The sampling design effect on partial least squares algorithm

Hugo Serrato González1

Universidad Iberoamericana, México

hserrato@up.edu.mx

Ignacio Méndez Ramírez 2

IIMAS-Universidad Nacional Autónoma deMéxico, México

nacho@sigma.iimas.unam.mx

Odette Lobato Calleros 3

Universidad Iberoamericana, México

odette.lobato@ibero.mx

Abstract

The objective of this article is to analyze the effect of the probability sampling's selection on the estimated results in Structural Equation Modeling (SEM) using the Partial Least Squares (PLS) algorithm.

The idea leading this work is to estimate the satisfaction level of government service users in a large and dispersed population, for which a sample design with an equal selection probability is not a feasible option. This study is based on the analysis of the sampling distributions of estimators under different sampling designs.

It is shown that the probability of selection of the units behind the sampling design affect the results of the PLS algorithm, both the scores of latent variables and the impacts between them.

To the author's knowledge, this issue has not been addressed before in the literature.

[1] Universidad Iberoamericana Ciudad de México, Departamento de Ingeniería. Prolongación Paseo de la Reforma 880, Lomas de Santa Fé, Distrito Federal, CP 01210, México, Tel: (52-55) 59504000 ext. 4934; Fax: (52-55) 91774450. hserrato@up.edu.mx

[2] Instituto de Investigación en Matemáticas Aplicadas y Sistemas *(IIMAS-UNAM).* nacho@sigma.**iimas**.unam.mx

[3] Departamento de Ingeniería, Universidad Iberoamericana, México. odette.lobato@ibero.mx

1. Introduction

PLS algorithm is available in some software alternatives, such as, PLS Graph, SmartPLS, PLS-GUI, SPAD-PLS, among others. The PLS algorithm was designed to test and measure cause and effect relationships between both, unobservable variables (latent variables) and observable variables (manifest variables). The PLS algorithm also allows to determine a score or measurement in latent variables.

The use of PLS algorithm is prevalent in areas of social sciences, such as psychology, where it is common to work with latent variables (IQ, level of depression). Another common use of PLS algorithm is in customer satisfaction indexes. Satisfaction indexes such as the American Customer Satisfaction Index (ACSI, 2005) or the European Customer Satisfaction Index (ECSI, in Chatelin et. al (2002)) use this algorithm to measure customer satisfaction, whether it is with products or services from the private or public sectors.

The origin of this work arises from a real problem, estimating user satisfaction level of beneficiaries of social assistance programs of the Mexican government based on Partial Least Squares Path Models, i.e. by Structural Equation Modeling and the PLS algorithm. The beneficiaries of these programs are widely dispersed throughout the country in hard to reach areas. They are people of low economic level (no phone) and the only way to contact them is by face to face interview. For the type of population described above, a sampling design with equal probability of selection, such as Simple Random Sampling (SRS), is not feasible, because it involves the interviewer's displacement to difficult places, sometimes just to survey one person, even forming subgroups in the population (stratified sampling). An unequal probability sampling design is a better option in these conditions because concentrates the sample in fewer regions.

During the development of this evaluation, we came to some questions on which there are not answers in the current literature. By using PLS path models, what about the results of the PLS algorithm when the sampling design does not give the same probability of selection to elements in the population? Will the results be severely altered? Just the impacts between variables are affected? Just the scores are affected? Both? The trend is towards sub (over) estimate? What we can be done to reduce the effect of sampling design? Is the use of weights (probability weights) convenient in this situation?

Analyze the situation from the theoretical point of view presents serious difficulties, as the PLS algorithm results are determined by a numerical iterative process for which formal developments hardly reach results. The theorem proofs we found in some publications, (Dijkstra 2009, Fornell 1982) are somewhat different to the demonstrations in other areas of mathematics (less formal). This is partly due to the difficulty caused by the fact that the PLS estimators are the result of an iterative process that complicates the subsequent handling in theoretical developments.

However, we have the option to analyze things from the perspective of sampling theory, simulating samples from a finite population known under different conditions (equal and unequal probability of selection). This procedure, also reaching answers to the above questions, provides an alternative methodology to analyze possible benefits or failures of PLS algorithm in other cases.

There are articles that tell us about certain advantages of estimations that result the PLS algorithm. For example, T. Dijkstra (2009) shows that the PLS algorithm has convergence from arbitrary points to unique solutions with probability approaching to one when the sample size is infinite. Hse, Chen and Hsieh (2006) mention that the PLS algorithm results are consistent in the long term, which means that their bias is reduced when the sample size and the number of manifest variables per latent variable increases. However, these results do not refer to aspects related to the sampling design, leaving the understanding that it is a simple random sampling and that the population is infinite.

