Quality Technology & Quantitative Management
Vol. 8, No. 2, pp. 167-181, 2011

QTQM

Errors in Customer Satisfaction Surveys and Methods to Correct Self-Selection Bias

Giovanna Nicolini and Luciana Dalla Valle

Department of Economics, Business and Statistics, University of Milan, Italy
(*Received* November 2009, *accepted* July 2010)

Abstract: This paper provides an overview of the main types of surveys carried out for customer satisfaction analysis according to the dimension of the firm and to the data collection method. Different errors can be generated by these methods, causing biases in the results. We focus on the self-selection error and we employ three methods to correct the bias associated to it. A simulation underlines the performances of the three methodologies.

Keywords: Customer satisfaction surveys, Heckman two-step procedure, hierarchical Bayesian approach, propensity score matching, self-selection bias.

1. Introduction

Statistical surveys need complex procedures borne out of organizational efforts as well as economical efforts that in many cases may appear too heavy for the firm. The survey may include every customer of the firm (census survey) or just a part of them (sample survey). The kind of survey chosen depends on many factors concerning the type of firm and its customers [11].

Firms define themselves through territorial presence and the market they operate in. Hence, these three types of firms:

- *big firms*: they operate nationwide in an oligopolist regime or more generally in "imperfect competition". Such firms, for example, are the ones providing telecommunications, electricity, transportation, banking, finance, insurance and health care services;
- *medium firms*: they generally operate in a smaller territory context, but if they operate nationwide, they are in "perfect competition" (local transports, tour operators, private clinics);
- *small firms*: they locally develop a wide range of services in a climate of "perfect competition".

Big firms often carry out sample surveys with big probabilistic samples, whilst medium or small firms prefer to carry out census surveys. If we focus our attention on the types of customers, we can certainly distinguish the firms who offer services to other firms (business to business) from the ones which offer services to the final consumer (business to consumer). In the first case, the list of customers (in general not a huge one) is known and the survey is quite often a census survey. Conversely, in the second case, the list of customers is not necessarily known and the survey can only be a sample survey. Considering these two factors, have a look at Table 1.

Table 1. Types of surveys for types of firms and customers.

Firm's dimension	B2B	B2C
small	C	C/S
medium	C	C/S
large	C	S

C= census survey; S= sample survey.

However, the usefulness of a survey (be it a census or a sample survey) depends on its reliability, namely the proximity of the results of the survey to the real ones. If the reliability is poor the results are inevitably biased and at the end the whole survey may be useless. Poor reliability may have many causes. One of these is the presence of errors. Literature identifies different kinds of errors. Some of these are related to the methods used to gather statistical data. Firms may use various methods according to the types of customers, the times and the costs involved. The final goal of this work is to highlight the features and the problems of the Customer Satisfaction Surveys (CSSs) regarding the types of firms that provide services, the types of customers, the data gathering methods carried out throughout the surveys and, consequently, the likely errors connected to them (second section). Our attention will be focused on a seldom studied kind of error (at least in this kind of survey) and on the methods provided by literature to correct it. These methods are generally borne out of different aims or in distant contexts and need to be aptly modified in order to be applied in CSSs (third section). The fourth section shows a simulation we have made in order to evaluate the performance of the applied methods and the changes we have introduced. Finally, we give concluding remarks.

2. Errors and Data Collection Methods in CSSs

According to the literature the reliability of the results of a survey depends on the existence of errors, that can be non-sampling errors or errors related to data collection methods.

2.1. Non-sampling Errors

The errors that might modify the results of a sample survey and a census survey are known as coverage errors, measurement errors, unit non-response errors and self-selection errors [9, 14]. In particular:

- *coverage error*. It is observed when the number of customers of a firm (target population) and the list of them (frame population), from which information are generally taken to involve the customers in the survey, do not coincide;
- *measurement error*. It is given by the difference between the real value of an item related to a surveying unit and the value observed. This kind of error has been frequently connected to the presence of an interviewer or to data charging;
- *unit non-response error*. We have it when the selected unit does not answer or does not fill in the questionnaire form. This non-response may be caused by the inability to reach the customer or by the refusal of the customer, who does not want to join the survey;
- *self-selection error*. It is observed when no selections have been made before the survey and the person independently decides to join the survey. If no selections are made the survey is a census survey. Though, self-selection determines a certain amount of non-responses. The self-selected who have provided answers form a non-probabilistic sample of the population.

