Graphical Tools for Analyzing Nonresponse and Nonresponse Adjustments in Survey Data

Ismael Flores Cervantes[1]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract
Advances in computing processing and the availability of open source statistical platforms have helped develop new graphical tools to analyze and explore data patterns. As a proof of concept, a graphical toolbox for nonresponse analysis in survey data was developed. The toolbox can be used for exploring nonresponse patterns in samples and for evaluating the effectiveness of nonresponse adjustments. The toolbox includes descriptive displays (mosaic plots, conditional density plots, and density plots) and analytical displays (*heat map fit* plots and *b-plots* for nonresponse and adjustment factors). The use of the toolbox is demonstrated by conducting several graphical analyses of nonresponse patterns in two published papers. Since this is a new approach for studying nonresponse, general guidelines need to be developed for this type of analysis.

Key Words: Nonresponse, graphical analysis, toolbox, b-plots, survey data

1. Motivation and Proof of Concept

Graphical Methods for Data Analysis (Chambers, Cleveland, Kleiner, & Tukey, 1983) was one of the first books on data visualization and graphical methods for data exploration published during the personal computer revolution more than 30 years ago. Since then and propelled by the continued development of high performance computers, graphical methods have become very popular as a way to analyze and discover patterns in data.

In contrast to the prevalence of this type of analysis in many statistical fields, graphical methods are not as common in survey methodology. Their current use is limited to the presentation of survey results in the form of *infographics* in recent years. One possible reason is that there is no easy-to-use software that offers these methods and handles the characteristics of survey data at the same time. Although there are some scattered functions in some software packages that can be used for this type of analysis; they are rarely easy to use and require programming knowledge.

In this paper, the proof of concept of developing graphical tools for survey data is evaluated; in particular, for the analysis of nonresponse. As a proof of concept, the feasibility and benefits of graphical analyses for exploring nonresponse and evaluating nonresponse weighting adjustments are examined. The tool created in this research is called *NR-Toolbox* (Nonresponse toolbox). Although it is a proof of concept, the toolbox is very complete. Different analyses on nonresponse and nonresponse adjustments on two published articles show how to do graphical analysis and how the toolbox can be used in practice.

There are several reasons for the development of graphical tools for the study of nonresponse. First, the most important reason is the need to produce timely nonresponse adjusted weights. Any method that expedites the creation of analysis weights is worth considering. The second reason is the need to improve the quality of the nonresponse adjustments. There are no commonly accepted methods for comparing one type of nonresponse adjustment to another. In the literature, nonresponse adjustments are compared using Monte Carlo simulations which cannot be implemented in practice. Graphical analyses do not require simulation because it can extract statistical signals from noise from a single dataset. Graphical analysis takes advantage of the innate human skill for summarizing large amounts of information and for detecting patterns when the information is presented visually. A third reason, but not less important, is the pedagogical value of a tool like this for teaching nonresponse and nonresponse adjustments in a way that is easier to understand. Finally, these tools offer a different way to look at nonresponse, opening the possibility of new ways to address this problem.

The final goal of this research is the development of a complete toolbox similar to the *R* package *VIMGUI: Visualization and Imputation of Missing Values* (Templ, Alfons, Alexander, & Bernd, 2013). This package presents new ways to visualize missing and imputed values. With the help of a graphical user interface (GUI), users can identify patterns of missing data and determine the appropriate method to impute missing values. In a similar way and with a GUI, the *NR-Toolbox* would allow users to explore nonresponse patterns, to determine the appropriate method to adjust for nonresponse, and to evaluate the results of the adjustment with minimal programming. A more ambitious goal is to include the different methods for adjusting weights as part of the package.

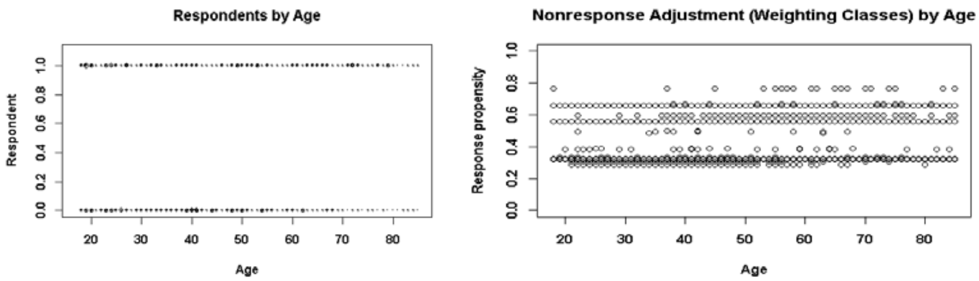
As in the package *VIMGI*, the initial version of the *NR-Toolbox* is based on *R*, an open source environment and programming language for statistical computing and graphics (R Development Core Team, 2013). The advantage of *R* is its availability under the GNU General Public License. This type of license allows users to freely run, copy, distribute, examine, and eventually modify the code. Since one the strengths of *R* is professional looking graphics, this programming language is well suited for the development of new ways to display data.

2. Nonresponse Analysis Graphical Toolbox

The *NR-Toolbox* is expected meet four objectives of any statistical graphical method (Jacoby, 1997): (1) exploring data, (2) displaying of patterns in the data, (3) checking assumptions of statistical models, and (4) communicating results of the analysis. The first obstacle in dealing with nonresponse is the way it is plotted. Response is a binary variable and any simple plot is not informative as shown in Figure 1. Figure 1a shows the response status of adults in a sample by age. Since respondents are indicated as 1's and nonrespondents as 0's, the graph shows two parallel lines where most dots overlap. If the computed response propensities are plotted as shown in Figure 1b, the plot consists of parallel lines that are also uninformative. In this example, the response propensities were computed as the inverse of the weighting class nonresponse adjustment factor, which is constant for records within the same class. To solve this problem, most of the graphical functions in the toolbox use smoothing non-parametric techniques.[1] These techniques are a standard way to visualize signals in noisy data. The signal can be empirical response propensities, modeled response propensities, nonresponse adjustment factor, or the weights after and before the adjustments. Displays of smoothed data help users focus on the distribution of the data instead of focusing on single points. In the *NR-Toolbox*, the main method for smoothing the data is the locally weighted scatter plot smoothing

[1] Nonparametric methods do not use parameters of probability distributions.