In this paper, the results related to the effect of the sampling design on PLS algorithm are presented; specifically, the selection probability behind the sampling design using PLS path models to measure citizen's satisfaction level.

The analysis is carried out with classical sampling theory that is, assuming that the population is finite and comparing two types of sampling designs: with equal and with unequal probability of selection.

2. Structural Equation Modeling

A PLS path model include two parts, a measurement or outer model and a structural or inner model. The measurement model establishes the way the manifest variables "measure" the value of the latent variables (a variable that cannot be measured directly). The structural model establishes relationships of cause and effect between different latent variables in a model; this includes a set of regression equations. The length or weight of a metal rod can be measured directly. In contrast, the concept of musical ability cannot be measured directly; it is then a latent variable. It can be postulated that musical ability has a direct effect on aptitude for mathematics, which is itself another latent variable. Both aptitude for mathematics and musical ability can be measured by activities or specific questions that can be directly assessed (manifest variables). A PLS path model usually is plotted in a diagram, for example,
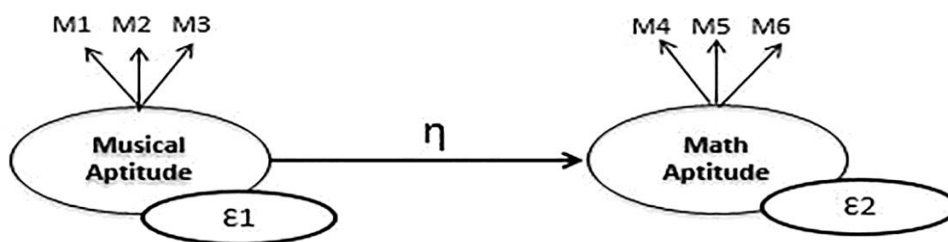
Figure 1.

Figure 1: A PLS Path Model

In Figure 1, it can be seen that musical ability is measured (measurement or inner model) through three manifest variables (M1, M2, M3), which determine a value Ɛ1 in this variable. Math aptitude is measured as well with three manifest variables (M4, M5, M6) and determines a value Ɛ2. In this part, PLS algorithm employs an iterative method to determine the values Ɛ1 and Ɛ2. Arrows out of the latent variables indicate that the reflective mode is used in the measurement model. Here there are two modes, reflective mode and formative mode. (Henseler et. al 2006). Details of the PLS algorithm can be found in several publications, and can be consulted in Henseler et al. (2009), Tenenhaus et al. (2005) or Haenlein et al. (2004), among others.

In the structural or inner model, it is postulated that musical aptitude affects the aptitude for mathematics, which is represented in the diagram by the arrow pointing from one variable to another. The effect or impact η between these latent variables is measured by common linear regression (simple or multiple).

Structural models can handle two or more latent variables, with as many manifest variables as deemed necessary or sufficient.

In Mexico, basing their work on the ACSI government model, Lobato et al (2011) constructed a PLS path model to estimate citizens' satisfaction with a social milk supply program named LICONSA from the Secretaría de Desarrollo Social (SEDESOL), a department of the Mexican government. In Mexico, this program sells milk at a preferential price to people that satisfy certain financial and social requirements. The LICONSA program is available in most of the country, as its´ beneficiaries (over 6500 000) are distributed widely. After developing a model, the authors, with the financial support of the Consejo Nacional de Ciencia y Tecnología (CONACYT), designed a measurement instrument and a sampling design. They determined a sample size of 1200. After eliminating questionnaires with contradictory answers, the actual sample size was 1140. It is this sample of 1140 beneficiaries that is considered as the total known population.