Amongst the enlisted errors the measurement one may appear, in CSSs, the least relevant because in many circumstances the questionnaires are filled in by the customer himself and because questions and answers are generally quite easy. More relevant is the coverage error which, along with the measurement one, may be attributed to the firm or to the way the survey has been carried out. The coverage error implies the existence of a frame. We meet this kind of error when a firm does not have a list of customers or does not have an updated version of it. The existence of this frame allows to make a census survey or a sample survey with a probabilistic sample. On the other hand, if we do not know the frame we will not be able to carry out a census survey, whilst the sample survey will generally use a non probabilistic sample (a quota sample for example). Though, we may also think of a nearly-probabilistic sample with a perfectly suited sampling scheme. Conversely, the last two errors can be attributed to the interviewee/customer. As far as the non-response is concerned, it is thought that this kind of error is associated to probabilistic samples. The researcher knows the features of the people who do not answer; as a matter of fact, the weighting methods suggested in presence of non-responses are based on models in which auxiliary variables are employed. These are quite known to both answering and non-answering people. What we do not know are the reasons behind the non-responses. The reasons behind non-responses in this type of surveys are different from those behind surveys about sensitive matters (as for example discriminatory diseases or private life facts), that may make some interviewees feel uneasy. Indeed, causes for non-response in CSSs can be sheer boredom and reluctance of people to fill in the questionnaire due to the number of CSSs one is asked to fill in everywhere (in hotels, after organized tours, at conferences, after the completion of a workshop/course/study programme, etc.), often very poorly designed questionnaires, and lack of obvious improvement in service processes following the implementation of a CSS. The self-selection error, instead, is associated to the independent choice of the customer who decides to take part in the survey. This is possible only if we carry out a census survey. As far as this survey is considered the questionnaire is sent to all the customers, who then independently choose to fill or not fill in the form. Self-selected subjects are a non random sample of the population, as well as not self-selected subjects are a non random sample of the population. The interview refusal in a sample survey is considered as unit non-response, while the interview refusal in a census survey is called not self-selection.

2.2. Errors Related to Data Collection Methods

Let us now see what are the most common data collection methods and the types of errors associated to them:

- *Face-to-Face Interviewing* (*F2FI*). This method is used when a survey has to be made because the list of customers is not known yet. It is frequently used by mobility services because for them it is possible to interview the customers only when they actually use the service. It is generally carried out in a sample survey context. This is a very expensive method; here the presence of the interviewer may produce a bias caused by measurement errors.
- *Computer Assisted Telephone Interviewing* (*CATI*). This method is frequently used and not very expensive. It is possible to employ this method if a list of customer already exists. It can be used in census surveys and sample surveys with probabilistic samples. It is possible to keep a close watch on measurement errors, whilst coverage errors (if the list of customer is not updated) and non-response ones are more frequent.
- *Computer Assisted Web Interviewing* (*CAWI*). It is the modern version of the now rarely used mail interview. This method is correctly applied if extended to all those customers who

have an e-mail address known to the firm. Being the survey a global one, it is possible to verify a self-selection error. Conversely, if not every customer has an internet access, the choice of this method implies the presence of coverage errors.

- *Web-Survey* (*WS*). Web-Surveys may be of various types [15] and get more widespread every day. In the same way the number of people having an internet access increases every day. The web approach most frequently employed in the CSSs are online questionnaires we can easily find on the firm website or questionnaires that pop up every x visitors on the firm website. In the first case (*WS*1), the filling in of the questionnaire, open to everyone, depends on the customer's will. Hence, we have a self-selection. In the second case (*WS*2), as a systematic sample is used, there might be a non-response error. If every customer of the firm have an internet access (e.g., an online bank), or if the survey is only destined to the visitors of the site, a coverage error is not present. On the other hand, if not every customer of the firm is an internet surfer the coverage error adds up to the errors we have already mentioned.
- *Open Surveys* (*OS*). The survey consists in giving the customer a paper questionnaire at the end of the service. The filling in of the questionnaire and the handing in of the filled in questionnaire to the firm depends on the customer. This survey is very easy, it is not supported by any kind of organization and the self-selection error is quite frequent.

A summary of the surveying methods with the errors associated to them is reported in Table 2.

Table 2. Data collection methods and errors.

	Measurement	**Coverage**	**Non-response**	**Self-selection**
F2FI	*		*	
CATI	*+	*+	*+	
CAWI		*+	*	+
*WS*1		+		+
*WS*2		*	*	
OS				+

* for sampling survey; + for census survey.

In this paper our focus is on census surveys in which self-selection errors are quite frequent. The presence of this error alters the survey itself, which becomes a sample survey. As a matter of fact, after the filling in procedure, we do not have N filled in questionnaires (N is the population size), but n questionnaires where $n < N$. What we have in our hands is a $n-$size non-probabilistic sample. It is well known that with this kind of sample it is not allowed to apply the inferential methods as far as the design-based approach is concerned. This may only be possible if the bias caused by the self-selection is in some ways eliminated. Alternatively, other inferential approaches are applied.

3. Methods to Correct Self-selection

A sample is biased (not representative of the population it is related to) when the sample distribution of the variables describing the structure differs from the same variables in the population. The bias of these variables generally implies the bias of the variables studied. It is likely for the self-selectioned sample to be biased because the self-selection process is not a

random one. This means that it does not indifferently involve every subject of the population. Conversely, it is thought to be quite typical of those people who have certain characteristics. For example, it may be thought that the gender and the age of a customer might influence the level of satisfaction of a service. If in a self-selected sample women and youths are mainly observed, when compared to the distribution of the population, the evaluation of the satisfaction of that service will turn out to be altered.