(LOWESS). The method is modified to handle sampling weights. Other methods to smooth data implemented in other functions in the toolbox are splines and kernel density estimation. Examples of the smoothed representations of the empirical response propensities, modeled response propensities, and other variables related to nonresponse are found in the following sections.



(a) (b)

Figure 1: Graphical representation of respondents and response propensities in a telephone survey

3. Displays in the NR-Toolbox

The current displays in the toolbox can be classified as descriptive displays (conditional density plot, spine plot, and mosaic plots) and analytical displays (*heat map fit* plot and *b-plots for response* and *b-plots for adjustment factors*).

3.1 Descriptive Displays

3.1.1 Simple Descriptive Displays

Simple descriptive plots or displays include those plots that present the distribution of the observations. For the toolbox, the focus is on plots that describe the distribution of both the population and sample by response status at the same time (i.e., respondents and nonrespondents). Simple plots available in the *R* core include histograms, cumulative plots, box plots, and density plots. However, many of these plots either lack options for handling sampling weights or handle other type of weights such as weights normalized to 1 (as in the function *density*). Contributed packages such as the package *survey* (Lumley, 2012) has graphical functions that handle weights but these are not as straightforward to use when combinations of weighted and unweighted plots are needed. In contrast, other contributed packages such as the package *weights* (Pasek, Tahk, Culter, & Schwemmle, 2014) and *ENmisc* (Neuwirth, 2013) have functions for weighted histograms and weighted box plots that are easier to use and can be easily integrated into the toolbox as dependencies. However, having many dependencies on many packages can be an issue if those packages are not maintained in future versions of *R.*

Figure 2 shows an example of a plot created combining simple displays. In this example, the plot combines the estimated population distribution by age and boxplots by respondent and nonrespondents in a telephone survey. Although the functions in the toolbox produce both the weighted and unweighted versions of the plots by default, Figure 2 shows only the weighted plot. The graph is created using sampling weights.

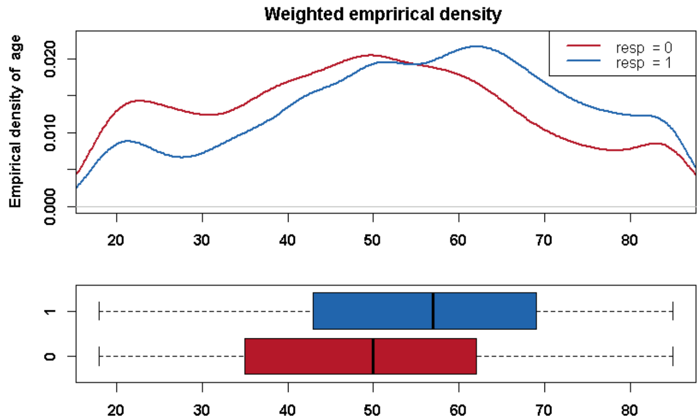
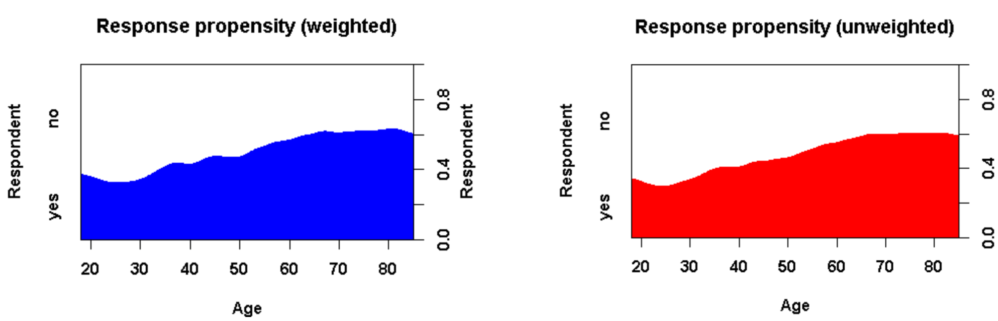


Figure 2: Empirical age population density and box-plots by response status

The utility of plot is self-evident after a quick inspection. The plot shows that respondents and nonrespondents have a different distribution and hence the need to adjust for nonresponse. The plot also shows the differences of response by age so age should be used when computing the nonresponse adjustments. Simple displays can show the effect of the nonresponse adjustments by displaying the similar plots created using the nonresponse adjusted weights.

3.1.2 Conditional Density Plots

Conditional density plots display the population distribution of a categorical variable by different values of a continuous (or close to a continuous) variable (Hofmann & Theus, 2005). This type of plot is ideal for studying the relationship and interaction between a categorical variable (i.e., response status) and a continuous variable. Without this plot, data is examined in a tabular form. The table would be produced by creating *n* categories of the continuous variable. Then a 2 x *n* table is produced for presenting numerically the distribution of the categorical variable (i.e., response status). There is no need to create these categories in conditional density plots. Figure 3 show the weighted and unweighted conditional plots of the response propensities by age in a telephone survey. The plot shows that younger adults tend to respond at a lower rate than older adults, and most older adults have a flat response pattern.



(a) (b)

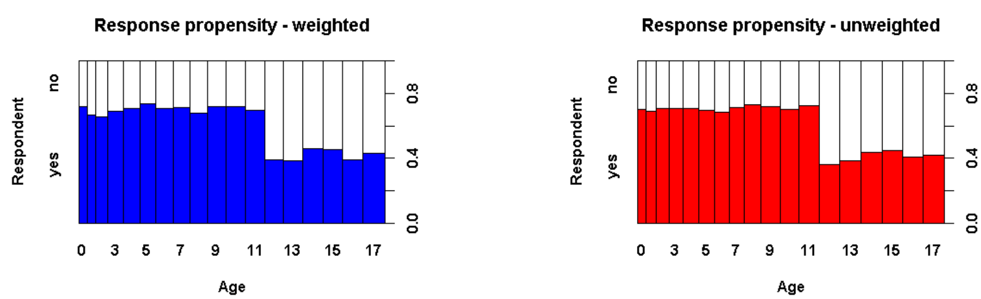
Figure 3: Weighted and unweighted conditional density plots by age

The conditional density plots function in the *R* core (*cdplot*) does not handle sampling weights. A modified version in the *NR-Toolbox* overcomes this limitation. However, additional research is needed for computing the optimal values of the parameters for the estimated density function. Mathematically, the density is approximated by a combination of Gaussian kernel densities over the range of the data. The default value of the parameter for the bandwidth *bw*, which affects how smoothed the lines are drawn, may not be the best for weighted data.