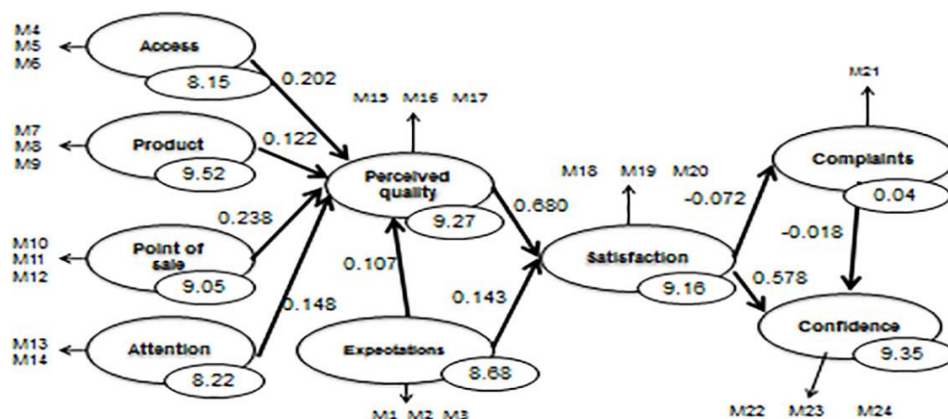The estimated satisfaction model for the milk supply program LICONSA is shown in

Figure 2.

Figure 2: LICONSA Satisfaction Model (Source: Lobato et. al. 2011)

The model is estimated using the PLS algorithm through SmartPLS software (developed by Ringle, Wende and Will, 2005) in a standardized metric. The Path weighting scheme was chosen from the three options offered by both the PLS algorithm and the SMARTPLS program. Henseler et al (2006) reported that, on final results, there are no substantial differences between the Path, Centroid and Factor Weighting Schemes.

Although the PLS algorithm casts a wide variety of data between their results, the salient points are the impacts between latent variables (numbers over the arrows in Figure 2) and the scores of each latent variable (values in small ellipses over the latent variables in Figure 2).

It can be seen, in figure 2, that nine latent variables were identified: Access, Product, Point of Sale and so on, and include each manifest variables, denoted by M1, M2,..,M24. All latent variables include three manifest variables, except the Complaints and Attention latent variables which include only one and two manifest variables, respectively.

Using PLS, it is estimated, for example, that the impact between Perceived Quality and Satisfaction is 0.680 (which is interpreted in the same way that any coefficient in a regression model) and that the Satisfaction of LICONSA beneficiaries toward the program is 9.16, on a scale from1 to 10.

3. Methodology

In order to study the effect of sampling design, the subsequent procedure was followed:

1) **Using a specific sample size, several samples are generated with two different sampling designs of a known finite population.** The known population is the group of 1140 beneficiaries of the program LICONSA. One of the sampling designs used equal selection probability (SRS) and the other, unequal selection probability. In this last case, stratified sampling (SS) is implemented. The 1140 LICONSA beneficiaries are classified into nine groups (strata) according to their academic level, taking sampling sizes into each stratum, non-proportional to the stratum size, but the squared stratum size, in order to assign an unequal probability to each population element. In this way, the elements in larger strata are more likely to be selected than elements in small strata.

For the stratified sampling (SS) design, the strata, stratum size and the stratum sample sizes are shown in Table I.

Table I: Academic Level of 1140 LICONSA Beneficiaries (Source: Lobato et. al. 2011)

Strata	Stratum Size	Stratum Sample Size
No school attendance	70	6
Elementary uncompleted	200	47
Elementary completed	242	68
Middle school uncompleted	95	10
Middle school completed	296	102
High school uncompleted	59	4
High school completed.	73	6
Technical career uncompleted	68	5
Technical career or Bachelor Degree	37	2
Total	1140	250

2) **Estimations of a particular model with each sample are calculated, using PLS algorithm**. The 1140 beneficiaries will be considered as the total population. The estimated results of impacts and scores are considered the "real values" or "population parameters" which are then estimated using two sampling designs, one of them with equal probability of selection (SRS) and other with different probability of selection (SS). The results obtained are studied in order to determine the effect that selection probability sampling design has on estimates of the PLS algorithm.

The sampling distribution of the estimates of the model parameters is useful for comparing these two different sampling designs. Therefore, to compare the level of accuracy, we check how near or far, on average, sample estimates are from the results of the total population.

3) **Comparison of estimations distributions from different sampling designs with the results of the total population**. The comparison was made through techniques of descriptive statistics, i.e., box plots, sample mean, sample standard deviations, variation coefficients.

4. Results

The results of the procedure described in the previous section are as follows:

1) **Using a specific same sample size, several samples are generated with two different sampling designs of a known finite population.** With a sample size of 250, there were generated 50 simple random samples. Then, under stratified sampling another 50 samples were selected. The sample sizes within each stratum were determined with the method described previously. This data can be found in Table 1. All samples were generated using

IBM SPSS 22.