The methods that will be shown here have a specific end: to differently correct the bias caused by self-selection. The first method suggests to evaluate the satisfaction of the non-answering people in probabilistic terms. The second considers two equations tied together by a latent factor that allows the missing data associated to the non-answering subjects to be correctly estimated. Finally, the third method proposes a linear model which is supposed to be able to show the relations between the variable examined and a group of variables thought to have generated the observed data. If the model is easily adaptable to the observed data, it will be possible to predict the values of the studied variable as far as those yet to be self-selected are concerned.

3.1. The Propensity Score Matching

The idea behind this method was born in the Seventies and was the work of Rubin [20]. Though, it has been tested in the Eighties by Rosenbaum and Rubin [18] in an experimental context to evaluate the effect of a health treatment. For this purpose we want to compare the outcomes of the interested variable observed on a group of individuals subjected to a treatment to the outcomes that should have been observed was the treatment not applied. Obviously, it is not possible to make such a comparison on the same group. It is then necessary to choose a control group and compare the two outcomes. Marked with T a dichotomous variable that takes a $t=1$ value (if the person receives the treatment) or a $t=0$ value (if the person doesn't receive the treatment), and marked with an x a vector of covariates, called pre-intervention variables, we call *propensity score* (*ps*) the probability of every person to receive or not receive the treatment conditioned by the vector of the covariates x. This probability is: $p(x)=Pr(t=1\,|\,x)$; that is, $1-p(x)=Pr(t=0\,|\,x)$ and is estimated with the multivariate logistic regression, with pre-intervention variables as regressors. *ps* [1] is the comparison term between the outcomes of the people from the two groups and allows us to do a "matching"[2] between them; thus, it will be possible to associate to a person of the control group the outcome observed on the person of the group that has been treated and has got the same vector x, that is to say the same value of *ps*. This method has been applied in different contexts, even recently, to correct the bias due to a self-selection error in the web-surveys [1, 2, 13]. In the context of the CSSs, the group undergoing a treatment is represented by those who, by self-selection, have filled in the questionnaire, whilst the control group is formed by those "not self-selected" individuals. Once calculated the propensity scores as far as the subjects of the two groups are concerned, the same standard of satisfaction of a self-selected person who has filled in the questionnaire and has got the same *ps* will be attributed to an individual of the control group who has not filled in the questionnaire. The propensity score matching method allows us to correct the self-selection error and to estimate the standard of satisfaction of those who have not filled in the questionnaire. It is based on the assumption that the subjects join the research independently and on the assumption that self-selection does not depend on the target variable (Conditional Independence Assumption). This implies that the pre-intervention variables, on which the

[1] The advantages connected to the use of *ps* is twofold: on one hand, a multidimensional problem is reduced to a one-dimensional one. On the other hand, the risk of finding more than one person with the same *ps* is eluded, since the propensity score is a monotone function of the discriminant score.

[2] The matching procedure occurs utilizing one of these procedures: *Nearest Neighbor Matching*, *Caliper Matching*, *Mahalanobis Metric Matching*, *Stratification Matching*, *Difference-in-Differences Matching* [18].

matching is influenced, must not affect the self-selection [18]. In the context of CSSs, pre-intervention variables could be different according to the type of service received.

3.2. The Heckman Two-step Procedure

Heckman [10] proposed a two-step procedure which allows to estimate the values of the variable of interest for not self-selected individuals.

Heckman's idea is based on a model made by two equations related to each other: the *substantial equation* and the *selection equation*.

The *substantial equation* for individual i (for $i = 1,..., N$) is:

$$Y_{1i} = X_{1i}\beta_1 + U_{1i}, \tag{1}$$

where Y_{1i} represents the continuous variable of interest, that in this case represents the satisfaction level. Supposed that one seeks to estimate equation (1) but data are missing on Y_1 for $N - n$ observations (the number of not self-selected individuals), we define the selection equation, that for an individual i is the following:

$$W_i = X_{2i}\beta_2 + U_{2i}, \tag{2}$$

where W_i is an unobserved latent random variable such that $W_i \geq 0$ corresponds to self-selected individuals, while $W_i < 0$ corresponds to not-self selected individuals. Equation (2) is the expression of a probit model, which can be alternatively written in the following form:

$$Prob(Y_{2i} = 1 \mid X_{2i}) = F(X_{2i}b_2), \tag{3}$$

where Y_{2i} is a dichotomous random variable taking value 1 for self-selected and 0 for not-self selected individuals.[3]

For both the previous equations we make the following assumptions: X_{1i} is the i-th vector $1 \times (K+1)$ of variables known for all N subjects, X_{2i} is then i-th vector $1 \times K$ of variables known for all N subjects, β_1 is a $(K+1) \times 1$ vector of parameters, β_2 is a $K \times 1$ vector of parameters, U_{1i} and U_{2i} are the vectors of residuals. $E(U_{ji}) = 0$, $E(U_{ji}U_{j'i'}) = \sigma_{jj'}$ for $i = i'$ and $E(U_{ji}U_{j'i'}) = 0$ for $i \neq i'$, where $i = 1,..., N$ and $j = 1,2$. The joint density of U_{1i} and U_{2i} is denoted by $h(U_{1i}, U_{2i})$ and we assume that it has a bivariate normal density.