The conditional density plots can also examine more than two response statuses such as response statuses for eligible respondents, eligible nonrespondents, ineligible, and unknown eligibility which are common in telephone surveys.

3.1.3 Sipinographs

Another useful plot for nonresponse analysis is the spine plot or spinogram. Spine plots are an extension of stacked bar plots (a type of histogram) with varying bar widths (Hummel, 1996). The width of the bar is a function of the number of observations represented by the bar. In weighted data, the width of the bar represents the estimate of the total population. Figure 3 shows the spine plot for the empirical propensities for persons 17 years old or younger[2] in a telephone survey. The plots show that the response for children (ages 0 to 11 years old) is very different than the response for adolescents. Furthermore, the response seems to be flat within these groups.



(a) (b)

Figure 4: Weighted and weighted Spine plot of response propensity by age.

The current *R* core function *spineplot* does not handle sampling weights. A modified version of this function in the toolbox overcomes this limitation. Although the spine plot looks similar to the cumulative density plot, the mathematical foundations are very different. The spine plot does not estimate the conditional density of the population. The spine plots are just simple histograms where the continuous variables are categorized before the bars are drawn.

A further improvement is drawing a band with different colors on the top of stacked bars to indicate the standard errors of the estimates for each bar. This can be useful when there is a need to determine if the response propensities are statistically different across the bars.

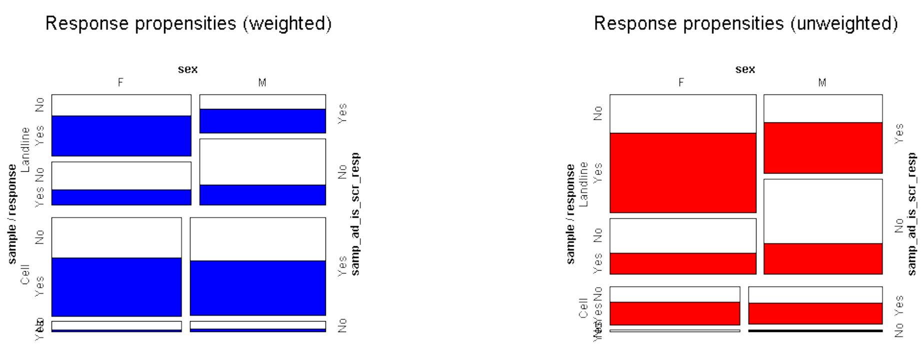
3.1.4 Mosaic Plots

Sometimes there is a need to examine the interactions between response and more than one categorical variable at the same time. Mosaic plots are ideal to examine these

[2] Children ages 0 to 11 are not interviewed directly. The interviewed person is the most knowable adult of the sampled child. On the other hand, adolescents age 12 to 17 are interviewed after permission is obtained from the parent or guardian.

interactions (Emerson, 1998; Friendly, 1994). A mosaic plot is a graphical representation of frequency tables (Meyer, Zeileis, & Hornik, 2006). The plot consists of tiles with areas proportional to the total number of records in the entry of the table. In the cases of a weighted frequency table, the areas correspond to the estimate of the population totals. The function *mosaic* in the package *vcd* is very complete and automatically handles the content of the tiles, split direction for each dimension, graphical parameters of the tiles' content, spacing between tiles, and labeling (Meyer, Zeileis, & Hornik, 2013). The function *mosaic* handles weighted counts as long they can be represented as a table object in *R*.

Figure 5 shows the weighted and unweighted mosaic plots of the extended interview response propensities of a telephone survey. In this example, the plots are created using auxiliary variables for sample type (landline phone sample, cell phone sample), sex (male, female), and the indicator for whether the screener respondent is the sampled adult or not. In this survey, the cell sample is small as shown by the area in the unweighted tiles. However, it represents a population larger than the population in the landline sample as shown by the weighed titles. Another piece of information that can be gathered from the plot is that males and females in the landline sample respond at different rates. In contrast, this difference is not observed in the cell phone sample. The plot also shows that there is an interaction between sex and the cases where the sampled adult was the screener interview respondent. These adults tend to be females and they have a larger response propensity than adults who were not the screener interview respondent. Although the mosaic plots are mainly used for categorical variables, they can also be created using categorized continuous variables.



(a) (b)

Figure 5: Weighted an unweighted mosaic plots for nonresponse

Another type of plot not described here is the contour plot. This plot explores the interaction between response and two continuous variables. The plot is computed using generalized additive methods for binary data so the smoothed propensities represented by the contours are bounded between 0 and 1.

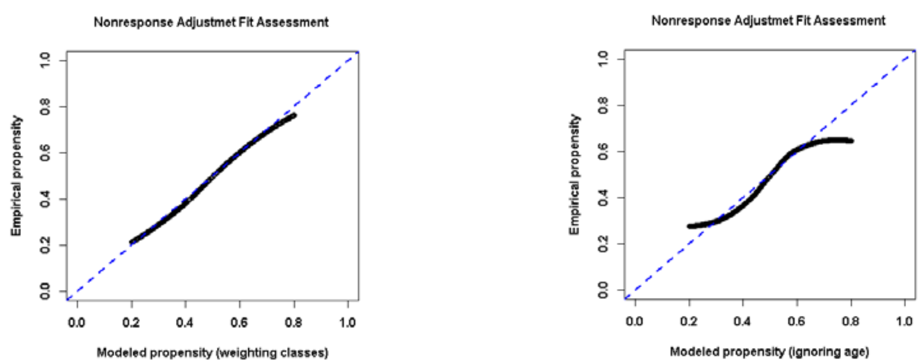
3.2 Analytical Displays

3.2.1 Heat map fit plot

The plots in previous sections are mainly descriptive and display the response patterns and their interactions with one or more auxiliary variables. The next plot is used to evaluate how well the response propensities are modeled. The *heat map fit* plot was proposed by Esarey and Pierce (2012a) as a diagnostics tool for logistic regression. The plot is used for assessing the fit quality and testing for misspecification in binary dependent variable models. Applying this plot in nonresponse analysis, the diagnostic contrasts the modeled propensities (computed using logistic regression) and the empirical propensities (computed using nonparametric methods). If the model is a good fit, the line formed by these propensities lie on top of the 45° angle line.