2) **Estimations of a particular model with each sample are calculated, using PLS Algorithm.** The structural model in Figure 2 was estimated using PLS algorithm for each of the 100 samples selected, 50 simple random samples (SRS) and 50 stratified samples (SS), recording the results associated with scores and impacts between latent variables.

3) **Comparison of estimation distribution from different sampling designs with the results of the total population model**. The above procedure yielded many results. To facilitate their interpretation, we present first the distribution of the estimates of one latent variable score, specifically, the Satisfaction score and one impact between the latent variables, specifically, the Point of Sale to Perceived Quality impact.

In Figure 3, we observe the sampling distribution for the 50 estimations of the score of the latent variable Satisfaction, with each sampling designs (SRS and SS).
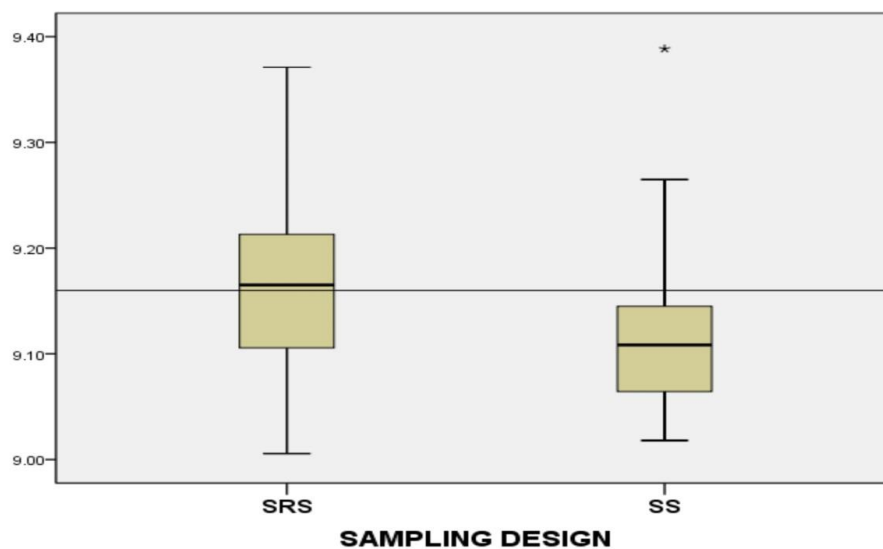

Figure 3: Satisfaction Scores Sampling Distribution

The horizontal line indicates the target value, which is also called the actual value of satisfaction. This value represents the real score in the latent variable Satisfaction (the population value for the 1140 beneficiaries) and, conveniently, the distribution of the estimates should be symmetrical around this value, to think in unbiased estimators.

In Figure 4, we observe the sampling distribution of 50 estimates of the impact Point of Sale-

Perceived Quality with each of the two sampling designs. The horizontal line, as previously mentioned, indicates the target value.
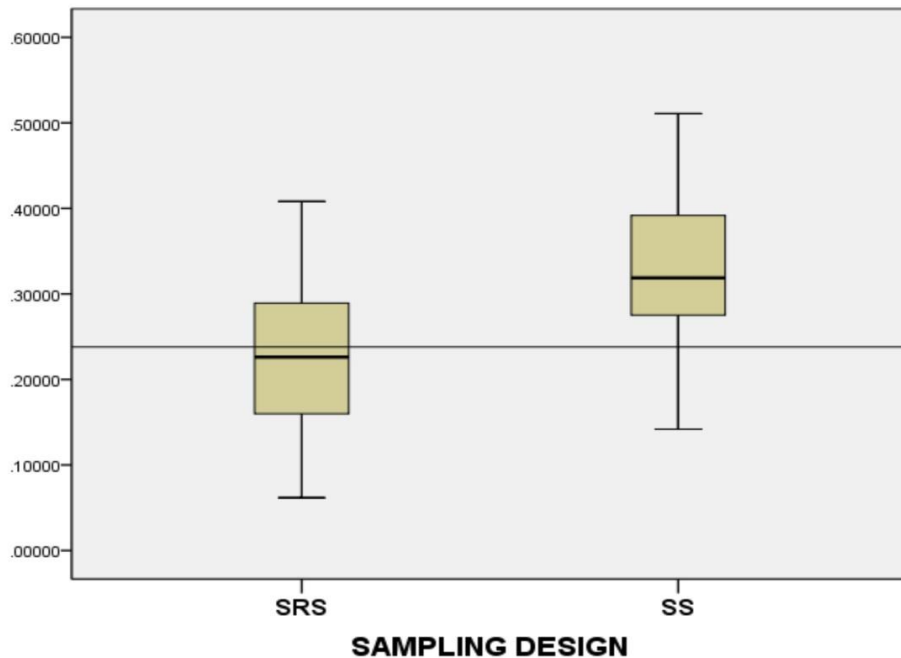