Hence, a relationship between equations (1) and (2) exists. This relationship is reflected in a non-zero correlation between the error terms of the equations for the same subject. If such correlation is present, we cannot get robust estimates of the substantial equation without taking the selection process into account [21]. Heckman's method allows the introduction of a *control factor*, which links equations (1) and (2) and expresses the underlying characteristics of customers in their choice of being self-selected [10].[4] In other words the method allows us

---

[3] If the latent variable $W_i \geq 0$ then $Y_{2i} = 1$ (corresponding to self-selected individuals), while if $W_i < 0$ then $Y_{2i} = 0$ (corresponding to not self-selected individuals).

[4] Summarizing the Heckman procedure, it basically takes the following steps.

- Estimate the parameters β_2 of the *selection equation* (2) using a probit analysis.
- The residuals of the *selection equation* U_{2i} are used to calculate λ_i, the *selection biased control factor*, computed as the inverse of Mill's ratio [12], the indicator of the hidden motivations of self-selection.

to estimate the level of satisfaction of the not self-selected customers with (1), that takes into account, through (2), the self-selection factor and the existence of a correlation between the residuals of the two equations. Though, the Heckman method assumes that the variable observed is a continuous one: this assumption is not verified in a customer satisfaction context because the level of satisfaction is taken with a categorical variable. Thus, to respect this assumption a transformation of the variable considered is needed. For this reason it is possible to use the Nonlinear Principal Component Analysis [8] or the satisfaction coefficients of the Rasch Analysis [4, 5, 16].

3.3. The Hierarchical Bayesian Approach

With this approach the value of the considered variable Y on the i-th unit of the population is not fixed but a realized value of a random variable $Y_i (i = 1,..., N)$.

Bayesian methods have become widespread in customer satisfaction surveys. This approach aims to get estimates for the variable of interest when the survey is unreliable because of data unavailability. This technique, like the previous ones, calculates estimates using information that are supposed to be correlated to the variable of interest. As it is well know in the literature, this technique was traditionally developed and employed in the small area estimation field, as explained in [17]. Indeed, this is a very powerful tool when dealing with lack of data [22]. Here we employ it successfully in CSSs, where we estimate the parameters of interest for the individuals who did not answer to the questionnaire through the information provided by covariates, known for every subject.

Among Bayesian methods, we employed the hierarchical Bayesian model as formulated by Fay and Herriot [6]. Suppose one is interested in estimating the characteristic Y_i for every subject, and that the auxiliary data are known for each individual i (for $i = 1,..., N$). Indeed, very often Y_i are not available for all the individuals, i.e. in sample surveys. The Fay-Herriot methodology consists of the *linking model* and the *sampling model*. The *linking model* is the following:

$$Y_i = X_i \beta + U_i, \tag{4}$$

where $U_i \sim N(0, \sigma^2)$, X_i is a $1 \times K$ vector of auxiliary variables and β is a $K \times 1$ vector of parameters. The linking model (4) is merely a mixed linear model where β are fixed effect coefficients, accounting for the effects of the auxiliary variables X_i, valid for the entire population, while U_i are random individual specific effects [19]. The *sampling model* is

$$\hat{Y}_i = Y_i + e_i, \tag{5}$$

where $\hat{Y}_i$ is the estimate of Y_i and $e_i | Y_i \sim N(0, \psi_i)$, being ψ_i the sampling variance, which is typically assumed to be known.

The hierarchical Bayesian model is described by the following equations

$$\hat{Y}_i | Y_i, \psi_i \sim N(Y_i, \psi_i). \tag{6}$$

$$Y_i \Big| \beta, \sigma^2 \sim N(X_i \beta, \sigma^2). \tag{7}$$

- The control factor λ_i is used as an additional covariate (the $(K+1)$-th covariate) in the *substantial equation* (1).

In absence of prior knowledge we opted for flat priors for $\beta_j \sim N(0, 10^{-6})$ (where $j = 1,..., K$) and $\sigma^2 \sim \Gamma^{-1}(0.001, 0.001)$, as suggested for example by Gamerman and Lopes [7]. In the hierarchical Bayesian model, inference on Y_i is straightforward and computationally feasible by using MCMC based standard methods and all the model uncertainty sources are simultaneously accounted for in the estimation process.