There is a package called *heatmapFit* that produces the heat map fit plot (Esarey & Pierce, 2012b). However, this package cannot be used for nonresponse modeling because it does not handle sampling weights. Furthermore, logistic regression is hard-coded in the source code of the package preventing its use for other types of adjustments such as weighting classes or calibration. The *NR-Toolbox* includes a function that produces this plot using weights. The input parameters of the modified function are the observed response indicators, the modeled propensities, and sampling weights. The empirical propensities are estimated using nonparametric LOWESS regression using the package *gam* (Hastie, 2013). At this stage of development of this function, the heat colors, which represent the level of significance for the goodness of fit of the model in the original plot, are not implemented. The Esarey and Pierce (2012b) plot uses bootstrap to generate colors of the levels of significance. Since sampling weights are used and these depend on the sample design, additional research is needed to determine if the same approach works for survey data.

Figure 6 shows two examples of these plots produced by the toolbox. In this figure, the goodness of fit of the nonresponse adjustment computed is assessed using weighting classes for the sample data in Figure 1b. In Figure 6a, all relevant variables were used to compute the weighting classes while the adjustment Figure 6b ignored age. Since the model is misspecified in the figure (e.g., there is an omitted variable), the modeled response propensities (i.e., computed as the inverse of the nonresponse adjustments) have an S-shaped curve.



(a) (b)

Figure 6: Heat fit plots of nonresponse adjustment using weighting classes

Notice that not only the *heat map fit* plot but all other plots used for diagnostics of goodness of fit for binary variables can be used to assess the fit of modeled response propensities, which is a binary dependent variable. Furthermore, the modified *heatmapFit* function with sampling weights can be used as a diagnostic plot for modeling any binary variables in survey data (i.e., variables such as indicators for whether the respondent is a smoker, obese, uninsured, etc.).

Another diagnostics plot in the toolbox for modeled response propensities not described here is the separation plot proposed by Greenhill, Ward, and Sacks (2011). One extension of the heat map fit plot is the evaluation of modeled propensities in domains. There are cases where the modeled propensities and the corresponding nonresponse adjustments work well for the whole population. Despite this, the model can have a bad fit for some domains. This type of checking is similar to diagnostics checks based on worm plots.

3.2.2. B-plots

A more complex plot is the *b-plot*, which displays in the same graph the propensity to respond and the distribution of the dependent variable in different vertical axes. The *b-plot* was initially created to visualize the interaction of the observed response propensities and the dependent variable in simulation studies in the literature. However, the *b-plot* can be also used for auxiliary variables. Figure 7a shows the basic form of the *b-plot*. The left axis (red) indicates the population density of the dependent variable and on the right axis (blue) indicates the empirical propensity to respond. In this example, the response propensity has a decreasing monotonic pattern and it is dependent of the variable *y* (e.g., larger values of *y* are less likely to respond that those with smaller *y* values). However, large values of *y* represent a small proportion of the population.

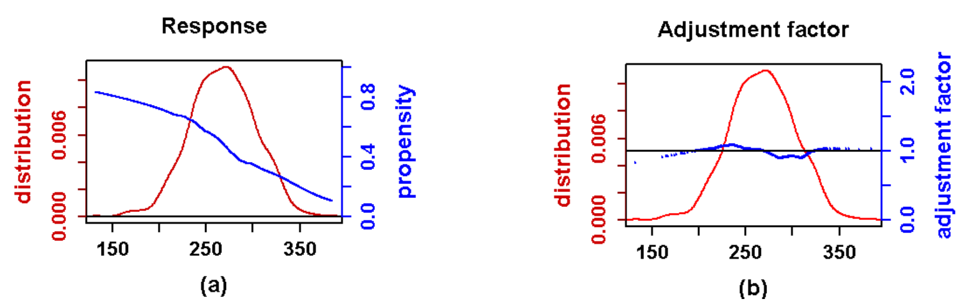


Figure 7: Response and adjustment factor *b-plots*

Based on this *b-plot*, strategies that model the propensity as a constant across the range of the variable *y* will not remove the bias. A better strategy is to model response as a continuous monotonic variable; however the final model selection depends on the available auxiliary variables. As a side note, if the nonresponse mechanism is *MCAR* (missing completely at random) the visual representation of the propensity is a horizontal line across the range of *y*. If nonresponse is *MCAR*, then a simple overall adjustment is enough to remove bias.

An extension of the *b-plot* can visually present the effect of the nonresponse adjustment factor for different values of the dependent or auxiliary variables. This effect is computed as the smoothed curve of the ratios τ_i's defined as $\tau_i = r_i/\widehat{\varphi}_i$, where r_i is the response indicator and $\widehat{\varphi}_i$, is the modeled propensity. If the modeled propensities are successful in removing the bias, the visual representation of this effect is a horizontal line crossing the vertical axis at 1. This plot is called *b-plot for adjustment factors* (in contrast to the *b-plot for response*). The theory behind this plot and the role of the ratios τ_i's will be presented in a future paper. Figure 7b is one example of this plot. This example displays the effect of adjustment factors shown in blue is close to horizontal line at 1; therefore, the adjustment is expected to remove the nonresponse bias. Note that different models for estimating propensities can achieve this flat pattern. However, in most cases, the model that produces propensities with smallest variability is expected to be more efficient (i.e., lowest variance). The variability of the adjustment factor is not observed in the current version of the *b-plot*. This would look like a cloud of dots hovering above and below the adjustment factor line. Future versions of this plot will have an option for displaying this cloud.

4. Examples of Nonresponse Graphical Analysis

This section describes two examples of nonresponse graphical analysis of two previously published articles using the *NR-Toolbox*. This type of analysis is new so there are no guidelines on how to proceed. However, these examples show what steps should be considered. In the first example, after analyzing the nonresponse pattern, alternative nonresponse adjustments are proposed based on the observations from the graphical analysis. In the second analysis, the graphical analysis is used to explain some unexpected results reported in the article.