Figure4: Impact Point of Sale to Perceived Quality Sampling Distribution

In the two previous figures, we can observe clearly the presence of sampling design effect: If the probability of selection is the same (SRS), the distribution estimates are more symmetrically around the objective value than with unequal selection probability (SS). The asymmetric distribution in Figure 3 shows that over 75% of the SS results are sub-estimating the real value of the Satisfaction score. Moreover, Figure 4 shows that over 75% of the SS results are over-estimating the real value of the impact Point of Sale-Perceived Quality. These two observed results imply that we have bias in the estimation process with unequal selection probability.

We can observe, in Figure 5, the sampling distributions of nine estimated scores in the complete model in figure 2, found in box plots that follow the same pattern as Figures 3 and 4.
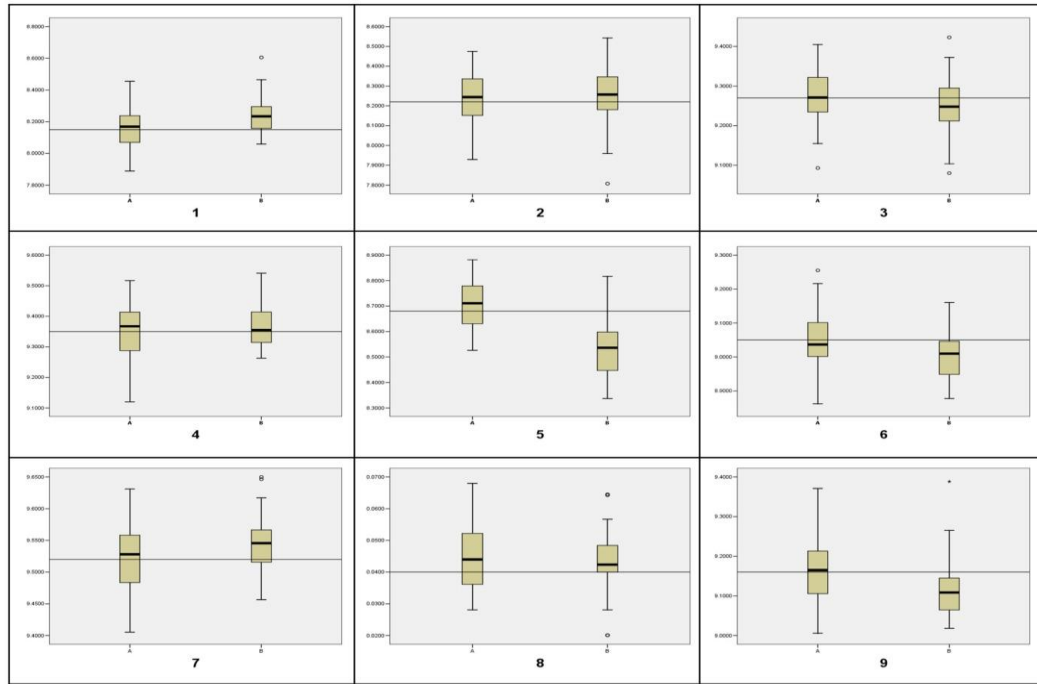
Figure 5: Scores Distributions: 1Access; 2 Attention; 3 Perceived Quality; 4 Confidence; 5 Expectations; 6 Point of Sale; 7 Product; 8 Complaints; 9 Satisfaction

The distribution of the ten estimated impacts can be found in Figure 6.

Figure 6: Impact Distributions: 1 Access- P. Quality; 2 Attention –P. Quality; 3 Point of sale-P. Quality; 4 Product-P. Quality; 5 Expectations-P. Quality; 6 P. Quality-Satisfaction; 7 Expectations-Satisfaction; 8 Complains-Confidence; 9 Satisfaction-Confidence; 10 Satisfaction-Complains