Once the parameters are estimated, for $i = 1,..., n$ we perform the hierarchical Bayesian model by means of MCMC methods and then, for $i = n+1,..., N$, we estimate Y_i through $\hat{Y}_i$ with the posterior predictive distribution.

4. Application

In this work we apply the Heckman, Bayesian and Propensity Score Matching (PSM) methods to a dataset of a Customer Satisfaction (CS) sampling survey of 4561 units. In particular, we use the data collected in 2004 through a F2FI survey by a real airline company, which here (to hide its identity) we call "Best Flight". The customers gave personal information as well as opinions about the service. The 4561 observed units, that were a sample for the Best Flight company, are considered as the population for our application. In the simulation (which is described in section 4.2), we choose a non random sample from this population. In order to implement the simulation, we assume that the data are of good quality.

Concerning personal information, we focus on the variables: *FREQCLASS*, *CLUB*, *NATION*, *AGE*, *GENDER*. Variable definitions are given in the Appendix.

Regarding the opinions about the service, we analyze 10 variables indicating the service dimensions and 4 variables indicating the overall customer satisfaction, for a total of 14 variables, each of them measured according to a Likert scale (1=extremely satisfied, 7=extremely dissatisfied). The 10 variables of the service dimensions are: *BOOKING*, *CHECKIN*, *TRANSFER*, *LOUNGE*, *DEPARTURE*, *CABIN*, *MEAL*, *ENTERTAINMENT*, *DUTYFREE*, *CREW*. The 4 variables of the overall customer satisfaction are: *EXPERIENCE*, *VALUE*, *RECOMMENDATION*, *REPEAT*. See the Appendix for the definitions of the variables.

We remark that the analysis of CSSs is different from other surveys for the types of variables employed, that are mainly categorical, often measured on a Likert scale. This has to be accounted for in the implementation of the statistical methodologies to the data.

4.1. Nonlinear PCA

In order to apply the Heckman method, we need to determine the dependent variable Y_i, indicating the overall customer satisfaction for each individual, thus summarizing the 4 variables of the overall satisfaction. In fact, one of the requirements of Heckman model is that the target variable Y_{1i} in the substantial equation (1) has to be a quantitative random variable. Since generally in CSSs the variables are categorical, our first step is the transformation of these variables into one quantitative indicator. In the Bayesian approach, instead, a quantitative dependent variable is not strictly required, since we could formulate a different hierarchical model even for categorical variables. The Propensity Score Matching can be implemented with categorical target variables and continuous ones as well. However, in order to make comparisons, we use a quantitative response variable for all the three models.

To compute the quantitative dependent variable Y, we use the Nonlinear Principal Component Analysis (Nonlinear PCA), belonging to the class of Nonlinear Multivariate

Analysis [8]. This technique is an exploratory analysis, suitable for reducing the dimensionality of ordinal variables. In our case, Nonlinear PCA computes optimal quantifications preserving categories order within each variable. We are interested just in one dimension, corresponding to the overall satisfaction variable Y. Nonlinear PCA computes the minimum of a loss function $l(S; q_1,...,q_m)$ under various normalization conditions, as described in [15]. The solution of this minimization gives us the Y_i overall satisfaction scores for each individual. Computed object scores are finally used as the dependent variable Y_i, where the lower the value of Y_i, the higher the satisfaction of the interviewed customer.

4.2. Results

Nonlinear PCA allow us to summarize the information brought by the 4 overall satisfaction qualitative variables into the quantitative variable Y. We employ this variable as the quantitative response in the Heckman, Bayesian and PSM methods, where the 10 service dimension variables play the role of independent variables in the first two methods, while they represent the pre-intervention variables in the third method.

In order to obtain one customer satisfaction index Y through the Nonlinear PCA technique, some conditions are to be verified. In particular, the first eigenvalue of the Nonlinear PCA solution has to be much higher than the others and Cronbach's α, a measure of the reliability of the scale lying between 0 and 1, has to be as close as possible to 1. Moreover, the component loadings should have all the same sign for each variable.

Table 3 shows that the first eigenvalue is much higher than the other (3.028), thus the one-dimension solution fits well the data. Cronbach's α is 0.893, very close to 1, denoting that the Nonlinear PCA is an appropriate methodology.

Table 3. Summary of the Nonlinear PCA model.

Dimensions	**Cronbach's α**	**Eigenvalue**
1	0.893	3.028
2	–1.041	0.562
Total	0.962	3.590

Table 4. Component loadings for two dimensions.