4.1 Example 1: Graphical analysis of Kang & Schafer (2007)

Kang and Schafer (2007) compared alternative strategies for estimating a population mean with incomplete data. The model used to simulate response is $\varphi_i = logit^{-1}(-z_{1i} + 0.5z_{2i} - 0.25z_{3i} - 0.1z_{4i})$ where $\boldsymbol{z_i} = (z_{1i}, z_{2i}, z_{3i}, z_{4i})^t$ is multivariate normally independently distributed with parameters $N(\boldsymbol{0}, \boldsymbol{I})$. However, the auxiliary variables $\boldsymbol{z_i}$'s are not observable. Instead, the variables $\boldsymbol{x_i} = (x_{1i}, x_{2i}, x_{3i}, x_{4i})^t$ are available for respondents and nonrespondents. The observed auxiliary variables are defined as $\boldsymbol{x_i} = (exp(z_{1i}/2), z_{2i}/(1 + exp(z_{1i})) + 10, (z_{1i}z_{3i}/25 + 0.6)^3, (z_{2i}+z_{3i} + 20)^2)^t$. As mentioned in Kang and Shafer (2007), "*except by divine revelation,*" it is impossible for an analyst to formulate a correct model for φ_i based only on the observed auxiliary variables $\boldsymbol{x_i}$'s. Kang and Shafer (2007) indicate that analysts would use a modeled propensities from logistic regression on the $\boldsymbol{x_i}$'s despite being incorrect. The reason is that the regular diagnostics including those model checks suggested by Hinkley (1985) make the model looked trustworthy and give no reasons to change it.

The first part of the analysis, we evaluate if a graphical analysis can determine if modeled propensities from a logistic regression on the $\boldsymbol{x_i}$'s produces biased estimates of the total without running simulations. In the second part, we evaluate better models for adjusting for nonresponse are proposed based on the graphical analysis.

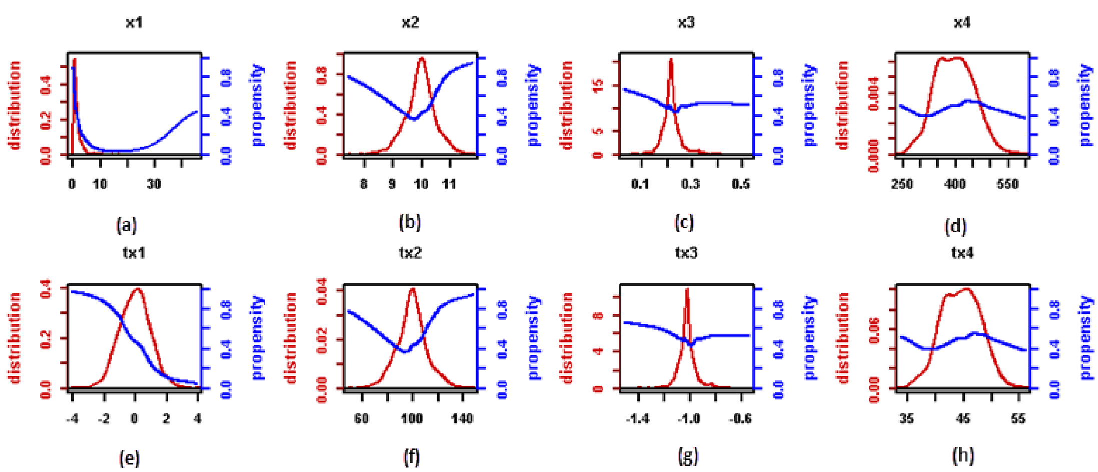


Figure 8: B-plot for simulation example in Kang & Schafer (2007)

The first step of the graphical analysis is to explore the response pattern of the observed variables. Figure 8a to 8d show the *b-plots* for the auxiliary variables $\boldsymbol{x_i}$'s. As shown in

Figure 8a and b, variables x_{1i} and x_{2i} are problematic since the relationship with the response is far from a smooth sigmoidal curve. Figure 9a shows the *heat map fit* plot of the propensities from a logistic regression on the $\boldsymbol{x}_i$'s. The plot shows that the model does not have a good fit because the curve does not match the 45° line; however, the model is not as bad as we initially thought despite of not observing the variables $\boldsymbol{z}_i$'s (Robins, Sued, Lei-Gomez, & Andrea, 2007; Samii, 2011). However, there is little room for improvement using the variables $\boldsymbol{x}_i$'s.

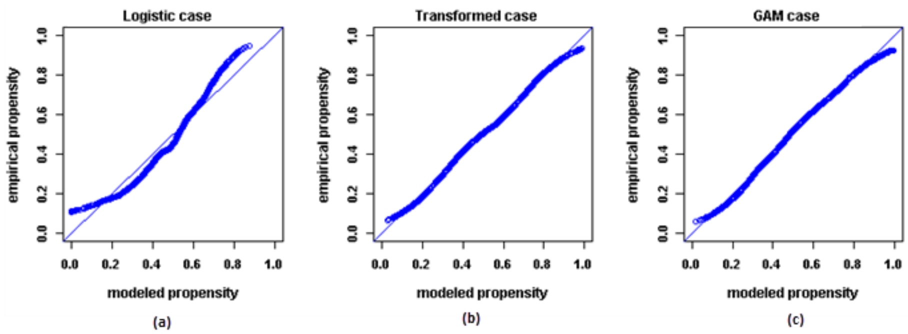


Figure 9: Heat fit plots for different nonresponse adjustments

For the second research question, a better logistic model can be created using transformed variables. New auxiliary variables were created using the Box-Cox power transformation. The *b-plots* for response of the transformed variables are shown in Figure 8e to 8h. The transformation has the largest impact on variable x_{1i} which is shown as tx_{1i} in Figure 8e. Because the response propensity is not monotonic for the transformed variable tx_{2i}, the logistic model included the interaction of the coefficient for tx_{2i} and the categorical variable created as $\delta(tx_{2i} > 9.3)$. The heat map fit plot for the modeled propensities using the transformed auxiliary variables is shown in Figure 9b.

A nonparametric model for modeling the propensities using the transformed variables was also developed. The nonparametric model uses LOWESS smoothing regression with a generalized additive model (GAM). The GAM modeled propensities are shown in Figure 9c. The last two models and the corresponding adjustment factors are expected to produce close to unbiased estimates of the total. The precision of these two nonresponse adjusted estimates is not evaluated in this paper.

4.2 Example 2: Graphical Analysis of Kreuter and Olson (2011)

Kreuter and Olson (2011) explored the relationship among the response, outcome variable, and auxiliary variables correlated to both response and outcome through a large simulation study. The article does not use survey data and their recommendations are based on the simulation results; hence, the importance of understanding the relationships among the variables that affect nonresponse in the simulation.