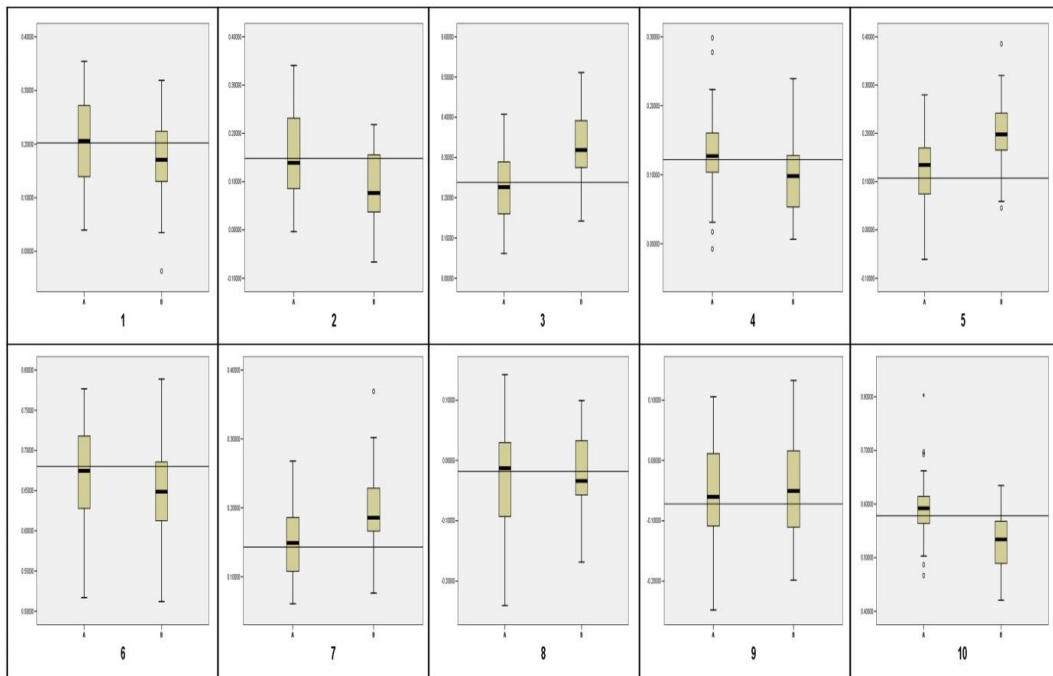For the estimate distributions with SRS, we can say that, in general, they are symmetrical around the objective value in all the cases (scores and impacts). We believe, under SRS, unbiased estimators have been used. Under a stratified sampling the picture changes, the unequal probability of selection present in the stratified sampling induces an apparent bias in the distribution of the estimates, both in impacts and scores. The conclusion cannot be neither under estimation or over estimation is induced, as both cases are presented.

Finally, table II and table III show results of descriptive statistics of estimates in impacts and scores, respectively.

Table II: Descriptive Results: Scores

Latent Variable	Sampling Design	N	Mean	Std. Deviation	Objective Value	Absolute Error	Percentage Error	Coefficient of variation
Access	SRS	50	8.165	0.141	8.147	0.018	0.002	1.73%
	SS	50	8.241	0.112	8.147	0.094	0.011	1.36%
Attention	SRS	50	8.236	0.131	8.218	0.018	0.002	1.59%
	SS	50	8.251	0.135	8.218	0.033	0.004	1.64%
Per. Quality	SRS	50	9.273	0.066	9.267	0.006	0.001	0.71%
	SS	50	9.249	0.063	9.267	0.018	0.002	0.68%
Confidence	SRS	50	9.352	0.087	9.351	0.001	0	0.93%
	SS	50	9.365	0.065	9.351	0.014	0.002	0.69%
Expectations	SRS	50	8.705	0.089	8.682	0.023	0.003	1.03%
	SS	50	8.531	0.1	8.682	0.151	0.017	1.17%
Point of Sale	SRS	50	9.053	0.077	9.048	0.005	0.001	0.86%
	SS	50	9.009	0.072	9.048	0.039	0.004	0.80%
Product	SRS	50	9.524	0.051	9.519	0.005	0.001	0.53%
	SS	50	9.544	0.043	9.519	0.025	0.003	0.45%
Complaints	SRS	50	0.044	0.01	0.045	0.001	0.024	22.80%
	SS	50	0.043	0.01	0.045	0.002	0.044	23.26%
Satisfaction	SRS	50	9.164	0.074	9.162	0.002	0	0.81%
	SS	50	9.116	0.066	9.162	0.046	0.005	0.72%