	Component Loadings Dimension 1	**Dimension 2**
EXPERIENCE	0.880	0.200
VALUE	0.812	0.510
RECOMMENDATION	0.926	–0.201
REPEAT	0.858	–0.470

Table 4 lists the component loadings of Nonlinear PCA. For each variable, the corresponding value for the first dimension is always positive, indicating a good fit of the methodology to the data. These component loadings represent the weight of the variable in the determination of the overall customer satisfaction indicator. The higher is the value of the component loading, the higher is the weight of the corresponding variable in the determination of the indicator. In this case, the values are all high and close to one, denoting that all the four considered variables are important and well represented in determining the

overall satisfaction dimension. Moreover, it is clear that the factor loadings of the second dimension are low and show different signs. This result confirms that the Nonlinear PCA one dimension solution is suitable for the data. The histogram of the target variable Y calculated through the Nonlinear PCA is displayed in Figure 1.

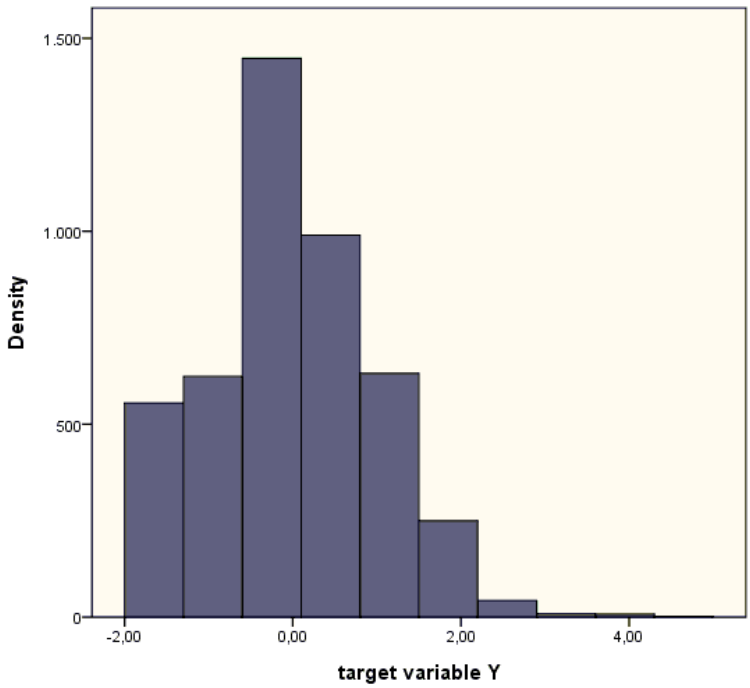

Figure 1. Histogram of the target variable Y obtained through the Nonlinear PCA.

The five personal information variables (*FREQCLASS*, *CLUB*, *NATION*, *AGE* and *GENDER*) are used in all three approaches as self-selection variables. First of all, we form different groups of individuals based on the categories of personal information variables. For example, the variable *GENDER* generates two sets of subjects: females (38.1%) and males (61.9%). We repeat this process for each personal information variable. We run some simulations where each time one category of one personal information variable corresponds to missing data for our target variable Y. More clearly, subjects corresponding to one category are considered as the group of people who do not choose to fill in the questionnaire (i.e. a simulation consists in an application of the models to a dataset where everyone but the first class passengers were supposed to be self-selected). Different categories of personal information variables correspond to different self-selection (and not self-selection) percentages, as explained in Table 5.

Therefore, in each simulation we estimate the overall satisfaction variable for missing data $\hat{Y}_i$ (not self-selected individuals) with the three methodologies presented. For example, we suppose that the first class passengers did not fill in the questionnaire (not self-selected), then we estimate the overall satisfaction of these subjects through the Heckman, Bayesian and Propensity Score Matching models, as explained in section 3.

Table 5. Personal Information variables used for self-selection.

Variables	Self-selection %	Not self-selection %	Classes (not self-selected)
FREQCLASS	96.6	3.4	First
	76.55	23.45	Concorde, First, Clubworld
	41.83	58.17	Concorde, First, Clubworld, WTPlus, Worldtraveller
CLUB	91.5	8.5	Gold
NATION	91.5	8.5	French, German, Japanese
	47.9	52.1	British
AGE	69	31	12-34
	41.1	58.9	12-44
	19.1	80.9	12-54
GENDER	61.9	38.1	Female

Once employed the three methodologies, the outputs give us the target variable $\hat{Y}_i$ estimates (for $i = 1,..., N-n$), expressing the global customer satisfaction. We make a comparison between the estimated results $\hat{Y}_i$ and the observed results Y_i aiming to calculate how close is our estimator from the real value of not self-selected individuals (i.e. first class passengers). To reach this goal, we calculate the estimated Mean Squared Error ($\widehat{MSE}$), through the following formula:

$$\widehat{MSE} = \frac{1}{N-n} \sum_{i=1}^{N-n} (\hat{Y}_i - Y_i)^2, \tag{8}$$

and we calculate the *bias* of $\hat{Y}_i$ as a predictor of Y_i as:

$$bias = \left[E(\hat{Y}_i) - \bar{Y} \right]. \tag{9}$$

The results of the simulations are reported in Table 6. Here we see the different percentages of self-selection (people who answered to the CS questionnaire), as illustrated in Table 5 and the results in terms of $\widehat{MSE}$ and corresponding bias value for the Heckman, the Bayesian and the PSM approaches.