Kreuter and Olson (2011) can be seen as two separate analyses. In the first analysis, the bias of the unadjusted mean is studied under different relationships among the propensity to respond, dependent variable, and auxiliary variables. In the second analysis, the effectiveness in removing the bias of different nonresponse adjusted estimators of the mean is evaluated in terms of bias and relative root mean squared error (RMSE) using the RMSE of the unadjusted mean as a reference. All nonresponse adjustment estimates of the mean are computed using estimated propensities from logistic regression with different omitted variables.

In the Kreuter and Olson (2011) simulation, the propensity to respond is modeled as $\varphi_i = logit^{-1}(1 + \gamma_1 z_{1i} + \gamma_2 z_{2i})$ where $\boldsymbol{z_i} = (z_{1i}, z_{2i})^t$ is the vector of auxiliary variables independently distributed as $MN\left(\binom{0}{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. The dependent variable y_i is a linear function of $\boldsymbol{z_i}$'s defined as $y_i = 10 + \beta_1 z_{1i} + \beta_2 z_{2i} + u_i$ where $u_i \sim N(0,1)$. The auxiliary variables $\boldsymbol{z_i}$'s influence response through the parameters γ_1 and γ_2 and affect the population y through the parameters β_1, and β_2.The simulation created 54 different populations corresponding to parameters $\beta_1 = 0.1$, 2, and 4; β_2.=0.1, 2, 4, -0.1, -2, and -4; and $\rho = 0, 0.2,$ and -0.2. For each artificial population, nine response patterns were generated using the parameters γ_1 =0.1, 1, and 3; and $\gamma_2 = 0.1$, 1, and 3. A total of 486 scenarios were simulated. For each scenario, four estimates of the mean of y were computed using the estimators in Table 1. Three estimators were computed using modeled propensities from logistic regression with different omitted auxiliary variables (e.g., both z_{1i} and z_{2i}, either z_{1i} and z_{2i}).

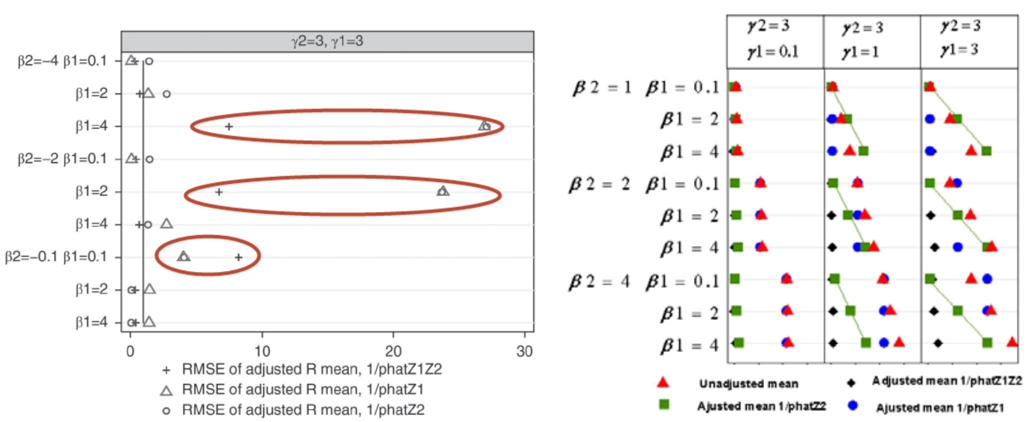
Table 1: Estimator of mean used in Kreuter and Olson (2011)

Estimators	*Description*
UADJ	No adjustment (or adjusted by $\hat{\varphi}_i^0$ =1/r)
PHAT_Z1	Adjusted using variable z_{1i} in $\hat{\varphi}_i^{z_1} = logit^{-1}(\hat{\gamma}_0 + \hat{\gamma}_1 z_{1i})$
PHAT_Z2	Adjusted using variable z_{2i} in $\hat{\varphi}_i^{z_2} = logit^{-1}(\hat{\gamma}_0 + \hat{\gamma}_2 z_{2i})$
PHAT_Z1Z2	Adjusted using variables z_{1i} and z_{2i} in $\hat{\varphi}_i^{z_1 z_2} = logit^{-1}(\hat{\gamma}_0 + \hat{\gamma}_1 z_{1i} + \hat{\gamma}_2 z_{2i})$

The bias of the unadjusted mean can be approximated by the formula $Bias(\bar{y}_{uadj}) \approx Cov(y, \varphi)/\bar{\varphi}$ (Lessler & Kalsbeek, 1992). Kreuter and Olson (2011) evaluated the bias of the unadjusted mean through the simulation. However, an algebraic expression to examine the bias can be obtained using the Taylor series approximation. The algebraic expression of the numerator is $Cov(y, \varphi) \approx e^{\gamma_0}(\beta_1\gamma_1 + \beta_2\gamma_2)/(1 + e^{\gamma_0})^2$. This expression makes it easier to understand the conditions where the unadjusted mean is unbiased.

Relative Ratio of Root Mean Squared Error

Estimated Means



(a) (b)

Figure 10: Estimated means and RMSE for the response pattern $\gamma_1 = \gamma_2 = 3$ (Source: Kreuter and Olson, 2011)

Kreuter and Olson (2011) report that when the influence on response is strong (i.e., $\gamma_1 = \gamma_2 = 3$) and the population parameters are such that $\beta_1 = -\beta_2$, the unadjusted mean is unbiased. In this case, the effect on nonresponse is said to cancel each other out because of the opposite signs between β_1 and β_2. In these scenarios, the unadjusted mean is very efficient compared to the other estimators as shown by the relative RMSE marked in red in Figure 10a. Although this is correct, there are many more conditions where the unadjusted mean is also unbiased. The unadjusted mean is unbiased when $Cov(y, \varphi) = 0$, that is when $\beta_1\gamma_1 = -\beta_2\gamma_2$. When $\beta_1 = -\beta_2$, the unadjusted mean is unbiased independently from the strength of the influence γ_1 and γ_2 are as long as $\gamma_1 = \gamma_2$. In other words, their influence on response does not necessarily need to be strong to yield an unbiased unadjusted mean. Figure 11 shows the *b-plots for response* for other situations where the unadjusted mean is also unbiased. In Figure 11a, the influence on response is not strong ($\gamma_1 = 1$ and $\gamma_2 = 1$). In Figure 11b, where $\beta_1 \neq -\beta_2$, the effects on the population do not cancel each other out. In Figure 11c, influences on response and the population are mixed. The *b-plots for response* of these situations show a flat response pattern.