Table III: Descriptive Results: Impacts

Impact	Sampling Design	N	Mean	Std. Deviation	Objective Value	Absolute Error	Percentage Error	Coefficient of Variation
Access-P. Quality	SRS	50	0.202	0.082	0.202	0.000	0.20%	40.62%
	SS	50	0.174	0.073	0.202	0.028	13.98%	41.95%
Attention-P. Quality	SRS	50	0.152	0.086	0.148	0.003	2.25%	56.72%
	SS	50	0.088	0.076	0.148	0.06	40.71%	86.36%
Point of Sale-P. Quality	SRS	50	0.226	0.082	0.238	0.012	4.98%	36.28%
	SS	50	0.323	0.083	0.238	0.085	35.68%	25.70%
Product-P. Quality	SRS	50	0.129	0.057	0.122	0.007	5.98%	44.47%
	SS	50	0.103	0.062	0.122	0.019	15.47%	60.19%
Expectations-P. Quality	SRS	50	0.122	0.081	0.107	0.014	13.46%	66.75%
	SS	50	0.201	0.07	0.107	0.094	87.08%	34.83%
P. Quality-Satisfaction	SRS	50	0.672	0.064	0.68	0.007	1.08%	9.45%
	SS	50	0.649	0.059	0.68	0.031	4.51%	9.09%
Expectations-Satisfaction	SRS	50	0.15	0.052	0.143	0.006	4.50%	34.99%
	SS	50	0.198	0.054	0.143	0.055	38.07%	27.27%
Complaints-Confidence	SRS	50	0.026	0.082	-0.018	0.008	42.31%	317.03%
	SS	50	0.025	0.069	-0.018	0.007	37.47%	276.00%
Satisfaction-Confidence	SRS	50	0.592	0.057	0.578	0.014	2.45%	9.62%
	SS	50	0.531	0.054	0.578	0.047	8.13%	10.17%
Satisfaction-Complaints	SRS	50	0.056	0.081	-0.072	0.016	22.76%	145.81%
	SS	50	0.048	0.079	-0.072	0.024	33.53%	164.58%

In these tables, it is shown that the percentage error and coefficient of variation is greater when estimating impacts than when latent variable scores are estimated with two different sampling designs (SRS vs. SS).

By comparing the variability in estimates of each, score and impact, with the two different sampling designs, we note that are no significant differences.

5. Conclusions

We can now establish some conclusions and comments related to the effect of the probability sampling's selection on the estimated results in Structural Equation Modeling (SEM) using Partial Least Squares Algorithm.

Principally, we conclude that, when using sampling design with equal selection probability, the estimators of impacts and scores of the latent variables are, apparently, unbiased. Otherwise, when a sampling design with unequal selection probability is used it has been found that these estimators are, in most cases, biased. As it mentioned above, the conclusion cannot be neither under estimation or over estimation is induced, as both cases are presented.

Although we can argue about the effect that structural model can have on subsequent results, as mentioned previously, the model used is similar to ACSI and ECSI satisfaction models and, although we can do a similar analysis to this with a simpler or more complex model, the idea is to study the sampling design effect in models that are the most similar possible to the models that are currently used to evaluate satisfaction indexes.

No one questions that real data have the advantage over simulated data, as being more "realistic data", since the simulated data are generated with artificial patterns that do not necessarily apply in the real world.

The PLS algorithm determines first the scores of the latent variables through an iterative procedure. After that, it determines the impact among latent variables using ordinary regression. Correcting the sample selection bias within the PLS algorithm represents an issue not previously addressed. Already there are procedures for the correction of this bias in regression models, (Skinner 2012, Magee 1998) which could be applied to try to correct the bias in impacts between the latent variables. However, previously we should have also corrected the bias present in the scores of the latent variables. For the latter, the alternatives are not clear.

The idea of incorporating weights in the estimation process related to the probability of selection, as does the estimator of Horvitz-Thomson (Horvitz and Thomson 1952), seems quite risky because the iterative procedure that determines the scores in the PLS algorithm,

could be modified seriously. It is not the same to estimate a population total, as in the Horvitz-Thomson estimator, than an iterative result. However, it is precisely in reference to this aspect that we can work in the future, specifically, which type of weighting could give us better results in a specific sample size. Nevertheless, commercial software that execute the PLS algorithm might well consider the alternative of incorporating weights (as the inverse of the selection probability) within the algorithm.

When estimating structural models using the PLS algorithm, there are no specific formulas to determine a sample size to satisfy precision conditions. Some criteria are mentioned in literature based mainly on empirical reasons (Henseler et al. 2009, Morgeson 2011).

