimation Software.* University of Michigan.

WesVar - *Software and Analysis of Data from Complex Samples.* http://www.westat.com/wesvar/