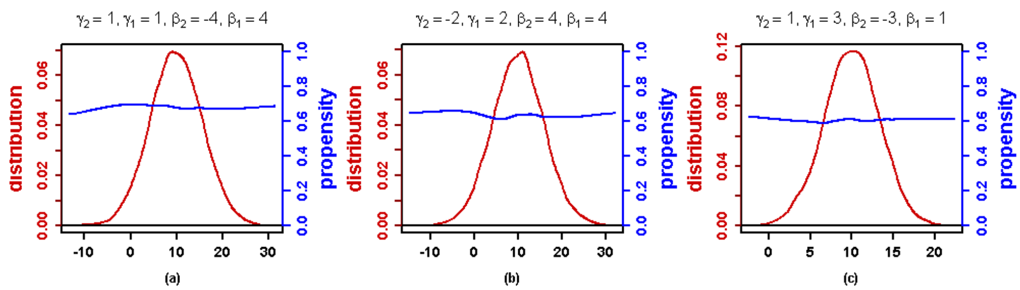


Figure 11: Other scenarios where the unadjusted mean is unbiased

In general, the unadjusted mean is unbiased when influences on response and population expressed by the terms of the products $\beta_1\gamma_1$ and $-\beta_2\gamma_2$ have the same value but opposite signs so both terms cancel each other out. This situation is unique and such a balance in situations with many influences may be difficult to occur in actual surveys. Here, the graphical analysis let us discover other scenarios not evaluated in the simulations where the unadjusted mean is unbiased.

In the second part of the graphical analysis, we examine other scenarios that are difficult to explain where the behavior of some nonresponse adjusted estimators of the mean seem counterintuitive. In the scenarios, "*…when either z variable is a moderate to strong predictor of the Y variable and a moderate to strong predictor of propensity, then including only one z variable in the adjustment model substantially increases the bias of the estimate. That is, accounting for only one of the competing influences on propensity will actually damage (i.e., increase the bias) survey estimates,*" (Kreuter and Olson, 2011). Some estimates in these scenarios are represented by green squares in the second and third panels in Figure 10b. To observe the bias, a green line connecting the squares is drawn on the plot. The line shows how the bias increases as the value of β_1 changes from 0.1 to 4 for the estimator of the mean PHAT_Z2, which accounts for only one of the competing influences (e.g., γ_2).

To graphically explore these scenarios, two types of *b-plots* and the *heat map fit* plots for the auxiliary variable z_{2i} that is used to model the propensities for the adjustment were produced. These plots are shown in Figure 12 and 13 for the populations with parameters $\beta_2 = 4$, β_1 =0.1, 2, and 4; and response pattern with parameters $\gamma_1 = 3$ and $\gamma_2 = 3$. All these plots indicate that the estimates should be unbiased (all observed propensities are modeled correctly). The reason for the increasing bias is found in the *b-plots* of the dependent variable *y* shown in Figure 14. The *b-plots for response* show that the parameters of simulation population create a specific response pattern where a portion of the population does not have any practical chance to respond while others respond with certainty. Cases with no chance to respond are located to the left of the vertical line in the plots. The population shifts to the left with respect to the response propensity curve as the value of β_1 increases (see Figure 14 a, b, and c). The fraction of the population without a probability to respond becomes larger despite the response propensities on auxiliary variable z_{2i} being modeled correctly. In other words, the simulation is creating a form of undercoverage, which increases (i.e., larger bias) as β_1 increases. The *b-plots for the adjustment factors* in Figure 14 c, d, and e show the effect on the adjustment factors when we try to adjust for this undercoverage using logistic regression. The curves are far from horizontal lines at 1. Even cases that responded with certainty have adjustment factors larger than one. In these scenarios, there are violations of the assumptions for the nonresponse adjustment since the method requires the response propensities to be bounded and away from 0 (Robins, Rotnitzky, & Ping, 1994). Many scenarios in the simulation study do not have response propensities away from 0. Using modeled propensities from logistic regression is not the best way to adjust for nonresponse/undercoverage in these situations. It is uncertain if these scenarios were planned in the simulation in Kreuter and Olson (2011) but this explanation is not reported in the article. The graphical analysis enabled us to identify the reasons of these results. A graphical analysis can us help examine the simulations and determine if they produce extreme conditions not intended in the study.

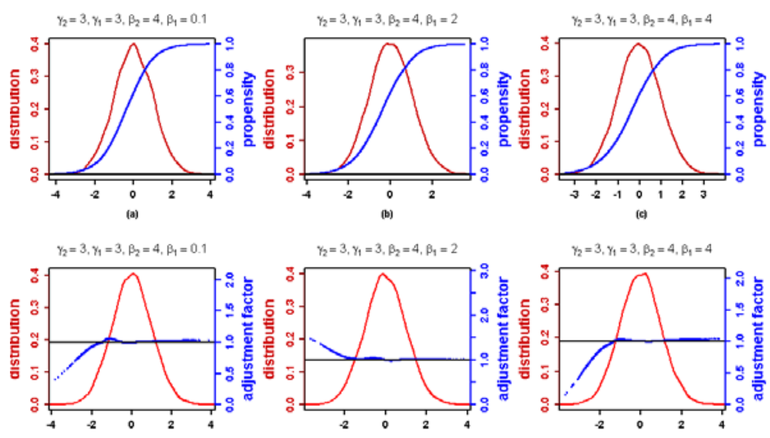


Figure 12: B-plots for auxiliary variable z_{2i}

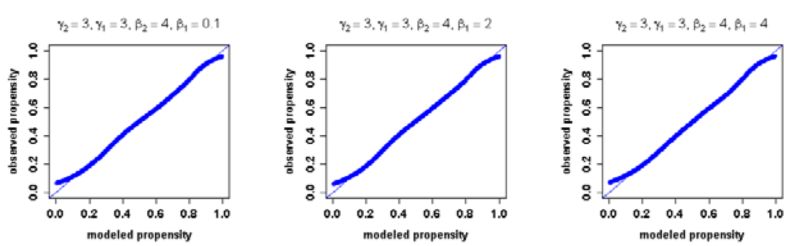


Figure 13: Heat map fit plots for propensities modeled using auxiliary variable z_{2i}