Once a sample size is determined by a criterion, as mentioned in Henseler et al. (2009), Morgeson (2011) or others authors, if an unequal selection probability sampling design is considered, a good choice would be to increase the sample size, say 150% or even 200%.

It is frequently mentioned in the literature (Chin and Newsted, 1999) that PLS requires smaller sample sizes than the LISREL method, but it is questionable to use small sample sizes under sampling designs with unequal probability of selection as it is shown here, as it may introduce bias.

A sample size of 250 was used in section 4, because it is commonly used. For example, the ACSI (Morgeson, 2011) uses this sample size in its evaluations.

We found that simple random sampling seemed to be the best option. We acknowledge that this is not always possible, due to the difficulties involved in giving the same probability to all possible population samples, or the costs and time involved in collecting data from a large and disperse population located over large geographical areas.

The results presented here correspond to a single case, a single structural model. It is likely the existence of cases with unbiased sample selection exist, however, this one case warns against the presence of a problem.

The fact that in this particular case the biased estimators arise should warn us of the existence of this same problem in similar cases when working with unequal selection probability.

Moreover, our previous work, we believe, is a good alternative to understand, in practice, what the concept of unbiased estimator is, which is a criterion often sought when carrying out parameter estimation.

6. Bibliography

American Customer Satisfaction Index (2005), American Customer Satisfaction Index. Methodology Report, Michigan Ross School of Business/American Society for Quality/CFI Group, Ann Arbor, MI.

Chatelin Y.M., V.E. Vinzi, and M Tenenhaus (2002) "State-of-art on PLS path modeling Through the available software", Mimeo

Chin, W. W., and Newsted, P. R. (1999). "Structural equation modelling analysis with small samples using partial least squares". In R. H. Hoyle (Eds.), Statistical strategies for small sample research. Thousand Oaks, CA: Sage. pp.307–341.

Djkstra, T.K. (2009), Latent variables and indices: "Herman Wold´s basic design and partial least squares", In: Esposito, V., Chin, W., Henseler, J., and Wang, H. (Ed), Handbook of partial least squares, Springer, Heidelberg. pp. 23-46.

Fornell, C. and Bookstein, F.L. (1982), "Two structural equation models: LISREL and PLS applied to consumer exit-voice theory", Journal of Marketing Research, (19:4), pp. 440-452.

Haenlein, M. and Kaplan A.M. (2004), "A beginner´s guide to Partial Least Squares Analysis", Understanding Statistics, (3:4), pp. 283-297.

Henseler, J., Ringle, C.M. and Sinkovics, R.R. (2009), "The use of Partial Least Squares Path Modeling in International Marketing", Advances in International Marketing, (20), pp. 277-319.

Horvitz, D. G.; Thompson, D. J. (1952) "A generalization of sampling without replacement from a finite universe", Journal of the American Statistical Association, 47, 663–685,

Hsu, S.H., Chen, W.H., and Hsieh, M.J. (2006), "Robustness testing of PLS, LISREL, EQS and ANN-based SEM for measuring customer satisfaction", Total Quality Management, (17:3), pp. 55–371.

Lobato O., Rivera H., Serrato H., Gómez M., León C., and Cervantes, P. (2011), "Reporte Final del IMSU-Programas Sociales Mexicanos. Programa de Abasto Social de Leche Liconsa – Modalidad de leche líquida", available at: http://www.2006-

2012.sedesol.gob.mx/es/SEDESOL/Evaluacion_de_la_Satisfaccion_de_los_Beneficiarios_. (Active link)

Magee, Lonnie (2005), "Improving Survey Sampling-Weighted Least Squeres Regression" Journal of The Royal Statistical Society, Series B (Statistical Mehodology) (60:1), pp. 115-126.

Morgeson III, F. V. (2011), "How much is enough?" Sample size, Sampling and the CFI Group Method. CFI Group internal document.

Ringle, C.M., Wende, S., and Will, A. (2005), SmartPLS, release 2.0 (beta), Hamburg, Germany: SmartPLS. http://www.smartpls.de.

Skinner, C. (2012) "Weighting in the regression analysis of survey data with a

cross-national application" Canadian Journal of Statistics, manuscript.

Tenenhaus, M., Vinzi, V.E., Chatelin, Y.M., and Lauro, C. (2005), "PLS path modeling", Computational Statistics & Data Analysis, (48:1), pp. 159-205.

Thomson, S.K. (2012), Sampling, John Wiley & Sons, Inc. Hoboken, New Jersey.