As we can see from Table 6, the $\widehat{MSEs}$ computed through the Bayesian method are very close to the one calculated according to the Heckman procedure, while the PSM $\widehat{MSEs}$ are higher. However, if we focus on the bias, we note that the PSM performs better than the other two methods in the majority of cases.

Focusing on the comparison of the Heckman-Bayesian models, we see that the Heckman approach works well only for self-selection percentages higher than 50%. The Bayesian approach instead works well even with a lower percentage of self-selection, since it produces good estimates even with small samples. However, the propensity score matching outperforms the Bayesian methods in many cases, even where the sample size dimension is small.

Table 6. $\widehat{MSE}$ and bias for the three methods.

Variables selection	Self-selection %	Not self-selection %	Heckman $\widehat{MSE}$	Heckman *bias*	Bayesian $\widehat{MSE}$	Bayesian *bias*	PSM $\widehat{MSE}$	PSM *bias*
FREQCLASS	96.6	3.4	0.3198	0.0007	0.3271	0.0199	1.1862	0.0188
	76.55	23.45	0.3495	0.0533	0.3479	0.0938	1.5653	0.0369
	41.83	58.17	0.4039	0.1303	0.4307	0.1021	1.8219	0.2451
CLUB	91.5	8.5	0.4001	0.0082	0.3991	0.0211	1.4304	0.0568
NATION	91.5	8.5	0.3895	0.0003	0.392	0.0521	1.0687	0.0158
	47.9	52.1	0.4279	0.1246	0.4296	0.0933	1.6830	0.0527
AGE	69	31	0.4300	0.0428	0.4303	0.1001	1.4446	0.0279
	41.1	58.9	0.4359	0.1466	0.4366	0.1399	1.1419	0.0064
	19.1	80.9	0.4261	0.5790	0.4317	0.4212	0.8431	0.0031
GENDER	61.9	38.1	0.4124	0.0399	0.4150	0.1192	0.9418	0.0006

5. Concluding Remarks

The literature referred to statistical surveys is widespread, as from the research methodology, as from the sample theory point of view. However, there is not a lot of information about CS surveys. This paper proposes indeed a novel analysis about CS surveys, considering the population involved (the whole population or a part of it), related to the type of company that supplies the service, to the knowledge of the company about its customers and to the data collection methods, that are associated to non-sampling errors. A type of error which is very common in statistical surveys is the unit non-response, when the interviewee refuses the interview. However, we need to distinguish between non-responses in sample surveys and non-responses in census surveys. The authors associate the non-response error to the sample survey, when the subject to be interviewed is selected, and the self-selection error to the census survey, where there is no selection of the interviewees. Therefore, the not self-selection in census surveys corresponds to the non-response in sampling surveys. In the literature the self-selection error has been less studied than the non-response error. We give an overview of the main methods able to correct self-selection bias (the propensity score matching, the Heckman two step approach and the hierachical Bayesian model). However, these methods have not been applied to the field of Customer Satisfaction Survey yet. For this reason we implement the methodologies to a dataset of CSSs, in order to illustrate the potential of these techniques in this context. The target variable is the global satisfaction level of the considered customers. It is derived by four categorical variables of overall satisfaction through the Nonlinear Principal Component Analysis technique. We run some simulations and we compare the estimated values of the target variable calculated through the three methodologies with the observed ones. The results comparison shows that the PSM gives smaller bias than the other two methods in the majority of simulations. The Heckman methodology performs better with big samples, while the Bayesian model performs better with small samples. Therefore, when the percentage of self-selected subjects exceeds 50% we suggest the use of the Heckman approach, since it is computationally more efficient than the PSM in terms of simulation time. However, when the percentage of self-selected subjects is lower than 50%, both the PSM and the hierarchical Bayesian methods are somewhat time consuming, but in our case the PSM gives the best results.

When the number of self-selected subjects is small the estimate of the level of customer satisfaction is very inaccurate. However, even if the number of self-selected subjects is high,

the estimate could still be inaccurate, since the higher the number of observations the larger are non sampling errors. Therefore, the employment of techniques aiming to eliminate non sampling errors is always suggested to obtain reliable estimates, although this implies a great effort from the firm.

Acknowledgements

We are grateful to the anonymous referees, for their insightful suggestions that significantly improved the paper. The research for this paper was supported by PRIN 2007.