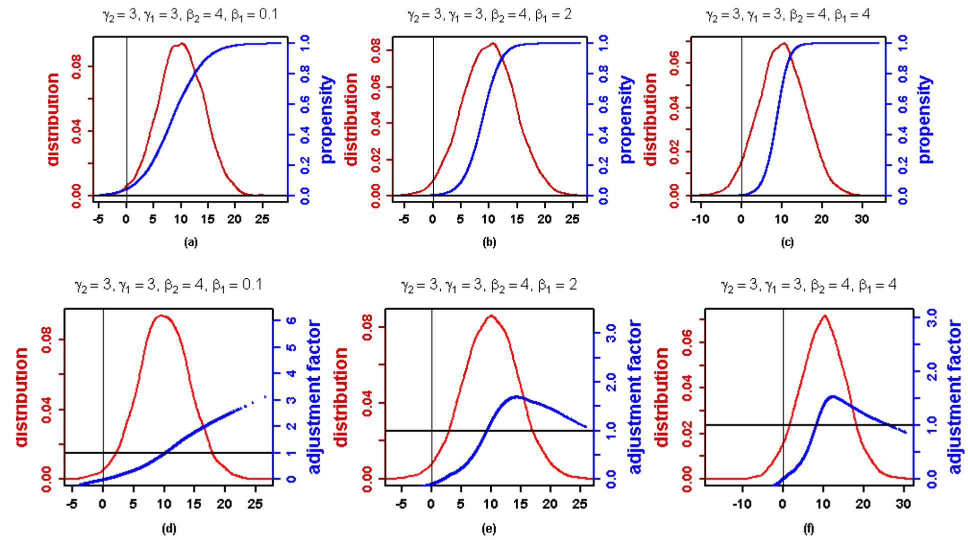


Figure 14: B-plots for dependent variable *y*

5. Summary

As a proof of concept, graphical tools can be implemented and used to analyze nonresponse in survey data. This type of graphical analysis has not been reported in the literature. The main limitation for a more extensive use of this type of analysis is the additional development work needed to complete the *NR-Toolbox*. This is connected to current time commitments and funding. There are also some theoretical and practical issues that need to be addressed. The theoretical issues include the computation of standard errors and their visualization in current displays, the application to domains in survey data, and the inclusion of new plots. Practical issues are the integration with current *R* packages, the survey data limitations of current package limitations, and the development of the GUI. Although the current version of the toolbox is not ready for primetime, its utility and potential has been demonstrated by examining the nonresponse patterns in two papers.

Acknowledgements

I would like to thank Dr. J. Michael Brick for his support and comments on this work.

References

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, A. (1983). *Graphical Methods for Data Analysis.* Belmont, CA: Wadsworth/ Duxbury Press.

Emerson, J. W. (1998). Mosaic displays in S-PLUS: A general implementation and a case study. *Statistical Computing and Graphics Newsletter* , *9*(1), 17–23.

Esarey, J., & Pierce, A. (2012a). Assessing Fit Quality and Testing for Misspecification in Binary Dependent Variable Models. *Political Analysis, 20*(04), pp. 480-500.

Esarey, J., & Pierce, A. (2012b). heatmapFit: Heatmap Fit Statistic For Binary Dependent Variable Models. R package version 1.0. Retrieved from ttp://CRAN.R-project.org/package=heatmapFit

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association, 89*, 190–200.

Greenhill, B., Ward, M. D., & Sacks, A. (2011). The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science, 55*(4), pp. 991-1002.

Hastie, T. (2013). gam: Generalized Additive Models. R package version 1.09. Retrieved from http://CRAN.R-project.org/package=gam

Hinkley, D. (1985). Transformation diagnostics for linear models. *Biometrika, 72*, pp. 487–496.

Hofmann, H., & Theus, M. (2005). Interactive graphics for visualizing conditional distributions,. *Unpublished Manuscript*.

Hummel, J. (1996). Linked bar charts: Analysing categorical data graphically. *Computational Statistics, 11*, pp. 23–33.

Jacoby, W. G. (1997). *Statistical Graphics for Univariate and Bivariate Data.* Thousand Oaks, CA: Sage.

Kang, J. D., & Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science, 22*(4), pp. 523-539. doi:10.1214/07-STS227

Kreuter, F., & Olson, K. (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research, 40*, 311-332.

Lessler, J., & Kalsbeek, W. (1992). *Nonsampling errors in surveys.* New York: John Wiley and Sons.

Lumley, T. (2012). survey: analysis of complex survey samples. R package version 3.28-2. .

Meyer, D., Zeileis, A., & Hornik, K. (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software, 17*(3), 1-48. Retrieved from http://www.jstatsoft.org/v17/i03/

Meyer, D., Zeileis, A., & Hornik, K. (2013). vcd: Visualizing Categorical Data. R package version 1.3-1.

Neuwirth, E. (2013). ENmisc: Neuwirth miscellaneous. R package version 1.2-7. Retrieved from http://CRAN.R-project.org/package=ENmisc

Pasek, J., Tahk, A., Culter, G., & Schwemmle, M. (2014). weights: Weighting and Weighted Statistics. R package version 0.80. Retrieved from http://CRAN.R-project.org/package=weights

R Development Core Team. (2013). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria. doi:http://www.R-project.org

Robins, J. M., Rotnitzky, A., & Ping, Z. L. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association, 89*(427), 846-866.

Robins, J., Sued, M., Lei-Gomez, Q., & Andrea, R. (2007). Comment: Performance of Double-Robust Estimators When Inverse Probability Weights Are Highly Variable. *Statistical Science, 22*(4), 544-559.

Samii, C. (2011). Weighting and Augmented Weighting for Causal Inference with Missing Data: New Directions. Retrieved August 5, 2014, from https://files.nyu.edu/cds2083/public/docs/methods/samii10_weighting100818.pdf

Templ, M., Alfons, A., Alexander, K., & Bernd, P. (2013). VIMGUI: Visualization and Imputation of Missing Values. Retrieved from http://CRAN.R-project.org/package=VIMGUI