References

1. Biffignandi, S. and Pratesi, M. (2003). Potentiality of propensity score matching in inference from web-surveys: a simulation study. *Working Paper, Dipartimento di Statistica, Informatica a Applicazioni*.
2. Biffignandi, S., Pratesi, M. and Toninelli, D. (2003). Potentiality of propensity scores methods in weighting for Web surveys: a simulation study based on a statistical register. *Proceedings of the ISI Conference*, Berlin.
3. Couper, M. P. (2000). Web survey: a review of issues and approaches. *Public Opinion Quarterly,* 64, 464-494.
4. De Battisti F., Nicolini, G. and Salini, S. (2005). The rasch model to measure service quality. *The ICFAI Journal of Services Marketing*, 3, 58-80.
5. De Battisti F., Nicolini, G. and Salini, S. (2010). The rasch model in customer satisfaction survey data. *Quality Technology and Quantitative Management*, 7, 15-34.
6. Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
7. Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall.
8. Gifi, A. (1990). *Nonlinear Multivariate Analysis*. John Wiley and Sons, New York.
9. Groves, R. M. (1989). *Survey Errors and Survey Costs*. John Wiley and Sons, New York.
10. Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-161.
11. Hothum, C. and Spintig, S. (1998). *Customer Satisfaction Research*, *in the Hesomar Handbook of Market and Opinion Research*, 4th edition. (Edited by McDonald and Vangelder), Esomar, 853-890.
12. Johnson, N. and Kotz, S. (1972). *Distribution in Statistics: Continuous Multivariate Distributions*. John Wiley and Sons, New York.
13. Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22, 329-349.
14. Lessler, J. T. and Kalsbeek, W. D. (1992). *Nonsampling Errors in Surveys*. John Wiley and Sons, New York.
15. Michailidis, G. and De Leeuw, J. (1998). The gifi system of descriptive multivariate analysis. *Statistical Science*, 13, 307-336.
16. Nicolini, G. and De Battisti, F. (2008). *Methods for Summarizing the Rasch Model Coefficients, in Metodi, Modelli e Tecnologie dell'Informazione a Supporto delle Decisioni*. (Edited by D'Ambra, Rostirolla and Squillante), 284-290.

17. Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley and Sons, New York.

18. Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

19. Rossi, P. E., Allenby, G. M. and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. John Wiley and Sons, New York.

20. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701.

21. Smits, J. (2003). *Estimating the Heckman two-step procedure to control for selection bias with SPSS*. Available at http://home.planet.nl/~smit9354/selbias/Heckman-SPSS.doc.

22. Trevisani, M. and Torelli, N. (2006). Comparing Hierarchical Bayesian Models for Small Area Estimation. *In Metodi statistici per l'integrazione di basi di dati da fonti diverse*, (Edited by Franco Angeli), 17-36.

Appendix

The definitions of personal information variables are the following:

FREQCLASS In which cabin do you most often travel when flying with Best Flight?
(Concorde, First, Club World, World Traveller Plus, World Traveller (economy), Club Europe, Euro Traveller (economy), Domestic);

CLUB If you are a member of the Best Flight Executive Club or any other frequent flyer scheme, please indicate which card you hold.
(Executive Club Gold, Executive Club Silver, Executive Club Blue);

NATION What is your nationality?
(American, British, French, German, Japanese, Other);

AGE What is your age?
(12-15, 16-21, 22-25, 26-34, 35-44, 45-54, 55-64, 65+);

GENDER Are you female or male?
(male, female).

The 10 variables of the service dimensions are the following:

BOOKING Overall, how satisfied were you with the ticket booking process?

CHECKIN Overall, how satisfied were you with the check-in process?

TEANSFER Overall, how satisfied were you with the ease of transferring flights?

LOUNGE Overall, how satisfied were you with the Best Flight lounge?

DEPARTUER Overall, how satisfied were you with the departure process?

CABIN Overall, how satisfied were you with the cabin environment?

MEAL Overall, how satisfied were you with the meal/refreshments service?

ENTERTAINMENT Overall, how satisfied were you with the in-flight entertainment?

DUTYFREE Overall, how satisfied were you with the choice of goods for sale?

CREW Overall, how satisfied were you with the cabin crew?

The 4 variables of the overall customer satisfaction are the following:

EXPERIENCE Overall, how satisfied were you with your experience of Best Flight today,

VALUE How satisfied are you with the value for money of this Best Flight flight?

RECOMMENDATION On the basis of your experience with Best Flight today, how likely are you to recommend Best Flight to a friend or colleague?

REPEAT On the basis of your experience with Best Flight today, how likely will you be to travel with Best Flight again?

Authors' Biographies:

Giovanna Nicolini has been Associate Professor of Sample Theory at the University of Milano Bicocca, now she is full Professor of Statistics at University of Milan. Her main research areas are the following: probabilistic models, robustness on power statistical tests, tests with balanced errors, permutation tests, sampling techniques, customer satisfaction measures, assessment of the quality of public services.

Luciana Dalla Valle earned a Ph.D. in Statistics from the University of Milano-Bicocca in 2007. Currently, she holds a Post-Doctoral position at the Department of Economics, Business and Statistics of the University of Milan. She is interested in Bayesian statistics, Markov Chain Monte Carlo methods, statistical models for financial risks, copula modeling for financial applications, statistical models for internationalisation and sampling techniques.